

Multimodal Biometric Recognition under Unconstrained Settings

João C. Monteiro

Supervisor: Jaime S. Cardoso

Programa Doutoral em Engenharia Electrotécnica e de Computadores

July, 2017

© João C. Monteiro: July, 2017

Faculdade de Engenharia da Universidade do Porto

Multimodal Biometric Recognition under Unconstrained Settings

João C. Monteiro

Dissertation submitted to Faculdade de Engenharia da Universidade do Porto
to obtain the degree of

Doctor Philosophiae in Electronic & Computer Engineering

President: Aurélio Joaquim de Castro Campilho (PhD)
Referee: António José Ribeiro Neves (PhD)
Referee: Luís Filipe Barbosa de Almeida Alexandre (PhD)
Referee: Jaime dos Santos Cardoso (PhD)
Referee: Luís Filipe Pinto de Almeida Teixeira (PhD)

July, 2017

A Engenharia.

Abstract

In recent years, it has been recognized that the path forward, regarding biometrics, concerns the development of systems that work in an accurate and robust way, independent of the conditions under which input data is acquired. With growing difficulties in performing biometric recognition under these new challenging scenarios, the use of multiple biometric trait representations, is also gaining growing attention by the research community. By combining multiple sources of biometric information, increased performance and robustness is achieved, when compared with most traditional systems based on a single representation.

In the present thesis, the idea of multimodal biometric recognition performed on data acquired under less controlled scenarios is explored. We start this analysis with a series of unimodal approaches based on face and periocular data. Both modalities are assessed for less than ideal acquisition scenarios, characterized by variable illumination and occlusion conditions, with the establishment of a recognition pipeline based on the universal background model (UBM) approach, with proven performance in the field of speaker verification. Furthermore, regarding face recognition, we propose a cognitively motivated hierarchical framework, based on the global-to-local flow of object recognition by the human brain. On the other hand, for periocular data, we explore usability in mobile devices, making use of transfer learning principles to achieve inter-operable biometric recognition across multiple devices. Face biometrics is further explored, from a multispectral point-of-view, using the novel Intel RealSense family of sensors to perform multimodal recognition in a series of alternative face image representations such as RGB, IR and depth.

With the work presented throughout this thesis, a series of state-of-the-art results were achieved, in a wide array of application scenarios, with a considerable margin for progression still left by the constant evolution of both technology, research trends within the field of biometrics and its applicability in society nowadays and in the near future.

Keywords: Machine Learning. Biometrics. Non-ideal Acquisition. Face Recognition. Periocular Recognition. Universal Background Model. Deep Learning.

Resumo

Recentemente, o caminho a seguir no campo de investição da biometria tem sido o do desenvolvimento de sistemas capazes de funcionar de forma precisa e robusta independentemente das condições sob as quais a aquisição de dados é realizada. Com os crescentes desafios resultantes de tais cenários, o uso de múltiplas representações de traços biométricos tem começado a ganhar crescente atenção por parte da comunidade científica. Através da exploração de múltiplas fontes de informação biométrica, um aumento tanto na performance como na robustez das metodologias tem vindo a ser observado, quando comparadas com alternativas baseadas numa representação singular.

Na presente tese, será explorada a ideia de reconhecimento biometrico multimodal utilizando informação adquirida em cenários menos controlados. Começamos esta análise com uma série de abordagens unimodais baseadas em dados de face e região periocular. Ambas as modalidades são testadas em cenários de aquisição longe dos ideais, com consideráveis variações tanto ao nível da iluminação como da oclusão, estabelecendo-se um algoritmo baseado na utilização de modelos universais de base (UBM - Universal Background Model), com resultados provados na área da biometria de voz. Considerando o reconhecimento da face, é proposta uma abordagem hierárquica, motivada pelos mecanismos de cognição humanos, baseado no fluxo de informação de representações mais globais para representações mais locais durante o processo de reconhecimento de objectos pelo cérebro humano. Por outro lado, para a região periocular, a sua possível utilização em dispositivos móveis é analisada, utilizando princípios de transferência de aprendizagem para atingir um processo de reconhecimento biométrico inter-operável entre uma multiplicidade de dispositivos. A biometria da face é também considerada de um ponto de vista multi-espectral, fazendo uso da recente família de sensores Intel RealSense, a partir dos quais é possível adquirir uma série de representações de imagens de face, como RGB, IV e profundidade, com as quais se torna exequível o desenvolvimento de alternativas multimodais para o processo de reconhecimento.

Com o trabalho apresentado ao longo da presente tese, uma série de resultados ao nível do estado da arte foram adquiridos, num leque considerável de aplicações, havendo no entanto ainda uma considerável margem para melhoria, motivada quer pela contínua evolução a nível tecnológico, quer pela constante adaptação dos desafios do ramo da biometria ao nível da sua aplicabilidade na sociedade.

Keywords: Aprendizagem Computacional. Biometria. Aquisição não-controlada. Reconhecimento Facial. Reconhecimento Periocular. Model Universal de Referência. Aprendizagem Profunda.

Acknowledgments

Há quase uma década atrás, um grupo de pessoas pelas quais vim a desenvolver o mais profundo dos respeitos, incutiu em mim um princípio que, passados todos estes anos, entra em directo conflito com esta, não raras vezes subvalorizada, parcela do presente documento. Centrava-se tal princípio na minimização da quantidade de vezes que o vocábulo "obrigado" é utilizado, como reacção a acções ou palavras que, originadas em terceiros, resultam num aumento não desprezável do bem-estar do receptor enquanto ser humano. Por detrás de tal ideia, aparentemente contraditória quando comparada com aquilo que são as mais basilares regras da vida em sociedade, encontra-se o racional que nenhuma acção deve ser levada a cabo como uma obrigação, esperando como resultado a recepção do tradicional "obrigado", mas sim como um fim em si mesma, realizada na noção de que todas as acções que levamos a cabo o devem ser com o simples cumprimento do objectivo que lhes deu origem, e nunca na necessidade de obter aprovação. Não agradecer implica o entendimento que a felicidade ou realização que palavras e acções de terceiros exercem sobre nós é algo intrínseco às pessoas, e não uma obrigação que careça de feedback comprovativo.

Ora, a ideia apresentada como preâmbulo a esta secção é significativamente comprometida quando a entidade alvo das acções e palavras sente a mais sincera necessidade de expressar o seu apreço com quem, sem qualquer necesidade de ouvir um "obrigado", mais o mereceria. E é neste ponto que me encontro, ao fim de uma década vivida na maior das plenitudes, rodeado de quem mais me fez crescer como pessoa e profissional, desejando agradecer a todos, mas não querendo nunca usar o vocábulo proibido, no receio que a sua proximidade semântica com "obrigação" menospreze toda a felicidade que com os alvos desta já extensa dissertação, partilhei. E é assim que utilizarei o nobre truque do sinónimo semanticamente distinto para endereçar a todos a mais sincera palavra de apreço, pelos momentos vividos, pelas conversas intermináveis, pelas longas noites e as complicadas manhãs subsequentes, pelos silêncios que falam mais que muitas palavras, pelas gargalhadas que ressoarão por anos na memória de quem as testemunhou, pelas lágrimas que nos fizeram crescer e continuar a viver, pelas histórias que durante anos levarei comigo, e das quais todos vós fazem, e para sempre farão, parte. E sendo assim, sentindo assim, como posso senão agradecer?

Como enumerar todos os que merecem ser incorporados nesta palavra de apreço? Se estas palavras vos chegaram às mãos é porque certamente as merecem, e sei que não seria um nome dactilografado a fazer a diferença. Continuem lá nas histórias que vamos escrevendo juntos e que nunca seja uma obrigação vossa aturar-me, mas sim uma necessidade minha continuar a contornar o vocábulo proibido para vos expressar o apreço por vos ter sempre ao meu lado.

João C. Monteiro

"Never forget what you are, for surely the world will not. Make it your strength. Then it can never be your weakness. Armour yourself in it, and it will never be used to hurt you."

George R. R. Martin

Contents

Agradecimentos **vii**

List of Figures **xvii**

List of Tables **xx**

List of Abbreviations **xxi**

I Introduction and Theoretical Background **1**

1 Introduction **3**
1.1 Contributions 4
1.2 List of Publications 5
1.3 Document Structure 6

2 Fundamentals of Biometrics **7**
2.1 Brief Historical Overview 7
2.2 Applications of Biometric Recognition 9
2.3 Main Standards and European Projects 11
2.4 System Architecture 14
2.5 Operating Modes 15
2.6 Biometric Traits 16
2.7 Multimodal Biometrics 18
2.8 Evaluation Metrics 21
2.9 Conclusion 25

3 Computer Vision and Machine Learning Tools **27**
3.1 Feature Extraction and Description 27
3.1.1 SIFT 27
3.1.2 LBP 28
3.1.3 HOG 29
3.1.4 GIST 29
3.1.5 Deep Neural Networks 30
3.2 Machine Learning Tools 31
3.2.1 Gaussian Mixture Models 31
3.2.2 Support Vector Machines 32
3.2.3 Stacked Denoising Autoencoders 34
3.2.4 Convolutional Neural Networks 34

3.3 Conclusion 38

II Scientific Contributions **39**

4 Unimodal Recognition **41**
4.1 Periocular Recognition 41
4.1.1 Related Work 42
4.1.2 Universal Background Model 43
4.1.3 Algorithm Overview 44
4.1.4 Hypothesis Modeling 45
4.1.5 Experimental Tests and Results 47
4.2 Periocular Recognition in Mobile Scenarios 54
4.2.1 Santos et al. Method 55
4.2.2 Algorithm Extensions 56
4.2.3 Results and Discussion 57
4.3 Hierarchical Face Recognition 63
4.3.1 Related Work 64
4.3.2 Algorithm Overview 66
4.3.3 Consideration on Hypothesis Modeling 66
4.3.4 Hierarchical Decision 68
4.3.5 Experimental Tests and Results 70
4.4 Conclusion 77

5 Transfer Learning for Cross-Sensor Biometrics **79**
5.1 Related Work 80
5.2 Cross-Sensor Methodologies 81
5.3 Transfer Learning Approaches 82
5.3.1 Transfer Learning supervised (TLs) 82
5.3.2 Source-Target-Source (STS) 83
5.3.3 Multi-source Source-Target-Source 83
5.4 Experimental Tests and Results 85
5.4.1 Experimental setup 86
5.4.2 Cross-sensor recognition performance 86
5.5 Conclusion 92

6 Multi-spectral Face Recognition **95**
6.1 An Overview of Multi-spectral Sensors 95
6.2 Related Work in 3*D* Face Recognition 98
6.2.1 2D Data Augmentation 99
6.2.2 Unimodal Approaches Based on 3*D* Models 100
6.2.3 Multimodal Alternatives 101
6.3 Related Work in IR Face Recognition 103
6.4 Multi-spectral RGB/3*D*/IR Datasets 105
6.5 RealFace Dataset 107
6.6 Experimental Tests and Results 110
6.6.1 Experimental setup for the RealSense dataset 110
6.6.2 EURECOM dataset 114
6.6.3 Performance analysis 115

6.7 Conclusion 115

7 Unsupervised Learning of Convolutional Neural Networks using Multi-spectral Data 119
7.1 Unsupervised CNN Learning Framework 120
7.1.1 Data Partitioning 120
7.1.2 Learning Process 120
7.1.3 Classification 123
7.1.4 Experimental Results 123
7.2 Conclusion and Future Work 125

III Conclusion 127

8 Conclusions and Future Lines of Research 129
8.1 Conclusions 129
8.2 Research Trends 133

Bibliography 139

List of Figures

2.1 Schematization of some of the measurements performed as part of the Bertillon system (Proença, 2006). 8
2.2 Example of pattern used in Sir Edward Henry's fingerprint classification system (International Biometrics Group, 2003). 8
2.3 Traditional framework of a biometric system. Adapted from (Jain et al., 2000). . 15
2.4 Operating modes of a biometric system: verification (top) and identification (bottom) (Griaulle Biometrics, 2014). 16
2.5 Some examples of commonly researched biometric traits (Jain et al., 2016). . . . 17
2.6 Scenarios in a multimodal biometric system (Ross and Jain, 2004). 19
2.7 Approaches to information fusion in mutimodal biometrics (Jain et al., 2005). . . 21
2.8 Evolution of the FAR and FRR error rates for growing decision threshold values. 23
2.9 Examples of receiver operating characteristic (ROC) curves for performance assessment of biometric recognition systems. 24
2.10 Generic example of a cumulative match characteristic (CMC) curve for biometric identification system performance assessment. 24

3.1 Gradient magnitude and orientation computation of a single keypoint using SIFT descriptor. Adapted from (Lowe, 2004). 28
3.2 Local binary pattern computation for a generic 3×3 neighborhood (adapted from (Wagner, 2011)). 29
3.3 Histogram of oriented gradients example (adapted from (Image Processing in Python, 2012)). 30
3.4 Visual example of the evolution of two GMM components for a growing number of iterations (from top-left to bottom-right). Adapted from (Becker, 2009) 32
3.5 Maximum margin classifier example for a simple two-dimensional dataset. Adapted from OpenCV's documentation (OpenCV, 2016). 33
3.6 Kernel trick for non-linearly separable scenarios (Liao, 2015). 34
3.7 Stacked denoising autoencoder architecture. Adapted from (Berniker and Kording, 2015). 35
3.8 Typical convolutional neural network architecture. Adapted from (Deep Learning, 2010). 35
3.9 Convolutional layer with a region of the image connected to a neuron. Adapted from (Deep Learning, 2010). 36
3.10 Convolutional layer with its respective multiple feature maps. Adapted from (Deep Learning, 2010). 37
3.11 Pooling layer outputting condensed feature maps. Adapted from (Deep Learning, 2010). 37

3.12 Fully connected layer with all feature map neurons connected to the output neurons. Adapted from (Deep Learning, 2010). 38

4.1 Example of periocular regions from both eyes, extracted from a face image Woodard et al. (2010). 42
4.2 Graphical representation of the main steps in both the enrollment and recognition (verification and identification) phases of the proposed periocular recognition algorithm. 45
4.3 Examples of noisy image from the UBIRIS.v2 database. 48
4.4 Examples of iris images in the MobBIO database. 48
4.5 Examples of face images in the MobBIO database. 49
4.6 Recognition rates obtained with the optimization subset for variable values of parameter M, number of mixtures in the trained GMMs. 51
4.7 ROC and CMC curves for: (a-b) UBIRIS.v2 database, setup (1); (c-d) UBIRIS.v2 database, setup (2) and (e-f) MobBIO database, setup (3);. ROC curves present the average results of cross-validation, whereas CMCs present the average value and error-bars for the first 10 ranked IDs in each setup. 52
4.8 Identification results for rank-4 in the UBIRIS.v2 database. The first column depicts the tested images while the remaining 4 images exemplify representative images from the 4 most probable models, after the recognition is performed. The blue squares mark the true identity. 54
4.9 Flow diagram of the main composing blocks of the methodology proposed in (Santos et al., 2015). 55
4.10 Examples of images from each subset of the CSIP database. From (a-j) respectively: $AR0$, $AR1$, $BF0$, $BR0$, $BR1$, $CF0$, $CR0$, $CR1$, $DF0$ and $DR0$. 58
4.11 Schematic representation of the proposed algorithm and its main blocks: (**a**) training of the universal background models using data from multiple individuals; (**b**) maximum *a posteriori* (MAP) adaptation of the universal background models (UBM) to generate individual specific models; and (**c**) testing with new data from unknown sources. 67
4.12 Example images from two subjects of the ORL database. 70
4.13 All images from a single subject enrolled in the Extended Yale B database. Images (**a**) to (**e**) correspond to Subsets 1 to 5, respectively. 71
4.14 Example images from one subject of the AR database. 72
4.15 Main results obtained with the proposed methodology for the sunglasses and scarf images of the ORL face database. Each plotted point represent a specific value of parameter θ_l, ranging from $[0,\infty]$. 72
4.16 Main results obtained with the proposed methodology for the sunglasses and scarf images of the Extended Yale B face database. Each plotted point represent a specific value of parameter θ_l, ranging from $[0,\infty]$. 73
4.17 Main results obtained with the proposed methodology for the sunglasses and scarf images of the AR face database. Each plotted point represent a specific value of parameter θ_l, ranging from $[0,\infty]$. 73

5.1 TLs for a simple example of a first layer feature transference. 83
5.2 Graphical representation of the MS-STS Rank-1 recognition rates obtained for all the no-flash subsets of the CSIP database using all the flash datasets as sources, plotted against the respective BL, TLs and STS results. 93

5.3 Graphical representation of the MS-STS Rank-1 recognition rates obtained for all the six possible orders of the chosen source datasets. Results concern to (a) *AR*0 and (b) *CR*0 as targets. 94

6.1 Visual example of the differences in depth map rendering and body joint detection between Kinect v1 (top) and Kinect v2 (bottom). Adapted from (Xu and McGorry, 2015). 96
6.2 Intel RealSense models: (a) R200 and (b) SR300. 98
6.3 Examples of data from some multimodal RGB/D/IR datasets: (a) Florence Superface, (b) IIIT-D, (c) Labeled Infrared-Depth and (d) FaceWarehouse. Adapted from (Fanelli et al., 2013), (Goswami et al., 2014) , (Cao and Lu, 2015) and (Cao et al., 2014) 106
6.4 RealFace dataset acquisition setup: (a) subject; (b) acquisition control software; (c) SR300 model and (d) R200 model. 108
6.5 Representative examples of the poses, occlusion, illumination and expression variations considered for each subject during the RealFace acquisition process. . . . 108
6.6 Multimodal data from a single acquisition in the RealFace dataset: (a) RGB; (b) Depth-map; (c) Point-Cloud and (d) Infrared. 109
6.7 Manual annotation on the RealFace dataset: (a) Frontal neutral; (b) Frontal occlusion; (c) Profile neutral and (d) Profile occlusion. 109
6.8 Comparison between the depth image quality of the R200 model of Intel RealSense and other low-cost alternative sensors. Adapted from (Song et al., 2015). 110
6.9 Example RGB (a)-(g) and depth (h)-(n) images from the multiple subsets of the EURECOM dataset. 114

7.1 Schematic representation of the unsupervised training and supervised test phases of the developed framework for multi-spectral face recognition. 121
7.2 Evolution of the outputs of CNN_{RGB} and CNN_{IR} from their original values, with a randomly initialized fc-loss layer, to the pre-trained versions after 1000 iterations. 124

List of Tables

2.1 Comparative data analysis of some common biometric traits (adapted from Jain et al. (2000) and Proença (2006)). 18
2.2 Type of errors observed in a typical biometric system: false matches or false genuines (FG) and false non-matches or false impostors (FI). The computation of the FAR and FRR can be easily performed based on such values. 22

4.1 Comparison between the average obtained results with both experimental setups for the UBIRIS.v2 database and some state-of-the-art algorithms. 51
4.2 Technical details concerning the acquisitions setups used for each subset of the CSIP database. 58
4.3 Rank-1 recognition rates obtained for each subset of the CSIP database for some variations of both tested methodologies. 60
4.4 Comparison between the obtained r_1 values for the ORL database and some state-of-the-art algorithms. 74
4.5 Comparison between the obtained r_1 values for the AR database and some state-of-the-art algorithms. 74
4.6 Comparison between the obtained r_1 values for the Extended Yale B database and some state-of-the-art algorithms. 74

5.1 Rank-1 recognition rates, in %, observed for the GMM-UBM algorithm for all possible cross-sensor scenarios in the CSIP database. 87
5.2 Rank-1 recognition rates, in %, observed for the SV-SDA methodology and TLs approach for all possible cross-sensor scenarios in the CSIP database. 87
5.3 Rank-1 recognition rates, in %, observed for the CNN methodology and TLs approach for all possible cross-sensor scenarios in the CSIP database. 88
5.4 Rank-1 recognition rates, in %, observed for the SV-SDA methodology and a single cycle of the STS approach for all possible cross-sensor scenarios in the CSIP database. 89
5.5 Rank-1 recognition rates, in %, observed for the SV-SDA methodology and the STS approach for all possible cross-sensor scenarios in the CSIP database. 89
5.6 Rank-1 recognition rates, in %, observed for the CNN methodology and a single cycle of the STS approach for all possible cross-sensor scenarios in the CSIP database. 90
5.7 Rank-1 recognition rates, in %, observed for the CNN methodology and STS approach for all possible cross-sensor scenarios in the CSIP database. 90

6.1 List of the most common sensors referenced in 3*D* face recognition works. 97
6.2 Summary of the most relevant works concerning unimodal 3*D* face recognition . 102
6.3 Summary of the most relevant works concerning multimodal face recognition . . 102

6.4 VGG-face network configuration as proposed in (Parkhi et al., 2015) 113
6.5 Performance comparison of shallow (S) and deep (D) features on the RGB and depth modalities of the EURECOM dataset. 117
6.6 Performance comparison of shallow (S) and deep (D) features on the RGB, depth and IR modalities of the RealFace dataset. 117

7.1 Rank-1 recognition rates observed for different iterations of the designed unsupervised learning approach. 125

List of Abbreviations

CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
GMM	Gaussian Mixture Model
HOG	Histogram of Oriented Gradients
IDSM	Individual Specific Model
MAP	Maximum *a Posteriori*
ML	Machine Learning
PCA	Principal Component Analysis
ROI	Region of Interest
SDA	Stacked Denoising Auto-encoder
SIFT	Scale-Invariant Feature Transform
UBM	Universal Background Model
VW	Visible Wavelength

Part I

Introduction and Theoretical Background

Chapter 1

Introduction

From times immemorial, mankind has relied on specific features, such as face or voice, to distinguish between individuals. Nowadays, the task of personal identification is starting to rely on more artificial strategies. The most common of such approaches can be easily cataloged as either token-based or knowledge-based. Token-based approaches take advantage of a personal item, such as a passport, driver's license, ID card, credit card or a simple set of keys to distinguish between individuals. Knowledge-based approaches, on the other hand, are based on pieces of information a specific user knows that, theoretically, nobody else has access to. Examples of these are passwords or personal identification numbers (PIN). Both of these approaches present obvious disadvantages: tokens may be lost, stolen, forgotten or misplaced, while passwords can easily be forgotten by a valid user or guessed by an unauthorized one (Jain et al., 2000). In fact, all of these approaches stumble upon an obvious problem: any piece of material or knowledge can be fraudulently acquired, making token and knowledge-based identification unsatisfactory means of achieving the security requirements set by our society's needs. A consumer survey released by the U.S. Federal Trade Commission (FTC) stated that 8.3 million people were victims of identity theft in 2005, with total associated losses of 15.6 billion dollars (Kahn and Liñares-Zegarra, 2015). Such numbers are alarming and are causing a renewed interest in the research of automated identity authentication systems that offer both high efficiency, as well as improved security against impostor attacks. Biometrics represents a return to a more natural way of identification. Several physical and behavioral traits in humans show a considerable inter-individual variability: fingerprints and palmprints, the shape of the ears, the pattern of the iris, among others. Biometrics works by recognizing patterns within such biological traits, unique to each individual, to achieve reliable recognition. Testing someone by who (s)he is, instead of relying on something (s)he owns or knows seems likely to be the way forward: it is clearly more difficult to change a fingerprint or iris pattern than acquiring a physical item or piece of knowledge. Taking advantage of specific unique biometric features to develop robust and reliable identification systems is, therefore, an important challenge for the years to come, to accompany the growth of technological innovation and the security challenges that such growth will undoubtedly carry.

With such motivation in mind, the work developed and presented throughout this thesis aimed

at considering a wide array of topics of interest to the biometrics research community. The final contributions of the present thesis range from approaches based on a single biometric trait, to alternatives based on the fusion of multiple representations of a single trait. Regarding individual traits we focus on both face recognition, as the face is a highly intuitive, unique and easily accessible biometric trait, and periocular recognition as an alternative to face in less ideal scenarios, such as heavily occluded face images. Analysis of periocular image is carried out in both single sensor, as well as cross-sensor applications, motivated by the growing availability and variability of mobile devices capable of performing the whole biometric recognition framework. Face recognition is also analyzed from a multi-spectral point-of-view, making use of the recent advances in depth imaging, in order to make use of multiple independent sources of information to improve performance. The multi-spectral motivation is further considered in a preliminary study on unsupervised learning of CNNs based on multimodal data. The achieved contributions are summarized in further detail throughout the next section.

1.1 Contributions

As referred above, the focus of the work presented in this thesis was divided to a set of topics that can be understood as follows:

- **Establish the GMM-UBM approach in image-based traits:** The GMM-UBM approach was first proposed by Reynolds et al. (Reynolds et al., 2000) for speaker recognition, and served as the basis for a multiplicity of works in the area of voice-based biometrics for years after its initial proposal. An extrapolation of the rationale behind this approach to image-based traits, resulting in a versatile framework for multiple applications, was developed and assessed for both face and periocular recognition.

- **Hierarchical framework for face recognition in occlusion scenarios:** Face recognition is severely compromised when the discriminative portions of the face are occluded. An hierarchical framework, based on the cognitive mechanisms of the human brain, was proposed, performing recognition on increasingly smaller regions of the face, when more global descriptions fail to achieve considerable discrimination between individuals.

- **Transfer learning approach for cross-sensor recognition:** With the increasing availability and ease of access to multiple devices capable of biometric trait acquisition, the questions of whether these sensors are inter-operable and approaches trained in one of them still manage to present acceptable performance in others become increasingly important. Based on the ideas of deep transfer learning, a cross-sensor approach was applied to periocular recognition on a set of 10 distinct sensors.

- **Cross-spectral face dataset:** a RGB-depth-infrared database was designed in order to allow the research of multimodal recognition based on multiple representations of the same

biometric trait. A baseline performance analysis using some state-of-the-art approaches was also carried out, allowing an easy comparison against newly developed methodologies.

- **Unsupervised learning of convolutional neural networks using multimodal data:** The use of convolutional neural networks in computer vision has seen a growing tendency in recent years, with excellent performance observed in a series of applications. The need of large amounts of labeled data represents the main limitation of these approaches in some scenarios. A preliminary approach to achieve unsupervised learning of CNNs, when the amount of accessible unlabeled data surpasses that of labeled data, was also explored.

1.2 List of Publications

The contributions outlined in the previous section resulted in three journal publications:

- João C. Monteiro and Jaime S. Cardoso. *A Cognitively-Motivated Framework for Partial Face Recognition in Unconstrained Scenarios.* Sensors 15.1, 2015: 1903-1924.
- Chetak Kandaswamy, João C. Monteiro, Luís M. Silva and Jaime S. Cardoso. *Multi-source deep transfer learning for cross-sensor biometrics.* Neural Computing and Applications, 2016, 1-15.
- João C. Monteiro, Tiago Freitas and Jaime S. Cardoso. *Multimodal Hierarchical Face Recognition using Information from 2.5 D Images.* U. Porto Journal of Engineering, 2016, 2.2: 39-54.

The following publications in international conferences were also the result of the research work presented throughout this thesis:

- João C. Monteiro and Jaime S. Cardoso. *A Novel Application of Universal Background Models for Periocular Recognition.* International Joint Conference on Biomedical Engineering Systems and Technologies. Springer International Publishing, 2015. p. 291-307
- João C. Monteiro, Rui Esteves, Gil Santos, Paulo Torrão Fiadeiro, Joana Lobo and Jaime S. Cardoso *A Comparative Analysis of Two Approaches to Periocular Recognition in Mobile Scenarios.* International Symposium on Visual Computing. Springer International Publishing, 2015. p. 268-280.
- Tiago Freitas, Pedro Alves, Cristiana Carpinteiro, Joana Rodrigues, Margarida Fernandes, Marina Castro, João C. Monteiro and Jaime S. Cardoso. *A comparative analysis of deep and shallow features for multimodal face recognition in a novel RGB-D-IR dataset* International Symposium on Visual Computing. Springer International Publishing, 2016. p. 800-811.

Furthermore the developed work was also presented in the following national conferences:

- João C. Monteiro and Jaime S. Cardoso. *Multimodal biometric recognition under unconstrained settings.* 18th Portuguese Conference on Pattern Recognition (RECPAD), 2012.
- João C. Monteiro and Eduardo Mota. *An assessment of the potential of distinct facial regions for biometric recognition.* 19th Portuguese Conference on Pattern Recognition (RECPAD), 2013.
- João C. Monteiro and Jaime S. Cardoso. *Partial Face Recognition Under Occlusion Using Universal Background Models.* 20th Portuguese Conference on Pattern Recognition (RECPAD), 2014.
- João C. Monteiro and Jaime S. Cardoso. *A Framework for Face Recognition in Occlusion Scenarios.* 1st Doctoral Congress in Engineering (DCE), 2015.

1.3 Document Structure

The rest of the thesis is structured as follows: Chapter 2 serves as an overview of the transversal topics regarding research in biometric recognition, with Chapter 3 focusing on the theoretical background behind the computer vision and machine learning tools used in the development of the present work. Chapter 4 focuses on the unimodal approaches developed for both periocular and face recognition. Chapter 5 presents the transfer learning approaches developed for cross-sensor periocular recognition, while Chapter 6 concerns the topic of cross-spectral face recognition, based on 3D and IR alternatives to the more traditional RGB approaches. Chapter 7 explores multimodality as an alternative to unsupervised learning of CNNs and their applications to cross-spectral face data, and, finally, Chapter 8 presents the most recent trends of research in the topics explored throughout this thesis, serving as a conclusion as well as a roadmap for follow-up future work.

Chapter 2

Fundamentals of Biometrics

Biometrics can be seen as a more natural strategy for personal identification. By attempting identification based on physiological and behavioral traits, we are testing someone by who (s)he is, instead of relying on something (s)he owns or knows. Such an approach seems likely to be the way forward. We start this section with a brief overview of the historical evolution of biometric recognition as a means of human identification.

2.1 Brief Historical Overview

The use of human physical and behavioral patterns for identification is as old as mankind itself. Some paintings in a cave in Avignon, France, dated 31.000 years, and depicting hunting scenes, are accompanied by palm prints, that archaeologists believe to have been used as some kind of a specific signature by the author (Renaghan, 1997; Bala, 2008). Event without this kind of knowledge it is trivial to affirm that mankind has always relied on faces and voices to distinguish between familiar and unfamiliar individuals. However, if written accounts of biometrics are needed to establish the birth of this field, one must turn to Portuguese writer João de Barros, who, in the 14th century, reported its first known application. According to his writings Chinese merchants stamped children's palm and foot prints on paper for identification purposes (Proença, 2006; Bala, 2008).

In 1869, the United Kingdom parliament enacted the Habitual Criminals Act, making it mandatory for a record of convicted felons to be kept alongside evidences of their identity (Jain et al., 2016). As a follow-up to this trigger, the first scientific and systematic method for human identification was created by the French anthropologist Alphonse Bertillon, who introduced the use of a fixed number of physical measurements to identify usual criminals (Proença, 2006; Jain and Kumar, 2010). This method consisted on several body measurements such as height, arm length, length and breadth of the head, length of fingers and length of forearms for person identification (Angle et al., 2005). As it can be deduced by the significant number of features, as well as their schematization in Figure 2.1, this was a very time consuming process, taking up to twenty minutes per person. This fact, combined with the introduction of human fingerprints early in the 1900s, turned the Bertillonage obsolete, eventually leading to its demise.

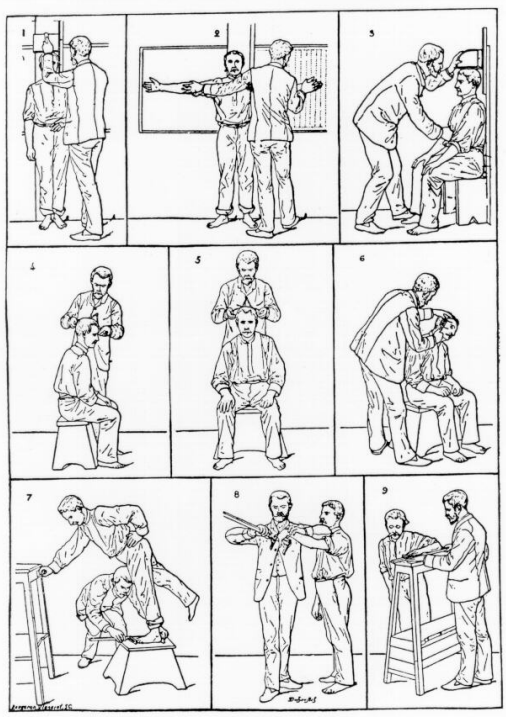


Figure 2.1: Schematization of some of the measurements performed as part of the Bertillon system (Proença, 2006).

In 1880, Henry Faulds and William James published an article in Nature describing the unique nature of fingerprints (Faulds, 1880). In 1891, the Argentinian police initiated the process of fingerprint acquisition of criminals, resulting in the first use of fingerprints as court evidence in 1892 (Jain et al., 2016). Soon after, Sir Francis Galton developed the first elementary fingerprint recognition system, soon improved by Sir Edward Henry who, for criminal identification purposes, established the Fingerprint Bureau in Calcutta in 1897 (Rajkumar et al., 2011). Some of the characteristic patterns considered in Henry's system are depicted in Figure 2.2. The success of Henry's method quickly disseminated throughout the world and led to the inauguration of the first fingerprint system in 1903 in the New York State Prison (Bouchrika, 2016). From that point onwards, fingerprint recognition grew on to become the standard biometric security system worldwide. Nowadays, virtually all law enforcement agencies make use of the Automatic Fingerprint Identification System (AFIS) (Jain and Kumar, 2010).



Figure 2.2: Example of pattern used in Sir Edward Henry's fingerprint classification system (International Biometrics Group, 2003).

With the constant evolution of technology, the role of fingerprints as the gold standard in biometric security for almost a century soon became less evident. With growing concerns about terrorist activities, security breaches and financial fraud, an increasing number of private and governmental companies, with either military or civil purposes, have been investing a considerable amount of human and financial resources in the attempt of developing biometric systems based on other physiological and behavioral human characteristics such as face, iris, palm print or signature. In recent years, biometric security began to find its way into a considerably heterogeneous group of applications, such as border crossing or visits to Walt Disney Parks, as it grew to become a mature technology with a wide margin of progression, both in terms of research as well as applications, in the years to come (Jain and Kumar, 2010; Proença, 2006). More recently, fully automated large-scale fingerprint recognition systems such as the US-VISIT program by the Department of Homeland Security, FBI's NGI program or India's Aashaar project are starting to spread worldwide (Jain et al., 2016). In the next section we will outline some examples of such applications already found in use throughout the world nowadays, as well as some open challenges that are being approached from a biometric recognition point-of-view.

2.2 Applications of Biometric Recognition

The use of biometric recognition systems for personal identification has been widespread to a multiplicity of environments. As of September 2015 the costumers of the Bank of America were able to use the fingerprint scanner on their mobile phones to manage logins to their account management app (Bank of America, 2015). In their announcement regarding this new feature the Bank of America states that "This is another example of how we are helping our mobile and online banking users simplify their financial lives.", further enforcing the notion of biometrics as an alternative to achieve highly intuitive automated identification. Besides the Bank of America, multiple other banking agencies have turned their focus towards biometrics as an alternative for costumer verification:

- Barclays through their Hitachi Vein ID fingerpint vein pattern software (Miyatake et al., 2015);
- The Royal Bank of Scotland introduced the technology for the first time for iPhone costumers in the UK (Thomson Reuters Canada, 2016);
- Corporate users of the Wells Fargo: Commercial Electronic Office (CEO) app were introduced to a multimodal face and fingerprint recognition system. The representative of the bank stated that "passwords simply don't cut it anymore for business customers." (Mendoza, 2015);
- Citibank received two Gartner Financial Services Cool Business Awards in 2015 for its voice biometrics authentication project. Using speaker recognition, customers are automatically

identified as they present an issue through the customer service over the phone (Citi Group, 2015).

- In South Africa, multiple banks have made use of the South African Home Affairs National Identification System (HANIS), launched in 2011, to verify customers through a second layer of biometric information, to ensure that customers are who they claim to be according to their identification documents. Capitec, Absa and FNB are some of the first banks to join this service, allowing them to verify or authenticate customer identity and control identity fraud (KPMG, 2016).

While bank account management seems to be the hot topic regarding real-world applications of biometric security, some other alternatives may already be found. Airport security is one of such examples. This application, in particular, can benefit from the fact that an automated system based on biometric traits would lead to a growing sense of comfort amongst costumers. In a study carried by ExpertFlyer, 79% of the subjects indicated that they would be willing to give up some of their privacy to partake in biometric screening, if such change lead to a cut in the time spent in airport security (Veridium, 2016). Some airports have already carried out preliminary attempts to integrate biometric data in the check-in process of passengers:

- A system developed by Chinese company Shanghai Doudian Technologies has already been implemented in the airport of Shanghai. The system acquires photos of passengers at security checkpoints within the airport, matching their faces to their boarding passes. When passengers board their planes, a final biometric scan in probed against all the previously acquired images, acting as an additional security check (Find Biometrics, 2016).

- Since 2007, Portugal has used the RAPID automatic border control system, developed by Portuguese company Vision-Box, to identify European over-18s holding electronic passports or national ID cards. Recently, Lisbon Airport has opened this automated border control system to a number of countries outside the European Union. RAPID uses electronic passports to store biometric data and facial recognition as the base for biometric matching. A face image of the passenger, obtained at the border, is compared with the one stored on the chip, resulting in a match or non-match decision (Planet Biometrics, 2016a).

- As a result of a drop in international tourism, due to the need of biometric enrollment prior to entering the country, the Department of Home Affairs of the South African government has started to implement automated acquisition and recognition systems in their main airports. The country plans to implement biometric sensors at four ports of entry every month, prioritizing those handling the highest volume of passengers (Planet Biometrics, 2016b).

- The Transportation Security Administration (TSA) of the United States of America has long been criticized for the long security lines at their airport's checkpoints. Recently, in association with the private company Clear, iris and fingerpint scans have been deployed in an attempt to mitigate this problem. Once biometrically cleared, passengers are led around

the line of people waiting to have their IDs checked by the TSA and right to screening. This preliminary step is seen as a means of eliminating the need for an ID, boarding pass or credit card during the check-in process (CBS News, 2016).

Besides the high-security scenarios presented throughout this section, biometric recognition can already be observed in a set of other applications. A recent report by Gartner shows that at least 30% of all companies from around the world will use biometrics on mobile devices by the end of 2016 (Gartner Technology, 2014). In addition, most smartphone manufacturers have started to embed biometric sensors on their devices, with fingerprints, facial recognition and voice biometrics representing the most popular choices (M2SYS, 2016). The simple and routine process of logging into a personal device is starting to shift from the gold-standard alternatives that PINs and passwords represented for years, to new biometric-based alternatives. With mobile technology representing such a vital part of nowadays society, it is estimated that the global biometrics market reaches 30 billion dollars annually in five years time, according to the market analyst firm ABI Research. A 2021 market value of 30 billion dollars represents a 118% increase with relation to the same values in 2015 (BioStore, 2016), and reflects the true expected impact of biometric technology in the short term. In the next section we analyze in more detail the ongoing European projects in the field of biometrics, as well as detailing the international standards regulating the collection, storage and sharing of biometric data.

2.3 Main Standards and European Projects

Due to the expected social impact of biometrics in the near future, referred in the previous section, the number of funded large-scale projects on the field has witnessed an increase in recent years. Within the Horizon 2020 EU Research and Innovation programme (European Comission, 2016a), a number of recently approved and ongoing projects can be found in relation to the topic of biometric recognition:

1. **PROTECT - Pervasive and UseR Focused BiomeTrics BordEr ProjeCT**: The goal of the PROTECT project (European Comission, 2016b) is the development of an enhanced biometric identification system to work robustly across a range of border crossings, using a set of user-centric features. The system will be deployed in Automated Border Control areas, supporting guards in providing smooth and non-intrusive crossing by travellers. The ability of the system to efficiently process low-risk travellers, combined with increased levels of accuracy, security and privacy standards, enabling the border guards to focus on higher-risk travellers, are the central motivations of the project. A multimodal enrollment and verification system is envisaged, taking into account current and next-generation e-Passports, mobile devices and person identification “on-the-move”. Research will be undertaken into optimization of currently deployed biometric modalities, application of emerging biometrics (including contactless finger vein, speaker recognition and anthropometrics), multi-modal biometrics and counter-spoofing.

- Start/End Date: 2016-2019
- Collaborators: University of Reading; ITTI Sp.zo.o.; Veridos GMBH; EURECOM; Intrepid Minds LTD; Wojskowa Akademia Techniczna im Jaroslawa Dabrowskiego; Université de Namur ASBL; Paris-Lodron-Universitat Salzburg; Home Office; Komenda Glowna Strazy Granicznej

2. **ARIES - reliAble euRopean Identity EcoSystem**: The main goal of ARIES (European Comission, 2016c) is to deliver a comprehensive framework for a reliable e-identity ecosystem comprising new technologies, processes and security features. The development of such framework is based on trustworthy security documents and biometrics for highly secure physical and virtual identity management, with the specific aim of tangibly achieving a reduction in levels of identity theft, fraud and associated crimes. IDs will be unequivocally linked to citizens' biometric features, increasing the level of identity assurance during the enrollment and authentication steps. High-security data protection standards will be followed to provide digital privacy-preserving features. The final outcome will be assessed in two use cases: secure eCommerce and identity virtualization for secure travel.

   - Start/End Date: 2016-2019
   - Collaborators: ATOS Spain SA; SAHER UK LTD; Service Public Federal Interieur; SONAE Center Serviços II S.A.; GEMALTO SRO; MORPHO; Police and Crime Comissioner For West Yorkshire; Gobierno Vasco - Departamento Seguridad; Universidad de Murcia

3. **OCTAVE - Objective Control for TAlker VErification**: OCTAVE's (European Comission, 2016d) goal is the integration of automatic speaker verification (ASV) systems with the latest environmental robustness and anti-spoofing technologies, to deliver a scalable and trustworthy biometric authentication service. While simultaneously relieving users from the inconvenience of dealing with text-based passwords, the OCTAVE platform will reduce the economic and practical burdens related to their loss and recovery. The developed system will support single text-dependent, text-prompted and text-independent operating modes, in addition to hybrid multimodal alternatives. The fact that authentication is performed by multiple operating modes of a single system will increase trust and privacy, avoid singular points of failure, as well as allow for rapid breach notification and remediation. Solutions will be installed in data-sensitive services and validated in two real commercial trials: banking services and physical access within an airport infrastructure. Flexibility will support wider exploitation in future applications such as customer care, telephone banking, e-commerce, logical and physical access control.

   - Start/End Date: 2016-2019
   - Collaborators: Fondazione Ugo Bordoni; Advalia SRL; Aalborg Universitet; APLCOMP OY; EURECOM; Società per Azioni Esercizi Aeroportuali S.E.A; ATOS Spain SA;

Validsoft UK Ltd; University of Eastern Finland: Itä-Suomen Yliopisto; Research and Education Laboratory in Information Technologies; The University of Hertfordshire Higher Education Corporation; Findomestic Banca S.p.A

4. **SpeechXRays - Multi-channel biometrics combining acoustic and machine vision analysis of speech, lip movement and face**: The SpeechXRays project (European Comission, 2016e) will develop and test in real-life environments, a user recognition platform based on voice acoustics analysis and audio-visual identity verification. The main goal associated with this project is to outperform available systems with regards to a set of points:

   - *Accuracy and Security*: achieve cross-over accuracies of $1/100$ i.e. twice the commercial voice/face solutions;
   - *Privacy*: allow for biometric data to be stored in mobile devices or in a private cloud under responsibility of the subject;
   - *Usability*: perform text-independent speaker identification with low sensitivity to surrounding noise;
   - *Cost-efficiency*: use of standard embedded microphone and cameras in mobile devices such as smartphones and laptops.

   The final output of the SpeechXRays project will be to provide a solution combining the convenience and cost-effectiveness of voice biometrics, achieving better accuracies by combining it with video, while also introducing anti-spoofing capabilities. The technology will be deployed on a set of 2000 users in 3 use-cases: workforce, eHealth and private consumers.

   - Start/End Date: 2015-2018
   - Collaborators: Oberthur Technologies; Horowitz Biometric Ltd; Realeyes OU; Foundation for Research and Technology Hellas; Elliniki Etairia Tilepikoinonion Kai Tilematikon Efarmogon AE; Tech Inspire Ltd; Institut Mines-Telecom; University College London; Siveco Romania SA; Institutul National de Cercetare - Dezvoltare Pentry Fizica si Inginierie Nucleara “Horia Hulubei”

All large-scale projects presented throughout the last topics are regulated by a set of international standards that, in the case of biometrics, mainly concern to the collection, storage and sharing of biometric data (Science and Technology Council, 2016). Going into more detail, the main areas of standardization in the field of biometric recognition can be defined as:

- **Technical Interfaces:** specify interfaces and interactions between biometric components and sub-systems, including the possible use of security mechanisms to protect stored data and data transferred between systems. Examples include ANSI INCITS 358-2002 BioAPI Specification v1.1 and ANSI INCITS 398-2005 [NISTIR 6529-A] Common Biometric Exchange File Format (CBEFF).

- **Data Interchange Formats**: specify the content, meaning, and representation of formats for the interchange of biometric data, as well as the notation and transfer formats that provide platform independence. Examples include ANSI INCITS 377-2004 Finger Pattern Based Interchange Format, ANSI INCITS 378-2004 Finger Minutiae Format for Data Interchange, and ANSI INCITS 379-2004 Iris Image Interchange Format.

- **Application Profile Standards**: specify one or more standardized profiles, and, where applicable, the identification of classes, subsets, options, and parameters necessary to accomplish a specific function or application. Examples include ANSI INCITS 383-2003 Biometrics-Based Verification and Identification of Transportation Workers, and ANSI INCITS 394-2004 Data Interchange and Data Integrity of Biometric-Based Personal Identification for Border Management.

- **Performance Assessment and Reporting**: specify biometric performance metric definitions and calculations, approaches to test performance, and requirements for reporting the results of such tests. Examples include ANSI INCITS 409.1-2005 Biometric Performance Testing and Reporting Part 1 - Principles Framework; ANSI INCITS 409.2-2005 Biometric Performance Testing and Reporting Part 2 - Technology Testing Methodology; and ANSI INCITS 409.3-2005 Biometric Performance Testing and Reporting Part 3 - Scenario Testing Methodologies.

We proceed this chapter with an analysis on how a typical biometric system is organized, in terms of functioning blocks, and how the identification process is carried out.

2.4 System Architecture

A biometric system can be viewed as a specific pattern recognition system that establishes the authenticity of a user based on characteristic physiological or behavioral traits. All biometrics systems, regardless of the chosen trait to serve as input, follow a well defined architecture, as schematized in Figure 2.3. Two main blocks compose the typical biometric recognition system: enrollment and identification. Enrollment concerns the acquisition of data from known individuals and its registration to a database of template data, to be queried during the identification process. Identification consists in the creation of a specific biometric signature extracted from acquired biometric data of an unknown source. This signature is compared to the previously enrolled templates, yielding, for each one, a similarity value. One assumes that two signatures come from the same person if this similarity value exceeds a specific threshold (Jain et al., 2006a; Proença, 2006).

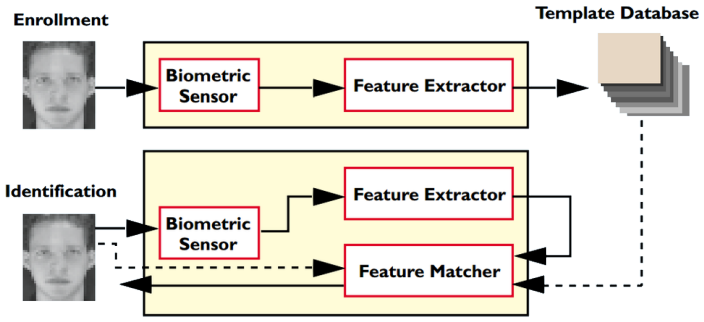


Figure 2.3: Traditional framework of a biometric system. Adapted from (Jain et al., 2000).

2.5 Operating Modes

Depending on the application, a biometric system may operate either in verification or identification mode. The generic structure of both operating modes are depicted in Figure 2.4. Verification, also known as positive recognition, assesses the validity of an identity claim by performing a comparison between the captured data and the enrolled biometric template of the claimed identity. The decision process is then reduced to the evaluation of the probability of whether the tested individual is who he/she claims to be. In an identification system, also known as negative recognition, the biometric signature of the tested individual is compared to the entire database of biometric templates, with the goal of discerning the subject's identity amongst a set of possibilities. In this mode it is usual to present a list of the k most probable identities for the tested individual. Both operating modes present a common limitation, as both work under the assumption that the analyzed subject went through the enrollment process before being exposed to the identification process, and is, thus, present in the biometric template database (Jain et al., 2000; Proença, 2006). Nevertheless, any system can be altered for the detection of non-enrolled individuals, either promoting their enrollment or classifying them as non-authorized individuals (Alice, 2003). Systems that have the option of a reject response are known as open-set, whereas systems forced to output an identity are referred to as closed-set (Jain et al., 2016).

On the basis of every biometric system is a single or a combination of multiple biometric traits. In the next section we define the main requisites that should rule this choice, as well as the most commonly observed alternatives in literature.

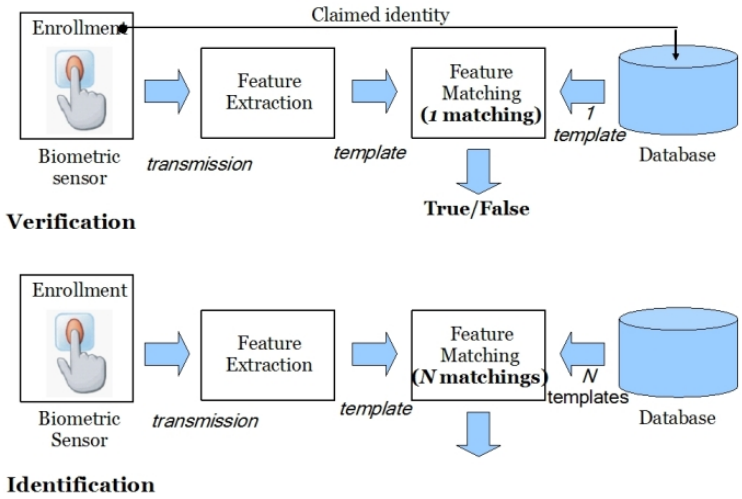


Figure 2.4: Operating modes of a biometric system: verification (top) and identification (bottom) (Griaulle Biometrics, 2014).

2.6 Biometric Traits

As referred before, a biometric trait may be defined as any anatomical of behavioural trait that can be used to differentiate between individuals. Biometric recognition works, thus, by identifying patterns within these biological traits, unique to each individual, to achieve reliable recognition. Amongst anatomical traits the face, the patterns of the iris or the shape hands may be considered. On the other hand, behavioral traits, such as the voice or the signature, relate to the conduct of a person and may not be directly related to any anatomical structure. Biometric traits can also be labeled as genotypic or phenotypic. Genotypic features are genetically defined from embryogenesis, whereas phenotypic features can be changed over time and depend on the surrounding environment (Proença, 2006). Figure 2.5 depicts some examples of the most commonly studied biometric traits.

While the potential of biometric recognition is well known and accepted, the growing need for reliability and robustness raises some expectations when translating this potential to automated systems. When developing a new system based on a specific trait, some conditions need to be met regarding the chosen trait:

1. **Universality**: every person must possess their specific variation of the trait;
2. **Uniqueness**: no two persons should share the same specific variation of the trait;
3. **Permanence**: the trait should neither change nor be alterable;
4. **Collectability**: the trait must be readily accessible to a sensor and easily acquirable.

Table 2.1 summarizes how each of the most commonly studied biometric traits ranks regarding the aforementioned characteristics. Depending on the focus of the designed application, one could

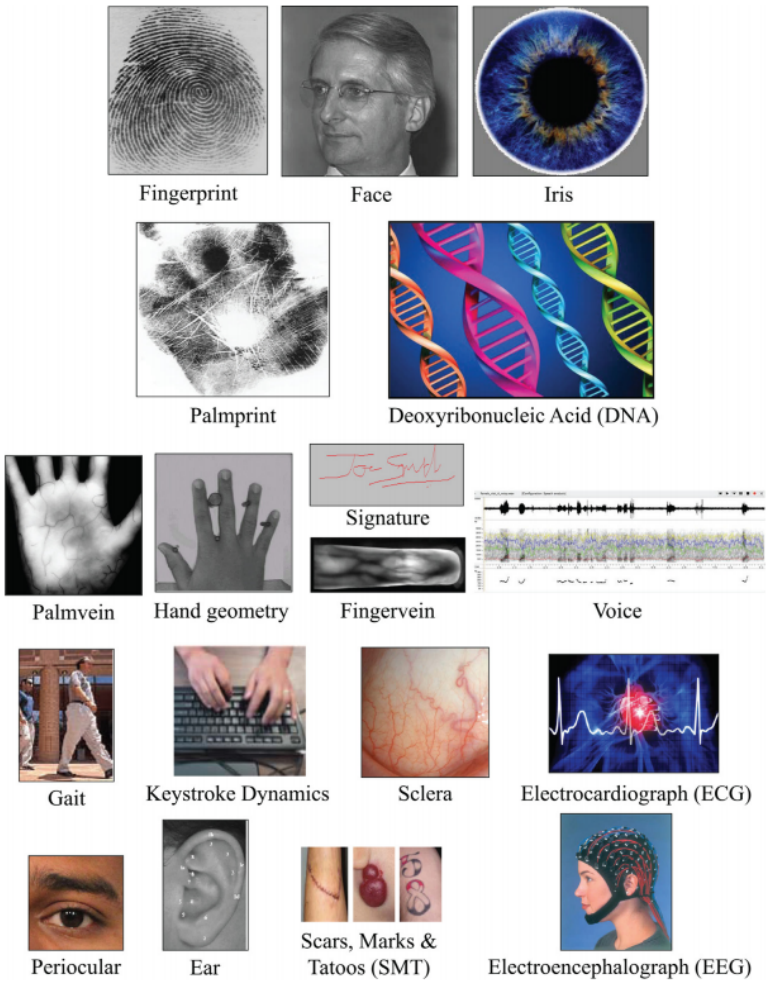


Figure 2.5: Some examples of commonly researched biometric traits (Jain et al., 2016).

restrain or constrain some of these conditions. A careful analysis of the advantages and disadvantages laid out in Table 2.1 leads to a series of conclusions regarding the choice of biometric traits for system development:

1. There is no "gold-standard" biometric trait, i.e. the choice of the best biometric trait will always be conditioned by the means at our disposal and the specific application of the recognition process;

2. Some biometric traits seem to present advantages that counterbalance other trait's disadvantages. For example, while voice's permanence is highly variable, due to external factors, the iris patterns represent a much more stable and hard to modify trait. However, iris acquisition in conditions that allow accurate recognition requires specialized NIR illumination and user cooperation, while voice only requires a standard sound recorder and even no need for direct cooperation of the individual.

Table 2.1: Comparative data analysis of some common biometric traits (adapted from Jain et al. (2000) and Proença (2006)).

	Requirements			
Traits	**Universality**	**Uniqueness**	**Collectability**	**Permanence**
DNA	High	High	Low	High
Ear	Medium	Medium	Medium	High
Face	High	Low	High	Medium
Facial Thermogram	High	High	High	Low
Hand Veins	Medium	High	High	Medium
Fingerprint	Medium	High	High	Medium
Gait	Low	Low	High	Low
Hand Geometry	Medium	Medium	High	Medium
Iris	High	High	Medium	High
Palm Print	Medium	High	Medium	High
Retina	High	High	Low	Medium
Signature	Medium	Low	High	Low
Voice	Medium	Low	Medium	Low

The observations presented above may be stated in an alternative way: even though there is no "best" biometric trait *per se*, marked advantages might be found by exploring the synergistic effect of multiple statistically independent biometric traits, so that each other's pros and cons counterbalance resulting in an improved performance over each other's individual accuracy. Biometric systems that include multiple sources of information for establishing an identity are known as multimodal biometric systems (Ross and Jain, 2004). The next section will enter into further detail regarding the main topics in the development of multimodal biometric systems.

2.7 Multimodal Biometrics

Fusion of multiple sources of information to improve biometric recognition was first proposed by Brunelli in 1995 (Brunelli and Falavigna, 1995). In this pioneer work, the authors proposed an approach based on the combination of face and voice features, setting most of the research topics for multimodal biometrics in the following years, namely normalization and fusion techniques. It is generally regarded, in many reference works of the area, that multimodal biometric systems might help cope with a variety of generic problems all unimodal systems generally stumble upon, regardless of their intrinsic pros and cons. These problems can be classified as:

1. **Noisy data**: when external factors corrupt the original information of a biometric trait. A fingerprint with a scar and a voice altered by a cold are examples of noisy inputs. Improperly maintained sensors and unconstrained ambient conditions also account for some sources of noisy data. As an unimodal system is tuned to detect and recognize specific features in the original data, the addition of stochastic noise will boost the probabilities of false identifications (Jain and Ross, 2004).

2. **Intra-class variations**: when biometric data acquired from an individual during authentication is significantly different from the data used to generate the template during enrolment (Jain and Ross, 2004). This may be observed when a user incorrectly interacts with a sensor (e.g. variable facial pose) or when a different sensor is used in two identification approaches (Ross and Jain, 2004).

3. **Inter-class similarities**: when a database is built on a large pool of users, the probability of different users presenting similarities in the feature space of the chosen trait naturally increases (Ross and Jain, 2004). It can, therefore, be considered that every biometric trait presents an asymptotic behaviour towards a theoretical upper bound in terms of its discrimination, for a growing number of users enrolled in a database (Jain and Ross, 2004).

4. **Non-universality**: when the biometric system fails to acquire meaningful biometric data from the user, in a process known as failure to enrol (FTE) (Jain and Ross, 2004).

5. **Spoof attacks**: when an impostor attempts to spoof the biometric trait of a legitimately enrolled user in order to circumvent the system (Jain et al., 2011).

It is intuitive to note that taking advantage of the evidence obtained from multiple sources of information will result in an improved capability of tackling some of the aforementioned problems. These sources might be more than just a set of distinct biometric traits. Other options, such as multiple sensors, multiple instances, multiple snapshots or multiple feature space representations of the same biometric are also valid options (Jain and Ross, 2004), as depicted on Figure 2.6.

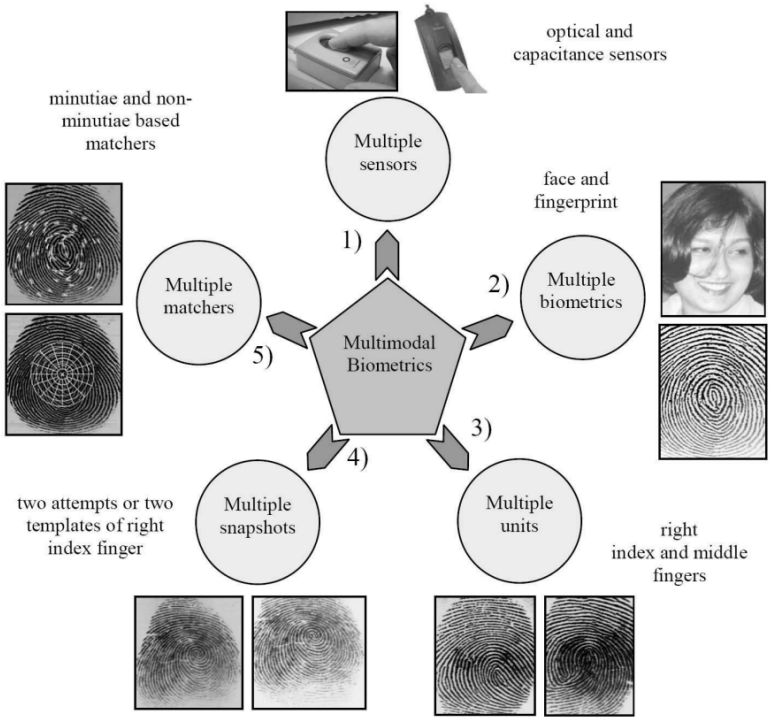


Figure 2.6: Scenarios in a multimodal biometric system (Ross and Jain, 2004).

The first consideration for multimodal system design concerns the choice of the block where integration of the multiple sources of information should be performed. For example, a face recognition system may either acquire multiple snapshots at different angles and try to reconstruct a 3D model of the face or work individually on each snapshot and perform a combined decision based on individual systems (Kaur and Kaur, 2013). The level at which fusion is performed has been extensively addressed in literature and a thorough summary of the most common strategies to fusion was compiled by Jain et al. (Jain et al., 2005) and is depicted in Figure 2.7. Post-classification fusion represents the great majority of the literature in multimodal biometrics. Even though four major classes of post-classification classifiers are presented in Figure 2.7, a simpler classification can be obtained based on the building blocks of a typical system architecture presented in Section 2.4.

In this way, fusion may be performed either by fusing individual matching scores and performing a single decision or by combining decisions of individual matchers (Gudavalli et al., 2012). Combination of decisions represents the simplest approach with trivial methods like majority voting being commonly used. However it has also been pointed that at decision level almost all of the initial raw biometric data has already been lost (Jain et al., 2005). In this alternative, the combined decision is performed more on an abstract entity (the individual decisions) than on the biometric data that formed the basis of the system, representing a more rigid and less reliable process. One could easily deduce that fusion at the sensor-level would be the most reliable in terms of maintaining the original features of the acquired biometric trait. However the process of performing fusion from two biometric traits that might not share any intrinsic similarity represents a non-trivial process that could prove more complex to implement for the benefit that would be observed in the resulting performance. For these reasons match score level fusion appears as the ideal alternative: it is as easy to get access to as decision-based approaches and gives meaningful information as to the similarity of the tested trait and the templates to which it is compared.

The theoretical framework that rules match score combination in multimodal biometrics was developed by Kittler et al. (Kittler et al., 1998) and works under the assumption that each individual unimodal system's matching score can be thought of as a posterior probability of a given user's identity claim being genuine. From this interpretation, given an input biometric signal Z, the process of decision can be viewed as a maximum *a posteriori* classification of Z to a set of m possibles classes (identities). Let $\vec{x_i}$ be a feature vector describing Z and processed by the i-th matching module (classifier) of the multimodal system. Each classifier outputs a matching score, also known as confidence level, that may be understood as the posterior probability of Z belonging to class ω_j, given input data x_i, i.e. $P(\omega_j|\vec{x_i})$. Considering this theoretical framework and assuming statistical independence between different biometric traits, a few combination rules can be defined:

1. **Product rule**: this rule is based on the assumption of statistical independence of the representations $\vec{x_1}, \vec{x_2}, ..., \vec{x_I}$. The input pattern is assigned to class c such that:

$$c = \arg\max_j \prod_{i=1}^{I} P(w_j|\vec{x_i}) \tag{2.1}$$

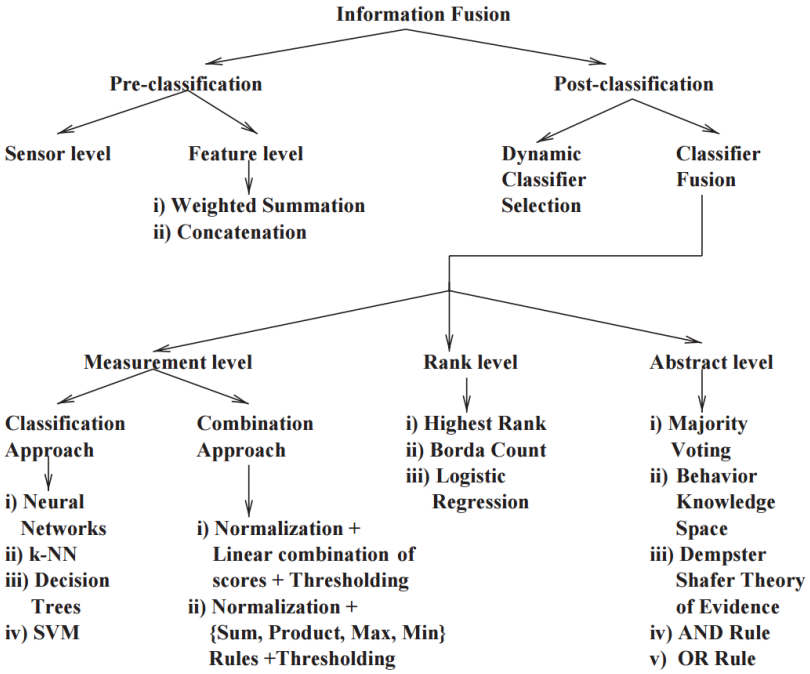


Figure 2.7: Approaches to information fusion in mutimodal biometrics (Jain et al., 2005).

2. **Sum rule**: apart from the assumption of statistical independence of the multiple representations used in the product rule, the sum rule also assumes that the posterior probabilities computed by the individual classifiers do not deviate much from their respective priors. This rule is applicable when there is a high level of noise leading to ambiguity in the classification problem. The sum rule assigns the input pattern to class c such that:

$$c = \arg\max_{j} \sum_{i=1}^{I} P(w_j|\vec{x}_i) \tag{2.2}$$

With system performance as an indispensable requirement for the development of biometric systems, a prominent question concerns how to quantitatively and objectively present this information. The next section will focus on common evaluation metrics, used to standardize performance assessment and facilitate comparison between methods.

2.8 Evaluation Metrics

A number of issues need to also be considered from a system design point-of-view (Jain et al., 2000):

1. **Performance**: a system's accuracy, speed, robustness, as well as its resource requirements, and operational or environmental factors that affect its accuracy and speed;

2. **Cost**: the total cost of the system, ruled by a trade-off between performance and the return cash-flow associated with its implementation.

3. **Interoperability**: the ability of the system to operate without the assumption that the same sensor, algorithms or operating conditions will be available during the entire length of its lifetime.

4. **Acceptability**: the extent to which people are willing to accept the integration of a particular biometric identifier in their daily lives;

5. **Circumvention**: how easy it is to fool the system through fraudulent methods.

Considering a known set of conditions, under which an identification process is being carried out by a biometric system, there are two classes of errors which can be observed regarding its decision: false matches, where a match is declared between two distinct biometric signatures, and a false non-match, where the system is unable to match two biometric signatures from the same subject (Jain and Kumar, 2010). Two quantitative measurements can be defined to assess the rate of false matches and false non-matches: the false match (or false accept) rate (FMR/FAR) and false non-match (or false reject) rate (FNR/FRR), respectively. The amount of false accepts and false rejects, as well as their true counterparts, can be summarized in a confusion matrix, as schematized in Table 2.2. From such values, the computation of both aforementioned error rates is trivial. Two other commonly used terms are specificity and sensitivity. Specificity corresponds to the ratio of authorized individuals that are correctly identified as such, whereas sensitivity measures the analogous metric for unauthorized individuals (Davis and Goadrich, 2006).

Table 2.2: Type of errors observed in a typical biometric system: false matches or false genuines (FG) and false non-matches or false impostors (FI). The computation of the FAR and FRR can be easily performed based on such values.

		True State		
		Genuine	**Impostor**	
Decision	**Genuine**	TG	FI	FRR = FI / (FI + TG) = 1 - sensitivity
	Impostor	FG	TI	FAR = FG / (FG + TI) = 1 - specificity

The FAR/FRR point at which we define a system to function is controlled by a decision threshold, T, the probability above which two biometric signatures are considered to belong to the same individual. For low T values, a low similarity between biometric signatures and templates will be enough to trigger an acceptance. The result is that the number of false match cases will increase, while the number of false rejects diminishes. The same rationale can be followed to understand the cases where T is set to a higher value. In this case the number of matches will be reduced to only the most similar of all cases, resulting in fewer falses matches, but also increasing the amount of false rejects. In an ideal situation the value of T should be set to achieve and optimal trade-off between false accept and false reject error rates (Proença, 2006; Jain et al., 2006b). Figure 2.8 summarizes the previous observations by plotting the trend of both FAR and FRR for growing values of T.

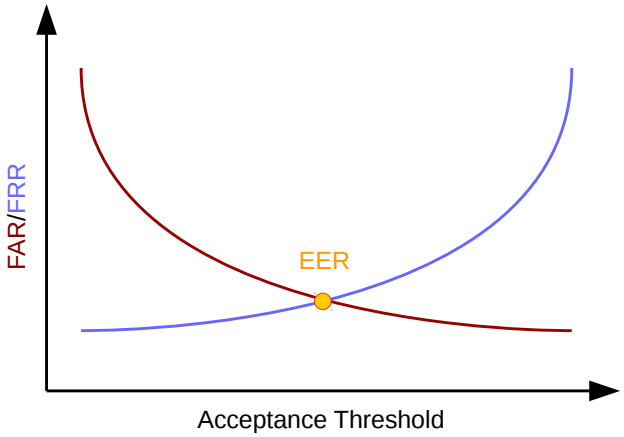


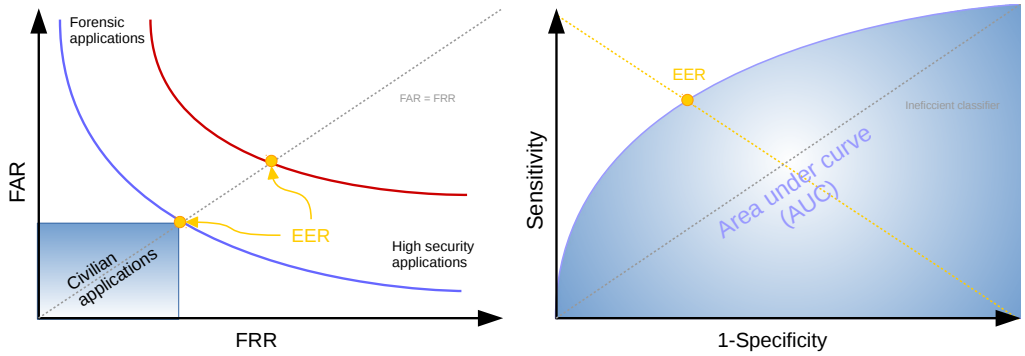
Figure 2.8: Evolution of the FAR and FRR error rates for growing decision threshold values.

It is readily observed the both the FAR and FRR curves intersect in a specific point, i.e. there is a single T value for which FAR and FRR have the same value. This point is commonly denominated as the equal error rate (ERR) and is a very commonly observed value for performance assessment of biometric systems (Jain et al., 2006b). Some biometric systems might work around this definition if the constraints under which they are designed are less prone to either false accepts of false rejects. For example, in personal bank account management, a biometric system should be designed to more heavily limit falsely accepted users rather than trying to reduce the number of falsely rejected non-users. While the latter might result in a tedious experience for the user of the biometric system, the alternative of allowing access to a non-authorized user is not acceptable given the nature of the application. On the other hand, the application of biometric systems for forensic applications will be more focused on the non-rejection of the true subject of interest, allowing more freedom in falsely accepting other subjects. With these two examples in mind, it is also common to analyze the FAR for a fixed FRR value or vice-versa, depending on the relative importance of both measures in the design of the system (Precise Biometrics, 2015).

Another typical technique for visualization of the FAR/FRR relation, with regards to the decision threshold value, is the receiver operating characteristic (ROC) curve. This representation is obtained by plotting the values of FAR against the corresponding values of FRR for each possible value of T, as exemplified in Figure 2.9a, or alternatively, the specificity against $1-$sensitivity (Figure 2.9b). The extreme examples referred in the last paragraph, regarding applications of diverse nature, were also considered in this representation. When comparing two systems by the analysis of their ROC curves, it is common to consider the area under the curve (AUC) to quantify how close the ROC curve is to the ideal behaviour of a biometric system[1] (Proença, 2006; Ooms et al., 2010).

FAR and FRR-based measurements are the most common indicators of performance for biometric systems meant to work in verification mode. When a system is designed for identification

[1] In the ideal case, the ROC curve would correspond to the axis defined by $FAR = 0$ and $FRR = 0$, i.e. $Sensitivity = 1$ and $1 - Specificity = 0$, maintaining the EER value at 0.



(a) (b)

Figure 2.9: Examples of receiver operating characteristic (ROC) curves for performance assessment of biometric recognition systems.

alternative measurements need to be considered. In these cases the most traditional metric is the rank-n identification ratio, which corresponds to the ratio of individuals whose true identity is present among the n most likely identities. From this generic description a visual depiction can be easily extrapolated, under the form of cumulative match curves (CMC), where the rank-n values are computed for all the possible values of $n \in [1...N]$, where N is the total number of enrolled individuals in the probed database. A generic example of a CMC is depicted in Figure 2.10. The rank-1 recognition rate is commonly used as a more compact indicator of performance in identification systems, as it simply quantifies the true identification ratio of a system. Intermediate metrics such as the rank-5 or rank-10 rates are also found in literature (Du et al., 2005; Layne et al., 2012), and might be considered to analyze the capability of a biometric recognition system to cluster the most likely identities amongst the highest ranked probabilities.

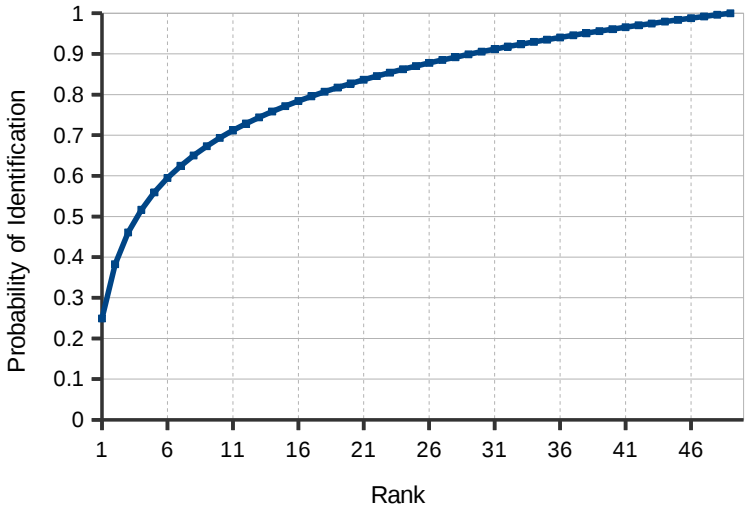


Figure 2.10: Generic example of a cumulative match characteristic (CMC) curve for biometric identification system performance assessment.

2.9 Conclusion

With all the considerations presented throughout this chapter, an overview of the computer vision and machine learning tools involved in the research of the scientific contributions of this thesis will be carried out in the next chapter.

Chapter 3

Computer Vision and Machine Learning Tools

Most of the work involved in the scientific contributions of the present thesis concerned image-based biometric traits. Their analysis and classification can be broadly classified as a computer vision and machine learning problem. This chapter serves as an overview to the techniques involved in the research work presented throughout this thesis. Image description by feature extraction will be the focus of the next section, while machine learning techniques for classification are considered afterwards.

3.1 Feature Extraction and Description

The first step of most computer vision applications concerns the extraction of an image content description. This step is essential as the memory consuming and redundant raw image data captured directly from cameras would render the whole process too slow and complex (Kadir and Brady, 2001). Feature extraction techniques focus on achieving reliable representations of image content while reducing the amount of data needed for this goal. Throughout this section some of such techniques will be outlined.

3.1.1 SIFT

The Scale Invariant Feature Transform (SIFT) local feature descriptor was introduced by Lowe (Lowe, 2004). Although many algorithms have been proposed after Lowe's, SIFT still prevails as one of the most stable feature description algorithms, regardless of the image acquisition conditions. Lowe proposed the SIFT descriptor for features detected with the Difference of Gaussians (DoG) operator, conferring the resulting representation an increased robustness to rotation and scale. Although the scale invariance is guaranteed by the stable set of candidate features points across different scales obtained with DoG, the orientation invariance is provided separately, as a method of assigning a repeatable orientation to the feature points is required. To that end, for each DoG

feature point, also known as keypoint, detected at location (x,y) and scale s, the gradient magnitude m and orientation θ are computed, as described by the following equations:

$$M(x,y) = \sqrt{I_x(x,y)^2 + I_y(x,y)^2} \tag{3.1}$$

$$\Theta(x,y) = tan^{-1}(I_y(x,y)/I_x(x,y)) \tag{3.2}$$

SIFT's final step is the concatenation of sub-regional histograms of gradients around each keypoint, resulting in a 128-dimensional descriptor that summarizes local gradient information. This descriptor is built by first computing a set of 8-bin orientation histogram in 4×4 sample regions, as presented in the rightmost image of Figure 3.1. All histograms are then concatenated into the $4 \times 4 \times 8 = 128$-dimensional descriptor. As the keypoint dominant orientation is known, a offset can be applied to the histograms in order to secure rotation invariance. Robustness to illumination is improved by normalization of the descriptor vector.

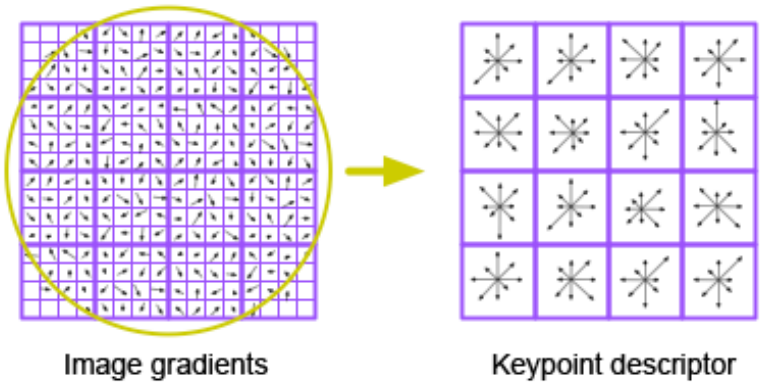


Figure 3.1: Gradient magnitude and orientation computation of a single keypoint using SIFT descriptor. Adapted from (Lowe, 2004).

Recently, an alternative approach to the SIFT descriptor has been proposed, with descriptors computed in fixed dense grids of equally separated keypoints, at a fixed scale value. Experimental evaluations in tasks such as object categorization, texture classification, image alignment, and biometrics, show that better performance is obtained using dense grids rather than the original DoG-based detection of sparse keypoints (Wang et al., 2010).

3.1.2 LBP

The basic local binary pattern (LBP) operator, introduced by Ojala et al. (Ojala et al., 1996), was based on the assumption that local texture may be decomposed in two complementary aspects: a local spatial pattern and its corresponding strength. The original LBP labels the pixels of an image by thresholding the 3×3 neighbourhood of each pixel with respect to its value. The resulting binary vector is then converted to decimal, as depicted in Figure 3.2, to achieve the LBP value for

the center pixel. Concatenation of such values results in the final texture image. A 256-dimension vector can be obtained as the histogram of the texture image.

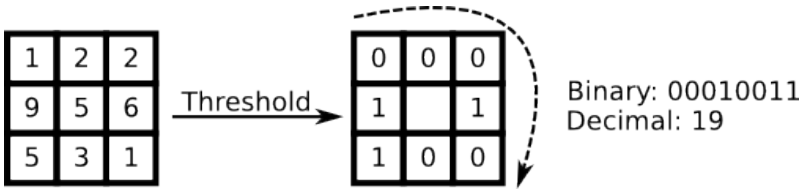


Figure 3.2: Local binary pattern computation for a generic 3×3 neighborhood (adapted from (Wagner, 2011)).

A variation of the traditional LBP, known as the uniform local binary pattern (ULBP), was proposed in (Ojala et al., 2002), capable of achieving a more compact version of the feature vector. Such alternative was motivated by the fact that some binary patterns occur more often in texture images than others. Uniform patterns are those where, at most, two bitwise transitions from 0 to 1, or vice-versa, take place. It was observed that that nearly 90% of all patterns in the 3×3 neighbourhood and around 70% for circular neighbourhoods with radius 2 and 16 sample pixels, are uniform. For example, in a 3×3 neighbourhood there are a total of 256 possible patterns, 58 of which are uniform. By joining all non-uniform patterns in a single LBP label, a total of 59 different labels (58 for uniform patterns plus 1 label for non-uniform pattern) is achieved.

3.1.3 HOG

Dalal and Triggs (Dalal and Triggs, 2005) presented a descriptor defined from a set of gradient orientation histograms, inspired by the discriminatory property of local position-dependent gradient orientation histograms computed over a grid in the image domain. Compared to SIFT, which is a local image descriptor, the Histogram of Oriented Gradients (HOG) descriptor is a regional image descriptor. As the HOG descriptor is not normalized with respect to orientation, it is not rotationally invariant, being, however, normalized with respect to image contrast. Dalal and Triggs showed that this descriptor could be used to achieve robust detection of humans in natural environments. HOG divides the image into small connected regions, which are called cells. For each cell, a histogram of gradient directions, or edge orientations, is computed for the pixels within that cell. Each cell is then discretized into angular bins according to its gradient orientation. The gradient magnitudes of the pixels in the cell are used to vote into the orientation histogram. Adjacent cells are grouped as a spatial region, which is called a block. Grouping the cells into a block is the basis for normalization of the histograms. A normalized group of histograms represents the block histogram, and the set of these block histograms is the called HOG descriptor.

3.1.4 GIST

Based on a series of that demonstrated humans' capability to integrate enough information about the meaning of a scene as quickly and accurately as a single object, GIST features were first

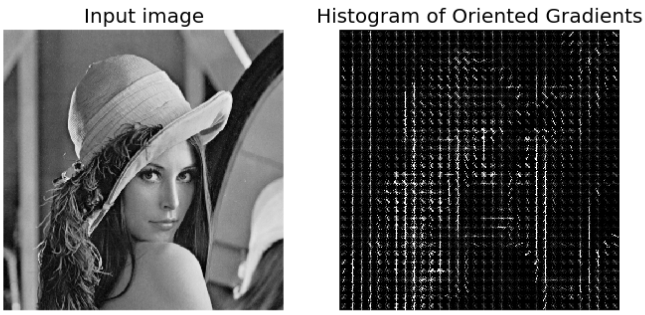


Figure 3.3: Histogram of oriented gradients example (adapted from (Image Processing in Python, 2012)).

proposed as an alternative to describe real world scenes (Oliva and Torralba, 2001). The GIST descriptor encodes 5 perceptual dimensions, using spectral analysis, that are used to design a low dimensional and holistic representation of the image (Bharadwaj et al., 2010):

- **Naturalness**: assesses the proportion of vertical and horizontal orientated edges, associating such observations with low degrees of naturalness;
- **Openness**: assesses the existence of a horizon line and the lack of visual references and boundary elements, conferring to such scenes a high degree of openness;
- **Roughness** assesses the fractal dimension of the scene and thus, its complexity.
- **Expansion** describes the depth of gradients within the image;
- **Ruggedness** accounts the deviation from the horizontal, by assessing contour orientation within the image.

The statistical analysis of the outputs given by the spectral transformation of the image allow the creation of a descriptor encoding all the aforementioned quantities.

3.1.5 Deep Neural Networks

Deep neural networks, like the Convolutional Neural Networks that will be explored in further detail in the next section, have also been used as feature extractors. By making use of the intermediate representations that output from the first layers of the deep network structure, powerful information can be obtained and used in a multiplicity of computer vision applications (Oquab et al., 2014). A more detailed analysis on the functioning of such networks will be carried out in Section 3.2.4. From this point onwards, the content of the following sections will concern the machine learning techniques used both in conjugation with the aforementioned feature extraction methodologies, or independently on raw image data.

3.2 Machine Learning Tools

Machine learning concerns the process of learning patterns within specific data representations, that can be later used to analyze and classify new samples (Bishop, 2006). While its applicability highly exceeds computer vision problems, its popularity and performance in this area is undeniable. In the present section we outline the machine learning tools explored in the development of the scientific contributions presented throughout the remainder of this thesis.

3.2.1 Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of individual Gaussian components:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \cdot g(\mathbf{x}|\mu_i, \Sigma_i) \tag{3.3}$$

where:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{(1/2)}} \cdot \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\} \tag{3.4}$$

By learning the weights, w_i, means, μ_i, and covariances, Σ_i, of each of the M components density functions $g(\mathbf{x}|\mu_i, \Sigma_i)$, the unsupervised learning of a mathematical model describing the training data can be achieved. The parameters that define such model are estimated from training data through the iterative Expectation-Maximization (EM) algorithm. For a sequence of T training vectors $X = \{\mathbf{x_1}, ..., \mathbf{x_T}\}$, the total GMM likelihood, assuming independence between the vectors, can be defined as:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\mathbf{x_t}|\lambda) \tag{3.5}$$

Direct maximization of this likelihood function is not possible due to the non-linearity of the GMM function with regards to the parameters of interest. The basic idea of the EM algorithm is to iterate a new model, $\bar{\lambda}$, from an initial model λ[1], such that, at each iteration, $p(X|\bar{\lambda}) > p(X|\lambda)$, with the new model $\bar{\lambda}$ serving as the initial model for the next iteration. The two steps of each iteration are the namesake of the algorithm: during expectation the projections of the training data points with regards to the current model λ are computed, with the parameters being adapted during the maximization step with respect to the previously computed projection. The model itself adapts iteratively to the training data until a set number of iterations is reached or the observed likelihood reaches a previously defined threshold (Reynolds, 2015). A visual example of how a simple two component GMM evolves and fits a set of two-dimensional data points during the iterations of the EM algorithm is depicted in Figure 3.4.

[1] Typically the initial model is estimated by a clustering technique like k-means (Bradley and Fayyad, 1998).

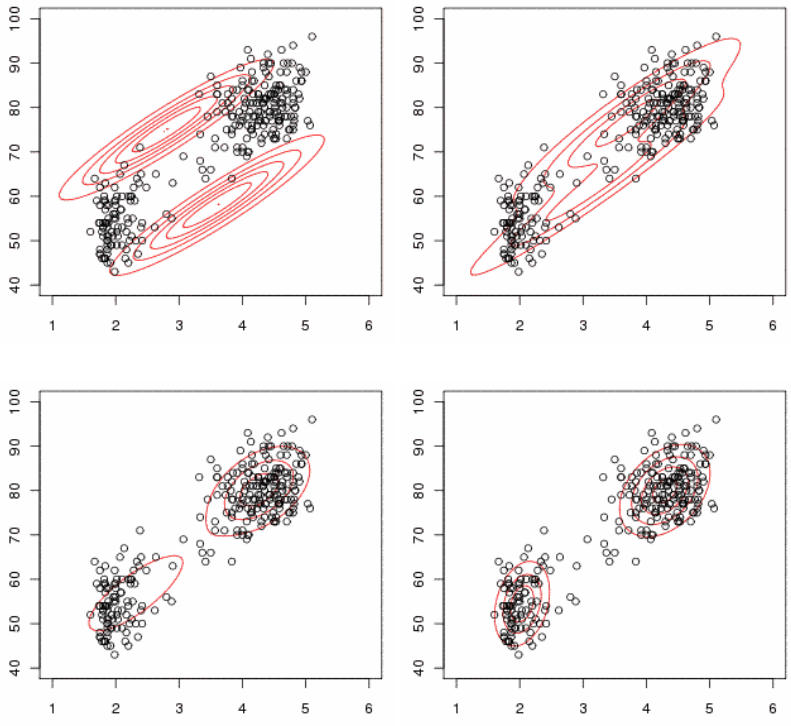


Figure 3.4: Visual example of the evolution of two GMM components for a growing number of iterations (from top-left to bottom-right). Adapted from (Becker, 2009)

3.2.2 Support Vector Machines

A support vector machine (SVM) is a binary classification method that works by optimizing a linear decision surface based on the concept of risk minimization. The boundary between classes is obtained by a weighted combination of specific elements of the training set, called support vectors. The input to a SVM algorithm is a set of feature vectors, $\mathbf{x_i}$, and associated labels, $y_i \in \{-1,1\}$, resulting in an output of N_S support vectors, s_i, coefficient weights, α_i, class labels, y_i for each support vector and a constant term b (Phillips et al., 1998). The linear decision surface can then be formalized as:

$$w \cdot \mathbf{x} + b = 0 \tag{3.6}$$

The optimization problem consists in finding the hyperplane, parametrized by the previous formulation, that optimally separates training data with regards to the two classes. From a geometric interpretation, the optimal hyperplane maximizes the sum of the distances to the closest positive and negative training samples (Jonsson et al., 2002). This sum is referred to as the margin of the separating hyperplane. It can be shown that the optimal hyperplane $w \cdot \mathbf{x} + b = 0$ is obtained by minimizing $1/2||w||^2$ subject to $y_i(w \cdot \mathbf{x_i} + b) >= 1, \forall i$, resulting in a quadratic optimization problem[2]. When the training samples are not linearly separable, slack variables ξ are introduced

[2]the vector represented by w is the normal vector of the hyperplane.

to the constraints, allowing the violation of the separation constraints to a certain degree. In this case the objective function takes the form of $1/2||w||^2 + C \cdot (\sum_i \xi_i)$, under the set of constraints $y_i(w \cdot x_i + b) >= 1 - \xi, \forall i$, where the parameter C tunes the trade-off between minimizing the sum of the slack errors and maximizing the margin (Gutschoven and Verlinde, 2000). A visual representation of the maximum margin and the respective support vector for a simple example is depicted in Figure 3.5.

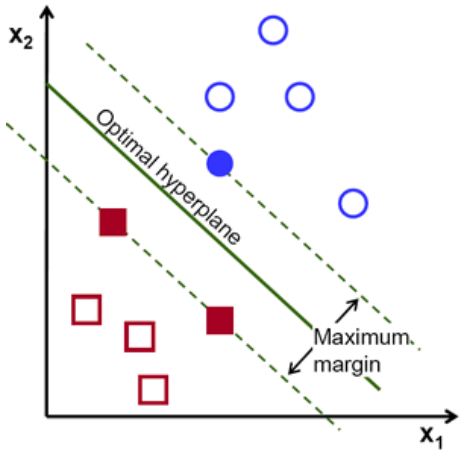


Figure 3.5: Maximum margin classifier example for a simple two-dimensional dataset. Adapted from OpenCV's documentation (OpenCV, 2016).

While the optimization presented above (known as primal problem) can be optimized directly, it is often solved via its Lagrangian dual problem, written in terms of multipliers, α_i, which are constrained so that $\sum_i \alpha_i y_i = 0$ and $0 < \alpha_i < C, \forall i$. Solving the dual has a computational complexity that grows proportional to the size of number of samples in the training set, instead of the feature space dimensionality. Furthermore, in the dual formulation, both data and slack variables become implicitly represented, with data represented by a kernel matrix, K, of all inner products between pairs of data points, that is, $K(\mathbf{x_j}, \mathbf{x_k}) =< \mathbf{x_j}, \mathbf{x_k} >$, and each slack variable associated with a Lagrangian multiplier (Biggio et al., 2014). Following the Lagrangian multipliers method, the dual problem can be formulated as:

$$\max_{\alpha_i} \sum_i \alpha_i - 1/2 \sum_{j,k} \alpha_j \alpha_k y_j y_k K(\mathbf{x_j}, \mathbf{x_k}) \tag{3.7}$$

constrained to $\sum_i \alpha_i y_i = 0$ and $0 < \alpha_i < C, \forall i$. The optimal hyperplane resulting from this optimization is given by:

$$\sum_{i=1}^{N_S} \alpha_i y_i K(\mathbf{s_i}, \mathbf{x}) + b = 0 \tag{3.8}$$

SVMs can be extended to nonlinear decision surfaces by manipulating the kernel function $K(\cdot, \cdot)$ presented in the dual formulation to an alternative that maps the original data vectors to a

higher dimensionality space where a linear separation achieves lower-margin solutions (Gutschoven and Verlinde, 2000). Some of the most commonly observed kernel functions are as follows:

- $K(\mathbf{x_i}, \mathbf{x_j}) = x_i \cdot x_j$: Linear SVM;
- $K(\mathbf{x_i}, \mathbf{x_j}) = (x_i \cdot x_j + 1)^p$: Polynomial SVM of degree p;
- $K(\mathbf{x_i}, \mathbf{x_j}) = \exp\left\{||x_i - x_j||^2/\sigma^2\right\}$: Radial Basis Function (RBF) or Gaussian SVM;

A visual example of the use of kernels to achieve linear separability in higher dimensions is depicted in Figure 3.6.

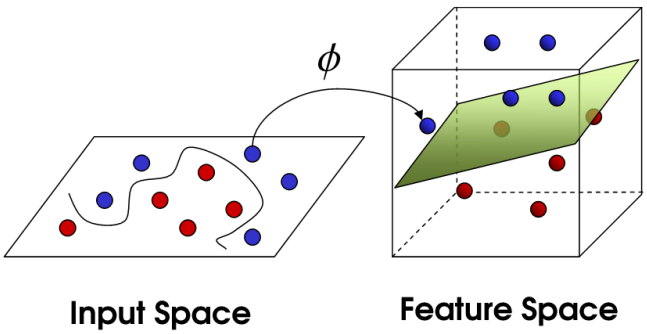


Figure 3.6: Kernel trick for non-linearly separable scenarios (Liao, 2015).

3.2.3 Stacked Denoising Autoencoders

An autoencoder is a simple neural network with one hidden layer designed to reconstruct its own input. The learning process is ruled by a minimization of the average reconstruction error between the original and reconstructed instances. An alternative to the original formulation is the denoising Autoencoder (dA) (Vincent et al., 2008), where corrupted "noisy" versions of the input are used to reconstruct the original instances. Stacking multiple dA's one on top of each other, gives the model an hierarchical nature, with features from lower and higher layers representing distinct degrees of abstraction of the original data (Bengio, 2012). This stacking process results in an architecture commonly known as Stacked Denoising Autoencoders (SDA). Their typical architecture is schematically represented in Figure 3.7.

3.2.4 Convolutional Neural Networks

In this section we give a brief explanation of how Neural Networks work and how the CNN architectures derive from them. CNN layers are then analysed to better understand how they apply to image processing tasks. In standard neural networks, each layer consists of simple neurons that are connected to one or more neurons in the next layer, until the final one, that can either be a classification model or a regression model, depending on the task at hand. For a given neuron in

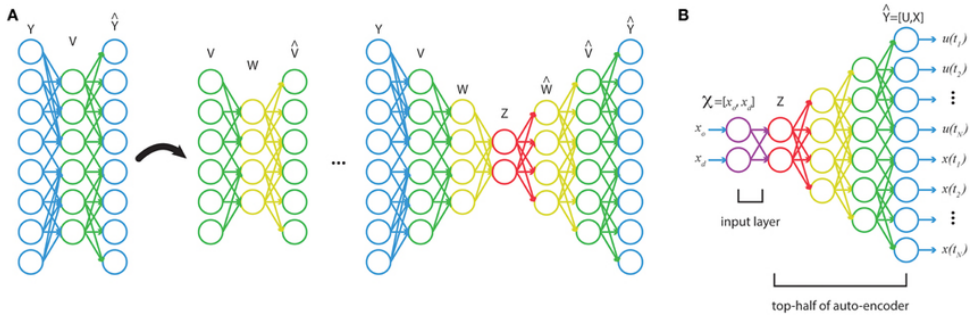


Figure 3.7: Stacked denoising autoencoder architecture. Adapted from (Berniker and Kording, 2015).

a given layer, with N input variables, the inputs to that neuron are first linearly combined in the form

$$a_j = \sum_{i=1}^{N} w_{ij} \cdot x_i + b_j \tag{3.9}$$

where a_j is known as the activation for the jth neuron, w_i is the weight for a specific input x_i and b_j is the bias of that neuron. Each activation is transformed using a nonlinear activation function $h(.)$, producing the output of the neuron

$$z_j = h(a_j) \tag{3.10}$$

The activation function is usually the logistic sigmoid or hyperbolic tangent (*tanh*) functions. The outputs of the neurons are then, again, linearly combined and transformed using an activation function, ultimately giving rise to the network outputs (Bishop, 2006; Mitchell, 1997).

CNN's follow a similar architecture, but are comprised of different types of *layers* that fulfill specific image related tasks. It's the way these layers are organised and how they work that makes CNN's so suitable for image processing problems. Convolutional Neural Networks, in their simplest architecture, are comprised of N pairs of convolutional and pooling layers, followed by one or more fully connected layers, similarly organized as multilayer neural networks, as can be seen in Figure 3.8.

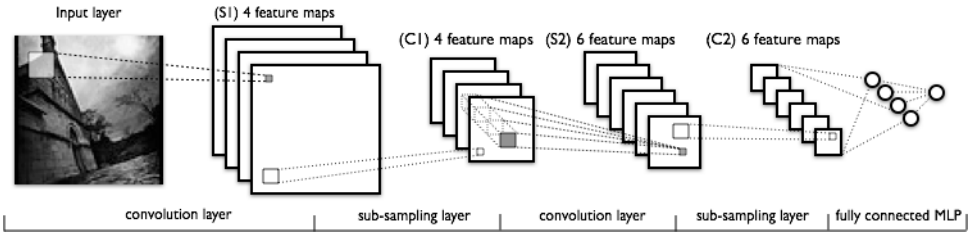


Figure 3.8: Typical convolutional neural network architecture. Adapted from (Deep Learning, 2010).

A detailed analysis regarding the nature and functionality of these types of layers will be carried out in the following sections.

3.2.4.1 Convolutional Layer

In a convolutional neural network, the input is a 2D matrix of values, or pixels when working with images. The first difference to standard neural networks is that all the pixels in the image are not fully connected to every neuron in the first hidden layer. This would lead to thousands of weight and bias parameters that would increase exponentially with larger, high-resolution images, leading to unbearable computational complexity. Instead, the neurons are locally connected to small regions of the input image, called the *local receptive field* for the hidden neuron. This field or window slides across the entire image and each region corresponds to a different neuron in the first hidden layer, like shown in Figure 3.9. If the input image is RGB, the window will slide the image for each color channel, where equal local receptive fields for each channel are combined in a single neuron in the hidden layer.

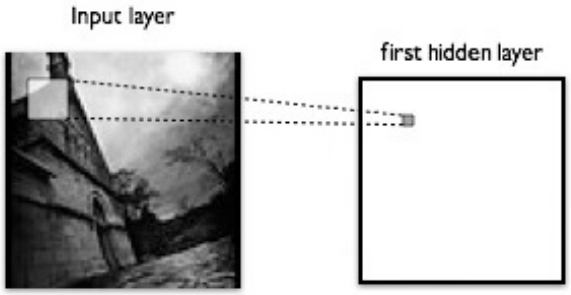


Figure 3.9: Convolutional layer with a region of the image connected to a neuron. Adapted from (Deep Learning, 2010).

Every connection between the input window's pixels and the corresponding neuron has a weight and bias, similarly to the neural networks connections seen before. However, the value of these parameters is the same for every neuron in the first layer. The image is being convoluted with the sliding window (or kernel), and since the parameters are the same for the entire image, it means that all the neurons are detecting the same feature, but at different locations in the image. Indeed, a feature, e.g. a horizontal edge, at a given location is likely to be at other places in the image. This means that a convolutional neural network is well adapted to translation invariance, because, for example, an image of a truck will still be the same even if it's moved. The first layer then may be viewed as a feature map of the image, created with a kernel defined by the shared weights and shared bias (Bishop, 2006; Murphy, 2012; Mitchell, 1997).

Since a feature map only detects a single kind of localized feature, to perform successful image recognition, the convolutional layer must consist of several different feature maps. Figure 3.10 shows an example of 4 feature maps defined by their own shared weights and bias, meaning that the network can detect 4 different kinds of features all across the entire image. Those feature maps

will then be transformed using the activation function of the neurons. While standard non-linear functions like $f(x) = (1+e^{-x})^{-1}$ or $f(x) = \tanh x$ could be used, they are much slower in terms of training time than the non-linear function $f(x) = max(0,x)$. Neurons with this activation function are called Rectified Linear Units (ReLUs) (Nair and Hinton, 2010), and present a training process several times faster than their equivalents with *tanh* units (Russakovsky et al., 2015).

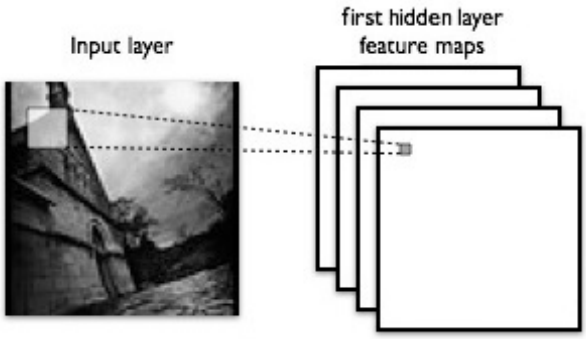


Figure 3.10: Convolutional layer with its respective multiple feature maps. Adapted from (Deep Learning, 2010).

3.2.4.2 Pooling Layer

This type of layer is used after convolutional layers to downsample the output information of the neuron. One common type of pooling is *max pooling*, but others, like average pooling, could also be used. In this layer, a $p \times p$ region of the feature map is considered and pooled, outputting the maximum value in that region, generating a more condensed version of the feature map, as shown in Figure 3.11. This kind of layers have the advantage of decreasing the amount of storage and processing of subsequent layers.

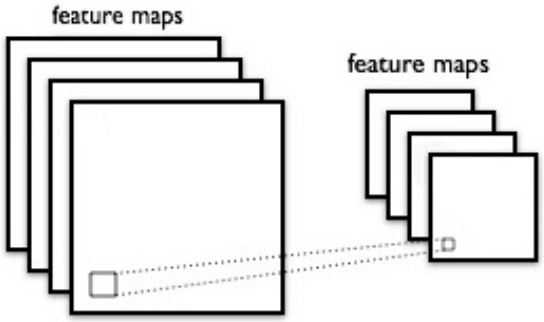


Figure 3.11: Pooling layer outputting condensed feature maps. Adapted from (Deep Learning, 2010).

3.2.4.3 Fully Connected Layer

After a set of convolutions, ReLU activations and downsampling, a fully connected layer, where all the neurons in the previous layer are connected to every single neuron in this layer, is inserted. Since data is not spatially related anymore, there can be no more convolutional layers after a fully connected layer. The purpose of this layer is to learn a non-linear combination of the features extracted in the previous convolutional layers that then will be used for classification, as schematized in Figure 3.12.

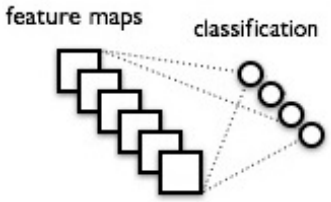


Figure 3.12: Fully connected layer with all feature map neurons connected to the output neurons. Adapted from (Deep Learning, 2010).

3.2.4.4 Loss Layer

The Loss layer is not considered, truly, as a separated new layer, but as the activation function of the fully connected neurons in the last layer, which will determine the output of the network. It specifies how the network training will penalize the deviation between the true and predicted labels. Various functions are appropriate for different tasks: Euclidean loss is used for real-valued labels, sigmoid cross-entropy to predict K independent probability values in the range $[0,1]$ and softmax loss is used for predicting a single class out of K mutually exclusive classes.

The design of a typical CNN model is achieved by the concatenation of layers in the typical structure $[Conv-Pool]\times N-FC-LR]$. Similar to a traditional neural network, parameter optimization is achieved by back-propagation of the classification errors. The number of layers, as well as the parameters involved in each type of layer, are variable, according to the problem at hand and should be optimized in accordance.

3.3 Conclusion

With all the considerations regarding biometric system design, as well as the computer vision and machine learning tools considered in this chapter, the scientific contributions of the present thesis will be detailed throughout the following chapters. We start this exploration by addressing unimodal strategies, i.e. approaches based on a single biometric trait. Further chapters will enter into further detail regarding multimodality and cross-sensor scenarios.

Part II

Scientific Contributions

Chapter 4

Unimodal Recognition

Over the past few years, the face and the iris have been on the spotlight of many research works in biometrics. The *face* is a easily acquirable trait with a high degree of uniqueness, while the *iris*, the coloured part of the eye, presents unique textural patterns resulting from its random morphogenesis during embryonic development (Bakshi et al., 2012). These marked advantages, however, fall short when low-quality images are presented to the system. It has been noted that the performance of iris and face recognition algorithms is severely compromised when dealing with non-ideal scenarios such as non-uniform illumination, pose variations, occlusions, expression changes and radical appearance changes (Bakshi et al., 2012; Boddeti et al., 2011). Several works have tried to explore alternative hypothesis to tackle this problem, either by developing more robust algorithms or by exploring new traits to allow or aid in the recognition process (Woodard et al., 2010). These two strategies will serve as the basis for each of the two main sections of the present chapter: we start by exploring an alternative trait in the periocular region and then return to a more traditional face recognition framework, exploring a more robust approach for certain types of non-ideal scenarios.

4.1 Periocular Recognition

Even though a true definition of the periocular region is not standardized, it is common to describe it as the region in the immediate vicinity of the eye (Padole and Proenca, 2012; Smereka and Kumar, 2013), as depicted in Figure 4.1. Periocular recognition can be motivated as a middle point between face and iris recognition. It has been shown to present increased performance when only degraded facial data (Miller et al., 2010a) or low quality iris images (Bharadwaj et al., 2010; Tan and Kumar, 2013) are made available, as well as promising results as a soft biometric trait to help improve both face and iris recognition systems in less constrained acquisition environments (Joshi et al., 2012).

The developed approach to periocular recognition under less ideal acquisition conditions will be detailed throughout the next sections. Such approach is based on the idea of maximum *a posteriori* adaptation of an Universal Background Model, as proposed by Reynolds for speaker

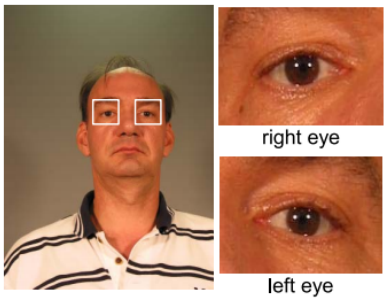


Figure 4.1: Example of periocular regions from both eyes, extracted from a face image Woodard et al. (2010).

verification (Reynolds et al., 2000). We evaluate the proposed algorithm on two datasets of color periocular images acquired under visible wavelength (VW) illumination. Multiple noise factors such as varying gazes/poses and heterogeneous lighting conditions are characteristic to such images, thus representing the main challenge for the developed algorithm[1].

4.1.1 Related Work

Periocular biometrics is a recent area of research, proposed by the first time in a feasibility study by Park et al. (Park et al., 2009). In this pioneer work, the authors suggested the periocular region as a potential alternative to circumvent the significant challenges posed to iris recognition systems working under unconstrained scenarios. The same authors analysed the effect of degradation on the accuracy of periocular recognition (Park et al., 2011). Performance assessment under less constrained scenarios is also the goal of the work by Miller et al. (Miller et al., 2010b), where factors such as blur and scale are shown to have a severe effect on the performance of periocular recognition. Padole and Proença (Padole and Proenca, 2012) also explore the effect of scale, pigmentation and occlusion, as well as gender, and propose an initial region-of-interest detection step to improve recognition accuracy.

Ross et al. (Ross et al., 2012) explored information fusion based on several feature extraction techniques, to handle the significant variability of input periocular images. Information fusion has become one of the trends in biometric research in recent years and periocular recognition is no exception. Bharadwaj et al. (Bharadwaj et al., 2010) proposed fusion of matching scores from both eyes to improve the individual performance of each of them. On the other hand, Woodard et al. (Woodard et al., 2010) place fusion at the feature level, using color and texture information.

Some works have explored the advantages of the periocular region as an aid to more traditional approaches based on iris. Boddeti et al. (Boddeti et al., 2011) propose the score level fusion of a traditional iris recognition algorithm, based on Gabor features, and a periocular probabilistic approach based on optimal trade-off synthetic discriminant functions (OTSDF). A similar

[1]The work presented in this section resulted in the conference paper: João C. Monteiro and Jaime S. Cardoso. *A Novel Application of Universal Background Models for Periocular Recognition.* International Joint Conference on Biomedical Engineering Systems and Technologies. Springer International Publishing, 2015. p. 291-307

work by Joshi et al. (Joshi et al., 2012) proposed feature level fusion of wavelet coefficients and LBP features, from the iris and periocular regions respectively, with considerable performance improvement over both singular traits. A recent work by Tan et al. (Tan and Kumar, 2013) has also explored the benefits of periocular recognition when highly degraded regions result from the traditional iris segmentation step. The authors have observed discouraging performance when the iris region alone is considered in such scenarios, whereas introducing information from the whole periocular region lead to a significant improvement.

Two other recent and relevant works by Moreno et al. (Moreno et al., 2013a,b) explore the well-known approach of sparse representation classification in the scope of the specific problem of periocular recognition. A thorough review of the most relevant method in recent years concerning periocular recognition and its main advantages can be found in the work by Santos and Proença (Santos and Proença, 2013).

The new approach to periocular recognition presented in this section is based on a general framework with proven results in voice biometrics. We explore a strategy based on the adaptation of a Universal Background Model (UBM) to achieve faster and more robust training of individual specific models. With such idea in mind we aim not only to design a high performance recognition system, but also to assess the versatility and robustness of the UBM strategy for biometric traits other than voice.

4.1.2 Universal Background Model

The universal background modeling strategy was initially proposed in the field of voice biometrics (Povey et al., 2008). The framework can be easily understood if the problem of biometric verification is interpreted as a basic hypothesis test. Given a biometric sample Y and a claimed ID, S, we define:

$$H_0\text{: } Y \text{ belongs to } S$$
$$H_1\text{: } Y \text{ does not belong to } S$$

as the null and alternative hypothesis, respectively. The optimal decision is made by a likelihood-ratio test:

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{accept } H_1 \end{cases} \tag{4.1}$$

where θ is the decision threshold for accepting or rejecting H_0 and $p(Y|H_i)$ is the likelihood of observing Y knowing that H_i is true. The goal of a biometric verification system can, thus, be accomplished by the computation of the likelihood values $p(Y|H_0)$ and $p(Y|H_1)$ for a given sample. It is intuitive to note that H_0 will correspond to a model λ_{hyp} that characterizes the hypothesized individual, whereas H_1 will represent the alternative hypothesis, that is the model of all of the alternatives to the hypothesized individual, $\lambda_{\overline{hyp}}$.

The computation of $p(Y|H_i)$ depends on the specific strategy chosen for data modeling. If H_0 and H_1 correspond to a pair of generative models, trained on sets of genuine and impostor data, respectively, then the $p(Y|H_i)$ values can be roughly expressed as the projections of biometric sample Y onto each of these models. This formulation motivates the need for a model that successfully covers the space of alternatives to the hypothesized identity. The most common designation in the literature for such a model is the universal background model or UBM (Reynolds, 2002). Such a model must be trained on a rather large set of data, so as to faithfully cover a representative user space.

Even though the UBM approach was initially proposed for verification mode, we extrapolate its rationale for identification systems. Instead of performing a single one *vs.* one likelihood-ratio test and checking the validity of the condition presented in Equation (4.1), a one *vs.* all approach may alternatively be considered. Given an unknown sample, the most likely identity, Id_{max}, will correspond to the highest likelihood-ratio value, amongst all enrolled users:

$$Id_{max} = \underset{i}{\text{argmax}} \left(\frac{p(Y|H_0^{(i)})}{p(Y|H_1)} \right) \tag{4.2}$$

where $H_0^{(i)}$ represents the model describing user i.

Such a decision step represents the second big advantage of the UBM approach. The ratio between the IDSM and the UBM probabilities of the observed data is a more robust decision criterion than relying solely on the IDSM probability. This results from the fact that some subjects are more prone to generate high likelihood values than others, *i.e.*, some people have a more "generic" look than others. The use of a likelihood ratio with a universal reference works as a normalization step, mapping the likelihood values in accordance with their global projection. Without such a step, finding a global optimal value for the decision threshold, θ, presented in Equation (4.1), would be a far more complex process. Defining an objective way of quantifying $p(Y|H_0)$ and $p(Y|H_1)$ becomes, thus, the true challenge when following this approach. In the following sections, we analyze in detail the strategies chosen to model both λ_{hyp} and $\lambda_{\overline{hyp}}$.

4.1.3 Algorithm Overview

The proposed algorithm is schematically represented in Figure 4.2. The two main blocks - *enrollment* and *recognition* - refer to the typical architecture of a biometric system, as seen previously in Section 2.4. During enrollment a new individual's biometric data is inserted into a previously existent database of individuals. Such database is probed during the recognition process to assess either the validity of an identity claim - *verification* - or the k most probable identities - *identification* - given an unknown sample of biometric data.

During the enrollment, a set of N models describing the unique statistical distribution of biometric features for each individual $n \in \{1,\ldots,N\}$ is trained by maximum *a posteriori* (MAP) adaptation the UBM. MAP adaptation works as a specialization of the UBM based on each individual's biometric data. The idea of MAP adaptation of the UBM was first proposed by

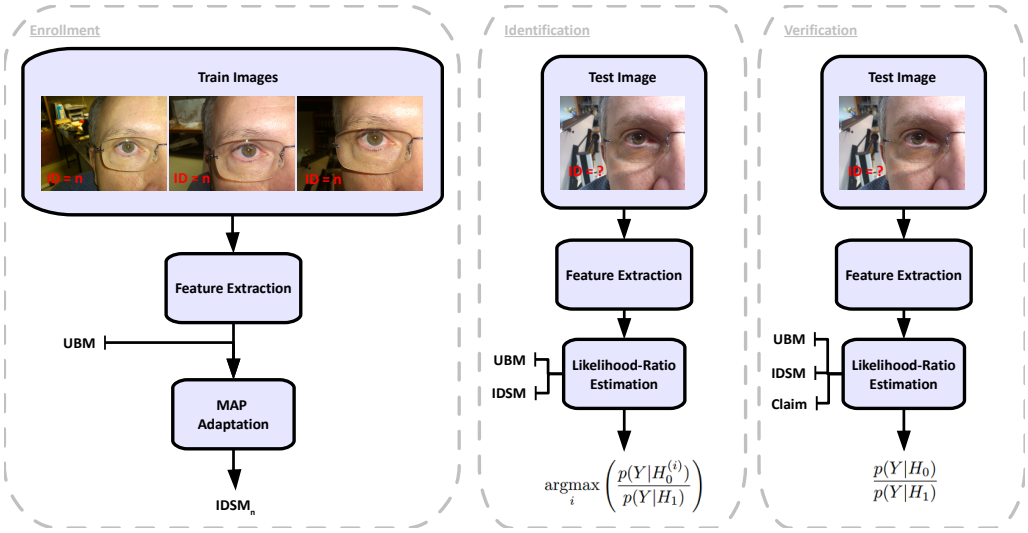


Figure 4.2: Graphical representation of the main steps in both the enrollment and recognition (verification and identification) phases of the proposed periocular recognition algorithm.

Reynolds (Reynolds et al., 2000), for speaker verification, and will be further motivated in the following sections.

The recognition phase is carried out through the projection of the features extracted from an unknown sample onto both the UBM and the individual specific models (IDSM) of interest. A likelihood-ratio between both projections outputs the final recognition score. Depending on the functioning mode of the system - verification or identification - decision is carried out by thresholding or maximum likelihood-ratio respectively.

4.1.4 Hypothesis Modeling

In the present work, we chose Gaussian mixture models (GMM) to model both the UBM, *i.e.*, $\lambda_{\overline{hyp}}$, and the individual-specific models (IDSM), *i.e.*, λ_{hyp}. Such models are capable of capturing the empirical probability density function of a given set of feature vectors, so as to faithfully model their intrinsic statistical properties (Reynolds et al., 2000). The choice of GMM to model feature distributions in biometric data is extensively motivated in many works of related areas. From the most common interpretations, GMMs are seen as capable of representing broad "hidden" classes, reflective of the unique structural arrangements observed in the analyzed biometric traits (Reynolds et al., 2000). Besides this assumption, Gaussian mixtures display both the robustness of parametric unimodal Gaussian density estimates, as well as the ability of non-parametric models to fit non-Gaussian data (Reynolds, 2015). This duality, alongside the fact that GMM have the noteworthy strength of generating smooth parametric densities, confers such models a strong advantage as generative models of choice. For computational efficiency, GMM models are often trained using diagonal covariance matrices. This approximation is often found in the biometrics literature, with no significant accuracy loss associated (Xiong et al., 2006).

All models are trained on sets of Scale Invariant Feature Transform (SIFT) keypoint descriptors (Lowe, 2004). This choice for periocular image description is thoroughly motivated in literature (Ross et al., 2012; Park et al., 2011), mainly due to the observation that local descriptors work better than their global counterparts when the available data presents non-uniform conditions. Furthermore, the invariance of SIFT features to a set of common undesirable factors (image scaling, translation, rotation and also partially to illumination and affine or $3D$ projection), confers them a strong appeal in the area of unconstrained biometrics. For all the reasons stated above we chose to also use the dense variation of SIFT for feature description in the face recognition contributions presented throughout the next section.

Originally, SIFT descriptors were defined in 128 dimensions. However, we chose to perform a principal component analysis (PCA), as suggested in (Shinoda and Inoue, 2013), reducing the dimensionality to 32. Such a reduction allows not only a significant reduction in the computational complexity of the training phase, but also an improved distinctiveness and robustness to the extracted feature vectors, especially as far as image deformation is concerned (Ke and Sukthankar, 2004). The fact that we are training all GMMs on diagonal covariance matrices also benefits from the feature decorrelation effect of PCA. We computed the principle components from the same set of keypoint descriptors used to train the UBM.

4.1.4.1 H_1: UBM Parameter Estimation

To train the universal background model, a large amount of "impostor" data, *i.e.*, a set composed of data from all the enrolled individuals, is used, so as to cover a wide range of possibilities in the individual search space (Shinoda and Inoue, 2013). The training process of the UBM is simply performed by fitting a k-mixture GMM to the set of PCA-reduced feature vectors extracted from all of the "impostors".

If we interpret the UBM as an "impostor" model, its "genuine" counterpart can be obtained by adaptation of the UBM's parameters, $\lambda_{\overline{hyp}}$, using individual specific data. For each enrolled user, n, an IDSM, defined by parameters λ_{hyp_n}, is therefore obtained. The adaptation process will be outlined in the following section.

4.1.4.2 H_0: MAP Adaptation of the UBM

IDSMs are generated by the tuning of the UBM parameters, in a maximum *a posteriori* (MAP) sense, using individual-specific biometric data. This approach provides a tight coupling between the IDSM and the UBM, resulting in better performance and faster scoring than uncoupled methods, as well as a robust and precise parameter estimation, even when only a small amount of data is available (Shinoda and Inoue, 2013). The adaptation process consists of two main estimation steps. First, for each component of the UBM, a set of sufficient statistics is computed from a set of M individual-specific feature vectors, $X = \{\mathbf{x}_1, ..., \mathbf{x}_M\}$:

$$n_i = \sum_{m=1}^{M} p(i|\mathbf{x}_m) \tag{4.3}$$

$$E_i(\mathbf{x}) = \frac{1}{n_i}\sum_{m=1}^{M} p(i|\mathbf{x}_m)\mathbf{x}_m \tag{4.4}$$

$$E_i(\mathbf{x}\mathbf{x}^t) = \frac{1}{n_i}\sum_{m=1}^{M} p(i|\mathbf{x}_m)\mathbf{x}_m\mathbf{x}_m^t \tag{4.5}$$

where $p(i|\mathbf{x}_m)$ represents the probabilistic alignment of $\mathbf{x}_m$ into each UBM component. Each UBM component is then adapted using the newly-computed sufficient statistics and considering diagonal covariance matrices. The update process can be formally expressed as:

$$\hat{w_i} = [\alpha_i n_i/M + (1-\alpha_i)w_i]\,\xi \tag{4.6}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \alpha_i E_i(\mathbf{x}) + (1-\alpha_i)\boldsymbol{\mu}_{\boldsymbol{i}} \tag{4.7}$$

$$\hat{\Sigma}_i = \alpha_i E_i(\mathbf{x}\mathbf{x}^t) + (1-\alpha_i)(\boldsymbol{\sigma}_{\boldsymbol{i}}\boldsymbol{\sigma}_{\boldsymbol{i}}^t + \boldsymbol{\mu}_{\boldsymbol{i}}\boldsymbol{\mu}_{\boldsymbol{i}}^t) - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\,t} \tag{4.8}$$

$$\boldsymbol{\sigma}_{\boldsymbol{i}} = \mathrm{diag}(\Sigma_i) \tag{4.9}$$

where $\{w_i, \boldsymbol{\mu}_{\boldsymbol{i}}, \boldsymbol{\sigma}_{\boldsymbol{i}}\}$ are the original UBM parameters and $\{\hat{w_i}, \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{i}}\}$ represent their adaptation to a specific speaker. To assure that $\sum_i w_i = 1$, a weighting parameter ξ is introduced. The α parameter is a data-dependent adaptation coefficient. Formally, it can be defined as:

$$\alpha_i = \frac{n_i}{r+n_i} \tag{4.10}$$

The relevance factor r weights the relative importance of the original values and the new sufficient statistics. In the present work we set $r = 16$.

4.1.5 Experimental Tests and Results

In this subsection we start by presenting the datasets and the experimental setups under which performance was assessed. Further sections present a detailed analysis regarding the effect of model complexity and fusion of color channels in the global performance of the proposed algorithm.

The proposed algorithm was tested on two noisy color iris image databases: UBIRIS.v2 and MobBIO. Even though both databases were designed in an attempt to promote the development of robust iris recognition algorithms for images acquired under VW illumination, their intrinsic properties make them attractive to study the feasibility of periocular recognition under similar conditions. The following sections detail their main features as well as the reasoning behind their choice.

4.1.5.1 UBIRIS.v2 database

Images in UBIRIS.v2 (Proença et al., 2010) database were captured under non-constrained conditions (at-a-distance, on-the-move and on the visible wavelength), with corresponding realistic noise factors. Figure 4.3 depicts some examples of these noise factors (reflections, occlusions, pigmentation, etc.). Two acquisition sessions were performed with 261 individuals involved and a total of 11100 300×400 color images acquired. Each individual's images were acquired at variable distances with 15 images per eye and per season. Even though the UBIRIS.v2 database was primarily developed to allow the study of unconstrained iris recognition, many works have explored its potential for periocular-based strategies as an alternative to low-quality iris recognition (Bharadwaj et al., 2010; Joshi et al., 2012; Padole and Proenca, 2012).

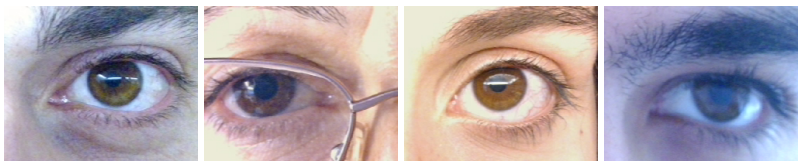


(a) (b) (c) (d)

Figure 4.3: Examples of noisy image from the UBIRIS.v2 database.

4.1.5.2 MobBIO database

The MobBIO multimodal database (Sequeira et al., 2014) was created in the scope of the 1st Biometric Recognition with Portable Devices Competition 2013, integrated in the ICIAR 2013 conference. The main goal of the competition was to compare various methodologies for biometric recognition using data acquired with portable devices. We tested our algorithm on the iris modality present on this database. Regarding this modality the images were captured under two alternative lighting conditions, with variable eye orientations and occlusion levels, so as to comprise a larger variability of unconstrained scenarios. Distance to the camera was, however, kept constant for each individual. For each of the 105 volunteers 16 images (8 of each eye) were acquired. These images were obtained by cropping a single image comprising both eyes. Each cropped image was set to a 300×200 resolution. Figure 4.4 depicts some examples of such images.



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

Figure 4.4: Examples of iris images in the MobBIO database.

The MobBIO database presents a face modality which has also been explored for comparative purposes in the present work. Images were acquired in similar conditions to those described above for iris images, with 16 images per subject. Examples of such images can be observed in Figure 4.5.



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

Figure 4.5: Examples of face images in the MobBIO database.

4.1.5.3 Evaluation Metrics

Besides the EER and rank-1 recognition rate, already presented as alternatives for evaluation metrics in Section 2.8, the decidability index (DI) was also considered for evaluation in this work. The DI quantifies the separation of "genuine" and "impostor" likelihood score distributions, as follows:

$$DI = \frac{|\mu_g - \mu_i|}{\sqrt{0.5(\sigma_g^2 + \sigma_i^2)}} \tag{4.11}$$

where (μ_g, σ_g) and (μ_i, σ_i) are the mean and standard deviation of the genuine and impostor score distributions, respectively. We chose this additional metric as it was the standard evaluation metric for the NICE.II competition (Proença, 2009), allowing an easy and objective comparison between our methodology and its top ranked algorithms.

4.1.5.4 Experimental Setups

Our experiments were conducted in three distinct experimental setups, two of them regarding the UBIRIS.v2 database and the remaining one the MobBIO database:

1. In the first setup, for the UBIRIS.v2 images, six samples from 80 different subjects were used, captured from different distances (4 to 8 meters), with varying gazes/poses and notable changes in lighting conditions. One image per individual was randomly chosen as probe, whereas the remaining five samples were used for the UBM training and MAP adaptation. The results were cross-validated by changing the probe image, per subject, for each of the six chosen images.

2. Many works on periocular biometrics evaluate their results using a well-known subset of the UBIRIS.v2 database, used in the context of the NICE II competition (Proença, 2009). This dataset is divided in train and test subsets, with a total of 1000 images from 171 individuals.

In the present work we choose to use a test subset composed by 904 images from 152 individuals. Only individuals with more than 4 available images were considered, as 4 images were randomly chosen for training and the rest for testing. Results were cross-validated 10-fold. The train dataset composed by the remaining 96 images from 19 individuals was employed in the parameter optimization step described in further sections.

3. Concerning the MobBIO database, 8 images were randomly chosen from each of the 105 individuals for the training of the models, whereas the remaining 8 were chosen for testing. The process was cross-validated 10-fold. For comparative purposes a similar experiment was carried out on face images from the same 105 individuals, using the same $8+8$ image distribution.

As both databases are composed by color images, each of the RGB channels was considered individually for the entire enrollment and identification process. For the parameter optimization described in the next section, images were previously converted to grayscale.

4.1.5.5 Parameter Optimization

A smaller dataset, for each database, was also designed to optimize the the number of GMM mixtures of the trained models. For the UBIRIS.v2 we chose to work with the well-known train dataset from the NICE II competition (Proença, 2009), composed by 96 images from 19 individuals. For the MobBIO database we chose a total of 50 images from 10 individuals to perform the previously referred optimization. The obtained performance was cross-validated using a leave-one-out strategy. The chosen metric to evaluate performance was the rank-1 recognition rate (R_1). The evolution of performance with the optimization parameter can be observed in Figure 4.6. With such results in mind, the recognition performance for the experimental setups presented in the last section was assessed for a number of mixtures $M = 128$ and $M = 64$ for the UBIRIS.v2 and MobBIO databases respectively. We choose both these values as the values where a performance plateau is achieved in the graph of Figure 4.6. We chose the lowest possible values for the parameter M so as to minimize the computational complexity of the UBM training, which constitutes the limiting step of the process, without a significant loss in performance.

4.1.5.6 Performance Analysis

The results obtained for both databases and experimental setups are represented through ROC and CMC curves on Figures 4.7a to 4.7f. A comparison with some state-of-the-art algorithms in the UBIRIS.v2 database is also presented in Table 4.1. In this table results are grouped according to the experimental setup of each reported work and also the studied trait: P - Periocular, I - Iris or $P+I$ - Fusion of both traits.

Besides testing each of the RGB channels individually, a simple sum-rule score-level fusion strategy (Kittler et al., 1998) was also considered. It is easily discernible, from the observation of Figure 4.7, that the fusion of information from multiple color channels translates in a significant

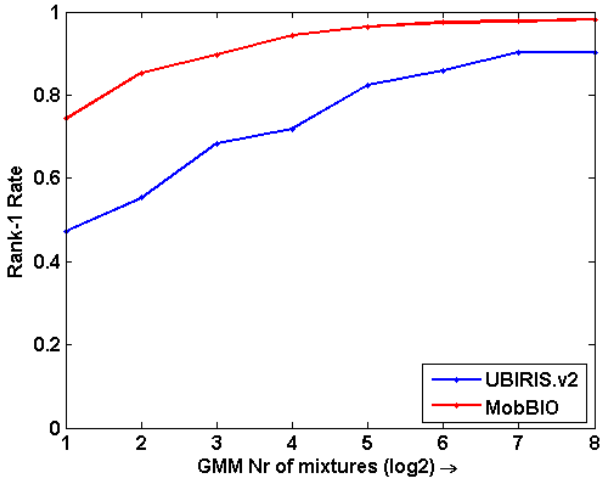


Figure 4.6: Recognition rates obtained with the optimization subset for variable values of parameter M, number of mixtures in the trained GMMs.

improvement in performance for all the tested datasets. When comparing the results obtained with this approach with some state-of-the-art algorithms a few points deserve further discussion. First, the proposed algorithm is capable of achieving and even surpassing state-of-the-art performance in multiple experimental setups. Concerning the most common of such setups (2), it is interesting to note that a few works attempted to explore the UBIRIS.v2 dataset for iris recognition. The obtained performance has been considered "discouraging" in the work by Kumar et al. (Kumar and Chan, 2012). Comparing the rank-1 recognition rate obtained with our algorithm (88.93%) with the 48.1% reported in the former work, we conclude that the periocular region may represent a viable alternative to iris in images acquired under visible wavelength (VW) illumination. Such acquisition conditions are known to increase light reflections from the cornea, resulting in a sub-optimal signal-to-noise ratio (SNR) in the sensor, lowering the contrast of iris images and the robustness of the system (Proença, 2011). More recent works have explored multimodal approaches, using combined information from both the iris and the periocular region. Analysis of Table 4.1 shows that none of such works reaches the performance reported in the present work for

Work	*Setup*	*Traits*	R_1	*EER*	D_i
Proposed	1	P	**97.7%**	**0.045**	4.98
Proposed	2	P	**88.9%**	**0.072**	**3.61**
Moreno et al. (Moreno et al., 2013b)	1	P	97.6%	0.142	–
Tan et al. (Tan and Kumar, 2013)	2	$P+I$	39.4%	–	–
Tan et al. (Tan et al., 2011)	2	$P+I$	–	–	2.57
Kumar et al. (Kumar and Chan, 2012)	2	I	48.0%	–	–
Proença et al. (Proença and Santos, 2012)	2	I	–	≈ 0.11	2.85

Table 4.1: Comparison between the average obtained results with both experimental setups for the UBIRIS.v2 database and some state-of-the-art algorithms.

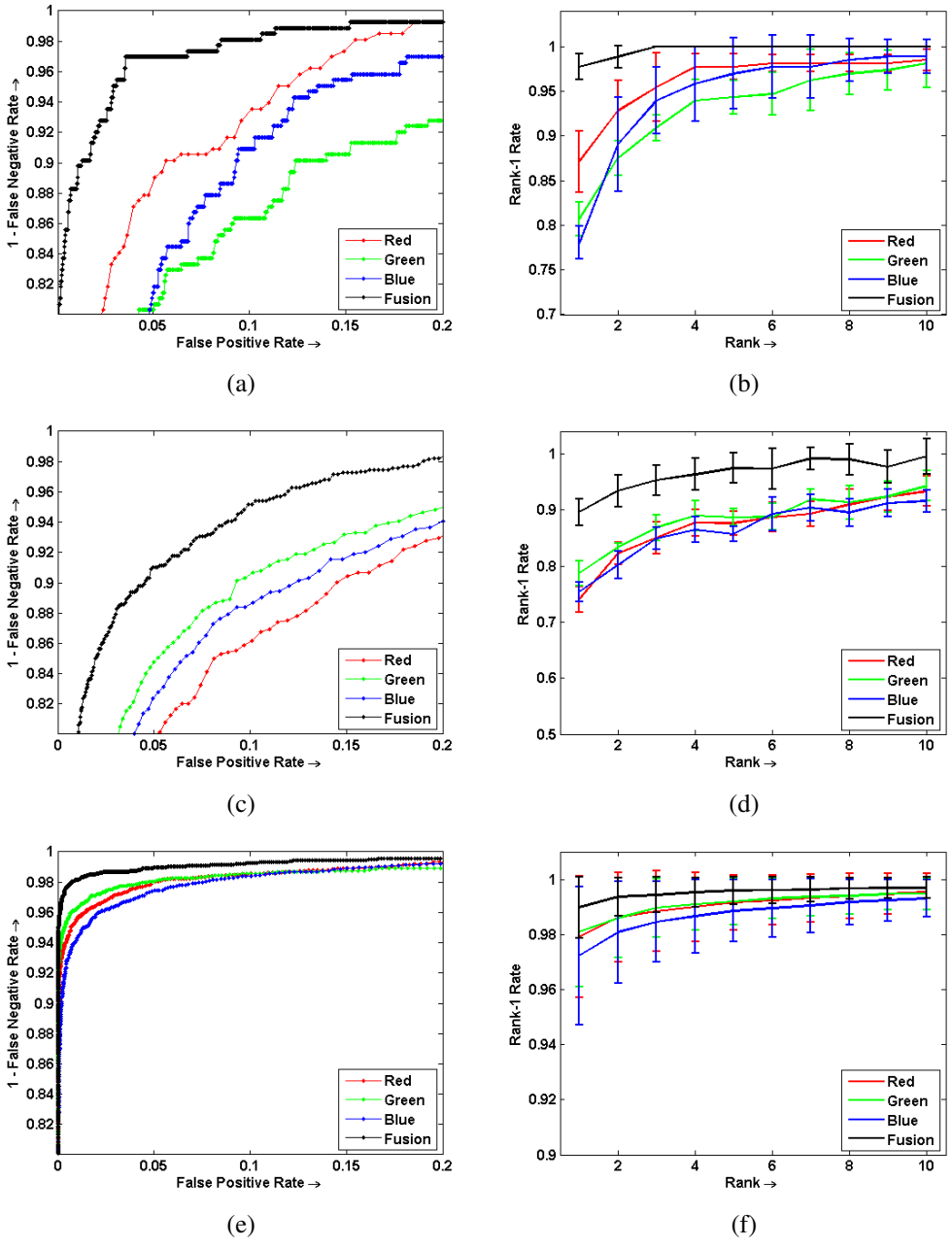


Figure 4.7: ROC and CMC curves for: (a-b) UBIRIS.v2 database, setup (1); (c-d) UBIRIS.v2 database, setup (2) and (e-f) MobBIO database, setup (3);. ROC curves present the average results of cross-validation, whereas CMCs present the average value and error-bars for the first 10 ranked IDs in each setup.

the same experimental setup. Such observation might indicate that most discriminative biometric information from the UBIRIS.v2 images might be present in the periocular region, and that considering data from the very noisy iris regions might only result in a degradation of the performance obtained by the periocular region alone.

Concerning the MobBIO database, an alternative comparison was carried out to analyze the

potential of the periocular region as an alternative to face recognition. The observed performance for periocular images was considerably close to that using full-face information, with rank-1 recognition rates of 98.98% and 99.77% respectively. These results are an indication that, under more ideal acquisition conditions, there is enough discriminative potential in the periocular region alone to rival with the full face in terms of recognition performance. In scenarios where some parts of the face are purposely disguised (scarves covering the mouth for example) this observation might indicate that a non-corrupted periocular region can, indeed, overperform recognition with the occluded full-face images. Such conditions were not tested in the present work but might be the basis for an interesting follow-up. Even though the observed results are promising, it must be noted that the noise factors present in the MobBIO database are still far from a highly unconstrained scenario. Furthermore, it must be noted that the total number of individuals present in this dataset is still very limited, and performance of the presented algorithm might be compromised when large-scale datasets are considered.

The robustness of the likelihood-ratio decision step was also assessed. We compared the performance observed for the scores obtained with Equation (4.1) and the scores obtained using only its numerator, i.e. only the likelihood of each test image without the UBM normalization. For the experimental setup (2) we obtained an average rank-1 recognition rate of 43.6%, whereas the MobBIO experimental setup (3) resulted in 90.5% for the same metric. It is easily noted that performance is less compromised in the MobBIO database. Considering only the numerator of Equation (4.1) is the same as considering a constant denominator value for every tested image. As the denominator represents the projection of the tested images on the UBM, this alternative decision strategy might be interpreted as assuming a constant reference background for every tested image. From the observed results we might conclude that such assumption fits better the images from the MobBIO database. We also note that for more challenging scenarios, where the constant background assumption fails, the use of background normalization produces a significant improvement in performance.

A few last considerations regarding the discriminative potential of the proposed algorithm may be taken from the observation of Figure 4.8. On each row we analyze the 4 highest ranked models for the images presented in the first column. The first two rows depict correct identifications. It is interesting to note how each of the 4 highest ranked identities in the second row correspond to individuals wearing glasses. Such observation seems to indicate that the proposed modeling process is capable of describing high-level global features, such as glasses. Furthermore, the fact that the correct ID was guessed also demonstrates its capacity of distinguishing between finer details separating individual models. The third and fourth rows present some test images whose ID was not correctly assessed by the algorithm. In the third row we present a case where even though the correct ID and the most likely model were not correctly paired, the correct guess still appears in the top ranked models. We note that even a human user analyzing the four highest ranked models would find it very difficult to detect significant differences. The fourth row presents the extreme case where none of the top ranked models correspond to the true ID. It is worth noting how the test images presented in the third and fourth rows are very similar to a large number of images

present in other individual's models. This observation leads to the hypothesis that *some users are easier to identify than others inside a given population*, an effect known as the *Doddington zoo effect* (Ross et al., 2009). It also shows that the proposed algorithm is capable of narrowing the range of possible identities to those subjects who "look more alike".

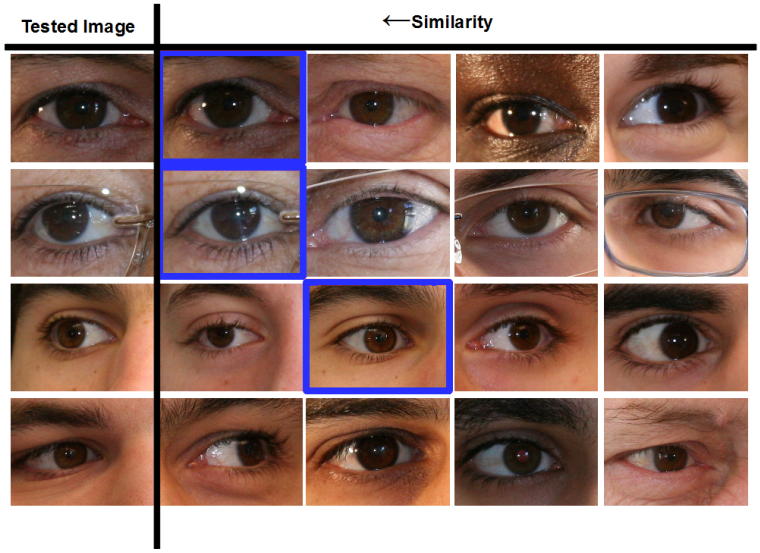


Figure 4.8: Identification results for rank-4 in the UBIRIS.v2 database. The first column depicts the tested images while the remaining 4 images exemplify representative images from the 4 most probable models, after the recognition is performed. The blue squares mark the true identity.

4.2 Periocular Recognition in Mobile Scenarios

We further explore the potential of the periocular region as a biometric trait of choice for mobile applications. Even in such scenarios, the periocular region does not require rigid capture or complex imaging systems, thereby making it easy to acquire even by an inexperienced user. Nevertheless, several problems arise when attempting to perform periocular biometrics in mobile environments. The wide variety of camera sensors and lenses used in mobile devices produce discrepancies in working images, as they might be acquired with both color distortions and multiple resolutions. On-the-go acquisition by inexperienced subjects will result in demanding pose, illumination, and expression changes, thereby yielding variable acquisition angles and scales, or rotated images. All these limitations are intrinsic to the nature of mobile devices and must, thus, be handled by the recognition algorithm.

Based on the GMM-UBM methodology explored in the previous sections, we perform a comparative analysis with another state-of-the-art alternative, proposed by Santos et al. (Santos et al.,

2015)[2]. A detailed analysis of this framework will be presented in the following subsection. For this point onwards we will refer to the GMM-UBM approach as Method 1 and to Santos et al.'s as Method 2.

4.2.1 Santos et al. Method

This algorithm, proposed in (Santos et al., 2015), may be divided into four main blocks: normalization using a device-specific color correction and region-of-interest (ROI) definition; feature encoding using information from both the iris and the periocular region; feature matching and score-level fusion. The flow of information through the aforementioned blocks is schematically depicted in Figure 4.9.

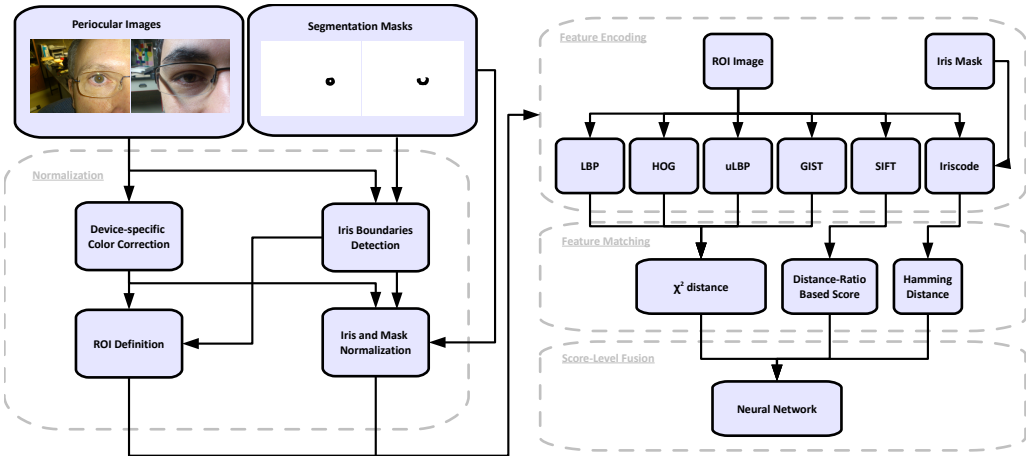


Figure 4.9: Flow diagram of the main composing blocks of the methodology proposed in (Santos et al., 2015).

During the normalization block a device-specific color-correction was applied so as to compensate for possible chromatic distortions observed in real-life scenarios. Another source of variability commonly observed in data acquired with mobile devices is variable scale. In order to overcome such problem, and making use of a state-of-the-art iris segmentation algorithm (Tan et al., 2011), the authors propose a segmentation of the iris boundary, to serve as a reference for the periocular region. Using the previously calculated radius of the iris, r_i, the periocular ROI was defined as 35 square patches that formed a 7×5 grid, where each patch had an area equivalent to $1.4r_i^2$.

Periocular data was encoded using a similar feature extraction scenario as the one described for the previous methodology. SIFT, HOG, uLBP (as well as the original LBP) and GIST, were also tested independently and used in a conjugated manner. For the iris region, in addition to these descriptors, a fifth approach was also explored, using the original iriscode algorithm proposed by

[2]The work presented in this section resulted in the conference paper: João C. Monteiro, Rui Esteves, Gil Santos, Paulo Torrão Fiadeiro, Joana Lobo and Jaime S. Cardoso *A Comparative Analysis of Two Approaches to Periocular Recognition in Mobile Scenarios.* International Symposium on Visual Computing. Springer International Publishing, 2015. p. 268-280.

Daugman (Daugman, 1993). For a single image a total of 11 feature descriptors is, therefore, extracted: 5 for the periocular region and the same 5 plus the iriscode for the iris region.

Matching was carried out by comparing the 11 pairs of feature descriptors extracted from a pair of images, using the matching algorithm specified for each of them (χ^2 for histogram-based algorithms, distance-based score for pairs of SIFT keypoints and Hamming Distance for iriscode), resulting, thus, in 11 individual scores. Performance can then be evaluated either for each descriptor individually or by exploring more complex fusion strategies. In the original algorithm, a multi-layer perceptron artificial neural network was used to achieve such fusion.

4.2.2 Algorithm Extensions

Besides the comparative analysis between the two methodologies described in the previous sections, we also present some extensions to their original formulations. Such modifications are as follows:

1. **Pre-processing:** As no pre-processing strategy was included in the original formulation of the GMM-UBM algorithm, we aimed to assess if its presence could translate in a significant improvement in performance. We chose the Discrete Cosine Transform (DCT), as proposed by Chen et al. (Chen et al., 2006) for illumination normalization in face images, as it yielded the best overall performance. This technique is based on the removal of low-frequency coefficients of the DCT, in order to compensate for the variations in lighting conditions, that are known to lie, mainly, on such frequency band. We also tested the device-specific color correction technique proposed in Method 2.

2. **Periocular segmentation:** As the GMM-UBM approach presented no preliminary periocular segmentation in its original form, and in order to achieve an uniform set of conditions for performance comparison, we chose to perform segmentation with the same methodology used for Method 2, as described in Section 4.2.1.

3. **Feature Descriptors:** Similar to segmentation, we chose to explore the performance of multiple features using Method 1, as we believed that multiple sources of information might offer the algorithm an increased robustness when dealing with more complex and realistic datasets. Similarly to Method 2 we chose to test the GMM-UBM with the LBP, HOG and GIST descriptors, besides the original SIFT formulation.

4. **Fusion strategies:** Fusion scenarios can contribute, in some complex situations, to an overall improvement of system performance. On the present work, two fusion strategies at score level were evaluated: performance-weighted score-level fusion and neural network score-level fusion. Both are novel to Method 1, which did not include any score-level fusion in its original formulation, while the simpler performance-weighted strategy is tested as an alternative to the original version of Method 2.

- *Performance-weighted score-level fusion:* the fusion score, s_f, is obtained by a weighted sum of the individual scores obtained for each feature individually:

$$s_f = w_{feat_1} \times s_{feat_1} + w_{feat_2} \times s_{feat_2} + ... + w_{feat_N} \times s_{feat_N} \tag{4.12}$$

  where s_{feat_n} is the individual score obtained for feature $n \in 1...N$ and w_{feat_n} is its corresponding weight in the final score. The weight of each feature is computed in relation to its individual performance:

$$w_{feat_n} = \frac{p_{feat_n}}{\sum_{i=1}^{N} p_{feat_i}} \tag{4.13}$$

  where p_{feat_n} is the individual performance obtained with a specific metric for feature n and the denominator term is introduced so that $\sum_i w_{feat_i} = 1$.

- *Neural Network score-level fusion:* the final recognition score, s_f, is obtained by a multilayer perceptron artificial neural network (MLP-NN), trained on a small data partition which is not included in the test phase. NN-based methods have been applied widely to classification problems because of their high learning capacities and good generalization (Santos et al., 2015). In the present study, two hidden layers NN were trained using back-propagation. The architecture of the NN was as follows: the first hidden layer had the same number of neurons as the number of individual scores derived from the matching stage, i.e. 4 for Method 1 and 11 for Method 2; the second hidden layer had 2 and 6 neurons for Methods 1 and 2 respectively, while the final (output) layer presented a single layer outputting the final s_f value.

Besides these strategies, we also analyzed the effect of fusing information from different channels of the *RGB* colorspace. Integration of information, in this case, was performed by treating each color channel individually and computing three independent recognition scores for each one: r_R, r_G and r_B. The final score was obtained by simple averaging of these three values.

4.2.3 Results and Discussion

The present section will serve as a detailed analysis of the comparison carried out between the two methodologies detailed throughout the last section. We start by offering some insight regarding the specific details of the multi-sensor periocular database on which both algorithms were assessed, as well as the experimental setups and performance metrics used for such assessment. We then present and discuss the main results regarding both recognition performance as well as processing time and possible limitations and advantages of each algorithm in real-world scenarios.

4.2.3.1 CSIP Database

The CSIP database, created for the assessment of the original version of Method 2 (Santos et al., 2015), is a recent and publicly available dataset, designed with the main goal of gathering perioc-

ular images from a representative group of participants, acquired using a variety of mobile sensors under a set of variable acquisition conditions. Given the heterogeneity of the camera sensors and lens setups of consumer mobile devices, 10 different setups were used during the dataset acquisition stage: four different devices, some of which had both frontal and rear cameras, and LED flash. This variety of sensors confers a strong appeal to the CSIP database regarding its potential use for the assessment of algorithms under a highly heterogeneous set of conditions. A summary of the technical details concerning each of such setups may be observed in Table 4.2 while a visual example of an image for each subset of the same individual is depicted in Figure 4.10. Each participant was imaged using all of the test setups.

Table 4.2: Technical details concerning the acquisitions setups used for each subset of the CSIP database.

Setup ID	AR0	AR1	BF0	BR0	BR1	CF0	CR0	CR1	DF0	DR0
Device	*A*		*B*			*C*			*D*	
Manufacturer	Sony Ericsson		Apple			ThL			Huawei	
Model	Xperia Arc S		iPhone 4			W200			U8510	
O.S.	Android 2.3.4		iOS 7.1			Android 4.2.1			Android 4.3.3	
Camera	Rear		Frontal	Rear		Frontal	Rear		Frontal	Rear
Resolution	3264×2448		640×480	2592×1936		2592×1920	3264×2448		640×480	2048×1536
Flash	No	Yes	No	No	Yes	No	No	Yes	No	No



Figure 4.10: Examples of images from each subset of the CSIP database. From (a-j) respectively: $AR0$, $AR1$, $BF0$, $BR0$, $BR1$, $CF0$, $CR0$, $CR1$, $DF0$ and $DR0$.

To simulate the variable noise associated with on-the-go recognition, participants were not imaged at a single location, but instead they were enrolled at multiple sites with artificial, natural, and mixed illumination conditions. In total, 50 participants were enrolled, all Caucasian and mostly males (82%), with ages ranging between 21 and 62 years (mean = 31.18 ± 9.93 years). For each periocular image acquired by the mobile devices, a binary iris segmentation mask was also produced. The masks were obtained automatically using the state-of-the-art iris segmentation approach proposed by Tan et al. (Tan et al., 2011), which is particularly suitable for uncontrolled acquisition conditions, as demonstrated by its first place ranking at Noisy Iris Challenge Evaluation - Part 1 (NICE.I) (Proença and Alexandre, 2007).

4.2.3.2 Experimental Setup

In order to achieve a fair comparison between both tested algorithms, a uniform experimental setup was defined and adapted to fit the specificities of each method. With that in mind, the set of

all images of the CSIP dataset was divided as follows: 50% of the images per individual and per subset were kept to either train the models in Method 1 or to serve as reference for each identity in Method 2; the remaining 50%, apart from a small independent set used to train the fusion neural networks, were used to assess the performance of both methodologies. Performance assessment was adapted to fit the nature of the originals algorithms:

- **Method 1:** given an input image I of an unknown source and an associated identity claim S, the score, r_s, associated with this image/claim pair is computed by the likelihood-ratio, $r_s = {proj(desc(I),IDSM_S)}/{proj(desc(I),IDSM_{UBM})}$, where $desc(I)$ represents the feature descriptor extracted from image I and $proj(X,GMM)$ represents the projection of feature vector X onto a specific GMM (either the claimed ID's IDSM or the UBM). This process is repeated for every possible ID, so that for image I the assessment block outputs a total of N scores, with N being the total number of individuals enrolled in the database.

- **Method 2:** given an input image I of an unknown source and an associated identity claim S, the score, r_s, associated with this image/claim pair is computed by the averaging of image/image pair similarities, $r_s = {\Sigma_{k=1}^{n_{ID}} score(I,I_{ID,k})}/{n_{ID}}$, where $score(I,I_{ID,k})$ is the comparison score obtained using Method 2 between the unknown image I and the k-th reference image from a specific known ID. The averaging is made in relation to the total number of reference images for the given ID, n_{ID}. This process is repeated for every possible ID, so that for image I the assessment block outputs a total of N scores, with N being the total number of individuals enrolled in the database.

After the N scores are extracted for each image, using both methodologies, performance is assessed for either identification or verification modes. For identification we chose to use the *rank-1 recognition rate* metric, while, on the other hand, for verification, we computed the *equal error rate*, as referred in Section 2.8.

4.2.3.3 Performance Comparison

The main results obtained for the setups outlined in the previous sections are summarized in Table 4.3. All results concern the average performance observed with 10-fold cross-validation for a specific methodology (GMM-UBM or Santos et al. (Santos et al., 2015)), pre-processing strategy (device specific color correction - DS-CC - or discrete cosine transform - DCT), and single feature (LBP, SIFT, GIST or HOG) or fusion of multiple features (performace-weighted sum-rule - PW-SR - or multilayer perceptron artificial neural network - NN).

Careful observation of the values presented in Table 4.3 allows for some interesting conclusions to be achieved. Regarding both tested metrics, there is no significantly better algorithm for all the tested subsets. While Method 2 achieves the best average identification performance (Table 4.3) amongst the whole set of tested subsets, the GMM-UBM algorithm still manages to achieve significantly better performances for two of such subsets - $CF0$ and $DR0$ - while managing to achieve values in a very similar range for five other subsets - $AR0$, $AR1$, $BR0$, $BR1$ and $CF0$.

Table 4.3: Rank-1 recognition rates obtained for each subset of the CSIP database for some variations of both tested methodologies.

				CSIP Subset									
	Pre-Proc.	**Feat(s)**	**Trait(s)**	**AR0**	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DF0**	**DR0**
M.1	DS-CC	PW-SR	P	88.4	97.8	73.5	86.5	93.1	85.0	72.1	92.1	45.5	81.7
	DCT	PW-SR	P	94.9	97.7	75.0	86.5	91.7	91.3	83.7	93.2	55.5	**91.7**
	DCT	RGB PW-SR	P	**97.4**	**100**	83.8	93.2	95.8	**92.5**	82.6	94.3	50.0	95.0
$r_1(\%)$	DCT	NN	P	94.3	99.4	83.8	89.3	96.8	84.2	79.8	94.2	54.8	82.9
M.2	DS-CC	LBP	P	89.8	**100**	75.6	81.7	91.8	75.3	79.6	96.8	59.4	72.0
	DS-CC	SIFT	P	81.8	**100**	73.2	84.2	**100**	61.8	61.3	96.8	82.8	57.3
	DS-CC	GIST	P	96.6	**100**	87.8	93.0	98.8	84.3	**92.5**	94.6	75.0	84.0
	DS-CC	HOG	P	68.2	92.6	56.1	64.6	72.9	42.7	55.9	78.5	40.6	54.7
$r_1(\%)$	DS-CC	NN	P+I	95.5	**100**	**92.7**	**95.3**	95.3	85.4	90.3	97.9	**76.6**	80.0
	DS-CC	PW-SR	P+I	93.2	**100**	90.2	91.4	98.8	87.6	81.7	**98.9**	73.4	82.7
M.1	DS-CC	PW-SR	P	3.3	3.0	8.7	**4.0**	1.4	6.2	8.0	4.5	17.4	7.5
	DCT	PW-SR	P	2.5	2.9	7.9	4.3	1.7	4.6	**5.1**	4.2	**14.8**	5.0
	DCT	RGB PW-SR	P	**1.7**	**0.5**	7.5	4.4	1.5	**4.5**	5.3	4.0	19.3	**4.1**
$EER(\%)$	DCT	NN	P	2.1	0.7	**7.3**	5.3	1.3	5.9	9.6	4.0	24.0	11.3
M.2	DS-CC	LBP	P	9.6	4.3	14.7	17.4	10.1	13.5	11.0	5.7	23.9	16.0
	DS-CC	SIFT	P	9.9	0.6	16.5	14.5	**0.1**	16.0	19.2	**2.1**	28.3	18.1
	DS-CC	GIST	P	5.2	3.3	11.1	12.0	6.3	10.3	8.9	5.3	20.2	12.0
	DS-CC	HOG	P	17.1	7.4	24.4	23.7	13.6	21.8	20.8	11.8	28.9	22.5
$EER(\%)$	DS-CC	NN	P+I	6.3	0.6	9.1	11.7	7.1	10.1	9.7	4.0	19.1	13.1
	DS-CC	PW-SR	P+I	7.5	0.8	10.4	12.2	5.9	9.3	10.5	3.5	20.5	12.6

In fact, only for the $BF0$, $CR0$ and $DF0$ subsets does the difference in performance between the two algorithms become significant. This non-uniformity in the relative behaviour of performance between the two methodologies might indicate that even though the sources of information and the feature descriptors used for its encoding are very similar, the modeling strategies used in both works adapt better for some acquisition scenarios. By visual observation of the images it is readily understandable that the $BF0$ and $DF0$ images are the ones that present lower resolution and overall image quality (Figure 4.10). The UBM modeling strategy might, therefore, not be able to aptly train GMMs capable of correctly and in a robust way describe such low quality data. As Method 2 uses direct matching algorithms that are optimized for each feature descriptor, the results in low quality data might be improved. On the other hand, when the quality of input images is a bit higher, as far as images acquired with mobile devices can go, Method 1 either reaches the same ranges of the results obtained with Method 2, or even exceeds it for moderately more complicated scenarios such as $CF0$ and $DR0$. A different set of observations can, however, be carried out by the analysis of the EER values used to assess performance in verification scenarios (Table 4.3). Here the GMM-UBM algorithm consistently achieves better performance regardless of the tested subset. This variable behavior might indicate that while Method 2 presents a higher discriminative power between individuals, the GMM-UBM approach, probably due to the inherent score normalization nature of the method, is more fit to distinguish between classes (genuine and impostor users) in an identity check application.

A behavior that is easily observable, regardless of the methodology and subset that we choose to focus on, is that images acquired with flash illumination present considerably better results than their non-flash counterparts. This observation was somewhat expected, as flash illumination might serve as a solution to overcome the variable lighting conditions that were referred as a natural limitation of mobile device acquisition in Section 4.2. To the extent of our knowledge no ocular health problems are commonly associated to overexposure to flash illumination in mobile devices, and with the growing technological advances in the manufacturing of such devices, flash illumination might play a crucial role in the implementation of image-based biometrics in mobile environments in the near future. Concerning pre-processing, Table 4.3 shows that DCT normalization far exceeds the performance obtained with the device-specific color-correction proposed in the original work with the CSIP database. Using a fixed transformation rather than a device-specific approach, that relies on the definition of new transformation matrices for each new device, presents a more robust and reliable alternative as far as the integration of periocular recognition in real-life applications is concerned. Furthermore, it can be seen that the results obtained using information from the three available color channels also results in a non-negligible increase in performance for a variety of subsets. Even though the recognition performance is increased, it must be noted that processing time is increased three times, as the same algorithm must be run in three separate instances. Even though these instances could be ran in parallel, the technological burden for such approach in mobile devices might exceed the current limits.

One more topic to take into consideration regards the alternative fusion strategies that were tested. From Santos et al.'s results, it can be observed that the use of neural networks over a simpler

performance-weighted approach, results mostly in non-significant variations in the performance. A similar set of conclusions can be drawn for the GMM-UBM results.When comparing this results with the ones obtained for individual features, however, the positive effect of fusion, regardless of its details, is readily discernible. The choice of the fusion strategy should, therefore, be constrained by the specific scenario of application and on how each strategy performs.

On a final note, regarding the processing time of each tested methodology, some considerations can be taken. Given an unknown image I and a single identity check, either by likelihood-ratio or average image similarity, the single-image processing time was computed and averaged for all test images. It was observed that Method 1 using DCT normalization and performance-weighted sum-rule fusion spent an average of 0.018 seconds on this process, whereas Method 2 spent an average of 0.130 seconds for an analogous computation, using device-specific color correction and neural network fusion. This discrepancy is easily explained by the larger amount of features used by Method 2 (11 vs. 4). Real-life applications based on periocular recognition are expected to work as fast as possible, so as to accurately replicate real-time operation. With that in mind, the GMM-UBM approach, with its uniform and fast matching algorithm based on GMM projections, seems to present an interesting alternative for further research. Even though the performance obtained for more unconstrained scenarios (for examples $DF0$) is still far from acceptable, future work on more robust representations or future improvements to the intrinsic architecture of the algorithm might help overcome such limitations. The same thought is applicable to Method 2, where a more efficient matching strategy might bring about considerable decreases in processing time, with no significant nefarious effect over performance.

While all results obtained from works developed on periocular recognition seem to show interesting promise with regards to identification potential as well as real-world applications, throughout the next section we return to a more traditional trait and detail a novel hierarchical framework to face recognition. The aim of such research was to achieve high robustness against some of the most nefarious scenarios for face recognition systems, namely occlusions and markedly low illumination.

4.2.3.4 Cross-sensor Recognition

Some preliminary experiments were also carried out in cross-sensor scenarios. In this alternative to the biometric recognition problem, enrollment and recognition are carried out using data acquired using different sensors. Some of the most interesting results obtained were that, for both methods, only some specific setups showed considerable interoperability. For example, using the GMM-UBM approach, and using the AR1 set for training and the BR1 set for testing, a recognition rate of 86.1% was achieved, whereas the single setup scenarios yielded 100% and 95.8% respectively. This relatively small loss in performance, when compared to other cross-sensor scenarios, might relate to similarities in the hardware of the rear cameras of devices A and B, as well as to the more uniform conditions in lighting, as a result of flash illumination. A similar behavior was observed for a few other pairs of setups - BR1/CR1, AR0/BR0 and BR0/DR0 for example - but, in general, a significant drop in performance is observed for cross-sensor scenarios, regardless of

the tested methodology. Further research is, therefore, needed to achieve stability in performance when enrollment and testing are carried out in highly variable acquisition conditions.

4.3 Hierarchical Face Recognition

It has been noted that the performance of face recognition algorithms is severely compromised when dealing with non-ideal scenarios, such as non-uniform illumination, pose variations, occlusions, expression changes and radical appearance changes (Min et al., 2014a). Whereas technological improvements in image capturing and transmitting equipment managed to attenuate most noise factors, partial face occlusions still pose a genuine challenge to automated face recognition (Li et al., 2014).

Facial occlusions may occur due to a multiplicity of deliberate or unintentional reasons. Whereas accessories, such as sunglasses and scarves, and facial hair represent quite common sources of occlusion in daily life, they can also be explored by bank robbers and shop thieves in an attempt to avoid recognition. Furthermore, the use of some accessories might be enforced in restricted environments (such as medical masks in hospitals and protection helmets in construction areas) or by religious or cultural constraints (Min et al., 2014a). The fact that humans perform and rely on face recognition routinely and effortlessly throughout their daily lives leads to an increased interest in replicating this process in an automated way, even when such limitations are known to frequently occur (Karande and Talbar, 2014).

Even though there is no consensus in the cognitive science field as to how the human brain recognizes faces, either based on their individual local features or, more holistically, on the basis of their overall shape (Tanaka and Farah, 1993), several works have shown that both levels of information play a non-negligible role in human face perception (Schwaninger et al., 2007; Gold et al., 2012). Whereas holistic representations provide a global summary of the spatial arrangement of contours and textures in an image, local features provide a more detailed regional description of the parts that compose it (Cho et al., 2014).

In the present section, we detail a robust alternative to face recognition under partial occlusions and variable illumination. An innovative hierarchical decision framework, incorporating both holistic and local descriptions, is proposed. The global precedent hypothesis for human perception (Navon, 1977) is the basis of this new decision strategy. Such a hypothesis claims that face recognition is performed by the human brain in a global-to-local flow, with holistic information gaining precedence over a more detailed local analysis. By following this rationale, we aim to replicate the cognitive process of face recognition by the human brain in an automated way. We evaluate the proposed algorithm on three widely-studied databases—the ORL, AR and the Extended Yale B databases—characterized by a variety of occlusions, small pose variations, facial expressions and illumination conditions [3].

[3]The work presented throughout this section resulted in the following journal publication: João C. Monteiro and Jaime S. Cardoso. *A Cognitively-Motivated Framework for Partial Face Recognition in Unconstrained Scenarios.* Sensors 15.1, 2015: 1903-1924.

4.3.1 Related Work

Face recognition has been a widely-studied research topic in the last few decades. The earliest research on automated recognition of human faces is dated to the mid-1960's, by Woodrow W. Bledsoe and his team at Panoramic Research (Bledsoe and Chan, 1965). While this work represented a semi-automatic process, with multiple steps requiring human supervision, it served as the basis for the most well-known approaches to be developed in the decades afterwards. Some of these traditional approaches, like eigenfaces (Turk and Pentland, 1991), Fisherfaces (Belhumeur et al., 1997) and active appearance models (Cootes et al., 1998), have become highly popular and laid out the foundations for a variety of commercial off-the-shelf systems. However, all of the previously mentioned techniques stumble upon the limitations presented in the last section: when non-ideal conditions are present during the acquisition step, recognition performance is severely compromised. The need to improve the state-of-the-art in face recognition to encompass a set of more realistic applications has, thus, been catalyzing research in the area to a set of new directions.

The study of invariant features to diminish the effects of occlusion, illumination and other nefarious sources of noise represents one of the most significant focuses of recent research. Liao *et al.* (Liao and Jain, 2011) tried to overcome the need for face alignment that characterizes most holistic approaches employing a multi-keypoint descriptor representation. In this work, any type of face image, either holistic or partial, can be probed for recognition, regardless of the global content. Nallammal and Radha (Nallammal and Radha, 2013) propose a non-negative matrix factorization (NMF) variation to explore the potential of the eyes and the bottom face regions for recognition when the probe images present a high degree of occlusion. An alternative approach is followed by Oh *et al.* (Oh et al., 2012), where random horizontal and vertical patches of face images are used as templates for cancelable identity verification. Such a technique intentionally distorts biometric information, in a repeatable, but non-reversible manner, to better deal with the compromising of biometric templates. Karande and Talbar (Karande and Talbar, 2014) address the problem of face recognition with large rotation angles and variable illumination conditions through the use of edge information for independent component analysis (ICA). The work by Geng and Jiang (Geng and Jiang, 2011) explores some of the known limitations of the widely-studied SIFT approach, for a specialized application in face image description. An alternative keypoint detection and a partial descriptor are both proposed in an attempt to adapt the traditional algorithm to non-rigid and smooth objects. Cho *et al.* (Cho et al., 2014) propose a two-step approach to face recognition, with principal component analysis (PCA) used at a coarser level and Gabor filtering at a finer level. This finer analysis in only carried out if the coarser recognition results do not present a high degree of reliability. Recently, Facebook's Deep Face Project claimed 97.25% accuracy, where humans achieve an accuracy of 97.53%. Their approach, published by Taigman *et al.* (Taigman et al., 2014), was based on deep neural networks, allowing the effective use of highly complex statistical models trained for large volumes of data.

Recently, approaches based on sparse representation classification (SRC) have shown impres-

sive performance in unconstrained face recognition and became one of the hot research topics in the area. The first reported use of SRC for face recognition, by Wright *et al.* (Wright et al., 2009), approached the problem of partial occlusion by representing face images as a linear combination of the whole face gallery and a vector of residuals at the pixel level. Classification was then achieved by l_1 minimization of the vector of residuals for each possible identity. Zhou *et al.* (Zhou et al., 2009) further improved this methodology by enforcing spatial coherence of occluded pixels through the use of Markov random fields (MRF). The spatial continuity of occlusions in face images was also explored by Qian *et al.* (Qian et al., 2014). Their methodology takes advantage of the low-rank error images that are originated in occluded images by traditional SRC methodologies, to perform effective and robust recognition when such conditions are observed. Besides spatial coherence of occlusions, some other topics have been explored in recent works as possible improvements to the original SRC proposal. Wang *et al.* (Wang et al., 2014) propose an Adaptive SRC (ASRC) approach capable of selecting the most discriminative samples for each representation, using joint information from both sparsity (l_1 minimization), as well as correlation (l_2 minimization). Shen *et al.* (Shen et al., 2014) propose a variation of SRC for implementation in Android and iOS mobile devices. Their proposal optimizes the computation of residual values with significant gain in computational efficiency and no considerable losses in recognition accuracy. Jian *et al.* (Jian et al., 2014) also center their attention on the computational speed limitations of the SRC approach. Their proposal, based on the orthogonal matching pursuit (OMP) algorithm, achieves fast and robust face recognition, even though the best results are only achieved through a preliminary occlusion detection block. Even though considerable work has been performed in the area, the main drawback regarding the SRC approach is still posed by the need for an extensive and diverse library of well-aligned face examples.

Another focus of research in recent years concerns the use of prior knowledge regarding occlusions in face images. Zhang *et al.* (Zhang et al., 2007) proposed an estimation of the probability distribution of occlusions in feature space using the Kullback–Leibler divergence (KLD). In a mixed approach regarding both previous detection of occlusions and SRC face recognition, Li *et al.* (Li et al., 2014) present a two-step SRC approach, where SRC is both used to first discriminate occluded pixels from unoccluded regions and then to perform face recognition. The use of downsampled images allows a significant improvement in processing speed. Min *et al.* (Min et al., 2014a), on the other hand, perform occlusion detection using MRF to promote the spatial coherence of the detected occlusion regions. The recognition step is then carried out solely on the non-occluded regions. Even though the *a priori* detection of occlusions may significantly improve the accuracy of local face recognition, the introduction of a new block in the recognition system may bring about a new set of problems. An increase in the computational cost of the process, as well as the creation of a new source for errors that may condition the recognition process from its earlier steps can be counted among such challenges.

We propose a robust approach to face recognition when non-ideal conditions, such as partial occlusions and severe illumination variations, affect the acquisition environment of the system. In an attempt to tackle most of the limitations presented in the works outlined in the last para-

graphs, we designed a new hierarchical recognition framework. This innovative approach allows a considerable reduction in the computational cost of the whole recognition process, while also allowing an intuitive integration of multiple region-based details. The proposed algorithm is able to achieve accurate face recognition, even when a limited set of images with small variations is used for model training.

4.3.2 Algorithm Overview

The proposed algorithm is schematically represented in Figure 4.11. Figures 4.11a and 4.11b depict the enrollment process in the proposed approach. During enrollment, a new individual's biometric data is inserted into a previously existent database of individuals. In the present work, a hierarchical ensemble of M partial face models is trained for each enrolled subject. The M individual-specific models are built by maximum *a posteriori* (MAP) adaptation of the corresponding set of M universal background models (UBM) using individual-specific data. The UBM is a representation of the distribution that a biometric trait presents in the universe of all individuals. MAP adaptation works as a specialization of the UBM based on each individual's biometric data.

The database is probed during the recognition process to assess either the validity of an identity claim (verification) or the k most probable identities (identification) given an unknown sample of biometric data. In the present work, we propose an innovative approach to the recognition process based on the global precedence hypothesis of face perception by the human brain. Recognition is performed hierarchically, as depicted in Figure 1c, with global models taking precedence over more detailed ones. Partial models are hierarchically organized into levels. Each level is composed by a set of non-superimposing subregions, I_l, of equal size (Levels 2–3 and 4–5 were hierarchically ordered in an arbitrary order, even though their composing regions are of equal size. Previous knowledge of expected types of occlusion could be explored when specifying this order.). Subregions at the same level sum to the full-face image, I_0. During recognition, a test image from an unknown source follows the hierarchical flow depicted in Figure 1c, until a decision can be made with a significant degree of certainty. The significance of a decision carried out at a single level is defined through the analysis of the likelihood-ratio values obtained for each possible identity claim. Decisions are made independently for each subregion at the same level, and only the most significant one is kept.

4.3.3 Consideration on Hypothesis Modeling

Contrarily to the periocular experiments described in the previous sections, the set of partial face models were trained on densely-sampled sets of scale-invariant feature transform (dSIFT) keypoint descriptors. Images were first exposed to an illumination normalization procedure using the Weber-face approach (Wang et al., 2011). Even though traditional SIFT descriptors present invariance to a set of common undesirable factors (image scaling, translation, rotation), thus conferring them a strong appeal in unconstrained biometrics, the fact that they fail to adapt to heterogeneous

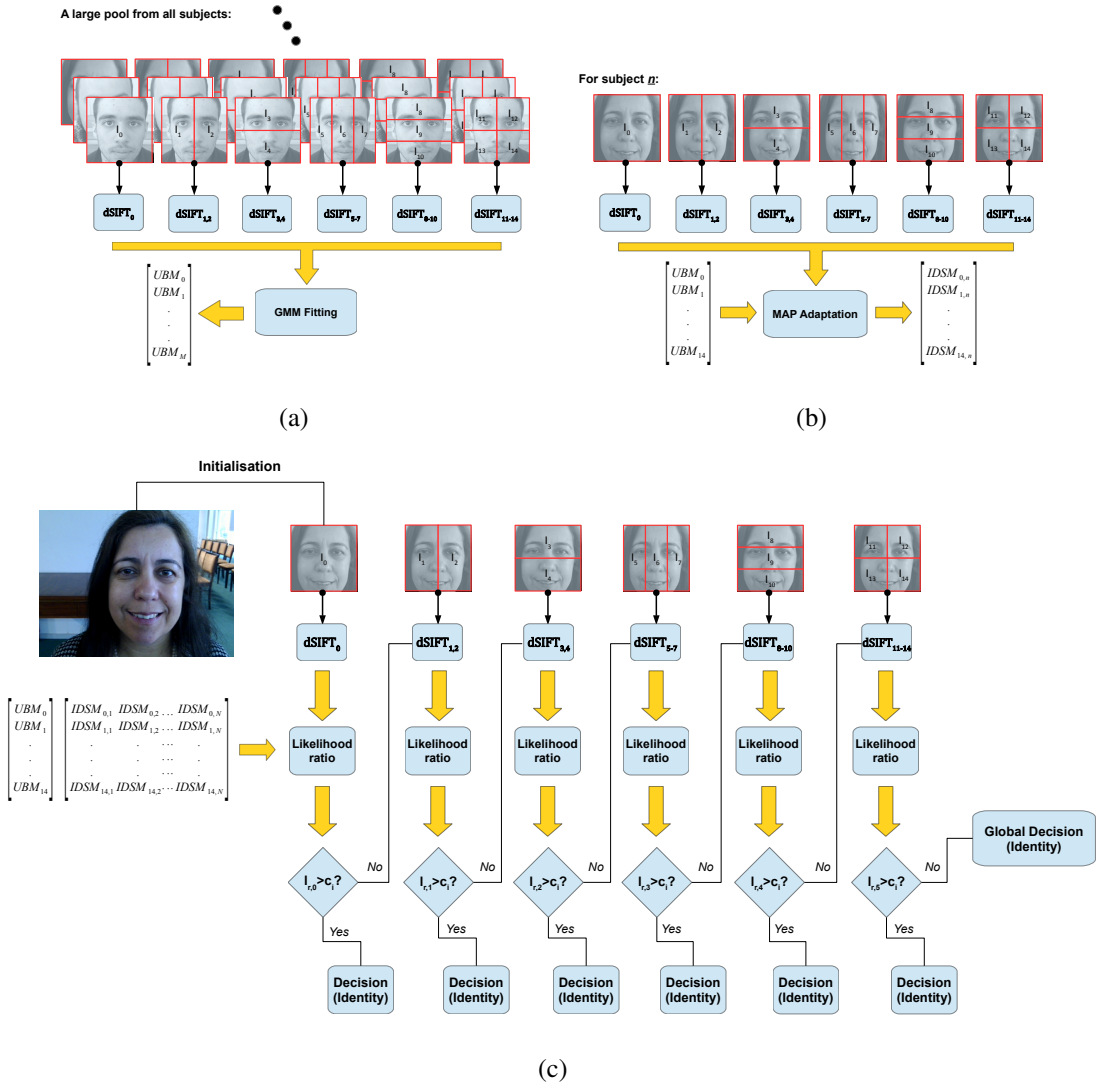


Figure 4.11: Schematic representation of the proposed algorithm and its main blocks: (**a**) training of the universal background models using data from multiple individuals; (**b**) maximum *a posteriori* (MAP) adaptation of the universal background models (UBM) to generate individual specific models; and (**c**) testing with new data from unknown sources.

illumination conditions severely hinders their practical use in real-life applications. The normalization step is, therefore, of the utmost importance.

Dense SIFT is a variation of the traditional SIFT methodology (Lowe, 2004), where keypoint descriptors are extracted in a roughly equivalent manner to running SIFT on a dense grid of locations at a fixed scale and orientation (Bosch et al., 2006). Dense sampling mitigates the potential errors introduced by the detection of an unreliable set of interest points in its sparse counterpart (Simonyan et al., 2013). In the present work, we train GMMs on the set of all densely-sampled keypoint descriptors from all individuals (UBM) and adapt individual models using data from specific subjects alone (IDSM), achieving a stable summary of the image content for every enrolled user.

4.3.4 Hierarchical Decision

The whole training process is repeated at most M times for each of M facial subregions, defined *a priori*, and is identical, for each region, to the GMM-UBM modeling strategy presented above in Section 4.1.4. We considered the $M = 14$ regions previously exemplified in Figure 4.11.

Traditionally, the recognition phase with new data from an unknown source is a fairly simple process. The new test data, $X_{test} = \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,N}\}$, where $\mathbf{x}_{t,i}$ is the i-th PCA-reduced SIFT vector extracted from a given subregion m of test subject t, is projected onto both the UBM and either the claimed IDSM (in verification mode) or all such models (in identification mode). The recognition score, $s_{t,m}$, is obtained as the average likelihood-ratio of all keypoint descriptors $\mathbf{x}_{t,i}$, $s_{t,m} = \frac{1}{N}\sum_{i=1}^{N} s_{t,m}^{(i)}$. The decision is then carried out by checking the condition presented in Equation (4.1), in the case of verification, or by detecting the maximum likelihood-ratio value for all enrolled IDs (Equation (4.2)), in the case of identification.

In an attempt to integrate meaningful information from the M facial subregions and to deal better with partial or missing data situations originated by occlusions, an hierarchical recognition framework is proposed. The rationale behind the methodology described below can be easily understood if some studies of the human brain's cognitive mechanisms of perception are taken into consideration. One such work, proposed in 1977 by David Navon (Navon, 1977), describes the aforementioned mechanism as a hierarchical process, where holistic representations precede more detailed local features. If we interpret this perception mechanism in the scope of human face recognition, we can conclude that an attempt to classify the face at a global level is the starting point to the whole recognition process, whereas increasingly detailed descriptions are only taken subsequently if necessary. This conceptual description of the human perception mechanism serves as the basis for the proposed hierarchical recognition algorithm, whose flowchart is presented in Figure 4.11c. The main steps of the proposed hierarchical identification algorithm are as described below:

1. **Initialization:** Starting with the full-face image from an unknown user, I_0, a densely-sampled grid of SIFT keypoint descriptors is extracted; the likelihood values, $l_{t,0}^{IDSM_t}$ and $l_{t,0}^{UBM_0}$, are then computed for the IDSM of every enrolled user $t \in \{1..T\}$ and the UBM of

the tested region; the recognition scores for every possible identity, $\boldsymbol{s}_{t,0} = \{s_{1,0}, ..., s_{T,0}\}$, are then computed through Equation (4.1).

2. **Certainty index computation:** The certainty index of a given region, c_m, measures how likely it is that the obtained vector of likelihood ratios, $\boldsymbol{s}_{t,m}$, corresponds with the ideal case of a single correct identity match. If no false positives corrupt the vector of likelihood ratios obtained for a single image, a significant difference will be observed between the highest value, $\boldsymbol{s}_{t^*,m}$, (true identity) and the average of all other values, $\frac{1}{T-1}\sum_{t=1,t\neq t^*}^{T}\boldsymbol{s}_{t,m}$, (average impostor). The certainty index can thus be interpreted as a degree of separability between these two quantities:

$$c_m = \boldsymbol{s}_{t^*,m} - \frac{1}{T-1}\sum_{t=1,t\neq t^*}^{T}\boldsymbol{s}_{t,m} \tag{4.14}$$

3. **Decision to go to next level:** If the c_m value exceeds a previously optimized threshold, θ_l, the maximum likelihood-ratio decision is accepted. When $c_m < \theta_l$, however, the algorithm will consider that an analysis at a more detailed level is necessary to achieve a decision with a higher degree of confidence. At this point, the algorithm proceeds to the next level, working on subregions I_{1-2}, the second in the hierarchical chain depicted in Figure 4.11. When one level is composed by multiple subregions, each one of them is treated independently, and only the maximum c_m value among them is considered for the decision criterion:

$$c_m = \begin{cases} c_0 & \text{if } l = 1 \\ \max\{c_1, c_2\} & \text{if } l = 2 \\ \max\{c_3, c_4\} & \text{if } l = 3 \\ \max\{c_5, c_6, c_7\} & \text{if } l = 4 \\ \max\{c_8, c_9, c_{10}\} & \text{if } l = 5 \\ \max\{c_{11}, c_{12}, c_{13}, c_{14}\} & \text{if } l = 6 \end{cases} \tag{4.15}$$

4. **Repeat:** Steps 1 to 3 are hierarchically repeated for every level until $c_m > \theta_l$. If all L levels are considered and none is able to achieve a significant decision, the decision corresponding to the highest value of c_m, $\max_m(c_m)$, amongst all L levels is considered.

As the proposed algorithm is capable of performing recognition without the need of processing all subregions, computational speed is significantly improved over simpler approaches that explore the fusion of all local recognition scores. Furthermore, the proposed algorithm is capable of automatically deciding if a more detailed exploration of local features is necessary or if the information obtained up to a certain point is enough to make a decision. This autonomy alongside the intuitive notion of the global precedence hypothesis are the most notorious strengths of the proposed recognition algorithm.

An alternative to deciding non-classified images at the end of the hierarchical chain was also considered. In this new setting, if no level is capable of making a decision according to the aforementioned criteria, images are kept as "doubtful", and no decision is made, in a process similar to classification with a reject option (Herbei and Wegkamp, 2006). This approach may be thought of as a viable alternative for real-life applications, where feedback to the user can be explored to adjust the environmental conditions in severely unconstrained scenarios.

4.3.5 Experimental Tests and Results

The algorithm was tested on the Database of Faces (formerly "The ORL Database of Faces"), the Extended Yale Face Database B and the AR face database. All tested databases are widely known for their diversity of pose, illumination and occlusion conditions, respectively. The next sections outline the main features of each database, as well as the experimental setup chosen for training and testing with each of them.

4.3.5.1 ORL

The Database of Faces (formerly "The ORL Database of Faces") (Samaria and Harter, 1994) contains 400 images from 40 subjects, divided equitably with a total of 10 images per individual. Images were taken at different points in time with variable lighting, expression and pose conditions. For performance assessment, we use a single sample from each individual for training and the remaining nine images for testing. This process is repeated for each possible training image, and the performance is computed as the average of the 10 runs. An alternative approach is also explored by using multiple templates per subject instead. In this new setup, the first five samples per subject are used for training, while the remaining five samples are used for testing. Example images from two subjects may be observed in Figure 4.12.



Figure 4.12: Example images from two subjects of the ORL database.

4.3.5.2 Extended Yale B

The Extended Yale Face Database B (Georghiades et al., 2001) is composed of 2432 images corresponding to a total of 38 individuals. All images are frontal faces acquired under varying illumination conditions. The database is divided into five subsets, numbered 1 to 5, according to the ranges of angles between the light source direction and the camera axis. An example of all images from a single subject may be observed in Figure 4.13.

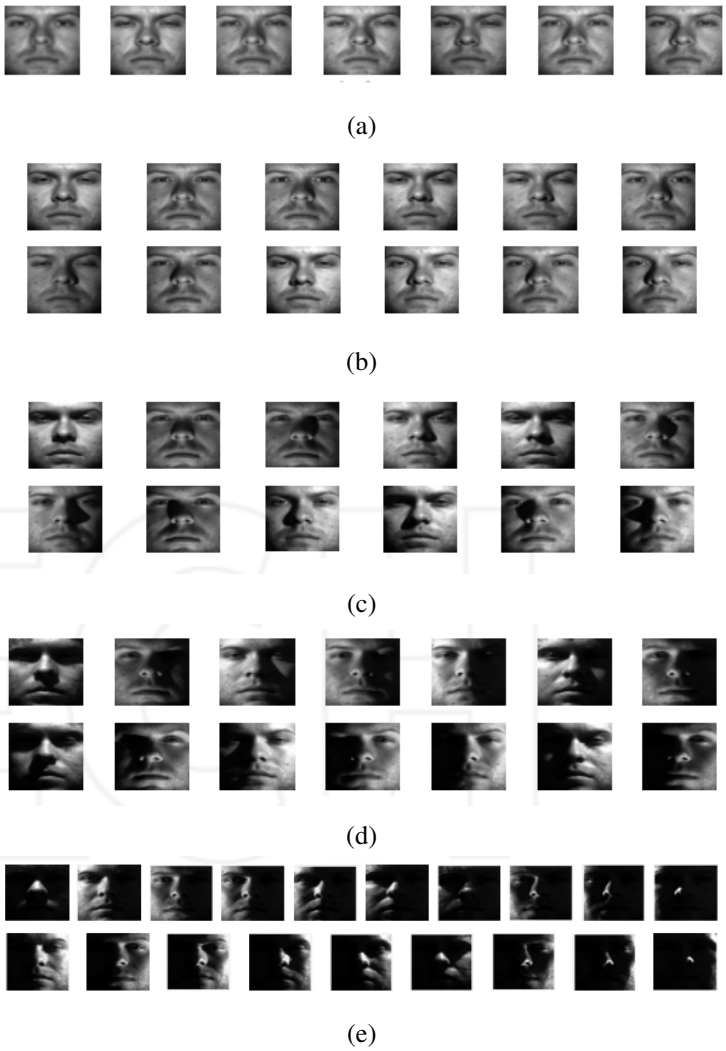


Figure 4.13: All images from a single subject enrolled in the Extended Yale B database. Images (**a**) to (**e**) correspond to Subsets 1 to 5, respectively.

All images from Subset 1 were used for model training, while all other subsets were tested independently, so as to better assess the robustness of the proposed algorithm to a variety of illumination conditions.

4.3.5.3 AR

The AR database (Martinez, 1998) contains over 4000 frontal face images from 126 individuals, acquired under variable illumination, expression and occlusion. Occlusions can be divided into two main categories: sunglasses and scarf. An example of all images from a single individual is presented in Figure 4.14.

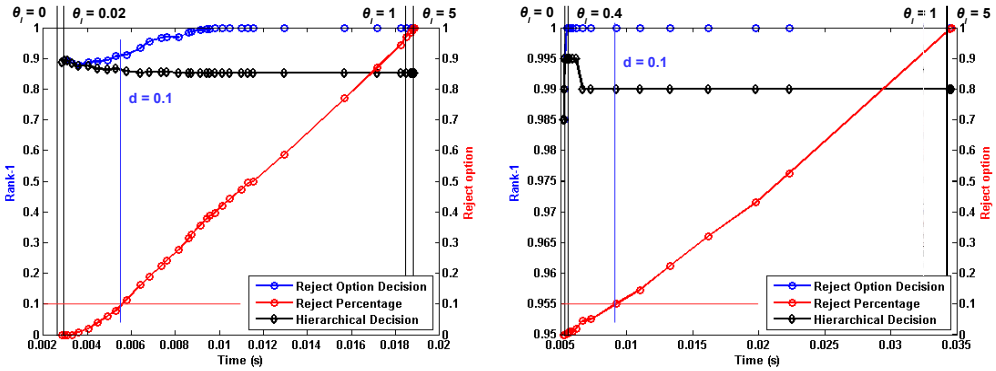
All unoccluded images from every individual are chosen to train both the local UBMs and IDSMs. The remaining scarf and sunglasses occluded images are tested separately, so as to better analyze the consistency of the algorithm when exposed to variable types of occlusion.



Figure 4.14: Example images from one subject of the AR database.

4.3.5.4 Performance Analysis

Figures 4.15–4.17 depict the most relevant results obtained by the proposed algorithm for the ORL, Extended Yale B and AR databases, using the previously mentioned training and testing setups. Furthermore, Tables 4.4–4.6 present a comparison between the proposed work and the reported performance in some recent works, performed under similar experimental conditions. We chose to assess the rate of correctly identified individuals, by checking if the true identity is present among the N highest ranked identities. The N parameter is generally referred to as rank. This allows us to define the Rank-1 recognition rate, r_1, as the recognition rate at $N = 1$.



(a) (b)

Figure 4.15: Main results obtained with the proposed methodology for the sunglasses and scarf images of the ORL face database. Each plotted point represent a specific value of parameter θ_l, ranging from $[0,\infty]$.

Each plotted point refers to a single θ_l value, ranging from $[0,\infty]$. For each tested θ_l value, we plotted a series of performance metrics against the corresponding average processing time per image. Using the $\theta_l = 0.1$ vertical line from Figure 7a as a reference, we can distinguish three metrics that are common to all tested setups. The black line represents the evolution of r_1 when "doubtful" images are classified through the aforementioned criterion, $\max_m(c_m)$. When no such *a posteriori* classification is performed, a set of images is left unclassified, as no level from the hierarchical chain was shown to present enough detail to perform an accurate recognition. The red line represents the evolution of the "doubtful" image percentage with increasing θ_l values. It is intuitive to note that lower θ_l values generate lower reject-option classification ratios, as a smaller

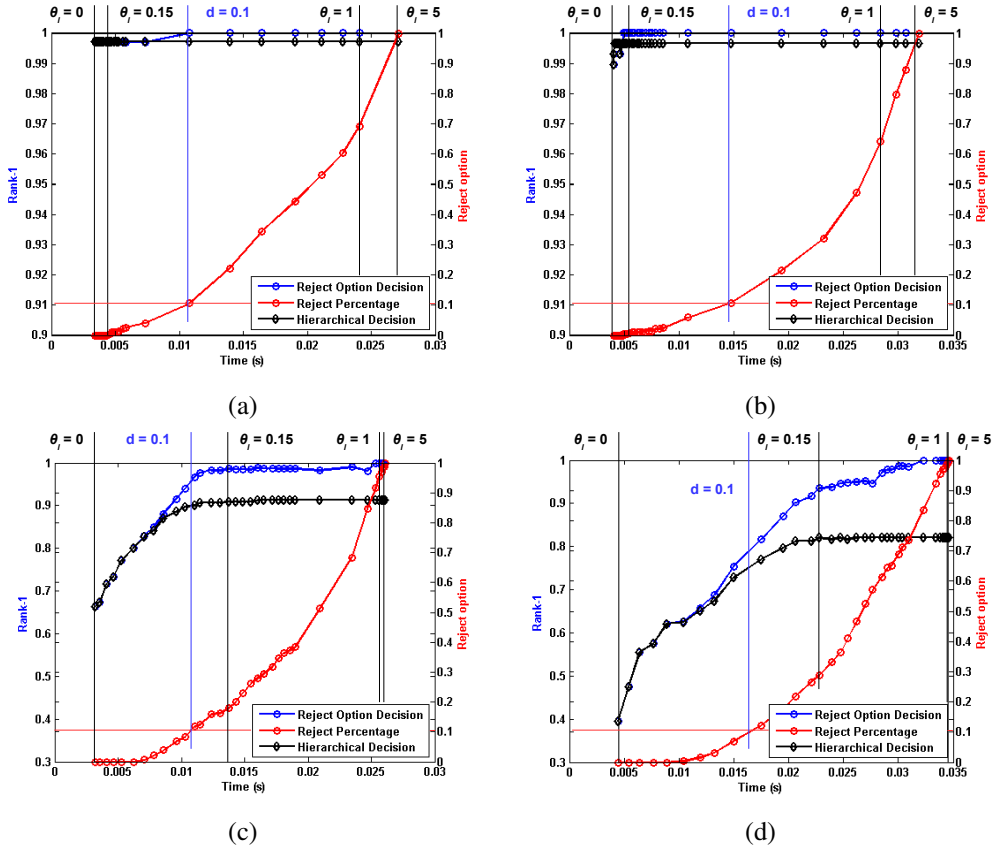


Figure 4.16: Main results obtained with the proposed methodology for the sunglasses and scarf images of the Extended Yale B face database. Each plotted point represent a specific value of parameter θ_l, ranging from $[0, \infty]$.

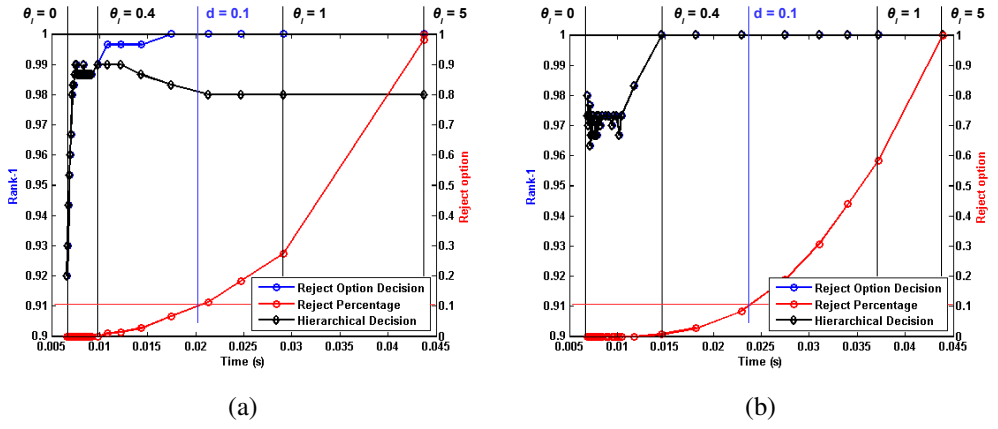


Figure 4.17: Main results obtained with the proposed methodology for the sunglasses and scarf images of the AR face database. Each plotted point represent a specific value of parameter θ_l, ranging from $[0, \infty]$.

dissimilarity between individuals will still trigger a "certain" decision. If no "doubtful" images are considered in the performance evaluation of the proposed algorithm, then the r_1 computation is performed only with respect to the non-rejected images. The blue line represents this alternative

Work	**Single Template $\mu \pm \sigma$**	**Multiple Templates**
Proposed$_{\theta=0.02\|0.15}$	**89.45**% $\pm$ **1.24**%	**99.5**%
Proposed$_{\theta\to\infty}$	85.79% $\pm$ 1.85%	99.00%
Proposed$_{d<0.1}$	*93.60% $\pm$ 1.85%*	*100.0%*
Geng and Jiang (Geng and Jiang, 2011)	89.0%	**99.5**%
Zhang *et al.* (Zhang et al., 2014)	–	85.0%
Xiao *et al.* (Xia et al., 2014)	–	97.50
Xu *et al.* (Xu and Lee, 2014)	–	97.10

Table 4.4: Comparison between the obtained r_1 values for the ORL database and some state-of-the-art algorithms.

Work	**AR Sunglasses**	**AR Scarf**	**AR Average**
Proposed$_{\theta=0.4}$	**100.0**%	99.00%	**99.50**%
Proposed$_{\theta\to\infty}$	**100.0**%	98.00%	99.00%
Proposed$_{d<0.1}$	*100.0%*	*100.0%*	*100.0%*
Min *et al.* (Min et al., 2014a)	75.00%	92.08%	83.54%
Li *et al.* (Li et al., 2014)	97.00%	52.00%	74.50%
Morelli Andrés *et al.* (Morelli Andrés et al., 2014)	97.46%	99.15%	98.31%
Qian *et al.* (Qian et al., 2014)	99.00%	**100.0**%	**99.50**%

Table 4.5: Comparison between the obtained r_1 values for the AR database and some state-of-the-art algorithms.

Work	**Yale B SS_2**	**Yale B SS_3**	**Yale B SS_4**	**Yale B SS_5**
Proposed$_{\theta=0.15}$	99.71%	99.66%	90.91%	81.99%
Proposed$_{\theta\to\infty}$	99.71%	99.66%	91.20%	81.99%
Proposed$_{d<0.1}$	*100.0%*	*100.0%*	*96.68%*	*81.66%*
Cho *et al.* (Cho et al., 2014)	**100**%	**100**%	99.2%	—
Yang *et al.* (Yang et al., 2014)	—	—	90.2%	47.9%
Jiang *et al.* (Jiang et al., 2014)	**100**%	99.12%	83.46%	57.84%
Jia *et al.* (Jia et al., 2013)	**100**%	**100**%	**97.15**%	**96.13**%

Table 4.6: Comparison between the obtained r_1 values for the Extended Yale B database and some state-of-the-art algorithms.

evaluation, and the corresponding reject percentage can be easily traced to the equivalent θ_l point in the red line. Taking, once again, the $\theta_l = 0.1$ vertical line from Figure 7a as a reference, we can see that the proposed methodology yields a r_1 value of roughly 98.00%. If the approximately 30% "doubtful" images are not considered for evaluation, the r_1 value increases to 100.00%, with respect to the non-rejected images.

For a better comprehension and deeper analysis of the method, we chose to assess performance in three specific points:

(1) Optimal $\boldsymbol{\theta_l}$ value: We consider the optimal θ_l value as the point where a visible performance

plateau is achieved in the time *vs.* performance plot. For the *AR* database, this value was set to $\theta_l = 0.4$, whereas for the Extended Yale B Face database, it was set to $\theta_l = 0.15$. For the ORL database, the θ_l value was optimized for each of the aforementioned experimental setups. For the single template approach, the value was set to $\theta_l = 0.02$, while for multiple templates, it was set to $\theta_l = 0.15$.

(2) $\boldsymbol{\theta_l \rightarrow \infty}$ **:** extreme behavior when the θ_l parameter is set to high values.

(3) $\boldsymbol{d < 0.1}$ **:** point where the ratio of non-classified images at the end of the hierarchical chain reaches 10% of all tested images.

A thorough analysis of Figures 4.15–4.17 leads to some interesting conclusions regarding the behavior of the proposed algorithm under variable image acquisition conditions. The performance metric r_1 was computed for a set of θ_l values, ranging from $[0,\infty]$. It is clear that lower values of θ_l will lead the proposed algorithm to an extreme case where all images are classified in the first level of information, *i.e.*, the full face, thus reducing the computational complexity and average processing time. On the other hand, when $\theta \rightarrow \infty$, all images will reach the end of the hierarchical chain without a certain decision having been made. In this opposite extreme behavior, all images are classified *a posteriori* using information from every level, with a significant increase in both computational complexity and performance. With such a wide variety of possibilities, we chose to analyze the global behavior of the proposed work by plotting the evolution of r_1 values against the average processing time, when the θ_l parameter goes from $[0,\infty]$.

Regarding the evolution of r_1 for variable θ_l values, we might point out the considerable drop in performance under less ideal conditions, when $\theta \rightarrow 0$. Whereas Subsets 2 and 3 from the Extended Yale B database show excellent performance from both aforementioned approaches, rivaling those obtained for $\theta \rightarrow \infty$, such behavior is highly compromised for the more challenging scenarios of Subsets 4 and 5. In such cases, the observed drop in performance is less significant for higher θ_l values. Furthermore, the results obtained for the AR database stress that the proposed methodology is capable of consistently presenting high performance regardless of the acquisition conditions and noise factors. Whereas the Extended Yale B database presented the challenge of heterogeneous illumination, which might be understood as a "natural" source of occlusion, the AR database presents the challenge of spatially coherent occlusion regions. For both cases, performance observed for the proposed hierarchical methodology rivals the state-of-the-art. Regarding the $\theta \rightarrow \infty$ case, the expectation that the best results would be consistently observed for this extreme scenario was fulfilled, at the expense of higher computation complexity and average processing time. Nevertheless, if no time constraints exist in the application scenario in which the proposed methodology is implemented, higher values of θ_l seem the ideal choice. The ORL database maintains a somewhat stable performance regardless of the chosen value for the θ_l parameter. As the proposed hierarchical methodology is contextually motivated for occlusion scenarios and no such conditions are present in this database, the observed behavior fits the expectations. Regardless of the absence of occlusions in the ORL images, the reject option classifier

still proves a useful tool at discriminating "doubtful" images and increasing the reliability of the decisions obtained for the "non-doubtful" ones.

Tables 4.4–4.6 present a comparative analysis between the proposed methodologies and some state-of-the-art works, in similar experimental setups, for the AR and Extended Yale B databases, respectively. It is readily observed that the proposed algorithm presents the most consistent and robust behavior, regardless of the nature of the present occlusions, for the AR face database. While the work by Li *et al.* (Li et al., 2014) presents higher performance for sunglasses occlusion, their performance regarding the alternative scarf occlusion is considerably lower than the one obtained with the proposed algorithm. Min *et al.* (Min et al., 2014a) present a work that suffers from exactly the opposite problem, with good performance observed for scarf occlusion, but lower results for sunglasses. The aforementioned works seem to suffer from overfitting to certain classes of images and lack the robustness to adapt to new cases. Such robustness to the nature and location of the occlusion can be observed for both the proposed methodology, as well as the works from Morelli *et al.* (Morelli Andrés et al., 2014) and Qian *et al.* (Qian et al., 2014). Both of these works show a similar trend to the proposed algorithm and high performance for the whole database. Alongside our work, and to the best of our knowledge, they represent the state-of-the-art performance for the AR database in unconstrained face recognition. A similar analysis can be made for both setups under which experiments were carried out with the ORL database. The best performance obtained with the proposed algorithm is on par with the best value found in the literature, in the work by Geng and Jiang (Geng and Jiang, 2011). This observation leads to the conclusion that the proposed methodology is capable of achieving state-of-the-art performance for a wider variety of scenarios besides occlusion. The aforementioned observation that performance seems to be independent of the chosen value for the θ_l parameter may, however, indicate that a simpler approach than the whole hierarchical chain might also output good results. Using simply the first level, *i.e.*, the full-face images, the r_1 values observed for the single and multiple template setups were $88.10\% \pm 1.05\%$ and 98.5%, respectively. These results differ very little from the optimal results presented in Table 4.4, thus corroborating the expectation of good performance for a less computationally complex approach.

Regarding the Extended Yale B Face database, our algorithm is shown to perform similar to state-of-the-art performance for all subsets, except Subset 5. Even for this subset, the obtained performance is still higher than a few recently published works, even though no direct comparison can be performed with the work by Cho *et al.* (Cho et al., 2014), whose authors claim that its images are "excluded from the experiment because the illumination conditions are so severe that some images are difficult to recognize even to the naked eye". Observing the behavior of the blue and red lines for high threshold values, we note that a significant amount of images in this subset exceeds the recognition capability of the proposed methodology. The fact that the blue and black lines diverge greatly, when compared to the results observed in other subsets, shows that posterior classification by the maximum c_m value may not be the most beneficial approach when working with such severely degraded images. In such cases, the reject option classification yields considerably higher performance than the posterior classification of doubtful cases, even

for low rejection ratios. It is easily noted how the blue and black lines start to diverge for a value of $d << 0.1$, as a result of this behavior.

4.4 Conclusion

All the contributions presented throughout this section refer to a single biometric trait acquired with the same sensor for both the enrollment and identification steps. When distinct sensors are considered for each of the steps, it is questionable whether an inter-operable behavior will be observed, with the aforementioned approaches being capable of adapting, with no need for alterations in their architecture, to the distinct acquisition characteristics of the newly observed data. The next chapter focuses on cross-sensor scenarios, where an attempt to achieve inter-operable behavior was performed, basing the work on the theoretical principles of transfer learning.

Chapter 5

Transfer Learning for Cross-Sensor Biometrics

With the increasing popularity and availability of mobile devices, capable of performing the whole biometric recognition framework, from data acquisition to final decision, a new obstacle is presented to the development of such systems: the need to adapt to the wide variety of available sensors and the resulting heterogeneity with regards to image quality. The question of whether or not sensors from different manufacturers show a high degree of interoperability allowing, for example, for an individual to be enrolled in a single system and, then, be successfully recognized in a vast variety of alternative devices, is of growing importance in the research field of biometrics. With recent studies showing that cross-sensor matching, where the test instances are verified using data enrolled with a different sensor, often lead to reduced performance, we attempt to overcome this challenge by making use of transfer learning principles and, thus, achieve state-of-the-art performance for a large variety of acquisition scenarios.

Transfer learning is an approach in which the knowledge acquired in solving one task is used to solve a new target task without having to redo the whole training procedure. It is anticipated that new tasks and concepts are learned more quickly and accurately by exploiting past knowledge. Such approach gains an increased interest given that the learning of each task in isolation represents an expensive process, requiring large amounts of both time and data. With this idea in mind, the focus of transfer learning falls on the leveraging of information from existing sources to train new models with increased efficiency. Transfer learning has found a series of practical applications in a vast array of research fields, such as text classification from one writing style to another (Dai et al., 2007), cross-domain video concept detection (Yang et al., 2007), customers sentiment classification over time (Blitzer et al., 2007), biomedical labeling for genes, proteins and biological entities (Dahlmeier and Ng, 2010) or sensor based location estimation (Pan et al., 2008).

Aiming to the issue of cross-sensor biometric recognition, we explore and extend the Source-Target-Source (STS) approach, first proposed in (Kandaswamy et al., 2015), while applying it to

the specific challenge of cross-sensor periocular recognition[1]. STS is a recent alternative that has shown increased performance in computer vision object recognition tasks, as well as a significant gain in processing speed. We start this chapter by assessing the state-of-the-art in transfer learning and cross-sensor biometric applications.

5.1 Related Work

The issue of learning from few training instances has motivated several works on classification in recent years. The NIPS-95 workshop on *Learning to Learn* unveiled the need for machine learning methods that retain and reuse previously learned knowledge, thus initiating the development of a vast array of transfer learning (TL) algorithms.

In (Caruana, 1997), the author introduced the notion of multitask learning where the knowledge from a task is improved by using the information contained in the training instances of other related tasks. This methodology enables the learning of multiple tasks in parallel, by taking advantage of shared representations, thus simultaneously improving generalization for all learnt tasks. The TL algorithm known as lifelong learning (Thrun, 1998) is based on the assumption that a learner faces multiple learning problems during its lifetime. Thus, when learning the n-th task, a learner can reuse knowledge gathered in the previous $n-1$ tasks to boost the generalization ability. The cross-domain learning or domain adaptation (see (Daume III and Marcu, 2006), (Glorot et al., 2011), (Ben-David et al., 2010), (Bruzzone and Marconcini, 2010)) algorithm, in which a machine learns to perform a task on training instances drawn from a source problem, and then performs the same task on a target problem, whose instances are drawn from a related distribution. Domain adaptation expects that the closer the distributions are the better the features trained on the source problem will perform on the target problem.

Deep Transfer Learning (DTL) is a framework that combines deep learning models with transfer learning algorithms (see (Ciresan et al., 2012), (Yosinski et al., 2014), (Kandaswamy et al., 2014)). Deep learning networks are trained on the source problem and then layer-by-layer feature transference is performed to solve a target problem in either a supervised (Ciresan et al., 2012) or unsupervised (Kandaswamy et al., 2014) setting. The advantage of DTL is that it offers a far greater flexibility in extracting high-level features that are transferred from a source to a target problem while, unlike the classical approach, not being affected by experts bias (Bengio et al., 2013).

The practical problem of cross-sensor biometrics has also been the focus of many works in recent years, highly motivated by the growing variety and availablity of mobile sensors. The most commonly found works concern mostly iris recognition. Connaughton et al. (Connaughton et al., 2011) performed a comparison between three commercially available iris cameras, with the aim of assessing the interoperability between them and the impact of some state-of-the-art recognition algorithms in both single as well as cross-sensor scenarios. The authors arrived at some crucial

[1] The work presented in this section resulted in the journal paper: Chetak Kandaswamy, João C. Monteiro, Luís M. Silva and Jaime S. Cardoso. *Multi-source deep transfer learning for cross-sensor biometrics.*

conclusions, namely the fact that the relative performance of a given algorithm in a variety of single-sensor scenarios does not relate reliably to the performance of the same algorithm when tested in cross-sensor scenarios. Furthermore, performance observed for all cross-sensor scenarios was consistently worse than their single-sensor counterparts. Another recent work on the field of iris recognition, proposed by Pillai et al. (Pillai et al., 2014), attempted to adapt iris instances acquired with one sensor to the characteristics of a new sensor, in an attempt to mitigate the performance-drop commonly observed in cross-sensor scenarios. Both Santos et al. (Santos et al., 2015) and Jilela and Ross (Jillela and Ross, 2014) propose methods based on information extracted from the periocular region. While Santos et al. propose a framework based on multiple descriptors to work on periocular data on multiple mobile sensors, Jilela and Ross attempt to match iris and face images from the same individual, acquired with distinct sensors, using periocular traits to help in the recognition process. The performance drop in cross-sensor scenarios is also corroborated by some simple experiments carried out in our work (Monteiro et al., 2015), described in more detail in Section 4.2.3.4, where two state-of-the-art recognition algorithms fail to cope with the variations introduced to their input data when faced with cross-sensor scenarios.

In the present chapter and throughout the next sections we will detail the baseline and transfer learning alternatives to the cross-sensor scenario. We chose to make use of the CSIP dataset due to its inherent cross-sensor nature and, thus, narrowed our attention to periocular recognition.

5.2 Cross-Sensor Methodologies

We set the GMM-UBM algorithm, described in Section 4.1, as the baseline for assessment of the proposed transfer learning alternatives in cross-sensor scenarios. It is important to note that this work was designed with single-sensor recognition in mind, i.e. the source and target data are drawn from the same senor. In the present work we assess its performance in cross-sensor scenarios, where training of models and classification are carried out on distinct data sources. Throughout this section we will more thoroughly analyze and discuss such results, as well as present a comparative analysis with alternative approaches more tuned for the cross-sensor application:

- **GMM Supervectors with SDA:** In the baseline work, recognition was carried out through the computation of a likelihood ratio between a target IDSM and the UBM. Recently, a significant amount of works have explored the use of an alternative GMM representation - GMM supervectors - as the input for classification algorithms, with some promising results being reported in the literature (Campbell et al., 2006). Super-vector notation consists on concatenating in a single vector all the parameters describing a GMM (weights, means and covariance matrices). For example, the mean values of the UBM can be concatenated to form a single mean super-vector, m, given by $m = [\mu_{T1}, \mu_{T2}, ..., \mu_{Tk}]$, where k is the total number of mixtures in the UBM (Ge et al., 2015). A similar representation can be extracted for the IDSM parameters or even for single images. On the present work we describe each training image t belonging to subject i, $Im_{t,i}$, by its supervector representation, obtained by

MAP adaptation of the UBM parameters using the feature data extracted solely from $Im_{t,i}$. SIFT keypoint descriptors are used for feature description and model training, as described in Section 4.1.4. We then perform training, validation and classification using the SDA methodology for both TLs and STS approaches, as detailed in Section 3.2.3.

- **CNN:** A CNN-based methodology was also carried out, as described in Section 3.2.4, using raw pixel intensity values. The architecture of our 5 layer CNN model has [Conv - Pool] x 3 -FC -LR. We first resize the images to 200×120 and then convert them to greyscale to serve as input for the designed CNN. This input is convolved with 30 different 1st layer filters, each of size 12×12, using a stride of 1 in both directions. The resulting feature maps are then pooled in (max within 2×2 regions, using stride 1) yielding 30 different 94×54 element feature maps. Similar operations are repeated with 60 and 90 different layer filters in 2nd and 3rd layers respectively. The 4th layer is fully connected, using features from the top convolutional layer as input, in vector form. The final layer is a c-way logistic regression classifier, where c corresponds to the number of classes.

5.3 Transfer Learning Approaches

Given a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ drawn from input space X and a set of labels Y, a classifier is any function $f(\boldsymbol{x}) : X \rightarrow Y$ that maps instances $\boldsymbol{x}_i \in X$ to labels. Our classifier is a deep network with K layers: $K-1$ hidden layers and one output classifier layer (logistic regression). The deep network is thus composed of a set of weights $W = \{\boldsymbol{w}^1, \boldsymbol{w}^2, \ldots, \boldsymbol{w}^K\}$ where $\boldsymbol{w}^k$ is an $M_{k-1} \times M_k$ weight matrix, M_k is the number of neurons in the k-th layer and M_0 is the number of inputs of the network. We define the baseline classifier (BL) as the one trained directly to solve the target problem without any type of transference, thus starting from a random initialization of the weight matrices. Classifier performance measures such as prediction accuracy or computation time are measured on a test set X_{test}.

5.3.1 Transfer Learning supervised (TLs)

We assume that the source dataset D_S, with input space X_S, and a set of labels, Y_S, is drawn from a distribution $P_S(X)$, and the target dataset D_T, with input space X_T and a set of labels Y_T is drawn from a distribution $P_T(X)$. In TLs we transfer fine-tuned source weights to the target network. Fig 5.1 depicts such transference between the first layers of both source and target networks. As a first step we randomly initialize each layer of the network using a uniform distribution (Glorot et al., 2011):

$$\boldsymbol{w}_S^k = \mathscr{U}\left[\frac{-\sqrt{6}}{\sqrt{M_{k-1}+M_k}}, \frac{\sqrt{6}}{\sqrt{M_{k-1}+M_k}}\right] \tag{5.1}$$

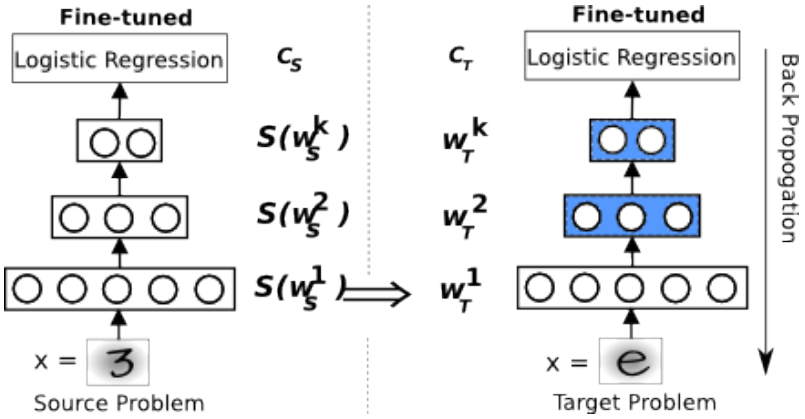


Figure 5.1: TLs for a simple example of a first layer feature transference.

Next, we perform training, $S(\cdot)$, on these unsupervised features for the source problem, using labeled instances as follows:

$$S(\boldsymbol{w}_S^1, \ldots, \boldsymbol{w}_S^K) \tag{5.2}$$

Feature transference is then applied by mapping the fine-tuned source features to the target network. Due to the possibility that the number of classes between the problems may differ, the logistic regression layer from the source problem cannot be reused. A new randomly initialized logistic layer is added instead of the original one, for the target problem. The entire deep network, with the new logistic regression layer, is then fine-tuned using target data as a multi-layer perceptron using back-propagation.

5.3.2 Source-Target-Source (STS)

In a previous work by Kandaswamy et al. (Kandaswamy et al., 2015) an alternative source-target-source (STS) approach was proposed. As previously discussed, the main idea of transfer learning is that the knowledge learned in a source problem may provide a good initialization for the learning task in a target problem, serving as an improvement over random initialization of the learning parameters for the target domain. In (Kandaswamy et al., 2015) the authors proposed an iterative learning between both domains. The intuition is that, like in typical metaheuristics in optimization, such as tabu search and simulated annealing, moving the learning process from one domain to the other will "shake" the current local optimal solution, allowing a continuous exploration of the space of solutions and, ideally, reaching a better solution in the process. The solutions reached in each iteration are kept and only the global best solution achieved is considered. The pseudocode of the STS approach may be consulted in Algorithm 1.

5.3.3 Multi-source Source-Target-Source

An extension of the STS transfer learning methodology presented in the last section was developed, by reusing knowledge learnt by a model trained on multiple sources. As previously men-

Algorithm 1 Pseudocode for STS

Initialize with trained features D_T:
Two datasets D_S and D_T, with associated labels Y_S and Y_T are drawn from P_S and P_T distributions.
Select D_S dataset to train
Baseline: train network A on D_S with target labels Y_S
Set maximum number of cycles, max cycles
Set list of error, errors list
for M in max cycles **do**
    Transfer: Transfer weights from network A to new network B
    Compute error and update error list
    if cycle = odd number **then**
        STS M = test error for Dataset D_S
    else
        STS M = test error for Dataset D_T
    end if
    if error $<$ avg(errors list) **then**
        BREAK
    end if
    Switch between dataset D_S and D_T
end for

tioned in Section 5.1, two different approaches have attempted to account for the reuse of multiple sources for TL: a) lifelong learning (Thrun, 1998) and b) multitask learning (Caruana, 1997). Both approaches are based on specific TL scenarios and assume that the data and the tasks are related.

Both previously mentioned approaches suffer from serious limitations. For example, if two tasks are negatively correlated, the learning process will cause degradation of the generalization performance of both tasks. In order to avoid such issues, a strong task selection is required in order to constrain the application of such methods to a limited set of positively correlated problems. The Multi-Source Source-Target-Source (MS-STS) approach improves generalization performance over multiple problems, with no need for prior task selection.

The MS-STS approach is briefly explained in the pseudocode presented in Algorithm 2. Consider a pool, *Pool*, containing multiple datasets from a similar problem or particular application, $Pool = \{D_A, D_B, \ldots, D_Z\}$ drawn from $P_A(X), P_B(X), \ldots, P_Z(X)$, where Z is number of datasets. We select a deep neural network architecture and initialize the weights of each layer of the network using a uniform distribution under the limits defined in step 1 of Algorithm 1. Initializing the weights through this method narrows down the gradient search parameter thus speeding up the training of the network (Glorot et al., 2011). Heuristically, we set the max number of cycles to 10. This value may suffer variations according to the nature of the problem. In step 2, we select a desired target dataset D_{dT} from the *Pool* for which we intend to have the best overall accuracy. During each of H cycles, a target dataset D_T is selected among the pool of datasets, for which we apply the Source-Target-Source approach, as discussed in (Kandaswamy et al., 2015), by selecting layers to transfer and/or to lock in the new network.

Algorithm 2 Pseudocode for MS-STS

Randomly initialize the weights of every layer of the network using uniform distribution:

$$\mathbf{w}^k \sim \mathscr{U}\left[\frac{-\sqrt{6}}{\sqrt{M_{k-1}+M_k}}, \frac{\sqrt{6}}{\sqrt{M_{k-1}+M_k}}\right]$$

Select a desired target dataset D_{dT} from the $Pool = \{D_A, D_B, \ldots, D_Z\}$.
for h in H cycles **do**
  for p in *Pool* **do**
    $D_T = p$ {set p as target dataset}
    Transfer: Transfer weights from network A to new network B and select which of the layers to transfer and which layers to lock, out of the K possible layers
    for k in K **do**
      $\boldsymbol{w}_S^k \Rightarrow \boldsymbol{w}_T^k$ {Transfer selected layers}
      $\boldsymbol{w}_S^k \Updownarrow \boldsymbol{w}_T^k$ {Lock selected layers}
    end for
    Train and test the new network with D_T
    Update the test accuracy list
    if accuracy of D_{dT} > avg(top 3 best test accuracy in D_{dT} accuracy list) **then**
      BREAK
    end if{Continue MS-STS step 3 till global optima is reached for the D_{dT}.}
  end for
end for

The new network is trained and tested as a regular deep network for the selected D_T. A list of best accuracies for each cycle is maintained for every dataset in the pool. If the current cycle test accuracy for the desired target dataset D_{dT} is greater than the average of the top test accuracies, for the desired target dataset D_{dT}, we break the cycle and store the final weights of the network. By training and testing serially on multiple datasets, we improve the domain generalization property of the approach, while by focusing on the desired target dataset we aim to improve the domain specialization property.

It is necessary to clarify exactly what is meant by MS-STS, as all of the above mentioned methods also take advantage of multiple sources to train the network. MS-STS proposes to extend the established STS methodology with multiple sources instead of a single source. The intuition is that, by providing multiple initialization points for exploring the space, we may reach a solution closer to the optimal. This search may result in an increase of the computational cost, but if a trade-off can be found between classification performance, system robustness and computational complexity, this might not end up as a significant limitation.

5.4 Experimental Tests and Results

The previously outlined methodologies were assessed on the Cross-Sensor Iris and Periocular (CSIP) dataset, already presented in Section 4.2.3.1. The following experimental setup was considered for performance assessment.

5.4.1 Experimental setup

5.4.1.1 Image pre-processing

Images from the CSIP database were converted to grayscale and resized so as to present a fixed number of pixels, necessary for the implementation of all the approaches based on the CNN methodology. Resizing was carried out in such a way that geometrical proportions were kept from the original images.

5.4.1.2 Data partitioning

In order to achieve a fair and meaningful comparison between the tested methodologies, a common experimental setup was designed. The set of all images of the CSIP dataset was divided as follows: 50% of the images per individual and per subset were kept for model training, 25% were chosen for validation of the trained models and the remaining 25% were used to assess performance. Train, validation and test subsets were randomly selected and all experiments were cross-validated 10 times.

5.4.1.3 Evaluation metrics

Performance was evaluated only for identification problems, where, given a biometric sample from an unknown source, the p most probable identities are assessed. For such problems the most commonly found performance metric is the rank-1 recognition rate, which refers to the ratio of correctly assessed identities, when $p = 1$.

5.4.2 Cross-sensor recognition performance

The main results obtained for the experimental setups detailed in the last section are summarized in Tables 5.1 through 5.7. Discussion of these results will be carried out, from this point onwards, starting with the BL and TLs approaches, followed by an analysis on how the STS strategy may improve performance in cross-sensor scenarios and, finally, on the effect that multiple sources of information may present.

5.4.2.1 Baseline and Transfer Learning

The baseline results for each tested methodology (GMM-UBM, SV-SDA and CNN) are presented in the diagonal values of Tables 5.1, 5.2 and 5.3 respectively. By the sole analysis of these results some conclusions may already be drawn. First of all it is easily discernible how the GMM-UBM methodology, specifically designed to solve the single-sensor periocular recognition problem, outperforms both alternatives in such conditions. Even for the CSIP subsets that, in theory, offer the least challenging conditions ($AR1$, $BR1$ and $CR1$), the performance drop observed is non-negligible. Taking $AR1$ as a specific example a relative performance drop of 26.6% and 18.3% is observed in the SV-SDA and CNN methodologies respectively. This effect is, however, reversed when cross-sensor scenarios are taken into consideration.

If we consider a single target dataset, it is readily observable that the variance in performance is a lot less pronounced for the CNN and SV-SDA methodologies than for GMM-UBM. Furthermore, it also notorious how the significantly better single-sensor scenario results of the GMM-UBM are severely degraded when a more complex challenge is presented to the algorithm. A trivial conclusion can be taken from such observations: even though the GMM-UBM presents the best baseline results, as expected from an algorithm tailored for that specific challenge, the application of transfer learning to both CNN and SV-SDA methodologies results in a considerably lower variance in the performance values observed for a single target dataset, regardless of the chosen source. A valid deduction, following such conclusions, is that improving the baseline performance of such methodologies will also result in an increased cross-sensor performance. Given that the challenge of cross-sensor biometric recognition is mostly concerned with the performance loss observed in such cases, the global behaviour of the tested methodologies seems to, at least, motivate further research seeking to improve the baseline performance.

In the next section we explore the effect that the source-target-source approach presents over the simpler TLs alternative.

Table 5.1: Rank-1 recognition rates, in %, observed for the GMM-UBM algorithm for all possible cross-sensor scenarios in the CSIP database.

	Target									
	AR0	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DR0**	
AR0	**94.4**	57.8	34.1	76.8	45.9	41.8	52.8	40.7	65.6	
AR1	64.6	**97.1**	26.5	47.9	83.6	25.5	35.4	66.8	33.7	
BF0	33.1	23.1	**78.2**	21.1	19.4	30.8	24.0	19.8	16.0	
BR0	67.4	39.6	19.7	**92.4**	54.2	36.5	42.3	34.3	67.7	
BR1	31.8	62.3	12.0	48.1	**95.5**	28.3	25.6	52.5	35.3	*Source*
CF0	36.4	29.1	34.7	36.5	30.8	**89.8**	55.8	39.8	46.3	
CR0	59.5	30.2	24.4	58.1	36.4	59.3	**80.3**	45.7	71.9	
CR1	42.6	64.9	21.2	47.8	70.8	47.5	50.5	**90.0**	49.0	
DR0	41.3	18.0	17.4	53.0	23.1	30.3	39.8	24.8	**88.7**	

Table 5.2: Rank-1 recognition rates, in %, observed for the SV-SDA methodology and TLs approach for all possible cross-sensor scenarios in the CSIP database.

	Target									
	AR0	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DR0**	
AR0	38.0	**81.8**	39.3	**33.4**	79.3	14.1	**27.7**	52.5	20.0	
AR1	35.4	76.7	22.7	24.6	77.3	15.9	20.9	**61.6**	17.6	
BF0	**43.8**	81.4	**41.4**	27.3	80.1	**22.3**	17.8	59.7	21.5	
BR0	35.2	79.0	35.8	25.6	77.3	14.6	21.8	45.9	22.1	
BR1	32.4	78.4	37.7	25.1	**82.1**	22.1	18.0	55.6	23.0	*Source*
CF0	32.4	79.8	36.4	24.1	75.3	12.1	16.3	50.6	**24.6**	
CR0	41.5	79.0	36.6	24.8	81.6	16.1	23.5	51.7	23.9	
CR1	28.7	79.0	36.6	22.1	78.0	16.4	18.2	57.9	22.4	
DR0	28.9	80.4	39.0	29.0	72.2	15.0	22.0	52.5	19.7	

Table 5.3: Rank-1 recognition rates, in %, observed for the CNN methodology and TLs approach for all possible cross-sensor scenarios in the CSIP database.

	Target									
	AR0	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DR0**	
AR0	61.5	84.2	**54.0**	**69.0**	91.0	**64.4**	63.9	92.0	**65.3**	
AR1	60.8	82.1	52.0	66.0	88.5	58.1	63.9	92.0	56.7	
BF0	**65.9**	84.5	50.0	68.5	89.5	62.6	67.4	92.3	**65.3**	
BR0	63.0	82.6	52.4	63.5	88.0	61.8	67.0	92.0	56.7	
BR1	63.0	85.0	52.4	60.5	85.5	60.4	64.3	90.3	57.3	*Source*
CF0	64.4	85.0	50.8	63.5	**91.5**	54.8	66.1	93.0	58.0	
CR0	62.2	**86.8**	51.2	66.5	90.5	60.7	67.9	**93.3**	62.7	
CR1	61.4	82.9	50.4	66.5	89.5	57.8	65.2	88.0	57.3	
DR0	60.0	83.2	53.2	62.0	86.5	55.2	**68.7**	90.0	53.3	

5.4.2.2 Source-target-source

As detailed in Section 5.3.2, we propose a cyclic source-target-source (STS) approach for classification using the CNN and SV-SDA methodologies. Tables 5.4 and 5.5 present the STS results observed for the SV-SDA methodology for a single cycle (STS_1) and for a total of 10 cycles, respectively. Analogous results for the CNN methodology may be observed in Tables 5.6 and 5.7. For a simpler analysis the baseline results are kept on the diagonal of each table, as in the last section.

The first observation to be taken from the analysis of the aforementioned tables is how even a single cycle of STS can significantly improve some of the baseline results. The $CF0$ baseline for example, presents a relative improvement of 18.2% for the CNN methodology, and most of the observed results already exceed those observed for the simpler TLs approach. This improvement is even more discernible when multiple STS cycles are carried out. The results presented in Tables 5.5 and 5.7 depict this behaviour. Here, and taking the same $CF0$ baseline result as referred before, the performance, comparing to the baseline, is increased to 29.0%. It is interesting to note how the stability observed in the last section, when a single target dataset is considered, is also observed in this approach, with the addition of significantly increased performance. The same conclusion can, thus, be achieved: if a stronger baseline performance is achieved, STS approaches to classification seem to present the capability of both improving the baseline performance, as well as guaranteeing the maintenance of such performance when different acquisition scenarios are considered.

Another consideration to be taken from the analysis of these results is how significantly worse the results using supervector-based SDA classification are when compared with their CNN counterparts. This observation can also be made from the analysis of Tables 5.1, 5.2 and 5.3 from last section. The most obvious explanation concerns the fact that the supervector representation based on the GMM modelling of SIFT keypoint descriptors might not present enough discriminative information for accurate SDA classification, except in some specific cases. For example, the datasets composed by higher quality images ($AR1$ and $BR1$) present considerably better per-

formance, even surpassing the performance of their CNN counterparts. These results show that some discriminative power exists, even though it seems severely compromised when the quality of the input images decreases. Regardless of that, the STS behaviours described above still remain relevant for the SDA methodology, and may earn some further research regarding the use of more adequate feature representation techniques.

As a final approach we also explored the effect of using information from multiple sources in order to improve the performance of the cyclic STS algorithm. The main results and observations regarding this approach will be outlined in the next section, in an attempt to summarize all the results and observations obtained in the present work.

Table 5.4: Rank-1 recognition rates, in %, observed for the SV-SDA methodology and a single cycle of the STS approach for all possible cross-sensor scenarios in the CSIP database.

	Target									
	AR0	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DR0**	
AR0	38.0	89.1	**54.4**	**31.1**	90.6	16.0	**27.7**	59.6	27.3	
AR1	40.8	76.7	42.1	29.2	92.2	**17.0**	25.4	**68.9**	**32.0**	
BF0	42.3	**91.6**	41.4	30.3	90.8	16.5	22.8	63.9	29.3	
BR0	**43.1**	88.7	48.2	25.5	90.6	14.0	27.9	62.5	25.7	
BR1	41.8	90.0	50.3	28.4	82.1	16.8	23.5	64.1	26.0	*Source*
CF0	39.2	88.0	47.4	30.0	89.4	12.0	25.4	63.9	25.0	
CR0	42.6	88.0	50.9	30.0	**93.1**	16.5	23.5	66.1	26.7	
CR1	37.4	89.6	50.0	28.9	90.0	18.3	25.1	57.8	26.0	
DR0	41.8	88.4	53.2	30.2	90.0	16.8	25.6	61.6	19.7	

Table 5.5: Rank-1 recognition rates, in %, observed for the SV-SDA methodology and the STS approach for all possible cross-sensor scenarios in the CSIP database.

	Target									
	AR0	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DR0**	
AR0	38.0	91.3	**54.4**	**38.7**	90.6	19.0	**30.9**	59.6	**32.7**	
AR1	40.8	76.7	42.1	29.7	93.9	**19.8**	25.4	**71.6**	32.0	
BF0	**48.5**	**91.6**	41.4	32.2	90.8	**19.8**	23.5	65.9	31.7	
BR0	43.1	88.7	48.2	25.5	90.6	17.3	27.9	62.5	31.7	
BR1	43.1	90.0	50.3	29.5	82.1	19.0	26.1	64.3	30.3	*Source*
CF0	40.0	88.0	47.4	33.8	89.4	12.0	26.1	63.9	28.7	
CR0	44.4	88.0	50.9	32.2	**93.1**	**19.8**	23.5	66.1	29.3	
CR1	37.4	90.0	50.6	28.9	92.2	18.3	25.1	57.8	30.0	
DR0	43.3	88.4	53.2	31.9	90.0	16.8	29.1	61.6	19.7	

5.4.2.3 Multiple Source STS

Figure 5.2 summarizes both the results obtained for the STS algorithm using multiple sources (MS-STS) as well as all the most relevant results presented in the last sections. The main goal of MS-STS is to achieve a high degree of domain generalization, in order to allow the trained

Table 5.6: Rank-1 recognition rates, in %, observed for the CNN methodology and a single cycle of the STS approach for all possible cross-sensor scenarios in the CSIP database.

	Target									
	AR0	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DR0**	
AR0	61.5	84.5	**55.2**	**70.0**	87.5	63.0	68.3	92.3	59.3	
AR1	64.4	82.1	50.4	66.5	88.5	59.3	68.7	91.0	60.0	
BF0	63.7	84.0	50.0	67.5	88.0	**64.8**	66.1	92.0	57.3	
BR0	62.2	84.0	54.0	63.5	**89.5**	64.1	67.0	92.0	61.3	
BR1	64.4	84.2	52.4	65.0	85.5	60.4	**70.9**	**93.3**	54.7	*Source*
CF0	61.5	84.5	53.6	64.0	88.0	54.8	69.6	90.7	62.0	
CR0	63.3	84.5	52.0	67.5	88.5	63.0	67.9	90.0	62.0	
CR1	**65.2**	85.3	52.8	68.5	86.5	64.1	68.7	88.0	**64.0**	
DR0	63.7	**85.5**	50.4	65.5	89.0	62.6	64.8	91.7	53.3	

Table 5.7: Rank-1 recognition rates, in %, observed for the CNN methodology and STS approach for all possible cross-sensor scenarios in the CSIP database.

	Target									
	AR0	**AR1**	**BF0**	**BR0**	**BR1**	**CF0**	**CR0**	**CR1**	**DR0**	
AR0	61.5	88.7	**58.4**	72.5	91.0	70.0	67.8	92.3	70.0	
AR1	71.9	82.1	55.2	69.5	**92.5**	66.7	70.9	92.0	66.7	
BF0	**73.3**	90.3	50.0	71.5	91.5	**70.7**	69.1	92.0	71.3	
BR0	72.6	89.0	57.6	63.5	90.5	69.6	70.4	92.7	**72.0**	
BR1	71.5	90.8	54.8	69.0	85.5	64.4	**72.2**	**93.3**	64.7	*Source*
CF0	72.6	90.5	57.2	72.5	91.5	54.8	70.0	92.3	69.3	
CR0	72.2	90.8	**58.4**	**73.5**	91.0	68.9	67.9	92.3	69.3	
CR1	70.7	**92.1**	56.4	72.5	90.5	65.9	70.4	88.0	68.0	
DR0	71.1	89.7	57.2	70.0	91.0	67.0	68.7	92.3	53.3	

classifiers to perform well for the widest possible variety of scenarios. For the multiple source examples we chose the flash subsets ($AR1$, $BR1$ and $CR1$) as the sources and all other no-flash datasets as the targets. This choice can be motivated by the fact that the three flash datasets consistently presented the best absolute performance among all the experiments that we carried out. Such observation seems to indicate that the intrinsic discriminative power of such datasets might be higher than the remaining alternatives, thus conferring them, at least in theory, a marked advantage as choice for source datasets. We also chose to work only with the CNN methodology, as the vast majority of the results observed in the last section seemed to point towards its better fit for the problem at hand.

So as to better visualize and understand the effect of the MS-STS approach over the approaches presented in the last sections we decided to present the results in the radial plot representation that can be observed in the 6 images from Figure 5.2. For each image a series of features can be observed:

- **Source and Target Datasets:** Each of the axis of the radial plot represents the rank-1 recognition rate (in %) for the chosen target (positive vertical axis) as well all the three

source datasets.

- **BL, TLs, STS_1 and STS:** The BL label represents the baseline performance as already presented in the diagonal values of Tables 5.1 to 5.3. TLs and STS_1/STS, on the other hand, represent the best results for each of the 4 depicted datasets (3 sources and 1 target), for each of their individual TLs and STS_1/STS experiments (bold values in Tables 5.2 to 5.7). STS_1 represents the performance after a single cycle of the STS approach, whereas STS_{10} refers to the best performance observed after 10 cycles.

- **STS (3 sources):** This label depicts the best performance observed for the target dataset, for its individual STS_1/STS experiment, considering only $AR1$, $BR1$ and $CR1$ as a possible source. We chose to include this label in order to achieve the fairest possible comparison between the MS-STS performance and the optimal single-source experiment.

- **MS-STS_1 and MS-STS:** The two polygons, in blue and red respectively, represent the first cycle and optimal performances, after 10 repetitions of the whole multi-source cyclic process, for each of the source and target datasets.

By the analysis of the plots, independent of the chosen target dataset, a few interesting conclusions can be drawn. First, there seems to be no significant performance change, regarding the target dataset, between the MS-STS (after 10 cycles) approach and the analogous results for STS using only the best single source from the $[AR1, BR1, CR1]$ set of sources. What the MS-STS offers is a way of achieving this optimal performance without the need of an empirical choice of the best source subset, thus conferring a more robust nature to the whole process. This is also the main advantage of the multiple source approach when compared to the optimal STS performance obtained when considering all 8 possible sources for a specific target: as the only way to achieve the best individual performance for a given target dataset is to extensively test all possible sources and, then, choosing the best, the real-word applicability of an approach based on STS will be limited by the amount of available data sources. By using the proposed multiple source approach we can achieve, with high confidence, a performance for the chosen target similar to the individual best observed among all the chosen sources. This observation, however, does not compensate the fact that by manually choosing a single optimal source, the performance observed for the chosen target dataset is consistently better or, in the worst case, in a similar range to the one observed for MS-STS. Further research is needed in order to optimize the choice of source datasets so as to reduce this performance gap.

Another interesting observation concerns the effect of the order in which the sources $AR1$, $BR1$ and $CR1$ are considered during the cyclic evolution of the MS-STS process. In order to assess whether this order had any discernible effect over the observed performance we chose to run, for each target dataset, a set of six variants of the original results, changing the order in which the three sources are organized during a single cycle: $[[A,B,C],[A,C,B],...,[C,B,A]]$. The performance plots from Figure 5.3 seem to point to the conclusion that the performance in all 4 datasets converges to a set of values in very similar ranges, regardless of the chosen organization

of source datasets along the MS-STS pipeline. This observation leads to the conclusion that, if the best sources are found, there is no need to optimize their order. Whereas the presented example was considerably simple, with a very small number of sources, in a practical application there is no guarantee that the number of combinations becomes unfeasible for a brute force optimization step of their organization. The observed results seem to indicate that this optimization process might be less relevant, especially in scenarios such as the tested, where all sources present a relatively similar nature (flash illumination in this specific case). It is still unclear, due to the preliminary nature of this study how increasing variability in the source dataset conditions would affect these observations. The focus of future research should, thus, fall on the optimal choice of sources so that the most complete domain generalization is achieved. With this in mind, the aim of future work would be to accurately and intelligibly perform classification under highly variable scenarios, especially using more heterogeneous sets of source information.

5.5 Conclusion

During this section we proposed a series of alternatives to help solve the cross-sensor biometric recognition problem. However all experiments were carried out in data of similar nature. Even though images acquired for each of the 10 subsets presented marked visual differences, they were still all RGB representations of the same regions. In the next section we expand the idea of recognition using data acquired from multiple sensors to the case where each sensor outputs independent and inherently distinct representations of the biometric trait, such as depth or infrared.

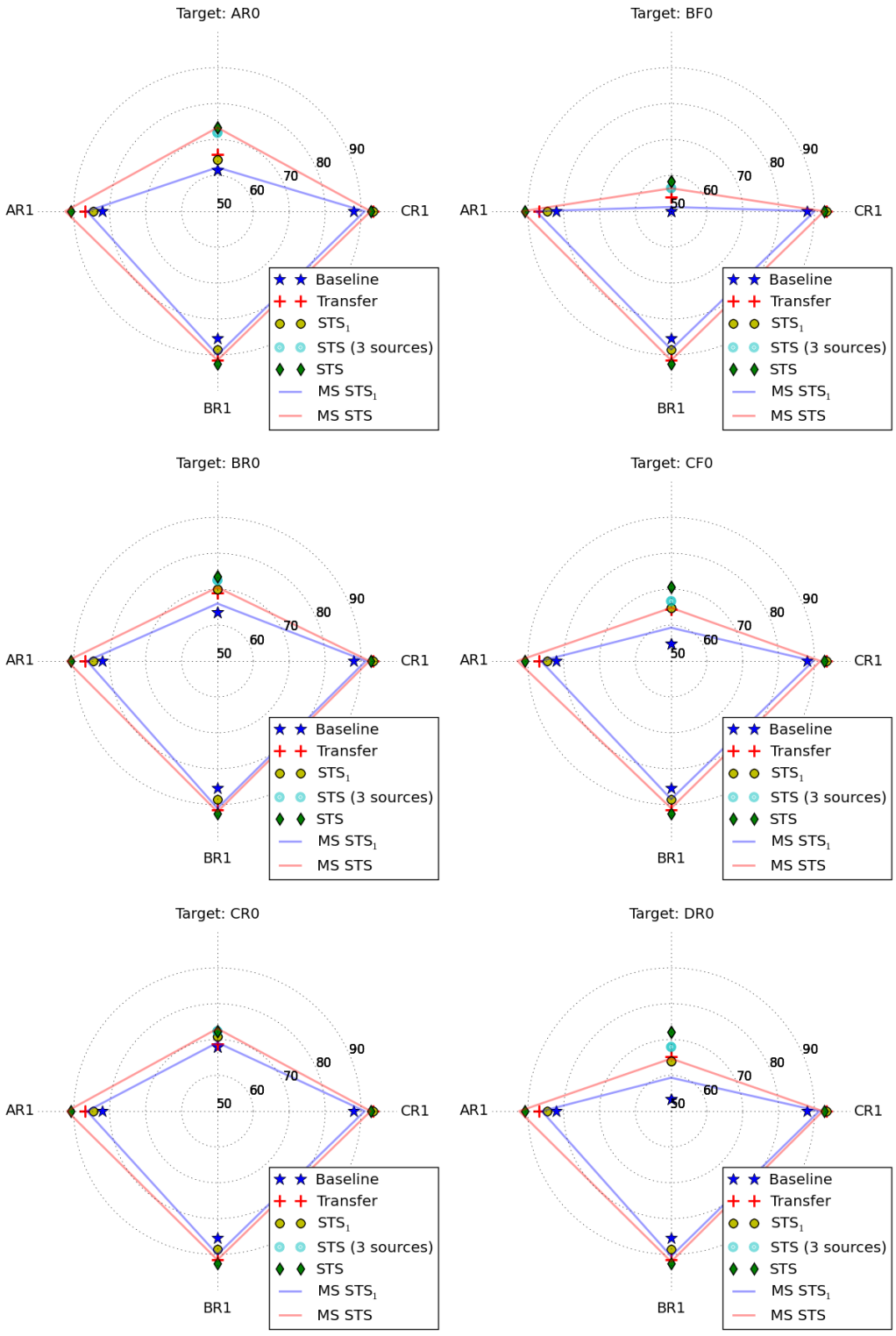


Figure 5.2: Graphical representation of the MS-STS Rank-1 recognition rates obtained for all the no-flash subsets of the CSIP database using all the flash datasets as sources, plotted against the respective BL, TLs and STS results.

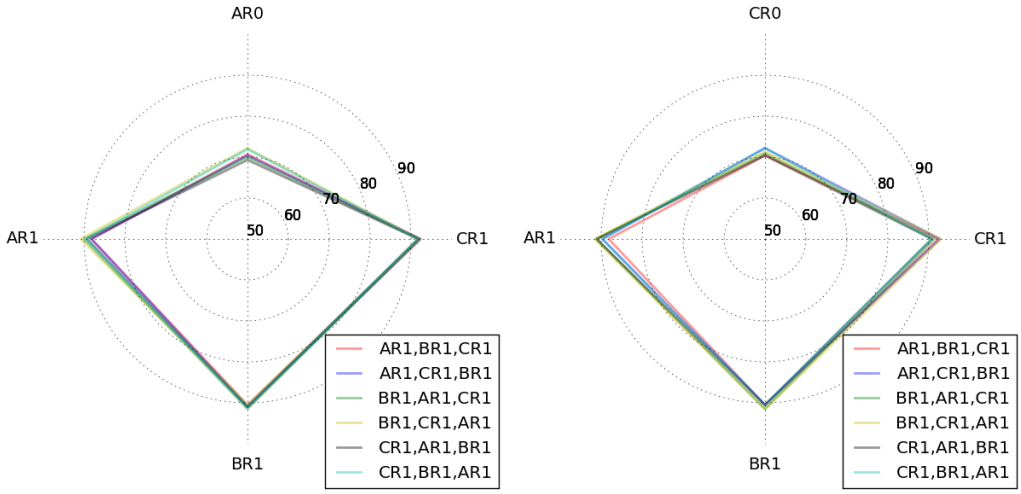


Figure 5.3: Graphical representation of the MS-STS Rank-1 recognition rates obtained for all the six possible orders of the chosen source datasets. Results concern to (a) *AR*0 and (b) *CR*0 as targets.

Chapter 6

Multi-spectral Face Recognition

Whereas technological improvements in image capturing and transmitting equipment managed to attenuate most noise factors, partial face occlusions, severe illumination changes and extreme pose variations still represent genuine challenges to automated face recognition. As referred in Chapter 4, approaching these issues will be a matter of either exploring new sources of data, to compensate the more traditional alternatives in less ideal scenarios, or designing more robust algorithms, capable of encompassing such limitations. While the last chapters outlined either a completely different biometric trait or algorithm-based strategies to tackle this problem, in this chapter the use of alternative representations of the same trait will be explored[1]. The main focus is on the use of $3D$ representations of faces, in order to introduce shape-related information to the classification framework. Nevertheless, infra-red images will also be considered, as the technological advances in sensor development are making it more readily available to have access to this type of data. We start this analysis by reviewing the evolution and the state-of-the-art alternatives regarding multi-spectral data acquisition, and then proceed to analyze the research work developed on both depth and infra-red (IR) face recognition.

6.1 An Overview of Multi-spectral Sensors

The focus of this overview will be on sensors that have been explored specifically in the scope of face recognition works. The recent trend within this area concerns the use of low-cost sensors that still allow the creation of facial recognition frameworks capable of a high recognition performance. Although, in the past, sensors with high precision like Minolta (Minolta, 2006), Inspeck (Savran et al., 2008), CyberWare (Cyberware, n.d.) and 3dMD (3dMD, n.d.) were used, their high prices led to the need for cheaper alternatives. Additionally, these systems are usually not fast and, thus, limit the potential of their application in real-world systems. These limitations led to the

[1]The work presented in this chapter resulted in the journal paper: João C. Monteiro, Tiago Freitas and Jaime S. Cardoso. *Multimodal Hierarchical Face Recognition using Information from 2.5 D Images.* and also in the conference paper Tiago Freitas, Pedro Alves, Cristiana Carpinteiro, Joana Rodrigues, Margarida Fernandes, Marina Castro, João C. Monteiro and Jaime S. Cardoso. *A comparative analysis of deep and shallow features for multimodal face recognition in a novel RGB-D-IR dataset*

development of low-cost 3*D* sensors that offered a cheap alternative, presenting both a mobile nature as well as the capability of working in real-time, at the expense of lower resolution in the acquired images. The question of how this trade-off affects performance of face recognition will be addressed in later sections regarding the state-of-the-art of this research topic.

3*D* sensors can be classified either as stereoscopic camera systems, structured light systems or laser range systems, which are capable of obtaining both 3*D*, RGB and, in some cases, also IR information. Some of the 3*D* sensors most commonly used in face recognition systems are summarized in Table 6.1. While the Minolta and Inspeck sensors are generic 3*D* sensors, CyberWare and 3dMD were designed specifically to 3*D* face scanning. All these sensors were used for 3*D* face recognition, but as referred before, have been replaced over time with low-cost alternatives.

The original Kinect (Microsoft, 2010), Kinect v1, is the most used sensor for depth acquisition. It consists in an IR laser emitter, an IR camera and a RGB camera. The latter captures color images directly, whereas a conjugation of the laser emitter and IR camera capture the depth information, resulting in a final RGB-D map. This depth map is obtained using a triangulation process based on these two sensors. Primarily the IR laser emitter, using a raster, projects a pre-designed pattern of spots in the scene, allowing the capture of the reflection of the pattern by the IR camera. Recently, a new version of this sensor was launched to replace the Kinect v1: the Kinect v2 (Microsoft, 2014) (or Kinect for XBOX One) operates with a different principle, the time-of-flight (TOF). With this methodology the depth images are obtained calculating the time between emitted IR light and its reflections (Dal Mutto et al., 2012). Recent experiments have resulted in an improved resolution and precision in this new sensor, as discussed in Amon et al. (2014) and Lachat et al. (2015). A visual example of the quality of the depth maps obtained with each of the Kinect models can be observed in Figure 6.1.

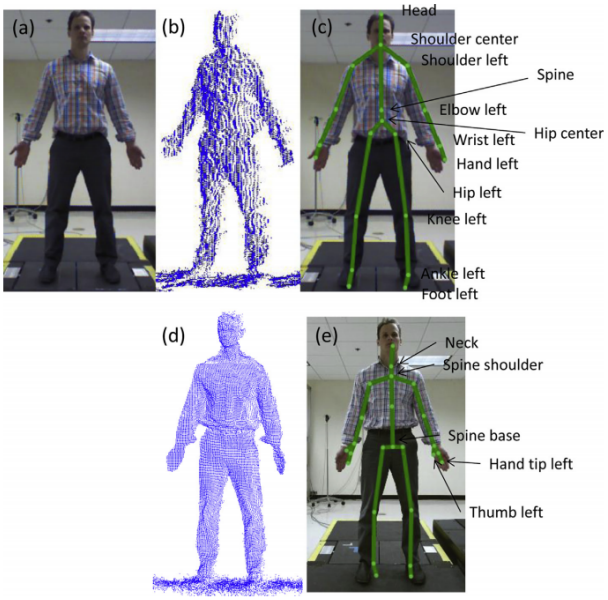


Figure 6.1: Visual example of the differences in depth map rendering and body joint detection between Kinect v1 (top) and Kinect v2 (bottom). Adapted from (Xu and McGorry, 2015).

Table 6.1: List of the most common sensors referenced in 3*D* face recognition works.

Sensor	Type	Resolution (mm)	Working Distance (m)	Price ($)
Minolta Sensors (Minolta, 2006)	3D Laser Scanning	0.041-0.22	$\sim$ 2.5	25000
3dMDface (3dMD, n.d.)	Vision Cameras	<0.2	——	10k - 20k
CyberWare 3030RGB/PS (Cyberware, n.d.)	Low-Intensity Laser Light Source	0.08 - 0.3	0.35	$\sim$ 72000
Inspeck Mega Capturer II (Savran et al., 2008)	Structured-Light	0.7	1.1	Not Available
Kinect v1 (Microsoft, 2010)	IR laser Emitter	$\sim$ 1.5 - 0.5	0.5 - 4.5	Not Available
Kinect v2 (Microsoft, 2014)	Time-of-Flight	-	$\sim$ 0.5 - 8	149.99
SoftKinetic DS325 (SofKinetic, 2007)	Diffused Laser	1.4 at 1 m distance	$\sim$ 0.15 - 1	259
Structure (Structure, 2013)	IR Structured Light	0.5 - 30	3.5	379
PrimeSense Carmine (I3DU, n.d.)	IR Laser Emitter	0.1 - 1.2	3.5	Not Available
ASUS Xtion Pro Live (Asus, 2011)	IR Laser Emitter	-	0.8 - 3.5	169.99
Intel RealSense (Intel, 2015a)	Structured Light	<1	$\sim$ 0.2 - >10	99 - 399

Other low-cost sensors have been developed to compete with Kinect in depth map acquisition. The SoftKinetic DS325 (SofKinetic, 2007), Structure sensor (mobile depth sensor in tablets) (Structure, 2013), Intel® RealSense™ (Intel, 2015a) , ASUS Xtion Pro Live (Asus, 2011) and PrimeSense (I3DU, n.d.) (recently bought by Apple (Guardian, 2013) and currently not available) have also been integrated in depth image acquisition systems. From these alternatives, one of the most promising is the Intel® RealSense™ family (Intel, 2015a). Intel® provides two models, the SR300 (previously named F200), for short range applications, and the R200 for long range acquisitions. The visual aspect of both sensors is depicted in Figure 6.2. Both sensors come with a Software Developer Kit (SDK), with pre-implemented modules for facial tracking and detection, and, similarly to Kinect v2, IR images are also obtainable. A mobile alternative in the ZR300 camera, developed specifically for smartphones, is also available. Both models present the same technology, consisting in 3 cameras that provide RGB images and stereoscopic IR, used to produce depth maps. With a laser projector, the sensors perform a scene perception and enhanced photography. The depth map can be $3D$ further filtered, allowing re-lighting, re-focusing and background segmentation. The main difference between these camera models is in the operating ranges. The R200 works from 0.5 to 3.5 meters and has an outside range up to 10 meters, while the SR300 model only operates from 0.2 meters to 1.2 meters.

The potential of this recent family of sensors in multi-spectral face recognition will be the main focus of Section 6.5, where we present what is, to the extent of the current state-of-the-art, the first multi-spectral face database developed using the Intel RealSense family of sensors. In the next section we will perform an overview of the state-of-the-art in $3D$ face recognition.



Figure 6.2: Intel RealSense models: (a) R200 and (b) SR300.

6.2 Related Work in $3D$ Face Recognition

Recently, a new trend has been observed in face recognition works, with information from the three-dimensional structure of the face being incorporated in recognition frameworks, in an attempt to grant higher robustness in scenarios such as low illumination, where the extraction of color information is severely limited, or extreme pose variations. In conjunction with the more traditional color images, $3D$ data can be used to develop more robust multimodal approaches. In this section an overview of some of the most relevant works in this field is carried out. We start with approaches that try to increase the amount of available $2D$ color images, by using virtual $3D$ approximations to generate additional poses. Then we explore approaches that make use solely

of the raw 3*D* data obtained diretly from 3*D* scanning sensors, while not considering color information. Finally, we explore alternatives that try to combine both sources of data in a multimodal fusion sense.

6.2.1 2D Data Augmentation

The use of 3*D* data to help improve the robustness and overall performance of more traditional image-based algorithms for face recognition found a multiplicity of variations throughout the years. Given the inherent discriminative power of color images alone, some authors explored the idea of using 3*D* shape as an aid to increase the amount and heterogeneity of available data, making use of virtual 3*D* models to simulate variations in pose and facial expression. One of the first works proposing such an approach was described by Blanz and Vetter (Blanz and Vetter, 2003), where a generic morphable model was adjusted to fit each input face image, estimating its tridimensional shape and texture. Such morphable models were based on a vectorized representation of gallery faces, in a convex combination of texture and shape. Using manual feature point selection, the system compares the faces using coefficients obtained from the generated models. Despite the manual selection, this paper presented a major step in face recognition field, introducing the potential of 3*D* models to increase the robustness of 2D systems.

Lu et al. (Lu et al., 2004) proposed another 3*D* generic face model, in conjugation with 2D face images, to generate shape and texture descriptors. Using also manual feature points, a depth model is created and variations in pose, illumination and facial expression are generated, to increase the variability of the training dataset. Face images are classified using minimum Euclidean distance between the two affine spaces, defined for the query face and each identity in the database. In the same year, Hu et al. (Hu et al., 2004) proposed a similar approach. For a single frontal face image, 83 key points are automatically detected and aligned, allowing the generation of a 3*D* face model, assisted by a generic tridimensional model. An orthogonal projection of the 2*D* intensity images on the generated 3*D* model is then performed. From these models different pose and facial expression images are generated. PCA and Linear Discriminant Analysis (LDA) were tested for dimensionality reduction, and the Nearest Neighbor method was used for classification.

The emergence of low-cost sensors motivated the work by Ajmera et al. (Ajmera et al., 2014), where the authors proposed the use of Speeded-Up Robust Features (SURF) descriptors in Kinect scans, assessing the performance of their methodology on both the EURECOM and CurtinFaces datasets. Using a Graph Based Iterative Hole filling interpolation, images presenting additional variations in pose are generated from the original depth image. In parallel, RGB images are processed with an Adaptive Histogram Equalization for contrast enhancement, a non-local means filter for pepper noise removal and a steerable filter. The SURF keypoint descriptors are then computed and matching is carried out by nearest neighbor computation. Hsu et al. (Hsu et al., 2014) proposed a 3*D* face reconstruction using RGB-D images, augmenting the gallery size through the generation of multiple synthetic 2D color face images. Additionally, a landmark detection to aid in the process of face alignment is considered. Sparse representation was used for classification.

6.2.2 Unimodal Approaches Based on 3*D* Models

Even though these works established the potential of 3*D* shape information, when combined with RGB data, for face recognition, they generally generated non-realistic models, also limited by the fact that all models were adapted from a single frame. An alternative approach to tackle such limitations consisted in the application of 3*D* scanning methods for more accurate, realistic and reliable models (Bowyer et al., 2006), while leaving the color information out of the recognition framework. Such methodologies present the advantage of preserving geometrical information about the face even when non-ideal illumination conditions are observed.

One of the first works exploring 3*D* models for facial recognition was introduced by Gordon (Gordon, 1991), where the author stated that some facial descriptors like the shape of the forehead, the jaw line, the eye corner cavities and the cheeks remain similar in a vast set of facial expressions. In this work, principal curvatures are calculated in depth images, and then used to detect feature points, with Euclidean distance between such points being used as the similarity metric. Years later, Tanaka et al. (Tanaka et al., 1998) proposed a different curvature-based approach, where principal curvatures and their respective orientations serve as the basis for feature extraction generating Extended Gaussian Images (EGI). EGIs are then used in the matching process, through Fisher spherical approximation. Similar to (Gordon, 1991), Chua et al. (Chua et al., 2000) argued that some rigid facial regions, such as the nose, eye socket and forehead, suffer much less deformation between different facial expressions. They propose the detection of such rigid regions by a point signature comparison among different facial expressions of the same person. Multiple works suggested the use of PCA as an improved representation of 3*D* face models. This idea was introduced by Hesher et al. (Hesher et al., 2002), who applied this transformation directly on depth maps. In two other works by Heseltine et al., PCA was used in combination with Eigenfaces (Heseltine et al., 2004a) and Fisherfaces (Heseltine et al., 2004b) representations of depth images, respectively.

In more recent works, Naveen et al. (Naveen and Moni, 2015) tested their framework on the FRAV3*D* dataset, using 2D-DWT (Discrete Wavelet Transform) and 2D-DCT (Discrete Cosine Transform) for spectral representation of high resolution depth images. Dimensionality of such features is reduced by PCA, and classification is carried out by minimum Euclidean distance. Fusion of the two techniques was then performed at the score-level. Tang et al. (Tang et al., 2015) performed landmark detection using the three main principal curvatures obtained from the construction of asymptotic cones, which described the local geometry of the mesh model. Local Normal Patterns (LNP) are extracted and then classified using a weighted sparse representation-based classifier (WSRC).

Recent technological advances have made it feasible to deploy low-cost alternatives to the more traditional high-cost 3*D* scanners, such as Minolta, Inspeck, CyberWare and 3dMD (Min et al., 2014b). The appearance of Microsoft Kinect has opened a wide array of opportunities to include three-dimensional information in computer vision solutions that were, otherwise, limited by the wider availability of color images. Cardia-Neto et al. (Cardia Neto and Marana, 2015)

proposed two variations of commonly used feature extraction methods: the 3*D*-LBP, an extrapolation of the traditional LBP algorithm specifically for Kinect depth maps, and the Histogram of Averaged Oriented Gradients (HAOG). Both descriptors are computed in parallel and an SVM classifier is applied independently to each resulting feature vector. The final decision is carried out by a weighted-sum fusion at score level. With a similar motivation Mracek et al. (Mracek et al., 2014) investigated the potential of low-cost sensors using Kinect v1 and SoftKinetic DS325. In this work the authors propose the use of a feature-preserving mesh denoising algorithm to depth images. All models were then aligned using Iterative Closest Point (ICP) (Besl and McKay, 1992) and converted to six representations of depth texture and curvature. An individual feature vector is then extracted using a PCA-reduced and z-normalized set of Gabor and Gauss-Laguerre filters.

Table 6.2 summarizes the feature extraction, classification techniques, datasets used for performance assessment and the obtained rank-1 recognition rate for each of the works outlined throughout this section.

6.2.3 Multimodal Alternatives

Even though the approaches detailed in the last section presented an interesting alternative for face recognition in less ideal acquisition scenarios, the fact that they do not take advantage of color information severely limits their true potential. The inclusion of two or more modalities has shown interesting results for a series of real-world systems working under unconstrained acquisition environments. In the case of face biometrics this is especially pertinent when high pose variations and low illumination are expected. The fusion of 2*D* color images and 3*D* scans has shown a consistent improvement of performance in systems exploring such strategies, when compared to the sole use of any of its unimodal counterparts (Abate et al., 2007).

Two of the pioneer works exploring multimodality using 2*D* and 3*D* face data, proposed by Chang et al. (Chang et al., 2003, 2005), investigated the benefits of 3*D* data integration in the more traditional 2*D*-based framework, using a Minolta Vivid 900 sensor. The authors state that both 2*D* and 3*D* modalities individually achieve similar performances, but when combined, using a simple weighted-sum fusion, a significant improvement is observed. Tsalakanidou et al. conducted two works where the potential of multimodal approaches was also assessed. In (Tsalakanidou et al., 2003), an Eigenfaces extension to depth scans is used. The Eigenfaces representations are computed in parallel for both 2*D* and depth representations, and the final score is obtained by the multiplication of both individual scores. Once again, the multimodal approach has shown significant improvements over both unimodal counterparts. In his second work (Tsalakanidou et al., 2005), an embedded hidden Markov model (EHMM) was used to achieve a combined decision using both depth and color images. In 2008, Mian et al. (Mian et al., 2008) proposed a new 3*D* keypoint detection using a PCA-based method, achieving results, in terms of keypoint repeatability, similar to SIFT. Once a 3*D* point is identified, a tensor representation is used to locally describe each keypoint. SIFT is used for an analogous representation in 2*D* images and fusion is performed at the score level, with significant performance increase being observed.

Table 6.2: Summary of the most relevant works concerning unimodal 3*D* face recognition

Author	Feature Extraction	Classifier	Dataset	r_1
(Gordon, 1991)	Distance Measures	Euclidean Distance	8 subjects (23 scans)	97.00
(Tanaka et al., 1998)	Curvature features	Fisher Spherical Approximation	37 subjects	100
(Heseltine et al., 2004a)	PCA on Eigenfaces	Euclidean Distance	York 3*D* Face	87.3
(Heseltine et al., 2004b)	PCA on Fisherfaces	Cosine Distance	York 3*D* Face	88.7
(Chua et al., 2000)	Point Signature Comparison	Ranked vote	6 subjects (24 scans)	100
(Cook et al., 2004)	Depth and Curvature	Gaussian Mixture Model	3*D*_RMA	97.33
(Naveen and Moni, 2015)	2D-DCT and 2D-DWT	Euclidean Distance	FRAV3*D*	96
(Tang et al., 2015)	Principal Curvatures + LNP	WSRC	FRGC v2	93.33
(Mracek et al., 2014)	Gabor and Gauss-Laguerre	Correlation Metric	Kinect (9 subjects), Kinetic (26 subjects)	<89
(Cardia Neto and Marana, 2015)	3*D*-LBP + HAOG	SVM	Eurecom	98

Table 6.3: Summary of the most relevant works concerning multimodal face recognition

Author	Feature Extraction	Classifier	Dataset	r_1
(Chang et al., 2003)	PCA	Mahalanobis Cosine Distance	275 subjects	98.8
(Tsalakanidou et al., 2003)	PCA	Euclidean Distance	295 subjects (3540 scans)	98.75
(Tsalakanidou et al., 2005)	2D-DCT	EHMM	50 subjects	79.2
(Mian et al., 2008)	Tensors + SIFT	4 Similarity Measurements	FGRC v2	96.6 - 99.9
(Hiremath and Manjunatha, 2013)	Gabor	Nearest Neighbor	Texas 3*D* + Bosphorus + CASIA 3*D*	99.5
(Elaiwat et al., 2015)	Curvelet Coefficients	Cosine Distance	FRGC, BU-3*D*FE, Bosphorus	99.2 / 95.1 / 91
(Naveen et al., 2015)	LPA + DFT	Euclidean Distance	FRAV3*D*	91.68
(Li et al., 2013a)	-	SRC	CurtinFaces	96.7
(Goswami et al., 2013)	Saliency + Entropy HOG	Random Decision Forest	III-TD and Eurecom	80.0 / 88.0
(Ajmera et al., 2014)	SURF	Nearest Neighbor	Eurecom and CurtinFaces	89.28 / 98.07
(Mracek et al., 2014)	Gabor and Gauss-Laguerre	Correlation Metric	Kinect v1, FRGC v2	<89
(Hsu et al., 2014)	-	SRC	Curtin Faces, Eurecom, AVL-RGBD Face	97.2 / ∼89 / ∼76
(Dai et al., 2015)	ELMDP + Gabor	Nearest Neighbor	CurtinFaces	∼ 95
(Sang et al., 2015)	HOG	Joint Bayesian	Bosphorus, BIWI, Eurecom	∼97 /∼84 / ∼93

In more recent works, Elaiwat et al. (Elaiwat et al., 2015) proposed a multimodal approach using 3*D* textured high-resolution face models, in which Curvelet coefficients are used to represent facial geometrical features. Using both depth and color information, each face is decomposed into multi-scale and multi-angle decompositions and a local surface descriptor is extracted from the neighborood of each keypoint. Naveen et al. (Naveen et al., 2015), used a Local Polynomial Approximation Filter (LPA) to obtain a directional representation of the face for each modality. A modified set of LBP features is computed and classification is carried out on the corresponding Discrete Fourier Transform (DFT) representation.

Multimodal works based on low-cost sensors have also been proposed. In 2013, Li et al. (Li et al., 2013a) used Kinect v1 to develop a face recognition system, characterized by invariance to pose, expression, illumination and disguise. Query faces were registered using a reference 3*D* model, obtained from high resolution scans. The RGB scans are transformed to the Discriminant Color Space and a Sparse Representation Classifier (SRC) is applied in parallel to both types of scans. Then, two sets of similarity scores are obtained, normalized and combined by sum-rule fusion for the final decision. Goswami et al. (Goswami et al., 2013) tested a new multimodal system in both the Eurecom and III-TD datasets. Saliency maps are computed for RGB images, while entropy maps are computed for both RGB and depth images. HOG descriptors are computed for each of these maps, and their concatenation results in the final feature vectors. A random forest classifier is used for identity assessment. Finally, Sang et al. (Sang et al., 2015) proposed a new multimodal system, using ICP for face alignment, while also proposing an alternative alignment technique for RGB images. HOG features are then extracted from each representation and a joint Bayesian classifier is used for classificaition.

Table 6.3 summarizes the feature extraction and classification techniques, the datasets used for performance assessment and the obtained rank-1 recognition rate for each of multimodal face recognition works outlined throughout this section.

All aforementioned approaches further consolidate the widespread conclusion of this research field, in that significant improvement in performance is consistently observed when comparing multimodal approaches with their unimodal counterparts. Even though a significant amount of the focus in multi-spectral face recognition is given to 3*D* data, infra-red (IR) data has become more easily available with the emergence of low-cost sensors, such as Kinect v2 and Intel® RealSense™, capable of performing IR acquisition in parallel with RGB and depth information. The next section will offer some insight regarding related work on IR face recognition and its combination with other modalities.

6.3 Related Work in IR Face Recognition

Even though it is generally not seen in recent multimodal face recognition approaches as often as RGB and depth, the use of IR images for face recognition as a stand-alone trait has seen a considerable amount of work in recent years, in an attempt to mitigate problems related to uncontrolled ambient lighting (Jain et al., 2016). With the most recent depth sensors allowing the

acquisition of IR images, it is of the utmost importance to understand the current state-of-the-art on IR face recognition and how this new representation might bring benefits over its more common alternatives. A thorough review of the evolution of IR face recognition methodologies can be found in (Farokhi et al., 2016), but a few of the most relevant works will be outlined in the next paragraphs.

One of the earliest works on IR data for face recognition was proposed by Zhao and Grigat (Zhao and Grigat, 2005). Discrete cosine transform (DCT) and support vector machines (SVM) were used for feature extraction and classification, respectively, and a recognition accuracy of 96.15% was obtained, when coupled with manual face segmentation. Li et al. (Li et al., 2005) propose an LBP alternative specifically tailored for IR images. This descriptor coupled with an AdaBoost classifier achieved an EER of 0.03% for a set of 100 cooperative users in frontal and non-occluded acquisition settings. Using a very similar approach but testing on a more challenging set of images, characterized by occlusion with sunglasses, a later work by Li et al. (Li et al., 2006) presented a 10% FRR at $FAR = 0.001$, as well as a 5% FRR at $FAR = 0.01$. It was also observed that this performance was severely compromised when changes in head position, noise or misalignment were observed.

Ke et al. (Pan et al., 2007) tackled the problem of pose variations by presenting a parts-based approach. In this work, each IR face image is decomposed into different facial parts, including the nose, eyes mouth. Combinations of single parts as well as the whole face are also considered. Classification is then performed on each part individually, using sub-window local binary pattern histograms (SLBPH) features boosted by AdaBoost learning. A set of score-level fusion alternatives were assessed to integrate information from each face part. A total of 3237 images from 35 individuals were considered and a FRR of 3.97% with $FAR = 0.001$ was observed. A moments-based approach was proposed by Farokhi et al. (Farokhi et al., 2013) as an alternative to the LBP-based works that preceded it. This approach addressed the limitation of rotation invariance by exploring an image description based on the absolute values of Zernike moments. Even though performance drop was shown to be mitigated in artificial rotations of up to 10°, in occluded scenarios robustness of the method was severely compromised. Wavelets and other orientation-based methodologies can also be found in literature. Two works by Zhang et al. (Zhang et al., 2010a) and Zheng et al. (l. Zheng et al., 2012) explored, respectively, a new feature extraction method, directional binary code (DBC), which encoded directional and spatial information of face images and a Gabor wavelet alternative for IR face recognition. In the first of such works a highly variable database of 35000 face images was acquired and a recognition rate of 97.6% was observed, whereas in Zheng's approach verification rates above 95% were observed for 3 distinct datasets.

Regardless of the methodology, it is interesting to observe that methodologies developed for IR images follow similar strategies than observed in many RGB-based works. The visual similarities between both types of images might account for this convergence of methodologies. A number of works that try to ascertain the similarity can also be found in literature. Goswami et al. (Goswami et al., 2011) analyze the inter-interoperability between RGB and IR data, by per-

forming the training and testing phases in each of the modalities separately. It is interesting to observe how both approaches present reasonable performance, further cementing the idea that information in both modalities may somehow be correlated in most controlled scenarios. The work by Kang et al. (Kang et al., 2014) explores the possibility of using IR data acquired during nighttime in a matching process against hight quality RGB images acquired in ideal illumination conditions. Reasonable performance of 80% is observed for distances of up to 60m to the sensor.

From the rationale ruling the works presented in the last paragraph, it is trivial to understand how IR recognition offers an interesting alternative to RGB when illumination is severely low and no discriminative features can be easily extracted using standard techniques. Even though color information is lost, the trade-off of being able to work in a scenario where RGB is completely fallible, renders IR data an alternative worthy of research effort. With the increased availability of low-cost sensors that allow easy access to this type of data, further motivation is given to this field. In the next section we carry a small overview of the publicly available RGB, depth and IR datasets and how there is still opportunity for further improvement in the creation of new alternatives for performance assessment in the scientific community.

6.4 Multi-spectral RGB/3D/IR Datasets

A number of datasets have already been built using multi-spectral sensors. The Aalborg University RGB-D Face Database (Hg et al., 2012) was one of the first available public datasets created. With 1581 samples from 31 different individuals, 17 different poses and facial expressions were captured for each subject using Kinect v1. The Florence Superface dataset (Berretti et al., 2012) includes RGB-D video for 20 subjects, with large pose variations. Additionally, this dataset also includes 3D high-resolution textured face scans obtained with the 3dMD scanner. The BIWI Kinect Head-Pose dataset (Fanelli et al., 2013) is a large database created for head-pose estimation, with more than 15000 scans from 20 individuals. CurtinFaces (Li et al., 2013a) contains over 5000 scans from 52 individuals, including variations in pose, illumination, facial expression and occlusions (sunglasses). The University of Western Australia Kinect Face database, described in (Hayat et al., 2015), contains 48 subjects, each with between 289 to 500 scans. Variations in facial expression and pose are included. FaceWarehouse (Cao et al., 2014) was built using facial scans captured with Kinect V1, with 20 different facial expressions. AVL-RGBD Face Database (Hsu et al., 2014) also explores pose variations, with 13 different poses (only in one plane) acquire at 5 different distances (with a maximum distance of 2 meters). EURECOM Kinect Face database (Min et al., 2014b) consists in multimodal RGB-Depth face images of 52 individuals (38 males and 14 females), with two sessions acquired with a time lapse of 5 to 14 days, captured with Kinect v1, including subjects from different ethnicities and 9 different facial expressions or partial occlusions. The IIIT-D face database (Goswami et al., 2014) includes 4605 scans (both RGB and depth) from 106 subjects, captured in two sessions with the Kinect v1 Sensor. The number of images per person ranges 11 to 254 images, including variations in pose and expression. The Labeled Infrared-Depth Face database (Cao and Lu, 2015) is, to the extent of our

knowledge, the most recent dataset and was acquired with Kinect v2. It does not provide RGB data, but it does provide data from the infrared (IR) stream, with the corresponding aligned depth data. It includes data from 17 individuals, with a total 918 scans. Each subject was scanned in 9 different poses with 6 different expressions. The manually labeled facial keypoints for each image are also provided. A few visual examples of the composing images of each of the aforementioned datasets can be observed in Figure 6.3.

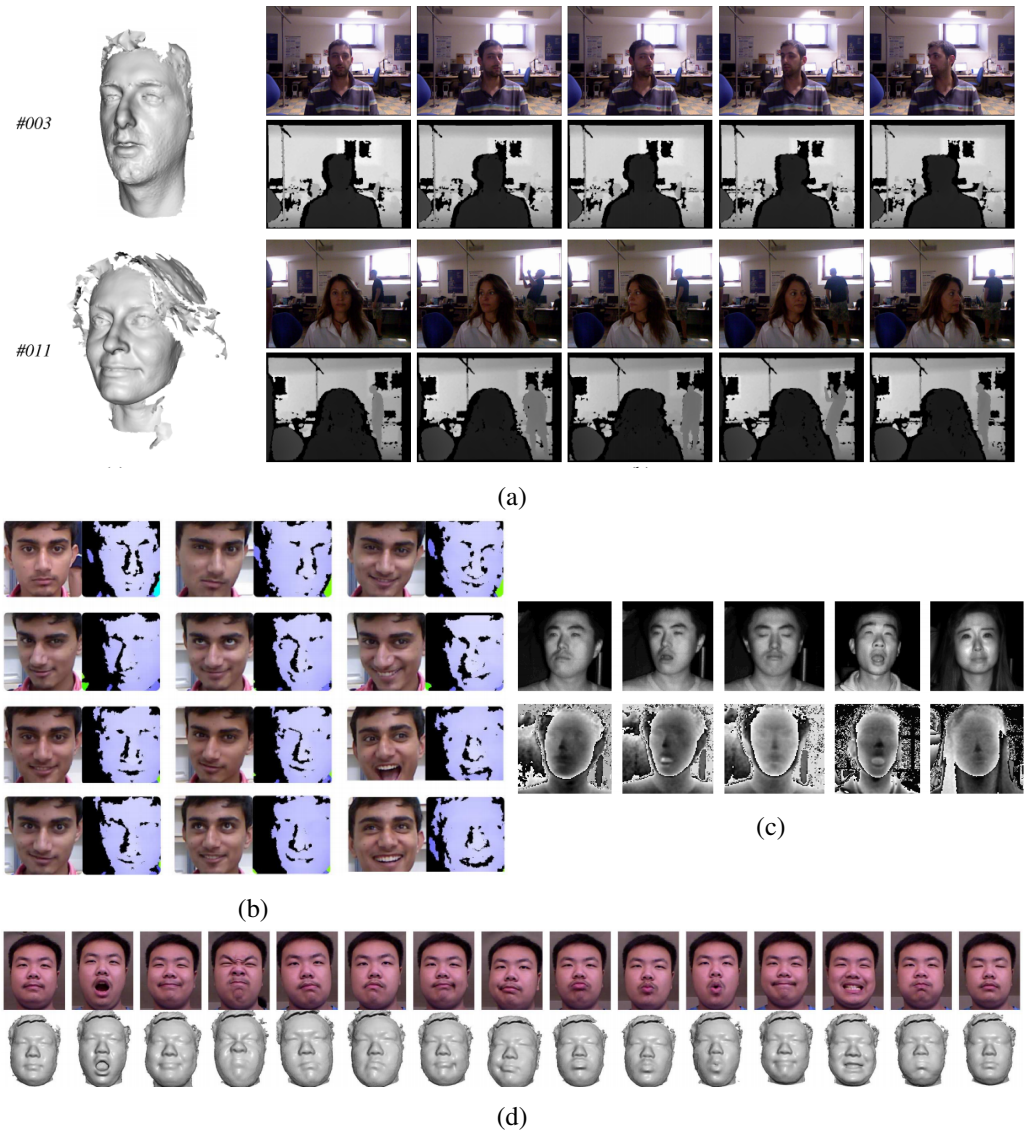


Figure 6.3: Examples of data from some multimodal RGB/D/IR datasets: (a) Florence Superface, (b) IIIT-D, (c) Labeled Infrared-Depth and (d) FaceWarehouse. Adapted from (Fanelli et al., 2013), (Goswami et al., 2014) , (Cao and Lu, 2015) and (Cao et al., 2014)

Some datasets have been built and made publicly available to the research community with regards to IR face images. The CASIA NIR face biometrics database (Li et al., 2013b) was compiled by the Institute of Automation, Chinese Academy of Sciences (CASIA) and was composed by 3940 IR facial images of 197 subjects at a resolution of 640×480, and characterized by moder-

ate facial expressions, head position and eyeglasses. The Hong Kong Polytechnic University NIR Face database (PolyU-NIRFD) (Zhang et al., 2010b) was developed by the Biometric Research Centre (UGC/CRC) at the Hong Kong Polytechnic University. A total of 335 subjects and 100 images per subject at a resolution of 768×576 were acquired, with variations in facial expression, head position, scale and time-lapse considered. Finally, the CSIST database (Xu et al., 2011), released by the Harbin Institute of Technology Shenzhen Graduate School contained two datasets, Lab1 and Lab2, characterized by a distinct set of challenges. The Lab1 database contains 500 VW and 500 NIR images of 50 subjects, at a resolution of 100, while the Lab2 database was composed by 1000 VW and 1000 NIR images of 50 subjects at a resolution of 200×200. Whereas the Lab1 images are acquired in controlled lightning conditions, their Lab2 counterparts present four different illumination conditions. Examples of variations in head position and facial expression are also present in both datasets.

While the aforementioned datasets present a wide variety of conditions, there is still not enough available relevant public data that uses the more recent sensors, capable of acquiring both depth and IR information, like the Intel® RealSense™ models presented in Section 6.1. Both models, similarly to Kinect v2, also provide IR images, besides RGB and depth representations. The prospect of designing a dataset that comprises all these modalities in conjunction with simulated real-world acquisition environments would certainly result in a strong contribution to the field. With this prospect in mind, the next section will serve as a detailed presentation of the RealFace multimodal face dataset, developed in the scope of the present thesis, with regards to both the acquisition setup as well as its final composition.

6.5 RealFace Dataset

The development of biometric recognition systems is generally limited by the shortage of large public databases acquired under real unconstrained working conditions. Database collection represents a complicated process, in which a high degree of cooperation from a large number of participants is needed. For that reason, nowadays, the number of existing public databases that can be used to evaluate the performance of biometric recognition systems in real-life acquisition conditions and making use of multiple sources of information is quite limited. Motivated by this need and the growing interest of the research community in both $3D$ and deep learning strategies for face recognition, we present a new database, named RealFace, acquired using the novel Intel® RealSense™ collection of sensors.

The RealFace dataset was acquired from a set of 42 volunteers, with different ethnicities, ages and genders. [2] Ages ranged from 18 to 40 years, gender distribution was 22 male and 20 female, while regarding nationality 41 were Portuguese and 1 was Venezuelan. After signing an agreement for the sole use of the images for scientific research purposes, each of these individuals carried out the acquisition protocol detailed below.

[2] All volunteers were gathered from the students and staff community of the Faculty of Engineering of the University of Porto.

The acquisition protocol followed in the present work was designed so that the environmental conditions presented to the sensor would closely simulate a realistic set of real-world unconstrained conditions. With this in mind variations in pose (frontal, left/right profile and left/right $\pm 45°$), facial expression (neutral and open mouth), occlusions (handkerchief and glasses) and illumination (natural, artificial and darkness) were considered in the acquisition setup. All combinations of occlusion (2) and facial expression (2) were replicated for every illumination (3) and pose (5, plus an extra neutral) conditions, and acquisition was made in a sequential way so as not to render the whole process too long and tedious for the volunteers. Due to the different optimal operating ranges, the whole process was repeated for each sensor, with the distance to the sensor being varied from 0.5 m, for the SR300 model, to 1.3 m, for the R200. The full acquisition setup is depicted in Figure 6.4 and the whole acquisition process took approximately 12 minutes per subject, resulting in a total of 72 conditions. Some representative examples are depicted in Figure 6.5.

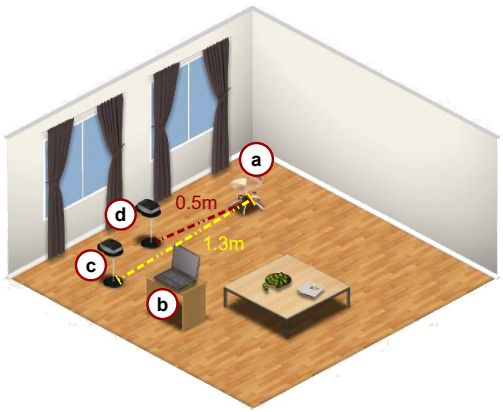


Figure 6.4: RealFace dataset acquisition setup: (a) subject; (b) acquisition control software; (c) SR300 model and (d) R200 model.



Figure 6.5: Representative examples of the poses, occlusion, illumination and expression variations considered for each subject during the RealFace acquisition process.

To take advantage of all data streams made available by the Intel® RealSense™ sensors, for each of the aforementioned conditions, we acquired an RGB image and its respective Point Cloud,

as well as the IR images provided by the integrated sensors (two for the R200 and one for the SR300) and the corresponding depth maps. This wide array of modalities and conditions confers our dataset a high versatility regarding its possible uses within the biometrics research community. A representative example of the data obtained with the SR300 model for a single acquisition is depicted in Figure 6.6.



Figure 6.6: Multimodal data from a single acquisition in the RealFace dataset: (a) RGB; (b) Depth-map; (c) Point-Cloud and (d) Infrared.

Additionally to the multiple data sources, for each RGB/Point Cloud/IR group, a set of facial keypoints were also manually annotated, to both facilitate the process of region-of-interest (ROI) segmentation, as well as allowing the dataset to be used as a benchmark for keypoint detection in any of the presented modalities. The amount and nature of the annotated keypoints depended on the variations that characterized each image. In frontal images, both eye centers, the nose tip and both mouth corners were annotated, except for the handkerchief occlusion scenario, where both nose tip and mouth corners were not considered. In profile and rotated pictures, the closest visible eye center, the nose tip and the closest visible mouth corner were considered, as well as the visible ear lobe. The handkerchief occlusion limitations were also verified here, with no nose tip and mouth corners being considered. When hair occlusion resulted in no visible ear lobe this point was also left out from the annotation. A visual example of the manual annotation in each of these scenarios is depicted in Figure 6.7.



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

Figure 6.7: Manual annotation on the RealFace dataset: (a) Frontal neutral; (b) Frontal occlusion; (c) Profile neutral and (d) Profile occlusion.

6.6 Experimental Tests and Results

6.6.1 Experimental setup for the RealSense dataset

In this section we set an experimental setup for performance assessment in the RealFace dataset. The baseline results obtained following the setup presented in this section will be presented afterwards. We will cover data partitioning and region-of-interest segmentation, as well as the shallow and deep feature representations chosen for the baseline performance assessment. We start this analysis by presenting some global considerations regarding some cases we chose to leave out of the present work, but could be the focus of future endeavors on this dataset.

6.6.1.1 Global considerations

We chose to work solely on frontal poses, leaving both the $\pm 45^{\circ}$ and the profile images out of our baseline analysis. We felt that including this kind of images would dilute the focus of the work at this point. The R200 model was also left out of this work due to the fact that we confirmed what it had been previously reported in literature: the quality of the depth images acquired with this model is still very low and unfit for object recognition problems (Song et al., 2015) as may be observed in the example depicted in Figure 6.8. Nevertheless, we propose a region-of-interest segmentation strategy for non-frontal images, and the whole setup is easily extrapolated for future work with the R200 images.

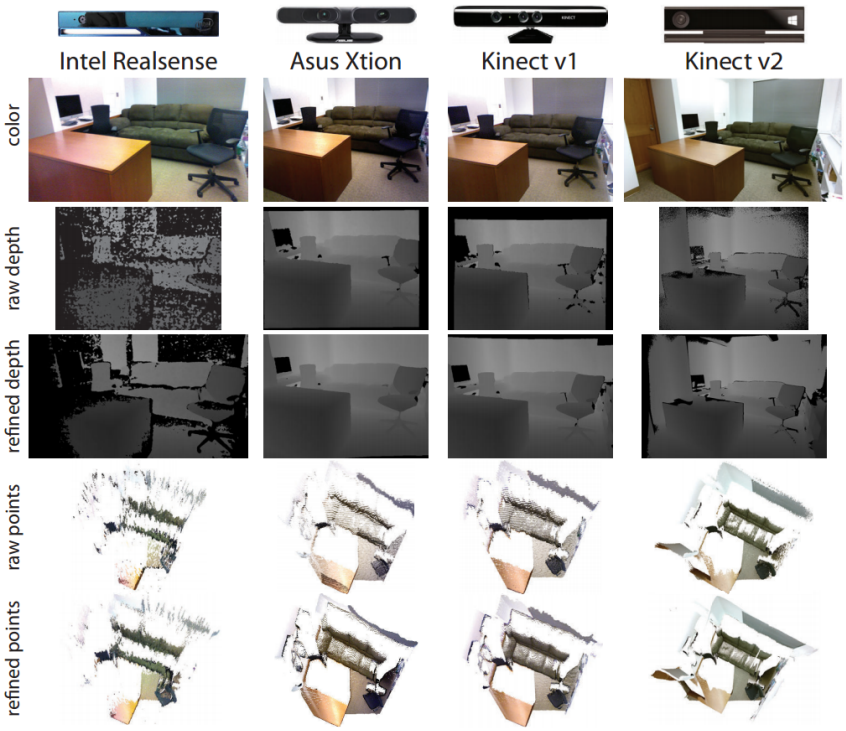


Figure 6.8: Comparison between the depth image quality of the R200 model of Intel RealSense and other low-cost alternative sensors. Adapted from (Song et al., 2015).

6.6.1.2 Pre-processing

The manually annotated keypoints were used to crop a region-of-interest (ROI) around the nose. In frontal images, a square ROI was considered, with the center corresponding to the nose tip and the side to twice the distance between the eye centers. For $\pm 45^{\circ}$ variations, on the other hand, we considered a square box centered on the horizontal line passing through the nose tip, and with side corresponding from $1.5\times$ the distance between the nose tip and the ear lobe. For the profile images a similar strategy was followed, but the side corresponded to $1\times$ the aforementioned distance.

6.6.1.3 Data partitioning

We chose images characterized by both neutral expression and artificial illumination to serve as training data for each individual. This decision is based on the fact that we want the most controlled scenarios to be used during training, and the most complicated ones be left for testing. It is intuitive to understand that training an algorithm to encompass all possible acquisition scenarios is unfeasible, when real-world applications are considered. Thus, by using the more stable images during training we aim to assess the capability of algorithms of presenting robust behaviour when more complicated challenges are presented to them. All other combinations of conditions were assessed individually during testing: illumination (natural - N; artificial - A; darkness - D), occlusion (S - scarf; G - glasses) and expression (N - neutral; OpM - open mouth).

6.6.1.4 Shallow features

From an extensive list of state-of-the-art feature extraction methods for face recognition using RGB images, the top performances were observed for PHOW, TPLBP and FHOG. A similar analysis was carried out for depth images, with both PHOW and FHOG presenting consistently better performances, while 3*D*-LBP finished the top-ranked descriptors. TPLBP (Three-Patch LBP) and 3*D*LBP are variants of traditional LBP. Presented in (Huang et al., 2006), 3*D*LBP was proposed as a variation of traditional LBP, for depth images, where depth differences are encoded in the final descriptor. In (Wolf et al., 2008) TPBLP was proposed as an upgrade of traditional LBP descriptor for face identification. Here, three patches are considered to produce a single bit value for each pixel. The Felzenszwalb's HOG (FHOG) descriptor has been described in (Felzenszwalb et al., 2010) as a variant of traditional HOG for object detection, where a feature pyramid is calculated for a finite number of scales, using repeated smoothing and sub-sampling. PHOW, presented in (Bosch et al., 2007) consists in a variation of dense-SIFT which is applied at multiple scales and combined with VLAD (Vector of Linearly Aggregated Descriptors) encoding.

6.6.1.5 Deep features

Using the pre-trained model provided by (Parkhi et al., 2015), and schematized in Table 6.4, we tested the VGG-Face CNN for all modalities, by extracting features from the $fc7$ layer and using them to train a logistic regression classifier, as described in the following paragraph. For depth

images, we also decided to train a new CNN from scratch. To serve as training we used data from 195 subjects obtained from all the available datasets presented in Section 6.4 (except EURECOM and RealFace, which were left for performance assessment). To avoid the class unbalance caused by the high degree of heterogeneity in the original number of samples per individual, we chose to generate synthetic depth maps, by flipping and rotating the original point clouds, until a total of 1000 samples per subject were obtained. The tested architecture consisted in 5 conv-relu-conv-relu-pool blocks followed by 2 fully connected layers. All conv layers include 3×3 filters, whereas the number of filters for each block is $8-16-32-64$, respectively. Finally the two fully connected layers consist of 256 units each. A batch size of 256 and a logarithmically decaying learning rate from 10^{-1} to 10^{-6} were considered, for a total of $50k$ iterations.

6.6.1.6 Classification

A set of logistic regression models was trained for classification, using each of the aforementioned shallow and deep feature descriptors. The model choice was motivated by its simplicity, leaving a considerable margin for improvement for future works on the dataset, as well as the good performance that it revealed when compared to other alternatives, such as SVM and k-nearest neighbors. The fact that class probabilities can be easily obtained was also a ruling factor of this choice, as it facilitated the multimodal fusion process described in the next section. Decision for a single feature representation and modality is carried out by maximum probability, with regards to all possible IDs.

6.6.1.7 Multimodal fusion

As referred earlier, the joint use of multiple data sources to solve the biometric recognition problem has shown improved performance in a multiplicity of recent works. To evaluate such effect in the RealFace dataset, we combine the individual logistic regression probabilities for a given ID from the RGB, depth and IR representations of a single test sample, $p_{mod}(ID|x_i)$, using a weighted-sum rule, $p(ID|x_i) = w_{RGB} \cdot p_{RGB}(ID|x_i) + w_d \cdot p_d(ID|x_i) + w_{IR} \cdot p_{IR}(ID|x_i)$, with $\sum_{mod} w_{mod} = 1$ and w_{mod} optimized by grid search. Decision is then carried out by maximizing the fusion probability $p(ID|x_i)$ with regards to all possible IDs. To overcome the loss of performance in the case of RGB in darkness conditions, a new method is proposed to deal with severely low illumination conditions. For all test images, the mean intensity of gray-scale converted RGB image, μ_i, is calculated and, depending on this value, a corrected weight for RGB-modality, w^*_{RGB}, is calculated, using a logistic function, $w^*_{RGB} = \frac{1}{1+e^{(-0.5(-\theta+\mu_i)}} \cdot w_{RGB}$, where θ was empirically set to 20 as it was observed to be the mean transition intensity between fair and poor illumination conditions. This adaptation allows the algorithm to self-adapt its performance, by adjusting the RGB weight to be higher in better illumination, and lower in less ideal low illumination conditions. The weight difference $w_{RGB} - w^*_{RGB}$ is then divided equally between the other modalities.

Following the experimental setup described throughout this section the baseline performance for the RealFace dataset will be presented next. Furthermore, some experiments were also carried

Table 6.4: VGG-face network configuration as proposed in (Parkhi et al., 2015)

layer	0	1	2	3	4	5	6	7	8	9	10	11	12
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu
support	-	3	1	3	1	2	3	1	3	1	2	3	1
filt dim	-	3	-	64	-	-	64	-	128	-	-	128	-
num filts	-	64	-	64	-	-	128	-	128	-	-	256	-
stride	-	1	1	1	1	2	1	1	1	1	2	1	1
pad	-	1	0	1	0	0	1	0	1	0	0	1	0

layer	13	14	15	16	17	18	19	20	21	22	23	24	25
type	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
support	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	256	-	256	-	-	256	-	512	-	512	-	-	512
num filts	256	-	256	-	-	512	-	512	-	512	-	-	512
stride	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	fc	relu	fc	relu	relu	softmx
support	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	-	512	-	512	-	-	512	-	4096	-	4096	-
num filts	-	512	-	512	-	-	4096	-	4096	-	2622	-
stride	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	0	0	0	0	0	0

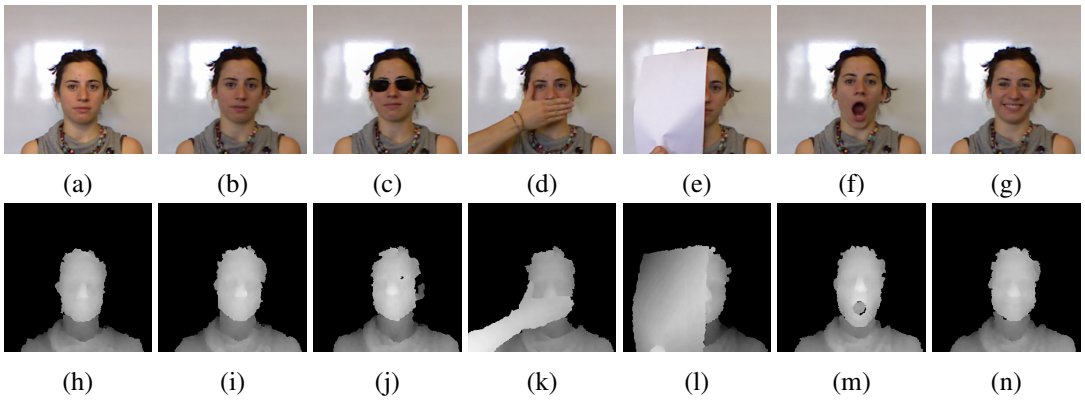


Figure 6.9: Example RGB (a)-(g) and depth (h)-(n) images from the multiple subsets of the EURECOM dataset.

out on the RGB-D EURECOM dataset, so as to better understand the challenges of our proposed dataset when compared with a state-of-the-art alternative. We start by giving some insight into the EURECOM dataset and the experimental setup used for performance assessment in this alternative. We then proceed with the discussion of the results obtained for each tested dataset, with regards to the specific challenges that each one poses.

6.6.2 EURECOM dataset

6.6.2.1 Composition

The EURECOM dataset (Min et al., 2014b), acquired with the Microsfot Kinect v1 sensor, is composed by a set of well-aligned 2D, 2.5D, 3*D* and video data. It includes scans from 52 subjects (38 males and 14 females) from two sessions interleaved from 5 to 14 days. Each session has nine types of scans that include: neutral face (N), open mouth (OpM), occlusion by scarf (S), strong illumination (LO), occlusion with sunglasses (OE), occlusion by hand (OM), occlusion by paper (OP), right face profile and left face profile. The acquisition environment is controlled in terms of luminosity, with the individuals always in a range from 0.7 to 0.9 meters to the sensor. A blank background was chosen to make the processing of the data easier. An example of the 2D and 2.5D images from a single individual is presented in Figure 6.9.

6.6.2.2 Experimental setup

We chose to follow a setup similar to the one we proposed for the RealFace dataset. The neutral images from both sessions were, therefore, chosen for training and all other subsets were considered individually for testing. ROI segmentation was carried out using the keypoints provided by the dataset, using a methodology analogous to the one described in Section 6.6.1.2 for both RGB and depth images. Resizing was carried out to 96×96 and 224×224, for depth and RGB data respectively. These dimensions were chosen to correspond to the inputs expected by the CNNs used in this work.

6.6.3 Performance analysis

The results for the EURECOM dataset are summarized in Table 6.5. In RGB images, shallow and deep features presented similar high performance for all tested conditions. PHOW with VLAD encoding was the shallow descriptor with better overall performance, and showed great versatility in image description by achieving the highest overall performance also for depth images. For such images, the overall performance drop comparatively to their RGB counterparts is clear. In this case, shallow features outperform deep features by a considerable margin. This drop in performance can be understood by the fact that only RGB images were considered during the training of the VGG-Face CNN. As there is no trivial visual similarity between the two types of images, it is logic to conclude that the filters learnt for the RGB problem are not directly applicable to depth inputs. This observation is also corroborated by the fair results presented by the proposed pre-trained CNN (PT_{CNN}) evaluated in the non-occlusion cases. Clearly, the learnt filters are able to achieve some discrimination, unlike the VGG-face alternative, but fail to adapt to non-ideal cases related to occlusions. When both modalities are combined, as referred in the state-of-the-art, the global performance is slightly increased, although not statistically relevant due to the already very high performances obtained by the RGB modality alone.

Table 6.6 presents the main results obtained for the RealFace dataset. As expected from being a more challenging dataset than EURECOM, the overall performance drop is evident. In RGB images, deep features clearly present a more robust behaviour, when presented to more variable illumination and occlusion conditions. The PHOW shallow descriptor, however, keeps the highest performance for depth images, proving to be an interesting alternative for object description in this type of data. The same observations regarding the VGG-Face and our proposed CNN for depth images can be made for this dataset, with the occlusion scenarios severely compromising global performance. In the IR modality some interesting observations can also be made. First, both PHOW and the deep descriptors from VGG-Face achieve the best overall performances. While the obtained performance is still significantly lower than the observed for RGB images, it is interesting to note how the filters learnt by VGG-Face still carry some of the discriminative power to this new modality. As referred above, for depth images, the visual similarity between RGB and IR images might translate into similar responses to the pre-trained filters, thus justifying the similar observed behaviour. The improvement caused by multimodal fusion in this dataset is more clearly noted than in EURECOM. It should be highlighted that no darkness conditions were evaluated for the RGB modality alone and, therefore, direct comparison of multimodal performance can only be carried out with the remaining modalities.

6.7 Conclusion

With so much attention being given in recent years to the potential of CNNs in a vast array of computer vision applications, one of the main limitations to their use still remains in the fact that a large amount of labeled data is necessary to perform a robust learning of their parameters. In the

next chapter we present some preliminary and exploratory work on an alternative unconstrained learning strategy for CNNs. With that goal mind, we make use of multimodal representations of the same object, such as the RGB, depth and IR representations of the same face, to develop a new unsupervised learning framework. The RealFace dataset presented throughout this chapter was used for feasibility and performance assessment.

Table 6.5: Performance comparison of shallow (S) and deep (D) features on the RGB and depth modalities of the EURECOM dataset.

		RGB							Depth						
	Feat/SS	LO	OE	OM	OP	OpM	Sm	G	LO	OE	OM	OP	OpM	Sm	G
S	FHOG	100	96.2	84.6	63.5	88.5	98.1	88.5	89.4	89.4	23.1	6.7	87.5	100	66.0
	PHOW	99.0	96.2	100	97.1	100	100	98.7	98.1	92.3	52.9	26.0	89.4	99.1	76.3
	LBP	100	95.2	95.2	88.5	95.2	98.1	95.7	–	–	–	–	–	–	–
	3*D*-LBP	–	–	–	–	–	–	–	90.4	93.3	12.5	4.8	89.4	99.1	64.9
D	VGG-F_O	100	98.1	95.2	96.2	100	100	98.2	34.6	12.5	18.3	13.5	16.4	38.46	19.1
	PT_{CNN}	–	–	–	–	–	–	–	85.6	82.7	6.7	2.9	49.0	84.6	51.9
MM	VGG+PHOW	100	98.1	96.2	96.2	100	100	98.4							
	PHOW+PHOW	100	99.1	100	97.1	100	100	99.4							

Table 6.6: Performance comparison of shallow (S) and deep (D) features on the RGB, depth and IR modalities of the RealFace dataset.

		RGB											
	Feat/SS	NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
	FHOG	54.8	39.3	26.2	66.7	96.4	42.9	90.5	–	–	–	–	59.5
	PHOW	42.9	33.3	14.3	42.9	100	54.8	97.6	–	–	–	–	55.1
	LBP	64.3	42.9	27.4	65.5	96.4	56.0	95.2	–	–	–	–	64.0
	VGG-F_o	98.8	96.4	81.0	89.3	100	88.1	95.2	–	–	–	–	92.7
		Depth											
	Feat/SS	NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
S	FHOG	71.4	46.4	28.6	79.8	78.6	33.3	76.2	90.5	66.7	32.1	70.2	61.3
	PHOW	76.2	66.7	46.4	66.7	88.1	48.8	65.5	88.1	77.4	36.9	54.8	65.0
	3*D*-LBP	54.8	47.6	26.2	66.7	76.2	21.4	61.9	83.3	60.7	16.7	57.1	52.1
D	VGG-F_o	17.9	13.1	6.0	15.5	11.9	9.5	14.3	19.1	9.5	9.5	15.5	12.9
	PT_{CNN}	48.8	44.1	15.5	66.7	72.6	16.7	56.0	58.3	51.2	13.1	53.6	45.1
		IR											
	Feat/SS	NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
	FHOG	88.1	75.0	45.2	89.3	94.1	53.6	89.3	95.2	79.8	44.1	82.1	76.0
	PHOW	97.6	94.1	60.7	90.5	100	66.7	96.4	98.8	91.7	48.8	88.1	84.9
	LBP	84.5	76.2	42.9	85.7	94.1	45.4	86.9	92.9	77.4	33.3	81.0	72.8
	VGG-F_o	100	96.4	76.2	78.6	96.4	75.0	73.8	98.8	95.2	61.9	71.4	84.0
		Multimodal											
	Feat/SS	NN	NOpM	NS	NG	AOpM	AS	AG	DN	DOpM	DS	DG	G
	VGG+PHOW+VGG	100	100	89.3	91.7	98.8	90.5	95.2	98.8	95.2	63.1	76.2	90.8
	VGG+PHOW+PHOW	98.8	96.4	81.0	89.3	100	88.1	95.2	100	90.5	58.3	90.5	89.8

Chapter 7

Unsupervised Learning of Convolutional Neural Networks using Multi-spectral Data

In recent years, deep learning methods have caused a significant impact in the field of visual object recognition. Since the publication of the breakthrough work by Kryzhevsky et al. (Krizhevsky et al., 2012), the success of deep learning approaches has been reported for a myriad of computer vision tasks. As mentioned throughout the last chapters, powerful feature representations can be learned from data automatically, surpassing, both in terms of performance and robustness, more traditional approaches based on hand-designed features. Following these successes, the focus of visual object recognition research has shifted from feature engineering to deep network design and optimization (Liu et al., 2016).

As a specific type of deep learning approach, CNNs have presented a considerable degree of success in multiple computer vision tasks. In the last chapter we observed how these networks are able to learn robust high-level feature representations, and maintain high performance for a variety of unconstrained image acquisition scenarios. However, both the task of biometric recognition as well as all the most commonly observed tasks for which CNN-based strategies have already been deployed can be described as discriminative supervised learning problems. In order to achieve a meaningful supervised learning of the considerable number of parameters associated with deep CNNs, a large set of labeled data is deemed necessary (Dosovitskiy et al., 2015). While the amount of data for computer vision tasks is becoming larger and more easily available, this need still represents a severe constraint for applications where this availability is still limited. Furthermore, for many applications based of problems other than classification, like object detection, the process of labeling huge sets of training images for regions-of-interest or object attributes is both cumbersome for the annotator, as well as ambiguous between different annotators (Oquab et al., 2015).

In the present chapter, a preliminary and exploratory work on how multi-spectral representations of the same biometric trait may be used to perform an unsupervised learning of CNNs, thus

relaxing the need for large amounts of labeled data, will be presented. We make use of the RGB and IR representations of the RealFace dataset, presented in Section 6.5, and alter the CNN learning framework in order to achieve a convergent high-level representation for a given unlabeled RGB-IR training pair.

7.1 Unsupervised CNN Learning Framework

7.1.1 Data Partitioning

The RealFace dataset is composed by 42 individuals. We chose to use all RGB-IR image pairs from 25 randomly selected identities for the unsupervised training process, and the images from the remaining 17 individuals for a supervised performance assessment. The rationale behind this choice is that the unsupervised learning carried on a set of identities should result in a model capable of extracting discriminative features in a non-coincident set of identities, even though the notion of identity is never introduced throughout training. Furthermore, we are interested in simulating an application scenario where the amount of available unlabelled data surpasses that of labelled data, allowing the learning process to adapt to both types and learn from the largest possible amount of data.

All RGB-IR images pairs are considered indiscriminately for the training process, resulting in a total of 1100 RGB-IR pairs of images for the said process. During performance assessment, the high-level feature descriptors, resulting from the learned CNN, are extracted for the 816 RGB-IR image pairs and randomly separated into train and test. The final classification is then carried out using both LR and RBF-SVM classifiers. Both the unsupervised learning process as well as the classification procedure will now be analyzed in further detail. A global overview of the whole training and testing procedure is schematically represented in Figure 7.1.

7.1.2 Learning Process

The proposed unsupervised learning process can be described as an iterative regression problem where two independent CNNs are tuned such that the Euclidean distance between their resulting outputs is minimized. These two independent CNNs are responsible for each of the two chosen modalities, RGB and IR. By forcing both CNNs to converge to a similar output we enforce that this common descriptor should represent a similar high-level abstract notion, such as the tested individual's identity. As each RGB-IR image pair present in the RealFace dataset represents the same face information acquired in two distinct spectra, forcing these two deep networks to converge to the same "identity" descriptor will allow the learning of the convolution filters that compose the intermediate layers, with no need of an associated identity label. The definition of the biometric recognition problem can, thus, be changed from "assessing the identity of a biometric trait" to "making two representations of the same biometric trait converge to the same identity". The main steps of the proposed unsupervised learning process can be described as follows:

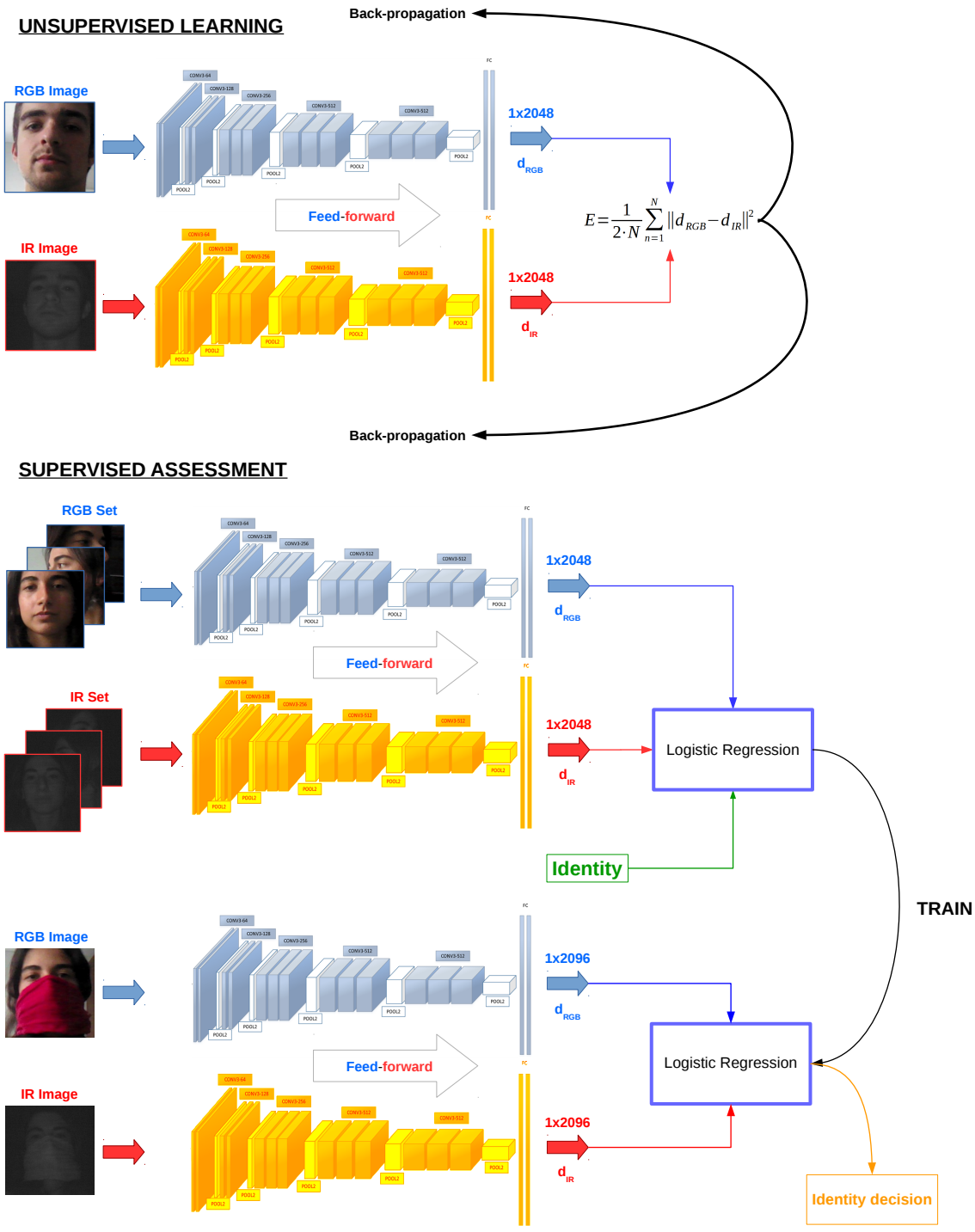


Figure 7.1: Schematic representation of the unsupervised training and supervised test phases of the developed framework for multi-spectral face recognition.

1. **Initialization:** The CNNs responsible for the processing of RGB and IR training data are initialized. From this point onwards we will use the notations CNN_{RGB} and CNN_{IR} to represent both networks respectively. We choose to initialize the filters of the intermediate layers using a pre-trained CNN, proposed by BLVC for generic object recognition, and publicly

available as a part of the Caffe library for deep learning (Jia et al., 2014). Although this network was not trained for the specific problem of face recognition, we chose to perform this exploratory work on a simpler and more generic network first, with a proof-of-concept goal in mind. All conv-relu-pool layers are kept from this original network, with the last fully connected layer, fc-loss, being randomly initialized using the uniform distribution presented earlier in Equation (5.1):

$$\boldsymbol{w}_S^k = \mathscr{U}\left[\frac{-\sqrt{6}}{\sqrt{M_{k-1}+M_k}}, \frac{\sqrt{6}}{\sqrt{M_{k-1}+M_k}}\right] \tag{7.1}$$

where M_k is the number of neurons in the k-th layer. While the original fc layers were designed for a specific classification problem, with the last layers responsible for outputting the final high-level descriptor and computing the class probabilities for the problem in question, the proposed regression-based approach requires a totally different representation in these last levels, thus justifying a random re-initialization of the last fc layers of the original models CNN^0_{RGB} and CNN^0_{IR}. We define that the abstract representation of the n-th RGB-IR pair's associated identity, d^n_{RGB} and d^n_{IR}, has a size of 1×2048. A more compact version with size 1×100 was also tested. After both CNNs are initialized, the first d^n_{RGB} and d^n_{IR} are computed for each of the $n \in [1...1100]$ RGB-IR image pairs by feed-forward.

2. **Loss computation:** An Euclidean-loss layer is introduced after the initialized fc layers. The generic loss function for this type of layer is of the type:

$$E = \frac{1}{2 \cdot N} \sum_{n=1}^{N} ||\hat{y}_n - y_n||^2 \tag{7.2}$$

where $\hat{y}_n$ represents the output of the last fc layer and y_n the target output, for each of the $n \in [1...N]$ train inputs. In the specific case at hand, we want the target outputs of both CNNs to converge to one another. Formalizing, we can define the Euclidean losses of CNN_{RGB} and CNN_{IR}, E_{RGB} and E_{IR} respectively, to be:

$$E_{IR} = E_{RGB} = \frac{1}{2 \cdot N} \sum_{n=1}^{N} ||d^n_{RGB} - d^n_{IR}||^2 \tag{7.3}$$

The obtained loss is then back-propagated independently for each of the individual CNNs and the process is repeated for a single iteration over every image-pair in the training set.

3. **Update targets:** After back-propagation, the convolution parameters for the layers of each CNN will present new values, according to the obtained Euclidean loss. After a single iteration over all training data is performed, the new outputs of the individual CNNs are re-computed and set as the new target descriptors for the next iteration of the algorithm. With this dynamic behavior, both RGB and IR outputs are updated towards an ideal scenario where they converge to a single abstract vector that codes identity information for

both modalities. After the targets are updated, loss is recomputed and back-propagation is used for a new parameter update, with the whole process running iteratively from this point onwards. The stopping criterion can be set to either a fixed number of iterations, as well as a fixed value of global Euclidean loss.

With the proposed methodology we aim to allow multi-spectral face recognition in scenarios where no knowledge is given regarding which modalities are made available to the sensor. As a common descriptor is obtained by the proposed training methodology, it is expected that regardless of the modality or modalities that are presented during the test phase, the model is capable of achieving a stable recognition performance. Furthermore, the fact that multiple representations of the same object might be used to learn robust CNN-based classification methodologies in an unsupervised would represent an interesting contribution to the field of computer vision.

7.1.3 Classification

After the aforementioned stopping criterion is achieved, two new CNN models, CNN^{*}_{RGB} and CNN^{*}_{IR}, are obtained. These two models accept as input either the RGB or IR representations of a face and output very similar 1×2048 descriptors that may be used to perform supervised identification on a new set of individuals. From each RGB-IR pair, two similar descriptors are acquired. The fact that both CNN outputs are set to converge, results in a proximity at feature-level that translates to a classification process where a single discriminative model can be used for both modalities, supplanting the need to train a specialized classifier for each representation. In this preliminary work we choose to perform classification using a logistic regression classifier, just to assess the discriminative potential of the learned common descriptor in a very simple classification scenario.

7.1.4 Experimental Results

We start by analyzing whether the convergence of both output towards a common descriptor is indeed observed. In fact, an optimal solution to the chosen loss function would be found if both descriptors collapsed to a vector of only zeros. In this case a global optimum would be achieved during the iterative search for the optimal parameters, even though the whole discriminative potential would be lost. While this is still a possibility, the fact that we initialize the intermediate layers of both proposed networks from a more stable pre-trained network should be enough to mitigate this nefarious effect. Figure 7.2 depicts the evolution of an example validation descriptor, in 3 distinct regions of the 1×2048 vector, between the randomly initialized fc layers (top half) until the final result (bottom half), achieved when the observed loss falls below a given threshold (1000 iterations). It is easily discernible how both descriptors converge to a similar shape. Visually the difference between both descriptors is represented by the shaded gray area, which suffers a considerable lowering after the unsupervised learning is carried out. While it is still arguable that the order of magnitude of the obtained descriptor values is significantly lower after training, the fact that the search process stabilizes in this solution and does not continue to converge to an all-zeros

alternative seems to indicate that the obtained descriptors do, indeed, represent some meaningful information.

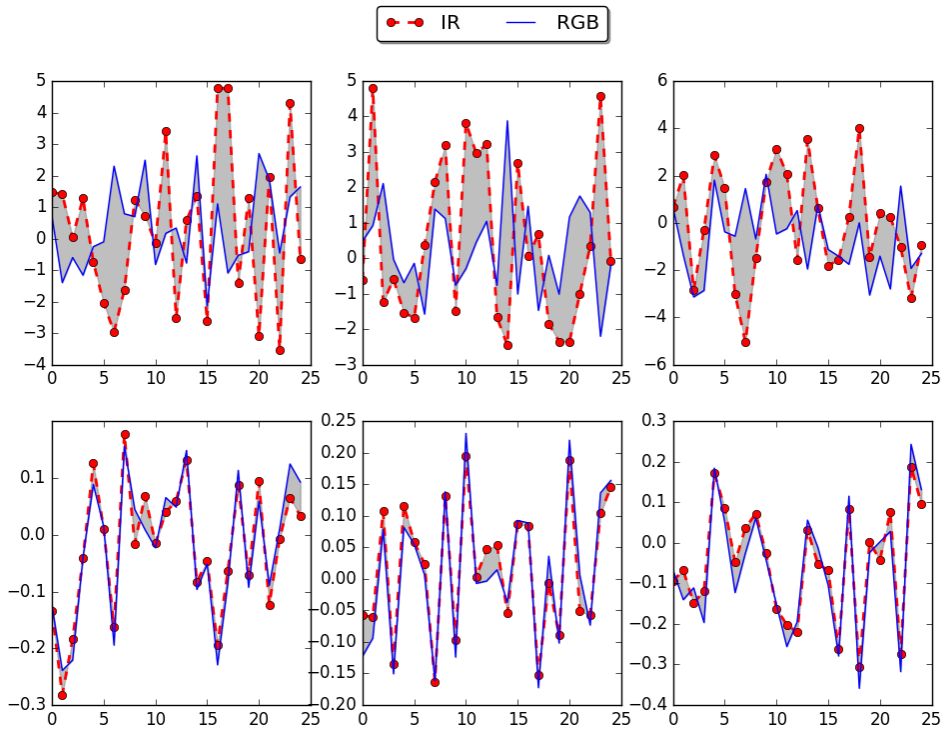


Figure 7.2: Evolution of the outputs of CNN_{RGB} and CNN_{IR} from their original values, with a randomly initialized fc-loss layer, to the pre-trained versions after 1000 iterations.

The observations presented throughout the last paragraph lead to the next step, where recognition performance in an independent supervised setting using the obtained descriptors was assessed. The main results are summarized in Table 7.1. The first value corresponds to the supervised classification obtained by the original BLVC reference network. In this case, the proposed fc-loss layer is not considered, and the output of the fc7 layer, i.e. the original descriptor, is considered for classification purposes. The baseline performance is, thus, defined by these values. The second value corresponds to the performance observed when the randomly initialized fc-loss layer is considered after the original fc7 layer. In this case, the 1×2048 common descriptors are used as the input for classification. We can observe that no significant variations occur on either of the tested modalities. While the last layer is randomly initialized, the fact that the resulting descriptor is a simple linear combination of the original fc7 output might account for the observed stability in performance.

The third value represents the recognition rate observed after 1000 iterations of the unsupervised learning of the fc-loss layer. A few interesting observations can be carried out based on these observations. First, the original performance is slightly degraded for the RGB modality (while it might be debated that this drop is not statistically relevant) and considerably degraded for its IR counterpart. The fact that the training process is not tuned in any way to reflect the identity of the

Table 7.1: Rank-1 recognition rates observed for different iterations of the designed unsupervised learning approach.

		Network		
		BLVC	**BLVC+fc-loss$_0$**	**BLVC+fc-loss_1000**
Modality	**RGB**	*67.8*	*67.8*	*65.8*
	IR	*55.6*	*54.7*	*45.3*

images composing the training set could be seen as an explanation for the observed performance drops. However, this observation was used as the motivation for the methodology: with no need for an identity, the whole learning process could be carried out on pairs of unlabeled representations of the same face image. By pointing that the performance is degraded by the same rationale that lead to the development of the methodology, the original design of the learning process can also be questioned in some key points. First, the chosen loss function might not translate the whole solution to the problem we are trying to solve. While the idea of learning a common descriptor for both modalities is well motivated, simply forcing both descriptors to converge to each other might not be the most interesting alternative. The fact that an optimal solution to the proposed loss function is found when both descriptors collapse to an all-zeros vector represents an obstacle to the whole process, *a priori*. Furthermore, the fact that the observed RGB and IR individual performances present a considerable disparity also needs some thought in light of the proposed methodology. While during training and validation the obtained fc-loss outputs presented a similar shape and order of magnitde, as observed by the example depicted in Figure 7.2, the same cannot be said for the test dataset. If the learned convergence did indeed translate to the test process, the same classification performance would be expected for both RGB and IR modalities, as the same descriptor would be outputted by their respective US-learned networks. The fact that a considerable disparity is observed between both modalities needs further research, as no new conditions are considered in the testing phase, with all occlusion and illumination scenarios considered for both the 25 training individuals as well as the 17 left for testing.

7.2 Conclusion and Future Work

While the idea that supports the developed methodology seems to present the theoretical soundness to represent an interesting alternative to the field of multi-spectral recognition, the preliminary results seem to indicate that the specific choices made throughout this first phase of research are still far from ideal. As this chapter still represents an exploratory and ongoing work it is of note to point where there is more room for improvement in order to make this methodology useful for the scientific community. First, the chosen loss function needs to be tuned, not only to force the training process to find a common descriptor for the chosen modalities, but also to mitigate the possibility that both descriptors collapse to low orders of magnitude with little to no discriminative power. The introduction of a second term to the computation of E_{mod}, that counters this collapse

will most likely be the ideal alternative:

$$E_{RGB} = w_{dist} \cdot \frac{1}{2 \cdot N} \sum_{n=1}^{N} ||d_{RGB}^{n} - d_{IR}^{n}||^2 - w_{mag} \cdot C_{RGB} \tag{7.4}$$

$$E_{IR} = w_{dist} \cdot \frac{1}{2 \cdot N} \sum_{n=1}^{N} ||d_{IR}^{n} - d_{RGB}^{n}||^2 - w_{mag} \cdot C_{IR} \tag{7.5}$$

where $C_{IR} \propto ||d_{IR}||$ and $C_{RGB} \propto ||d_{RGB}||$, so as to force the descriptors to maintain higher magnitude (thus the negative sign as the loss is designed to be minimized), as well as similar, values. The relative importance of the distance between descriptors and their order of magnitude is weighted by the constants w_{dist} and w_{mag} respectively, under the constraint $w_{dist} + w_{mag} = 1$.

Besides this change, further analysis is necessary regarding the fact that such a significant disparity is observed between the test performance of both modalities. Even if the learned common descriptor loses all its discriminative power, both modalities should converge to a single performance. Future work will focus on this observation in conjugation with the improved loss function suggested above to try and improve the usability of this preliminary methodology in the field that motivated its design.

The next chapter will work as a global conclusion to the present thesis, as well as a review of some relevant hot-topics in the research field of biometrics in the years to come.

Part III

Conclusion

Chapter 8

Conclusions and Future Lines of Research

This final chapter serves as both a summary of the main conclusions achieved as a result of the scientific contributions presented throughout the last chapters, as well as a setup for future lines of research, with a review of the current trends in the field of biometrics.

8.1 Conclusions

The main conclusions on each topic explored in the present thesis are as follows:

- **Unimodal periocular recognition:** On a first approach to this problem, an automatic modeling of SIFT descriptors, using a GMM-based UBM method, to achieve a canonical representation of individual's biometric data, regardless of the number of detected SIFT keypoints was developed and assessed on periocular images from two databases, achieving state-of-the-art performance for all experimental setups. Even though the developed GMM-UBM algorithm was meant for periocular recognition solely, the framework can be easily extrapolated for other image-based traits. To the extent of our knowledge, GMM-based UBM methodologies were only explored for speaker recognition so far. The proposed pipeline may, thus, represent the first of a series of experiments that explore its main advantages in the scope of multiple trending biometric topics. For example, the fact that any number of keypoints triggers a recognition score may be relevant when only partial meaningful data is available for recognition. The hierarchical framework presented in Section 4.3, that was assessed on faces that are purposely occluded, revealed another example where this observation may be relevant. Besides from the conceptual advantages, a few technical details may be improved in further work. Exploring further color channels besides the RGB space could bring benefits to the proposed algorithm. Regarding fusion, exploring individual specific parameters, instead of a global parametrization, would enable the algorithm to be trained to counter the Doddington zoo effect (Ross et al., 2009). As not all people are as easy to identify, fitting the properties of the designed classification block to adapt to different classes of

individuals seems like an interesting idea. Finally, and regarding the training setup, some questions might be worthy of a more thorough research. In the case of voice recognition it is common to train two separate UBMs for male and female speakers. Extrapolating this idea to image-based traits, multiple UBMs trained on homogeneous sets of equally or similarly zoomed images might improve the results when more realistic and dynamic conditions are presented to the acquisition system. In a related topic it is also not consensual whether the left and right eyes, due to the intrinsic symmetry of the face, should be considered in a single model or as separate entities.

Regarding the assessment in mobile device acquisition scenarios, we observed that each tested algorithm presents its advantages and disadvantages. Santos et al. is more fit to identification problems with less restrictions concerning processing time, presenting high performance for a wide variety of noise factors. On the other hand, the GMM-UBM approach presents a faster matching time, with better performance in verification scenarios. It would also be of relevance to assess how the presented methodologies behave, in their current state, when implemented in existing mobile devices, so as to better understand their current limitations in more realistic scenarios. While the cross-sensor performance observed for both tested methodologies was severely compromised when compared to the original single-sensor scenario, the work presented in Chapter 5 served as a very promising alternative for such cases.

- **Unimodal face recognition:** An algorithm for face recognition under unconstrained settings, such as heterogeneous illumination and severe occlusions was developed. By training models that only describe features of limited regions of the face, we confer our algorithm a robustness to events where all other regions are occluded. In an attempt to replicate the global precedent hypothesis of the human brain's cognitive mechanisms of perception, we designed an innovative hierarchical algorithm, where face recognition is performed locally only if a more global representation is not capable of achieving a decision with a high degree of certainty. Even though good performance was observed for a wide variety of non-ideal conditions, some ideas are worth noting for future research in the area. First of all, the choice of SIFT keypoint descriptors distributed in a dense grid is based on a series of assumptions, like constant scale and orientation, that might not always hold. Besides these assumptions, we performed no comparative analysis with other descriptors that might prove more appropriate for face description in less controlled acquisition scenarios. The feature extraction algorithm proposed by Miao and Jiang (Miao and Jiang, 2013), for example, is reported to be more efficient at extracting interest points from human face images than the SIFT approach followed in the developed work, and might prove to be an interesting alternative. Still, regarding face representation, no alternative pre-processing methods were considered besides the Weber-faces approach. Given that the most significant performance drop of the proposed algorithm was observed for images acquired under significantly low illumination, alternative normalization techniques might deserve further research. A different approach

might be considered, as previously mentioned, by introducing a feedback mechanism to the recognition system, with the user being notified when the reject option is triggered in the end of the hierarchical chain. This could allow a user to actively adapt the environmental conditions so as to facilitate a more certain decision.

The choice of the fittest θ_l value for each specific application might also motivate a more detailed study. The hypothesis that some users are easier to identify than others inside a given population, an effect known as the Doddington zoo effect (Ross et al., 2009) as referred earlier in the present discussion, suggests that parameter optimization in biometric applications might benefit from a more individual-specific approach. Besides this observation, we fixed the θ_l value for all levels of the hierarchical chain. We assumed that due to the normalization of the likelihood-ratio provided by the UBM, recognition scores at different levels share a similar range of values. Even though this assumption seems valid enough in the proposed framework, analyzing the effect of level-specific threshold values might bring about an increase in performance. In a final consideration, the assessment of the developed algorithm was carried out fully on frontal face images. Most real-life applications in unconstrained scenarios should not enforce such a rigid constraint, and thus, the proposed algorithm should be capable of coping with pose variations. Two alternatives to approach this problem could be the training of multiple models describing individual poses, or the learning of a single model built on images acquired under multiple poses. Further research into both alternatives would be needed before either of them can be considered as the most fit to expand the proposed framework.

- **Cross-sensor recognition:** Regarding this topic, we proposed an extended version of the source–target–source approach to deep transfer learning, making use of multiple sources of information. We successfully applied the developed algorithm to the specific problem of cross-sensor biometrics, a recent field of research that aims to mitigate the performance drop observed when training and testing acquisition conditions are considerably heterogeneous. We observed that, when compared to a state-of-the-art algorithm designed for single-sensor scenarios, the proposed STS and MS-STS approaches revealed a worse baseline performance, but managed to present a very interesting cross-sensor stability, regardless of the nature of the data used in the training process. It is trivial to deduce that an improvement in the baseline performance of any of the proposed methodologies – CNN or SV-SDA – would, necessarily, result in a stable increase of performance in all cross-sensor scenarios. Some ideas to achieve such improvement would necessarily consist on exploring alternatives to the SIFT description chosen for the supervector generation, or on the development of ensemble or joint strategies capable of making the most of the pros of both GMM-UBM (or any other state-of-the-art single sensor methodology) and STS strategies, to simultaneously achieve good baseline and cross-sensor performance. Achieving a tight coupling between both methodologies will, most certainly, represent a very significant step in the field of cross-sensor biometrics. Concerning the CNN methodology, it must be noted that,

traditionally, this strategy is explored when large datasets are available, so as to achieve the most robust modeling possible. In the current work, we used only two training instances from each source dataset, thus, theoretically, limiting the potential of achieving good results for the problem at hand. With this observation in mind, we can conclude that testing the proposed approaches on a larger cross-sensor periocular dataset, or extending the original one using jittering strategies (Hays, 2016), would probably result both in higher baseline and higher cross-sensor performances. One must note, however, that the availability of large amounts of data to perform the enrollment step is not guaranteed in real-life applications. This limitation should, therefore, be overcome in the long run if this strategy is expected to be implemented in more practical solutions.

Another focus of future work for CNN would be using information from all three RGB channels instead of gray scale transformation used on the present work. Regarding the MS-STS, we may conclude that even though the optimal STS performance managed to outperform its multi-source counterpart in almost all scenarios, the reasoning for this behavior is both expected and negligible for practical applications. As we are manually choosing the best-performing source when presenting the STS results, whereas in the MS-STS, we are fixing the same set of sources for all experiments, it is expected that optimal performance is not achieved in the situations where the best-performing single source is not included on the set of chosen sources. From a practical point of view, testing the universe of all possible sources to empirically choose the best one is not a viable possibility. The focus of future work should therefore fall on the automatic choice of the fittest sources to achieve the highest degree of domain generalization during the MS-STS learning process and, thus, cause the convergence of the MS-STS performance to the best possible STS result.

- **Multi-spectral face recognition:** Even though we only assessed the use of the RealFace dataset for biometric recognition, we acknowledge that its usability can extend to fields such as face alignment, gender and age prediction, as well as face detection in depth and IR images. The manually annotated keypoints, as well as the defined ROI segmentation methodologies, make the performed experiments easily replicable and confer the presented performance baseline a strong starting point for further research in the community. However, the number of enrolled subjects is still not as high as desirable, and an extended version of the dataset would be an interesting line of research in the future. The inclusion of more intermediate non-frontal poses would further extend the usability of the dataset for alternative applications such as pose quantification.

  Regarding the comparative analysis between deep and shallow features we observed that a few challenges are still unsolved. While the publicly available VGG-face network showed excellent performance in the RGB modality for all tested scenarios in both datasets, surpassing all alternative shallow feature alternatives, performance dropped considerably for depth images. The pre-trained CNN that we designed showed increased performance in some scenarios, but still stays below the results obtained with specific tailored features such as

PHOW and FHOG. The fact that the amount of data used to train VGG-Face is considerably higher than the amount of depth data used to train our CNN may account for these observations. With the appearance of more datasets based on depth representations of faces, and the consequent growth in the amount of available data, an improved version of the proposed CNN could also be easily obtained, namely by augmenting the training dataset to better deal with the presence of occlusions.

With the ideas outlined in the last paragraphs, it can be concluded that the present thesis tackled a set of interesting topics within the general idea of multimodal biometrics, with proven performance observed for non-ideal acquisition conditions, cross-sensor scenarios and multi-spectral applications. The presented contributions will surely allow for future work to build on its ideas and continue to accompany the never stopping evolution of the field in nowadays' society. The next section will outline some of the present hot topics of the research field of biometrics, as well as exploring what the main challenges for its evolution will be in the near future.

8.2 Research Trends

The future of biometric recognition and its main challenges for the years to come has been one of the focus of a recent review by Jain et al. (Jain et al., 2016). The authors consider a series of broad topics that will serve as a guideline for future follow-up work to the present thesis:

- **Biometric Trait Distinctiveness:** the lack of objective ways of quantifying the individuality of a biometric trait represents an obvious disadvantage with regards to more traditional methodologies. While a fixed-size PIN-number represents a fixed and known number of possible combinations, the same trivial computation is not possible for the ever changing set of biometric traits in the world's population. Techniques to estimate distinctiveness would work as a helpful tool for system design when, for example, there is a need to choose one or multiple traits for the application at hand. While no such technique is standardized, all choices will be carried out mostly in a trial-and-error point-of-view. In this field some interesting works can already be found. Ye et al. (Ye et al., 2016) assess the individuality and distinctiveness of finger vein patterns. A large-scale finger vein database was built, composed by a total of 718399 images from 363703 distinct fingers. Fusion at the decision level of two state-of-the-art feature extraction algorithms was carried out, and the distribution of similarity scores was modeled from a total of 83 billion matches. The obtained results showed that significant distinctiveness was obtained for sets of 300 thousand images, using the aforementioned fusion of state-of-the-art algorithms. Manabe et al. (Manabe et al., 2016) explore the distinctiveness of gait by statistical analysis of matching scores, coming to the conclusion that for datasets above 50 identities, its performance may be compromised. On a less usual approach, Neal et al. assess the distinctiveness of device applications, Wi-Fi and Bluetooth usage as a potential biometric behavioural trait. Data was collected from 200 subjects over a 19-month span, and a nearest neighbor classifier, trained on all source

of data, achieved a recognition rate of 85%. Application features were found to be more unique over the course a whole week, whereas Bluetooth and Wi-Fi features achieved higher performance when a single day period was considered. Globally, Wi-Fi features presented the highest distinctiveness, while maintaining permanence in time periods of up to a month. With the acceptability and demand of biometric systems for real-world applications expected to grow in years to come, research on empyrical estimation of distinctiveness will surely accompany this growth and become indispensable in the future of biometrics.

- **Biometric Trait Persistence:** the issue of how a biometric trait is affected by aging has been on the spotlight of research for a few years, but will certainly become a hot-topic as the performance of systems can be severely compromised over large time-periods if such effect is not taken into account. In a similar rationale to the one presented in the last paragraph, it is interesting to understand if there is some way to objectively quantify the permanence of a biometric trait. If a certain degree of change is known to occur, then system development can be adapted to expect a periodic and systematic update of a biometric template due to aging related changes. Most of the works already found in literature consist in modeling the process of aging from sets of pairwise time-lapsed face images (Park et al., 2010). Such models can, then, be introduced during the matching process to adapt the acquired data to potential disparities between probe and template data. Friedman et al. (Friedman et al., 2016) propose a technique of the assessment of temporal persistence of biometric traits based on the statistical analysis of time-lapsed biometric recognition matches. While all their tests were carried out on a set of 8 eye-movement and 2 gait related features, the authors hypothesize that this would likely be the case for any biometric modality where the features can be expressed as quantitative values. Jain et al. (Jain et al., 2015) analyze the specific problem of child vaccination tracking in sub-Saharan Africa. In such scenarios, access to any form of identification document is severely limited, making biometric identification a viable alternative to vaccination schedule tracking. Given that the main goal is to enroll children from young ages, but maintaining the vaccine record throughout their lifetime, the issue of persistence becomes of the utmost importance. The work proposed in this work is still ongoing, and reflects an application where the society might greatly benefit from research in this field.

- **Unconstrained Acquisition Surveillance Scenarios:** As presented throughout this thesis, the question of non-ideal acquisition scenarios is one of the main hot topics in biometric research in recent years, and will most likely maintain this status in the near future. Applications in the area of personal surveillance gave origin to most of the reasonings behind the need for further development in this field. The fact that sensor technology for long-range imaging results in images of poor quality, allied to the challenge of recognizing non-cooperative subjects and the high variability of pose, illumination and occlusion states, brings about a series of challenges that are still far from solved. The increasing availability of video data in such scenarios introduces still another layer of complexity, as the relation-

ship between successive frames of the same individual needs to be taken into account, as opposed to more traditional single probe recognition applications (Jain et al., 2016). Some recent works can be found on the use of surveillance biometrics. Klontz and Jain (Klontz and Jain, 2013) simulated the use of face recognition based on the surveillance videos from the Boston bombings, using as templates the database of mugshots from the Pinellas County Sherrif's Office. While one of the subjects was successfully identified with a considerable degree of certainty, this work highlighted the technical difficulties that variable acquisition conditions bring to the problem of face recognition. The use of soft biometric traits is gaining popularity is this field. Contrarily to the traits described in Section 2.6, known as hard biometrics, their soft counterparts can be defined as a collection of human describable, distinguishing visual features, like gender, height and physical complexion, that can be derived from salient and permanent body characteristics (Martinho-Corbishley et al., 2016a). Martinho-Corbishley et al. (Martinho-Corbishley et al., 2016a) published the publicly available SoBiR dataset built from surveillance camera images of moving individuals, coupled with a baseline soft biometric approach based on the Ensemble of Localised Features (ELF) descriptor (Gray and Tao, 2008) as well as a setup for performance analysis. Coupled with that work, the same authors promoted a crowdsourcing application for manual annotation of the proposed dataset (Martinho-Corbishley et al., 2016b). Proença and Neves (Proença and Neves, 2016) concluded that the quality of VW iris/periocular recognition in surveillance scenarios is still sub-par using more traditional pan-tilt-zoom (PTZ) cameras, but also refer 4K next-generation PTZ cameras as a promising alternative for the problem at hand. The effect of the subject-camera distance on the overall iris region quality, and the possible use of the periocular region were also addressed.

- **System security:** As any alternative for security applications, biometric-based systems present a number of particular vulnerabilities that need to be taken into account. One of the most commonly referred of such vulnerabilities is the problem known as spoofing, where a biometric sensor is presented with a counterfeit biometric trait. In non-supervised recognition scenarios, spoof detection represents a critical requirement (Jain et al., 2016). Due to the fact that spoof detection typically involves checking acquired data for sign of human vitality, the process is also known as liveness detection. A number of works can be found in recent literature concerning this problematic. Patel et al. (Patel et al., 2016) suggest a image distortion-based framework for spoof detection in face recognition applications for smartphones. Working on an alternative trait, Balaji et al. (Balaji et al., 2016) propose a multimodal system based both on fingerprint and fingernail patterns. The finger corresponding to the assessed fingernail/fingerprint pair is illuminated with white light and, depending on the resulting observed patterns, the decision of whether to consider the respective fingerprint as a real or fake is taken accordingly. Another hot topic regarding biometric system security is the protection of stored biometric templates against unauthorized access. While the risks of inappropriate access to such data can be mitigated, for many large-scale applications that

require a central template database, feasibility of such alternatives is questionable. With such limitations in mind, recent works have turned towards eliminating the need to store any raw biometric information in databases, being replaced instead by pseudo-random keys, re-generated everytime a new sample of the same biometric trait is presented to the system (Uludag et al., 2004). A more tractable alternative may concern the transformation of raw biometric template into more secured versions, satisfying a few main requirements (Jain et al., 2016):

- **Non-invertibility:** The process of recovering stored biometric features from the stored secure templates should be computationally hard;
- **Non-linkability:** Multiple independent secure templates should be possible to create from the same biometric data;
- **Discriminability:** The secure version of the template should not result in a degradation in recognition accuracy of the biometric system.

With this specifications in mind, a few recent works have explored the problematic of secure biometric templates. Verma et al. (Verma and Kant, 2016) propose an approach based on chaotic maps for face recognition applications, even though the methodology's requirements are not assessed in a rigorous way. Bringer et al. (Bringer et al., 2016) assess the non-linkability of secure templates built from two different iriscodes extracted from the same iris, as well as discussing the security of other biometric template protection methodologies based on Blooms filters. A multimodal approach was suggested by Yildiz et al. (Yildiz et al., 2016), with multiple biometrics from the same user being considered in a layered template generation process. With this approach, each biometric trait is "concealed" among the others, making it difficult to access one of them without accessing the full multimodal template. The non-linkability property was assessed at state-of-the-art low levels, while the discriminative power was mostly kept, with an EER of 2.1% against the state-of-the-art 1.9%. Another security-related issue that has been gaining increased attention in recent years concerns the high-degrees of public self-disclosure in online social networks. With a considerably large amount of photos posted online daily, on networks such as Facebook, there is a strong possibility that personally sensitive data can be deduced from the mining of such data combined with available off-the-shelf face recognition technology. Issues such as visual cryptography and privacy-preserving photo sharing are, thus, gaining some popularity as identity security tools in the biometrics research community (Jain et al., 2016). Kumari et al. (Kumari et al., 2016) present a visual cryptography methodology for face template security, where one biometric template of interest is encrypted in a set of public face images via the GEVCS (Gray-Level Extended Visual Cryptography Scheme) method (Lin and Tsai, 2003). Nandhinipreetha and Radha (Nandhinipreetha and Radha, 2016) propose a multimodal approach based on finger vein patterns and signature analysis, where both templates are masked by visual cryptographic scheme, and only made available when both

traits are present during recognition at the same time. Transversal to all the concerns referred throughout this section, personal privacy should always be kept in mind during the development of any biometric recognition system. With the growing ease of access to face and voice data, devices such as Google Glass have been raising ethical and privacy questions that should always be kept in mind. Venier [28] summarized some of the ethical aspects to be taken into consideration when developing biometric-based systems:

- **Respect for human dignity:** An individual's identity should never be reduced to a code or number, rather than their name. The risk of making the user feel objectified would represent a considerable obstacle in the widespread deployment of a biometric recognition systems;
- **Informatization of the human body:** As the amount of information acquired and stored about an individual becomes gradually higher, the risk of it becoming available to a wider population than intended also becomes more concerning. This point connects directly to all the data protection and privacy issues explored throughout the last paragraphs;
- **Respect for body integrity and intimacy:** As some biometrics are characterized by a heavy degree of intrusiveness, it should be taken into consideration if the user would be comfortable, physically as well as psychologically, during the enrollment and recognition process.

Some interesting recent works can be consulted for more views on the topic of ethics in the scope of biometric applications. Darelle van Greunen (van Greunen, 2016) explores this topic in light of the efforts carried by the South African Ministry of Social Development in children biometrics. Also in South Africa, but in the field of patient enrollment for clinical trials, Moodley et al. (Moodley et al., 2016) overview a series of privacy and ethics concerns that ruled the process of co-enrollment detection of participants in multiple non-independent clinical trials.

- **Technological Advances:** It is instinctive to understand that in the years to come technological advances will influence how biometric recognition systems evolve and are adapted to new social realities. One of such recent innovations in the availability of lab-on-a-chip products that opened up new frontiers in the use of DNA as a biometric trait in applications other than forensic sciences (Birdwell et al., 2016). Whereas in this specific field time constraints can be more relaxed, for most daily-life applications, the fact that the time for a DNA match is being reduced from days to just a few hours might end will certainly be the attention of research works in the near future. The development of sensors capable of detecting multiple traits at the same time will also facilitate the implementation of multimodal systems, and allow a better exploration of all the advantages of such strategies (Jain et al., 2016). Another direct consequence of the advance of technology concerns the computational power, in terms of both storage and processing, that allowed many known strategies to be reworked

under these new considerations. The topic of CNN, that was extensively referred throughout this thesis, can be viewed as an extrapolation of the knowledge of neural networks to the new era of processors that have been made available in recent years. While neural netowrks have been around for more than 30 years, their true potential, in the form of CNNs, for many computer vision problems only now starts to be unveiled. New trends in deep learning will surely appear in the years to come, and biometric recognition will certainly find a new set o algorithmic tools to keep improving the performance and robustness of developed systems.

Future lines of research, following the work presented in this thesis, will certainly shape to the number of challenges presented throughout this section. The final goal of achieving a biometric system capable of robust and accurate recognition of individuals at any single moment in any place will continue to capture the interest of multiple areas of applied knowledge, and will certainly represent a non-negligible part of the socio-economic panorama for the years to come.

Bibliography

3dMD, "3dMDface system," n.d., available at http://www.3dmd.com/3dmd-systems/#face, Accessed on 15 January 2016.

A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2D and 3D face recognition: A survey," *Pattern Recognition Letters*, vol. 28, no. 14, pp. 1885–1906, 2007.

R. Ajmera, A. Nigam, and P. Gupta, "3D face recognition using Kinect," in *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. ACM, 2014, p. 76.

I. Alice, "Biometric recognition: Security and privacy concerns," *IEEE Security & Privacy*, 2003.

C. Amon, F. Fuhrmann, and F. Graf, "Evaluation of the spatial resolution accuracy of the face tracking system for Kinect for Windows v1 and v2," in *Proceedings of the 6th Congress of the Alps Adria Acoustics Association*, 2014.

S. Angle, R. Bhagtani, and H. Chheda, "Biometrics: A further echelon of security," in *UAE International Conference on Biological and Medical Physics*, 2005.

A. Ansari, M. Abdel-Mottaleb *et al.*, "3D face modeling using two views and a generic face model with application to 3D face recognition," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2003, pp. 37–44.

Asus, "Xtion PRO LIVE," 2011, available at: https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/overview/, Accessed on 15 January 2016.

S. Bakshi, S. Kumari, R. Raman, and P. K. Sa, "Evaluation of periocular over face biometric: A case study," *Procedia Engineering*, vol. 38, pp. 1628–1633, 2012.

D. Bala, "Biometrics and information security," in *Proceedings of the 5th annual conference on information security curriculum development*. ACM, 2008, pp. 64–66.

A. Balaji, V. HS, and S. OK, "Multimodal fingerprint spoof detection using white light," *Procedia Computer Science*, vol. 78, pp. 330–335, 2016.

B. Bank of America, "Bank of America introduces fingerprint and touch ID sign-in for its mobile banking app," http://newsroom.bankofamerica.com/press-releases/consumer-banking/bank-america-introduces-fingerprint-and-touch-id-sign-its-mobile-ban, 2015, accessed: 2016-11-16.

A. T. Becker, "Expectation-maximization on Old Faithful," https://www.mathworks.com/matlabcentral/fileexchange/49869-expectation-maximization-on-old-faithful, 2009, accessed: 2016-12-12.

P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *Unsupervised and Transfer Learning Challenges in Machine Learning*, vol. 7, p. 19, 2012.

Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

M. Berniker and K. P. Kording, "Deep networks for motor control functions," *Frontiers in computational neuroscience*, vol. 9, 2015.

S. Berretti, A. D. Bimbo, and P. Pala, "Superfaces: A super-resolution model for 3D faces," in *ECCV Workshops*, 2012, pp. 73–82.

P. J. Besl and N. D. McKay, "Method for registration of 3D shapes," in *Robotics-DL tentative*. International Society for Optics and Photonics, 1992, pp. 586–606.

S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh, "Periocular biometrics: When iris recognition fails," in *4th IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–6.

B. Biggio, I. Corona, B. Nelson, B. I. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli, "Security evaluation of support vector machines in adversarial environments," in *Support Vector Machines Applications*. Springer, 2014, pp. 105–153.

biometria.sk, "Biometric principles," http://www.biometria.sk/en/principles-of-biometrics-pf.html, 2016, accessed: 2016-10-16.

B. BioStore, "Biometrics market worth 30 billion by 2021," http://biostore.co.uk/company/news-articles/biometrics-market-worth-30-billion-by-2021/, 2016, accessed: 2016-11-16.

J. D. Birdwell, N. Q. Haas, and S. F. Hansen, "Mobile biometrics information collection and identification," Jan. 12 2016, US Patent 9,235,733.

C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

W. W. Bledsoe and H. Chan, "A man-machine facial recognition system—some preliminary results," *Panoramic Research, Inc, Palo Alto, California., Technical Report PRI A*, vol. 19, p. 1965, 1965.

J. Blitzer, M. Dredze, F. Pereira *et al.*, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *ACL*, vol. 7, 2007, pp. 440–447.

V. N. Boddeti, J. M. Smereka, and B. V. Kumar, "A comparative evaluation of iris and ocular recognition methods on challenging ocular images," in *2011 International Joint Conference on Biometrics*. IEEE, 2011, pp. 1–8.

E. Bondi, P. Pala, S. Berretti, and A. Del Bimbo, "Reconstructing high-resolution face models from Kinect depth sequences acquired in uncooperative contexts," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, vol. 7. IEEE, 2015, pp. 1–6.

A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 517–530.

A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

I. Bouchrika, "On using gait biometrics for re-identification in automated visual surveillance," *Developing Next-Generation Countermeasures for Homeland Security Threat Prevention*, p. 140, 2016.

K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Computer vision and image understanding*, vol. 101, no. 1, pp. 1–15, 2006.

P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering." in *ICML*, vol. 98. Citeseer, 1998, pp. 91–99.

J. Bringer, C. Morel, and C. Rathgeb, "Security analysis and improvement of some biometric protected templates based on bloom filters," *Image and Vision Computing*, pp. –, 2016.

R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 10, pp. 955–966, 1995.

L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 770–787, 2010.

W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: a 3D facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2014.

Y. Cao and B.-L. Lu, "Intensity-depth face alignment using cascade shape regression," in *Neural Information Processing*. Springer, 2015, pp. 224–231.

J. B. Cardia Neto and A. N. Marana, "3DLBP and HAOG fusion for face recognition utilizing Kinect as a 3D scanner," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, 2015, pp. 66–73.

R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

C. CBS News, "Biometric screening service helps flyers "clear" long security lines," http://www.cbsnews.com/news/tsa-long-security-lines-clear-private-screening-company-eliminates-wait/, 2016, accessed: 2016-11-16.

K. Chang, K. Bowyer, and P. Flynn, "Face recognition using 2D and 3D facial data," in *ACM Workshop on Multimodal User Authentication*, 2003, pp. 25–32.

K. I. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multimodal 2D + 3D face biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 619–624, 2005.

W. Chen, M. J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 458–466, 2006.

H. Cho, R. Roberts, B. Jung, O. Choi, and S. Moon, "An efficient hybrid face recognition algorithm using PCA and Gabor wavelets," *International Journal of Advanced Robotic Systems*, vol. 11, no. 59, pp. 1–8, 2014.

C.-S. Chua, F. Han, and Y.-K. Ho, "3D human face recognition using point signature," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2000, pp. 233–238.

D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3642–3649.

C. Citi Group, "Citi wins two 2015 gartner "cool business" awards with its agent-based voice biometrics authentication initiative," http://www.citigroup.com/citi/news/2015/151118a.htm, 2015, accessed: 2016-11-16.

R. Connaughton, A. Sgroi, K. W. Bowyer, and P. Flynn, "A cross-sensor evaluation of three commercial iris cameras for iris biometrics," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2011, pp. 90–97.

J. Cook, V. Chandran, S. Sridharan, and C. Fookes, "Face recognition from 3D data using iterative closest point algorithm and gaussian mixture models," in *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission*. IEEE, 2004, pp. 502–509.

T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Computer Vision—ECCV'98*. Springer, 1998, pp. 484–498.

M. Cosenza, "OpenCV (open source computer vision)," 2016, available at: http://opencv.org/, Accessed on 15 January 2016.

Cyberware, "Cyberware model 3030 color 3D scanhead," n.d., available at: http://cyberware.com/products/scanners/3030Specs.html, Accessed on 15 January 2016.

D. Dahlmeier and H. T. Ng, "Domain adaptation for semantic role labeling in the biomedical domain," *Bioinformatics*, vol. 26, no. 8, pp. 1098–1104, 2010.

W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 210–219.

X. Dai, S. Yin, P. Ouyang, L. Liu, and S. Wei, "A multi-modal 2D+3D face recognition method with a novel local feature descriptor," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 657–662.

C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, *Time-of-Flight Cameras and Microsoft Kinect^{TM}*. Springer Science & Business Media, 2012.

N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition*, C. Schmid, S. Soatto, and C. Tomasi, Eds., vol. 2, June 2005, pp. 886–893.

J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE transactions on pattern analysis and machine intelligence*, vol. 15, no. 11, pp. 1148–1161, 1993.

H. Daume III and D. Marcu, "Domain adaptation for statistical classifiers," *Journal of Artificial Intelligence Research*, pp. 101–126, 2006.

J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

D. Deep Learning, "Graphical depiction of a LeNet model," http://deeplearning.net/tutorial/lenet.html#the-full-model-lenet, 2010, accessed: 2016-10-06.

A. Dosovitskiy, J. Tobias Springenberg, and T. Brox, "Learning to generate chairs with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1538–1546.

Y. Du, B. Bonney, R. Ives, D. Etter, and R. Schultz, "Analysis of partial iris recognition using a 1D approach," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2005, pp. 961–964.

S. Elaiwat, M. Bennamoun, F. Boussaid, and A. El-Sallam, "A curvelet-based approach for textured 3D face recognition," *Pattern Recognition*, vol. 48, no. 4, pp. 1235–1246, 2015.

E. European Comission, "ARIES: reliable european identity ecosystem," http://aries-project.eu/, 2016, accessed: 2016-12-19.

E. European Comission, "Horizon 2020: The EU framework programme for research and innovation," https://ec.europa.eu/programmes/horizon2020/en, 2016, accessed: 2016-12-19.

E. European Comission, "OCTAVE: Objective control for talker verification," https://www.octave-project.eu/, 2016, accessed: 2016-12-19.

E. European Comission, "PROTECT: Pervasive and user focused biometrics border project," http://projectprotect.eu/, 2016, accessed: 2016-12-19.

E. European Comission, "Speechxrays: Multi-channel biometrics combining acoustic and machine vision analysis of speech, lip movement and face," http://www.speechxrays.eu/, 2016, accessed: 2016-12-19.

G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, February 2013.

S. Farokhi, S. M. Shamsuddin, J. Flusser, U. U. Sheikh, M. Khansari, and K. Jafari-Khouzani, "Rotation and noise invariant near-infrared face recognition by means of Zernike moments and spectral regression discriminant analysis," *Journal of Electronic Imaging*, vol. 22, no. 1, pp. 013 030–013 030, 2013.

S. Farokhi, J. Flusser, and U. U. Sheikh, "Near infrared face recognition: A literature survey," *Computer Science Review*, 2016.

H. Faulds, "On the skin-furrows of the hand," *Nature*, vol. 22, p. 605, 1880.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

F. Find Biometrics, "Biometric airport security debuts in China," http://findbiometrics.com/biometric-airport-security-china-307082/, 2016, accessed: 2016-11-16.

L. Friedman, I. Rigas, M. S. Nixon, and O. V. Komogortsev, "Method to assess the temporal persistence of potential biometric features: Application to oculomotor, and gait-related databases," *arXiv preprint arXiv:1609.03948*, 2016.

G. Gartner Technology, "Gartner says 30 percent of organizations will use biometric authentication for mobile devices by 2016," http://www.gartner.com/newsroom/id/2661115, 2014, accessed: 2016-12-12.

Z. Ge, C. McCool, C. Sanderson, and P. Corke, "Modelling local deep convolutional neural network features to improve fine-grained image classification," *arXiv preprint arXiv:1502.07802*, 2015.

C. Geng and X. Jiang, "Face recognition based on the multi-scale local image structures," *Pattern Recognition*, vol. 44, no. 10–11, pp. 2565 – 2575, 2011.

A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 513–520.

J. M. Gold, P. J. Mundy, and B. S. Tjan, "The perception of a face is no more than the sum of its parts," *Psychological science*, vol. 23, no. 4, pp. 427–434, 2012.

G. G. Gordon, "Face recognition based on depth maps and surface curvature," in *San Diego,'91, San Diego, CA*. International Society for Optics and Photonics, 1991, pp. 234–247.

D. Goswami, C. H. Chan, D. Windridge, and J. Kittler, "Evaluation of face recognition system in heterogeneous environments (visible vs nir)," in *IEEE International Conference on Computer Vision Workshops*. IEEE, 2011, pp. 2160–2167.

G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On RGB-D face recognition using Kinect," in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems*. IEEE, 2013, pp. 1–6.

G. Goswami, M. Vatsa, and R. Singh, "RGB-D face recognition with texture and attribute features," *Information Forensics and Security, IEEE Transactions on*, vol. 9, no. 10, pp. 1629–1640, 2014.

D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

G. Griaulle Biometrics, "Types of biometric systems: Verification x identification x watch list," http://www.griaulebiometrics.com/en-us/book/understanding-biometrics/introduction/model/types, 2014, accessed: 2016-09-26.

T. Guardian, "Apple buys 3D sensor company PrimeSense for $350m," 2013, available at: http://www.theguardian.com/technology/2013/nov/25/apple-buys-primesense-microsoft-kinect, Accessed on 15 January 2016.

M. Gudavalli, S. V. Raju, A. V. Babu, and D. S. Kumar, "Multimodal biometrics – sources, architecture and fusion techniques: An overview," in *International Symposium on Biometrics and Security Technologies*. IEEE, 2012, pp. 27–34.

B. Gutschoven and P. Verlinde, "Multi-modal identity verification using support vector machines (SVM)," in *Proceedings of the Third International Conference on Information Fusion*, vol. 2. IEEE, 2000, pp. THB3–3.

M. Hayat, M. Bennamoun, and A. A. El-Sallam, "An RGB-D based image set classification for robust face recognition from kinect data," *Neurocomputing*, 2015.

J. Hays, "Introduction to computer vision - project 6: Deep learning," http://www.cc.gatech.edu/~hays/compvision/proj6/, 2016, accessed: 2016-12-22.

M. Heikkila, M. Pietikainen, and S. Member, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, p. 2006, 2006.

M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with center-symmetric local binary patterns," in *Computer vision, graphics and image processing*. Springer, 2006, pp. 58–69.

R. Herbei and M. H. Wegkamp, "Classification with reject option," *Canadian Journal of Statistics*, vol. 34, no. 4, pp. 709–721, 2006.

T. Heseltine, N. Pears, and J. Austin, "Three-dimensional face recognition: An eigensurface approach," in *International Conference on Image Processing*, vol. 2. IEEE, 2004, pp. 1421–1424.

T. Heseltine, N. Pears, and J. Austin, "Three-dimensional face recognition: A fishersurface approach," in *Image Analysis and Recognition*. Springer, 2004, pp. 684–691.

C. Hesher, A. Srivastava, and G. Erlebacher, "Principal component analysis of range images for facial recognition," in *Proceedings of the International Conference on Imaging Science, Systems, and Technology*, 2002, pp. 62–68.

R. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet, "An RGB-D database using Microsoft's Kinect for Windows for face detection," in *Eighth International Conference on Signal Image Technology and Internet Based Systems*. IEEE, 2012, pp. 42–46.

P. Hiremath and H. Manjunatha, "3D face recognition based on depth and intensity gabor features using symbolic pca and adaboost," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 5, pp. 1–12, 2013.

G.-S. J. Hsu, Y.-L. Liu, H.-C. Peng, and P.-X. Wu, "RGB-D-based face reconstruction and recognition," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2110–2118, 2014.

Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang, "Automatic 3D reconstruction for face recognition," in *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2004, pp. 843–848.

D. Huang, M. Ardabilian, Y. Wang, and L. Chen, "3D face recognition using eLBP-based facial description and local feature hybrid matching," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1551–1565, 2012.

Y. Huang, Y. Wang, and T. Tan, "Combining statistics of geometrical and correlative features for 3D face recognition." in *BMVC*. Citeseer, 2006, pp. 879–888.

I3DU, "PrimeSense™ 3D sensors," n.d., available at: http://www.i3du.gr/pdf/primesense.pdf, Accessed on 15 January 2016.

s.-i. Image Processing in Python, "Histogram of oriented gradients," http://sharky93.github.io/docs/gallery/auto_examples/plot_hog.html, 2012, accessed: 2016-10-13.

Intel, "Introducing the Intel® RealSense™ R200 camera (world facing)," 2015, available at: https://software.intel.com/en-us/articles/realsense-r200-camera, Accessed on 15 January 2016.

Intel, "Intel® RealSense™ development kit," 2015, available at: https://software.intel.com/en-us/RealSense/Devkit/, Accessed on 15 January 2016.

I. International Biometrics Group, "The Henry classification system," http://newmanboston.org/customized/uploads/henry%20fingerprint%20classification.pdf, 2003, accessed: 2016-09-21.

A. Jain, L. Hong, and S. Pankanti, "Biometric identification," *Communications of the ACM*, vol. 43, no. 2, pp. 90–98, 2000.

A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

A. Jain, R. Bolle, and S. Pankanti, *Biometrics: personal identification in networked society*. Springer Science & Business Media, 2006, vol. 479.

A. Jain, A. A. Ross, and K. Nandakumar, *Introduction to biometrics*. Springer Science & Business Media, 2011.

A. K. Jain and A. Kumar, "Biometrics of next generation: An overview," *Second Generation Biometrics*, vol. 12, no. 1, pp. 2–3, 2010.

A. K. Jain and A. Ross, "Multibiometric systems," *Communications of the ACM*, vol. 47, no. 1, pp. 34–40, 2004.

A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *IEEE transactions on information forensics and security*, vol. 1, no. 2, pp. 125–143, 2006.

A. K. Jain, S. S. Arora, L. Best-Rowden, K. Cao, P. S. Sudhish, and A. Bhatnagar, "Biometrics for child vaccination and welfare: Persistence of fingerprint recognition for infants and toddlers," *arXiv preprint arXiv:1504.04651*, 2015.

A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, 2016.

Q. Jia, C. Fang, D. Wen, and X. Ding, "Generating face images under multiple illuminations based on a single front-lighted sample without 3D models," in *International Conference on Biometrics*. IEEE, 2013, pp. 1–6.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

Z. Jian, H. Luxi, J. Jian, and X. Yu, "A fast iterative pursuit algorithm in robust face recognition based on sparse representation," *Mathematical Problems in Engineering*, vol. 2014, 2014.

Y. Jiang, Y. Wu, W. Li, L. Wang, and Q. Liao, "Log-domain polynomial filters for illumination-robust face recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 504–508.

R. Jillela and A. Ross, "Matching face against iris images using periocular information," in *IEEE International Conference on Image Processing*. IEEE, 2014, pp. 4997–5001.

K. Jonsson, J. Kittler, Y. Li, and J. Matas, "Support vector machines for face authentication," *Image and Vision Computing*, vol. 20, no. 5, pp. 369–375, 2002.

A. Joshi, A. K. Gangwar, and Z. Saquib, "Person recognition based on fusion of iris and periocular biometrics," in *12th International Conference on Hybrid Intelligent Systems*. IEEE, 2012, pp. 57–62.

T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

C. M. Kahn and J. M. Liñares-Zegarra, "Identity theft and consumer payment choice: Does security really matter?" *Journal of Financial Services Research*, pp. 1–39, 2015.

C. Kandaswamy, L. M. Silva, L. A. Alexandre, J. M. Santos, and J. M. de Sá, "Improving deep neural network performance by reusing features trained with transductive transference," in *Artificial Neural Networks and Machine Learning–ICANN 2014*. Springer, 2014, pp. 265–272.

C. Kandaswamy, L. M. Silva, and J. S. Cardoso, "Source-target-source classification using stacked denoising autoencoders," in *Pattern Recognition and Image Analysis*. Springer, 2015, pp. 39–47.

C. Kandaswamy, J. C. Monteiro, L. M. Silva, and J. S. Cardoso, "Multi-source deep transfer learning for cross-sensor biometrics," *Neural Computing and Applications*, pp. 1–15, 2016.

D. Kang, H. Han, A. K. Jain, and S.-W. Lee, "Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching," *Pattern Recognition*, vol. 47, no. 12, pp. 3750–3766, 2014.

K. J. Karande and S. N. Talbar, *Independent component analysis of edge information for face recognition*. Springer, 2014.

D. Kaur and G. Kaur, "Level of fusion in multimodal biometrics: a review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 2, pp. 242–246, 2013.

Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2004, pp. 506–515.

J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

J. C. Klontz and A. K. Jain, "A case study of automated face recognition: The Boston marathon bombings suspects," *Computer*, vol. 46, no. 11, pp. 91–94, 2013.

KPMG, "Biometrics make life easier and safer for customers – a south african case study," https://home.kpmg.com/za/en/home/insights/2016/01/biometrics-makes-life-easier-and-safer-for-customers-fs.html, 2016, accessed: 2016-11-16.

P. Krishnan and S. Naveen, "RGB-D face recognition system verification using Kinect and FRAV3D databases," *Procedia Computer Science*, vol. 46, pp. 1653–1660, 2015.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

A. Kumar and T.-S. Chan, "Iris recognition using quaternionic sparse orientation code (QSOC)," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 59–64.

D. Kumari, R. Sharma, and D. Kumar, "Framework for preservation of biometric information with facial template protection using gray-level extended visual cryptography scheme," *International Journal of Advanced Research in Computer Science*, vol. 7, no. 3, 2016.

Y. l. Zheng, Z. l. Xu, and G. J. Zhang, "Analysis on the longitudinal temperature and computation of ampacity of 10kv cables in pipes crossing unfavorable thermal regions," in *IEEE 10th International Conference on the Properties and Applications of Dielectric Materials*, July 2012, pp. 1–4.

E. Lachat, H. Macher, M. Mittet, T. Landes, and P. Grussenmeyer, "First experiences with Kinect V2 sensor for close range 3D modelling," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS)*, 2015.

R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes." in *BMVC*, vol. 2, no. 3, 2012, p. 8.

B. Y. Li, A. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *IEEE Workshop on Applications of Computer Vision*. IEEE, 2013, pp. 186–192.

S. Z. Li, L. Zhang, S. Liao, X. Zhu, R. Chu, M. Ao, and R. He, "A near-infrared image based face recognition system," in *7th International Conference on Automatic Face and Gesture Recognition*, April 2006, pp. 455–460.

S. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.

S. Z. Li, R. Chu, M. Ao, L. Zhang, and R. He, *Highly Accurate and Fast Face Recognition Using Near Infrared Images*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 151–158.

Y. Li, L. Meng, J. Feng, and J. Wu, "Downsampling sparse representation and discriminant information aided occluded face recognition," *Science China Information Sciences*, vol. 57, no. 3, pp. 1–8, 2014.

R. Liao, "Support vector machines," http://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial9.pdf, 2015, accessed: 2016-12-12.

S. Liao and A. K. Jain, "Partial face recognition: An alignment free approach," in *International Joint Conference on Biometrics*. IEEE, 2011, pp. 1–8.

C.-C. Lin and W.-H. Tsai, "Visual cryptography for gray-level images by dithering techniques," *Pattern Recognition Letters*, vol. 24, no. 1, pp. 349–358, 2003.

M. Y. Liu, A. Mallya, O. Tuzel, and X. Chen, "Unsupervised network pretraining via encoding human design," in *2016 IEEE Winter Conference on Applications of Computer Vision*, March 2016, pp. 1–9.

D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

X. Lu, R.-L. Hsu, A. K. Jain, B. Kamgar-Parsi, and B. Kamgar-Parsi, "Face recognition with 3D model-based synthesis," in *Biometric Authentication*. Springer, 2004, pp. 139–146.

B. o. B. T. M2SYS, "How biometrics on smartphones is changing our lives," http://blog.m2sys.com/biometric-resources/biometrics-on-smartphones/, 2016, accessed: 2016-11-16.

T. Mäenpää and M. Pietikäinen, "Classification with color and texture: jointly or separately?" *Pattern recognition*, vol. 37, no. 8, pp. 1629–1640, 2004.

Y. Manabe, K. Matsumoto, and K. Sugawara, "Multi-angle gait recognition based on skeletal tracking data," *Journal of Information Processing*, vol. 24, no. 3, pp. 451–459, 2016.

A. M. Martinez, "The AR face database," *CVC Technical Report*, vol. 24, 1998.

D. Martinho-Corbishley, M. Nixon, and J. Carter, "Analysing comparative soft biometrics from crowdsourced annotations," *IET Biometrics*, pp. 1–16, 2016.

D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, "Soft biometric retrieval to describe and identify surveillance images," in *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*. IEEE, 2016, pp. 1–6.

M. Mendoza, "Wells Fargo set to roll out biometric ID for corporate customers," http://www.bizjournals.com/twincities/blog/banking/2015/10/wells-fargo-biometric-id-mobile-corporate-accounts.html, 2015, accessed: 2016-11-16.

A. S. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2D-3D hybrid approach to automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1927–1943, 2007.

A. S. Mian, M. Bennamoun, and R. Owens, "Keypoint detection and local feature matching for textured 3D face recognition," *International Journal of Computer Vision*, vol. 79, no. 1, pp. 1–12, 2008.

Z. Miao and X. Jiang, "Interest point detection using rank order log filter," *Pattern Recognition*, vol. 46, no. 11, pp. 2890–2901, 2013.

Microsoft, "Kinect for Windows sensor components and specifications," 2010, available at: https://msdn.microsoft.com/en-us/library/jj131033.aspx, Accessed on 15 January 2016.

Microsoft, "Kinect for Xbox One," 2014, available at: http://www.xbox.com/en-us/xbox-one/accessories/kinect-for-xbox-one#fbid=AN5dWEs2zxt, Accessed on 15 January 2016.

P. E. Miller, J. R. Lyle, S. J. Pundlik, and D. L. Woodard, "Performance evaluation of local appearance based periocular recognition," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2010, pp. 1–6.

P. E. Miller, A. W. Rawls, S. J. Pundlik, and D. L. Woodard, "Personal identification using periocular skin texture," in *2010 ACM Symposium on Applied Computing*. ACM, 2010, pp. 1496–1500.

R. Min, J. Choi, G. Medioni, and J.-L. Dugelay, "Real-time 3D face identification from a depth camera," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1739–1742.

R. Min, A. Hadid, and J.-L. Dugelay, "Efficient detection of occlusion prior to robust face recognition," *The Scientific World Journal*, vol. 2014, 2014.

R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 11, pp. 1534–1548, 2014.

K. Minolta, "Konica Minolta 3D laser scanners," 2006, available at: http://www.3dscanco.com/products/3d-scanners/3d-laser-scanners/konica-minolta/, Accessed on 15 January 2016.

T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill, Inc., 1997.

T. Miyatake, E. A. Nagasaka, and K. Kumakiri, "Cross-industry deployment of core technology and associated ip management support," *Hitachi Review*, vol. 64, no. 6, p. 333, 2015.

J. C. Monteiro and J. S. Cardoso, "Partial face recognition under occlusion using universal background models," in *Proceedings of the 20th Portuguese Conference on Pattern Recognition (RECPAD)*, 2014.

J. C. Monteiro and J. S. Cardoso, "A cognitively-motivated framework for partial face recognition in unconstrained scenarios," *Sensors*, vol. 15, no. 1, pp. 1903–1924, 2015.

J. C. Monteiro and J. S. Cardoso, "A novel application of universal background models for periocular recognition," in *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer International Publishing, 2015, pp. 291–307.

J. C. Monteiro and E. Mota, "An assessment of the potential of distinct facial regions for biometric recognition," in *Proceedings of the 19th Portuguese Conference on Pattern Recognition (RECPAD)*, 2013.

J. C. Monteiro, R. Esteves, G. Santos, P. T. Fiadeiro, J. Lobo, and J. S. Cardoso, "A comparative analysis of two approaches to periocular recognition in mobile scenarios," in *Advances in Visual Computing*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 9475, pp. 268–280.

J. C. Monteiro, T. Freitas, and J. S. Cardoso, "Multimodal hierarchical face recognition using information from 2.5D images," *University of Porto Journal of Engineering*, vol. 2, no. 2, pp. 39–54, 2016.

J. Moodley, S. Naidoo, N. S. Morar, V. Govender, P. Charls, and G. Ramjee, "Ethical considerations in implementing a biometric co-enrolment prevention system in clinical trials in south africa," *South African Journal of Bioethics and Law*, vol. 9, no. 1, p. 39, 2016.

A. Morelli Andrés, S. Padovani, M. Tepper, and J. Jacobo-Berlles, "Face recognition on partially occluded images using compressed sensing," *Pattern Recognition Letters*, vol. 36, pp. 235–242, 2014.

A. B. Moreno, A. Sánchez, J. F. Vélez, and F. J. Díaz, "Face recognition using 3D surface-extracted descriptors," in *Irish Machine Vision and Image Processing Conference*, vol. 2003. Citeseer, 2003.

J. C. Moreno, V. B. S. Prasath, G. M. M. Santos, and H. Proença, "Robust periocular recognition by fusing sparse representations of color and geometry information," *CoRR*, vol. abs/1309.2752, 2013.

J. C. Moreno, V. Prasath, and H. Proença, "Robust periocular recognition by fusing local to holistic sparse representations," in *Proceedings of the 6th International Conference on Security of Information and Networks*. ACM, 2013, pp. 160–164.

S. Mracek, M. Drahansky, R. Dvorák, I. Provaznik, and J. Vána, "3D face recognition on low-cost depth sensors," in *International Conference of the Biometrics Special Interest Group*. IEEE, 2014, pp. 1–4.

K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

N. Nallammal and V. Radha, "Comparative analysis of partial occlusion using face recognition techniques," *International Journal of Image Processing (IJIP)*, vol. 7, no. 2, p. 132, 2013.

A. Nandhinipreetha and N. Radha, "Multimodal biometric template authentication of finger vein and signature using visual cryptography," in *2016 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2016, pp. 1–4.

S. Naveen and R. Moni, "A robust novel method for face recognition from 2D depth images using DWT and DCT fusion," *Procedia Computer Science*, vol. 46, pp. 1518–1528, 2015.

S. Naveen, S. S. Nair, and R. Moni, "3D face recognition using optimised directional faces and fourier transform," in *International Conference on Advances in Computing, Communications and Informatics*. IEEE, 2015, pp. 1856–1861.

D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive psychology*, vol. 9, no. 3, pp. 353–383, 1977.

B.-S. Oh, K.-A. Toh, K. Choi, A. Beng Jin Teoh, and J. Kim, "Extraction and fusion of partial face features for cancelable identity verification," *Pattern Recognition*, vol. 45, no. 9, pp. 3288–3303, 2012.

T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

D. Ooms, R. Palm, V. Leemans, and M.-F. Destain, "A sorting optimization curve with quality and yield requirements," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 983–990, 2010.

OpenCV, "Introduction to support vector machines," http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html, 2016, accessed: 2016-12-12.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? – Weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

C. N. Padole and H. Proenca, "Periocular recognition: Analysis of performance degradation factors," in *5th IAPR International Conference on Biometrics*, 2012, pp. 439–445.

K. Pan, S. Liao, Z. Zhang, S. Z. Li, and P. Zhang, "Part-based face recognition using near infrared images," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–6.

S. J. Pan, D. Shen, Q. Yang, and J. T. Kwok, "Transferring localization models across space." in *AAAI*, 2008, pp. 1383–1388.

T. Papatheodorou and D. Rueckert, "Evaluation of automatic 4D face recognition using surface and texture registration," in *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2004, pp. 321–326.

U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study," in *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, 2009, pp. 1–6.

U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 947–954, 2010.

U. Park, R. R. Jillela, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 96–106, 2011.

O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *Proceedings of the British Machine Vision*, vol. 1, no. 3, p. 6, 2015.

K. Patel, H. Han, and A. K. Jain, "Secure face unlock: spoof detection on smartphones," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2268–2283, 2016.

P. J. Phillips *et al.*, *Support vector machines applied to face recognition*. US Department of Commerce, Technology Administration, National Institute of Standards and Technology, 1998, vol. 285.

J. Pillai, M. Puertas, and R. Chellappa, "Cross-sensor iris recognition through kernel learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 73–85, Jan 2014.

P. Planet Biometrics, "Lisbon airport expands access to biometric immigration system," http://www.planetbiometrics.com/article-details/i/4756/desc/lisbon-airport-expands-access-to-biometric-immigration-system/, 2016, accessed: 2016-11-16.

P. Planet Biometrics, "South african airport begins biometric data capture," http://www.planetbiometrics.com/article-details/i/4687/desc/south-african-airport-begins-biometric-data-capture/, 2016, accessed: 2016-11-16.

D. Povey, S. M. Chu, and B. Varadarajan, "Universal background model based speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4561–4564.

S. Prabhakar, S. Pankanti, and A. K. Jain, "Biometric recognition: Security and privacy concerns," *Security & Privacy, IEEE*, vol. 1, no. 2, pp. 33–42, 2003.

P. Precise Biometrics, "Understanding biometric performance evaluation," http://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation.pdf, 2015, accessed: 2016-11-16.

H. Proença, "NICE.II: Noisy iris challenge evaluation - Part II," 2009, http://http://nice2.di.ubi.pt/.

H. Proença, "Non-cooperative iris recognition: Issues and trends," in *19th European Signal Processing Conference*, 2011, pp. 1–5.

H. Proença and L. Alexandre, "The NICE.I: noisy iris challenge evaluation - Part I," in *First IEEE International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 2007, pp. 1–4.

H. Proença, "Towards non-cooperative biometric iris recognition," *University of Beira Interior. Department of Computer Science*, 2006.

H. Proença and J. C. Neves, "Visible-wavelength iris/periocular imaging and recognition surveillance environments," *Image and Vision Computing*, 2016.

H. Proença and G. Santos, "Fusing color and shape descriptors in the recognition of degraded iris images acquired at visible wavelengths," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 167–178, 2012.

H. Proença, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre, "The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1529–1535, 2010.

J. Qian, J. Yang, F. Zhang, and Z. Lin, "Robust low-rank regularized regression for face recognition with occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 21–26.

R. Rajkumar, K. N. Devi, and K. Hemachandran, "Fingerprint identification system in library and information center," *International Journal on Computer Science and Engineering*, vol. 3, no. 12, p. 3681, 2011.

J. Renaghan, "Etched in stone," *Smithsonian Zoogoer Magazine*, 1997.

D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827–832, 2015.

D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2002, pp. IV–4072.

A. Ross and A. K. Jain, "Multimodal biometrics: An overview," in *12th European Signal Processing Conference*. IEEE, 2004, pp. 1221–1224.

A. Ross, A. Rattani, and M. Tistarelli, "Exploiting the "Doddington zoo" effect in biometric fusion," in *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 2009, pp. 1–7.

A. Ross, R. Jillela, J. M. Smereka, V. N. Boddeti, B. V. Kumar, R. Barnard, X. Hu, P. Pauca, and R. Plemmons, "Matching highly non-ideal ocular images: An information fusion approach," in *5th International Conference on Biometrics*. IEEE, 2012, pp. 446–453.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*. IEEE, 1994, pp. 138–142.

G. Sang, J. Li, and Q. Zhao, "Pose-invariant face recognition via RGB-D images," *Computational Intelligence and Neuroscience*, vol. 2016, 2015.

G. Santos and H. Proença, "Periocular biometrics: An emerging technology for unconstrained scenarios," in *IEEE Workshop on Computational Intelligence in Biometrics and Identity Management*. IEEE, 2013, pp. 14–21.

G. Santos, E. Grancho, M. V. Bernardo, and P. T. Fiadeiro, "Fusing iris and periocular information for cross-sensor recognition," *Pattern Recognition Letters*, vol. 57, pp. 52–59, 2015.

A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management*. Springer, 2008, pp. 47–56.

A. Schwaninger, S. Schumacher, H. Bülthoff, and C. Wallraven, "Using 3D computer graphics for perception: The role of local and global information in face processing," in *Proceedings of the 4th symposium on applied perception in graphics and visualization*. Tubingen, Germany: ACM, 2007, pp. 19–26.

N. Science and N. Technology Council, "Biometric standards," file:///home/jcsm/Downloads/biostandards.pdf, 2016, accessed: 2016-12-19.

A. F. Sequeira, J. C. Monteiro, A. Rebelo, and H. P. Oliveira, "MobBIO: a multimodal database captured with a portable handheld device," in *Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 3, 2014, pp. 133–139.

Y. Shen, W. Hu, M. Yang, B. Wei, S. Lucey, and C. T. Chou, "Face recognition on smartphones via optimised sparse representation classification," in *Proceedings of the 13th international symposium on Information processing in sensor networks*. IEEE Press, 2014, pp. 237–248.

K. Shinoda and N. Inoue, "Reusing speech techniques for video semantic indexing," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 118–122, 2013.

K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proceedings of BMVC*, vol. 1, no. 2, 2013, p. 7.

J. M. Smereka and B. Kumar, "What is a" good" periocular region for recognition?" in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2013, pp. 117–124.

SofKinetic, "SofKinetic DepthSense® 325," 2007, available at: http://www.softkinetic.com/Store/ProductID/6, Accessed on 15 January 2016.

S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.

Structure, "Structure sensor ST01," 2013, available at: http://structure.io/embedded, Accessed on 15 January 2016.

M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, Nov. 1991.

Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1701–1708.

C.-W. Tan and A. Kumar, "Towards online iris and periocular recognition under relaxed imaging constraints," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3751–3765, 2013.

T. Tan, Z. He, and Z. Sun, "Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition," *Image and vision computing*, vol. 28, no. 2, pp. 223–230, 2010.

T. Tan, X. Zhang, Z. Sun, and H. Zhang, "Noisy iris image matching by using multiple cues," *Pattern Recognition Letters*, 2011.

X. Tan and B. Triggs, "Fusing Gabor and LBP feature sets for kernel-based face recognition," in *Proceedings of the 3rd International Conference on Analysis and Modeling of Faces and Gestures*, ser. AMFG'07, 2007, pp. 235–249.

H. T. Tanaka, M. Ikeda, and H. Chiaki, "Curvature-based face surface recognition using spherical correlation. principal directions for curved object recognition," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 372–377.

J. W. Tanaka and M. J. Farah, "Parts and wholes in face recognition," *The Quarterly Journal of Experimental Psychology*, vol. 46, no. 2, pp. 225–245, 1993.

Y. Tang, X. Sun, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen, "3D face recognition with asymptotic cones based principal curvatures," in *International Conference on Biometrics*. IEEE, 2015, pp. 466–472.

T. Thomson Reuters Canada, "UK's RBS launches fingerprint technology for mobile banking app," http://ca.reuters.com/article/technologyNews/idCAKBN0LM00K20150218, 2016, accessed: 2016-11-16.

S. Thrun, "Lifelong learning algorithms," in *Learning to learn*. Springer, 1998, pp. 181–209.

F. Tsalakanidou, D. Tzovaras, and M. G. Strintzis, "Use of depth and colour eigenfaces for face recognition," *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1427–1435, 2003.

F. Tsalakanidou, S. Malassiotis, and M. G. Strintzis, "Face localization and authentication using color and depth images," *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 152–168, 2005.

M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain, "Biometric cryptosystems: issues and challenges," *Proceedings of the IEEE*, vol. 92, no. 6, pp. 948–960, 2004.

D. van Greunen, "Ethics, children, and biometric technology," *IEEE Technology and Society Magazine*, vol. 35, no. 3, pp. 67–72, 2016.

V. Veridium, "Biometrics: Bringing future tech to airport security," https://www.veridiumid.com/blog/biometrics-bringing-future-tech-airport-security/, 2016, accessed: 2016-11-16.

U. Verma and C. Kant, "Secure biometric template protection approach using chaotic maps," *International Journal of Advanced Research in Computer Science*, vol. 7, no. 3, 2016.

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

P. Wagner, "Local binary patterns," http://bytefish.de/blog/local_binary_patterns/, 2011, accessed: 2016-10-13.

B. Wang, W. Li, W. Yang, and Q. Liao, "Illumination normalization based on Weber's law with application to face recognition," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 462–465, 2011.

J.-G. Wang, J. Li, W.-Y. Yau, and E. Sung, "Boosting dense SIFT descriptors and shape contexts of face images for gender recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2010, pp. 96–102.

J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, "Robust face recognition via adaptive sparse representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2368–2378, 2014.

Q. Wang and S. You, "A vision-based 2D-3D registration system," in *Workshop on Applications of Computer Vision*. IEEE, 2009, pp. 1–8.

Y. Wang, G. Pan, Z. Wu, and S. Han, "Sphere-spin-image: A viewpoint-invariant surface representation for 3D face recognition," in *Computational Science-ICCS 2004*. Springer, 2004, pp. 427–434.

L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

D. L. Woodard, S. J. Pundlik, J. R. Lyle, and P. E. Miller, "Periocular region appearance cues for biometric identification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 162–169.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

S. Xia, G. Zheng, Y. Ma, and X. Ma, "A novel super-resolution approach based on supervised canonical correlation analysis," in *2014 International Conference on Computer, Communications and Information Technology (CCIT 2014)*. Atlantis Press, 2014.

Z. Xiong, T. Zheng, Z. Song, F. Soong, and W. Wu, "A tree-based kernel selection approach to efficient Gaussian mixture model–universal background model based speaker identification," *Speech communication*, vol. 48, no. 10, pp. 1273–1282, 2006.

W. Xu and E.-J. Lee, "Face recognition using wavelets transform and 2D PCA by SVM classifier," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, no. 3, pp. 281–290, 2014.

X. Xu and R. W. McGorry, "The validity of the first and second generation Microsoft Kinect for identifying joint center locations during static postures," *Applied Ergonomics*, vol. 49, pp. 47 – 54, 2015.

Y. Xu, A. Zhong, J. Yang, and D. Zhang, "Bimodal biometrics based on a representation and recognition approach," *Optical Engineering*, vol. 50, no. 3, pp. 037 202–1, 2011.

J. Yang, J. Qian, L. Luo, F. Zhang, and Y. Gao, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *arXiv preprint arXiv:1405.1207*, 2014.

J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 188–197.

Y. Ye, H. Zheng, L. Ni, S. Liu, and W. Li, "A study on the individuality of finger vein based on statistical analysis," in *International Conference on Biometrics*. IEEE, 2016, pp. 1–5.

M. Yildiz, B. Yanikoğlu, A. Kholmatov, A. Kanak, U. Uludağ, and H. Erdoğan, "Biometric layering with fingerprints: Template security and privacy through multi-biometric template fusion," *The Computer Journal*, 2016.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

B. Zhang, L. Zhang, D. Zhang, and L. Shen, "Directional binary code with application to PolyU near-infrared face database," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2337 – 2344, 2010.

B. Zhang, L. Zhang, D. Zhang, L. SHEN, and Z. GUO, "PolyU near-infrared face database," 2010.

D.-L. Zhang, J. Qiao, J.-B. Li, L.-Y. Qiao, S.-C. Chu, and J. F. Roddick, "Optimizing matrix mapping with data dependent kernel for image classification," *Journal of Information Hiding and Multimedia Signal Processingm*, vol. 5, no. 1, 2014.

W. Zhang, S. Shan, X. Chen, and W. Gao, "Local Gabor binary patterns based on Kullback–Leibler divergence for partially occluded face recognition," *IEEE Signal Processing Letters*, vol. 14, no. 11, pp. 875–878, 2007.

G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

S. Zhao and R.-R. Grigat, *An Automatic Face Recognition System in the Near Infrared Spectrum*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 437–444.

Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using Markov random fields," in *IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 1050–1057.

Z. Zvetco Biometrics, "Biometric knowledge center," http://www.zvetcobiometrics.com/Support/definitions.php, 2012, accessed: 2016-10-19.