UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

QUIS-CAMPI: Biometric Recognition in Surveillance Scenarios

João Carlos Raposo Neves

Tese para obtenção do Grau de Doutor em
Engenharia Informática
(3º ciclo de estudos)

Orientador: Prof. Doutor Hugo Pedro Martins Carriço Proença

Covilhã, junho de 2018

Thesis prepared at *IT - Instituto de Telecomunicações*, Pattern and Image Analysis Group, Covilhã Delegation, and submitted to University of Beira Interior for defense in a public examination session.

This work has been supported by '*FCT - Fundação para a Ciência e Tecnologia*' (Portugal) through the project 'UID/EEA/50008/2013', the research grant 'SFRH/BD/92520/2013', and the funding from 'FEDER - QREN - Type 4.1 - *Formação Avançada*', co-founded by the European Social Fund and by national funds through Portuguese 'MEC - *Ministério da Educação e Ciência*'.

Governo da República Portuguesa

UNIÃO EUROPEIA
Fundo Social Europeu

Acknowledgments

The research work presented in this thesis would not have been possible without the help and support of the individuals and organizations mentioned here.

First and foremost, I would like to express my utmost gratitude to my supervisor, Professor Dr. Hugo Proença for his constant guidance, trust, patience, and support. Without his motivation, encouragement, expertise, research insight, and invaluable help it would not have been possible to complete this thesis.

The support of the University of Beira Interior, and *Instituto de Telecomunicações* is also acknowledged, as well as the financial support of the *Fundação para a Ciência e a Tecnologia* (FCT) through the UID/EEA/500008/2013 Project and the grant contract SFRH/BD/92520/2013.

Also, I would also like to express my gratitude to all my colleagues of the Soft Computing and Image Analysis Lab (SOCIA-LAB) research group, for the enjoyable working environment that they contributed to. A special acknowledgment is also due to all the volunteers that took part in the acquisition of the QUIS-CAMPI dataset.

Last, but not least, I would like to thank all the people close to myself in the last years, for their strong support, encouragement, friendship, and love. I am grateful for their understanding for the time during which I was absent due to this research work.

List of Publications

Articles included in the thesis resulting from this 4-year doctoral research program

Published, accepted or submitted for publication in international journals

1. **"A Leopard Cannot Change Its Spots": Improving Face Recognition Using 3D-based Caricatures**

   João C. Neves and Hugo Proença

   IEEE Transactions on Information Forensics and Security, in press.

   DOI: dx.doi.org/10.1109/TIFS.2018.2846617

2. **QUIS-CAMPI: An Annotated Multi-biometrics Data Feed From Surveillance Scenarios**

   João C. Neves, Juan C. Moreno and Hugo Proença

   IET Biometrics, Vol. 7, No. 4, pp. 371-379, 2018.

   DOI: dx.doi.org/10.1049/iet-bmt.2016.0178

3. **Biometric Recognition in Surveillance Scenarios: A survey**

   João C. Neves, Fabio Narducci, Silvio Barra, and Hugo Proença

   Springer Artificial Intelligence Review, Vol. 46, No. 4, pp. 1-27, 2016.

   DOI: dx.doi.org/10.1007/s10462-016-9474-x

4. **A Master-slave Calibration Algorithm with Fish-eye Correction**

   João C. Neves, Juan C. Moreno and Hugo Proença

   Hindawi Mathematical Problems in Engineering, Article ID: 427270, 2015.

   DOI: dx.doi.org/10.1155/2015/427270

Presented and published in the proceedings of international conferences

1. **Exploiting Data Redundancy for Error Detection in Degraded Biometric Signatures Resulting From in the Wild Environments**

   João C. Neves and Hugo Proença

   *Proceedings of the International Workshop on Biometrics in the Wild, IEEE Conference on Automatic Face and Gesture Recognition (**FG**)*, Washington D.C., USA, 2017, pp. 981-986.

   DOI: dx.doi.org/10.1109/FG.2017.122

2. **ICB-RW 2016: International Challenge on Biometric Recognition in the Wild**

   João C. Neves and Hugo Proença

   *Proceedings of the International International Conference on Biometrics (**ICB**)*, Halmstad, Sweden, 2016, pp. 1-6.

   DOI: dx.doi.org/10.1109/ICB.2016.7550066

3. **Acquiring High-resolution Face Images in Outdoor Environments: A master-slave Calibration Algorithm**

   João C. Neves, Juan Moreno, Silvio Barra, and Hugo Proença

   *IEEE International Conference on Biometrics: Theory, Applications and Systems (**BTAS**)*, Washington D.C., USA, 2015, pp. 1-8.

   DOI: dx.doi.org/10.1109/BTAS.2015.7358744

4. **Dynamic Camera Scheduling for Visual Surveillance in Crowded Scenes using Markov Random Fields**

   João C. Neves and Hugo Proença

   *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance (**AVSS**)*, Karlsruhe, Germany, 2015, pp. 1-8.

   DOI: dx.doi.org/10.1109/AVSS.2015.7301790

5. **Quis-Campi: Extending In The Wild Biometric Recognition to Surveillance Environments**

   João C. Neves, Gil Santos, Silvio Filipe, Emanuel Grancho, Silvio Barra, Fabio Narducci and Hugo Proença

   *Proceedings of the International Conference on Image Analysis and Processing (**ICIAP**)*, Karlsruhe, Germany, 2015, pp. 1-8.

   DOI: dx.doi.org/10.1007/978-3-319-23222-5_8

6. **A Calibration Algorithm for Multi-camera Visual Surveillance Systems Based on Single-View Metrology**

   João C. Neves, Juan Moreno, Silvio Barra, and Hugo Proença

   *Iberian Conference on Pattern Recognition and Image Analysis (**IbPRIA**)*, Santiago de Compostela, Spain, 2015, pp. 552-559.

   DOI: dx.doi.org/10.1007/978-3-319-19390-8_62

7. **Evaluation of Background Subtraction Algorithms for Human Visual Surveillance**

   João C. Neves, and Hugo Proença

   *Proceedings of the IEEE International Conference on Signal and Image Processing Applications (**ICSIPA**)*, Kuala Lumpur, Malaysia, 2015.

Published as a book chapter

1. **Unconstrained Data Acquisition Frameworks and Protocols**

   João C. Neves, Juan Moreno, Silvio Barra, Fabio Narducci and Hugo Proença

   Human Recognition in Unconstrained Environments. M. de Marsico, M. Nappi, H. Proença (Eds.) Springer-Verlag book series, Communications Engineering/Computer Vision, pp. 1-30, 2017.

   DOI: dx.doi.org/10.1016/B978-0-08-100705-1.00001-4

Other publications resulting from this doctoral research program not included in the thesis

1. **IRINA: Iris Recognition (even) in Inacurately Segmented Data**

   Hugo Proença and **João C. Neves**

   *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**)*, Honolulu, Hawai, U.S.A, 2017, pp. 538-547.

2. **Soft Biometrics: Globally Coherent Solutions for Hair Segmentation and Style Recognition based on Hierarchical MRFs**

   Hugo Proença and **João C. Neves**

   IEEE Transactions on Information Forensics and Security, Vol. 12, No. 7, pp. 1637-1545, 2017.

   DOI: dx.doi.org/10.1109/TIFS.2017.2680246

3. **Fusing Vantage Point Trees and Linear Discriminants for Fast Feature Classification**

   Hugo Proença and **João C. Neves**

   Springer Journal of Classification, Vol. 34, pp. 85-107, 2017.

   DOI: dx.doi.org/10.1007/s00357-017-9223-0

4. **Visible-wavelength Iris / Periocular Imaging and Recognition in Surveillance Environments**

   Hugo Proença and **João C. Neves**

   Elsevier Image and Vision Computing, Vol. 55, pp. 22-25, 2016.

   DOI: dx.doi.org/10.1016/j.imavis.2016.03.015

5. **Joint Head Pose / Soft Label Estimation for Human Recognition In-The-Wild**

   Hugo Proença, **João C. Neves**, Silvio Barra, Tiago Marques and Juan C. Moreno

   IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 38, No. 12, pp. 2444-2456, 2016.

   DOI: dx.doi.org/10.1109/TPAMI.2016.2522441

6. **Segmenting the Periocular Region using a Hierarchical Graphical Model Fed by Texture/Shape Information and Geometrical Constraints**

   Hugo Proença, **João C. Neves**, and Gil Santos

   *Proceedings of the IEEE International Joint Conference on Biometrics (**IJCB**)*, Clearwater, USA, 2014, pp. 1-8.

   DOI: dx.doi.org/10.1109/BTAS.2014.6996228

Resumo

A crescente preocupação com a segurança dos indivíduos tem justificado o crescimento do número de câmaras de vídeo-vigilância instaladas tanto em espaços privados como públicos. Contudo, ao contrário do que normalmente se pensa, estes dispositivos são, na maior parte dos casos, usados apenas para gravação, não estando ligados a nenhum tipo de software inteligente capaz de inferir em tempo real informações sobre os indivíduos observados. Assim, apesar de a vídeo-vigilância ter provado ser essencial na resolução de diversos crimes, o seu uso está ainda confinado à disponibilização de vídeos que têm que ser manualmente inspecionados para extrair informações relevantes dos sujeitos envolvidos no crime. Como tal, atualmente, o principal desafio da comunidade científica é o desenvolvimento de sistemas automatizados capazes de monitorizar e identificar indivíduos em ambientes de vídeo-vigilância.

Esta tese tem como principal objetivo estender a aplicabilidade dos sistemas de reconhecimento biométrico aos ambientes de vídeo-vigilância. De forma mais especifica, pretende-se 1) conceber um sistema de vídeo-vigilância que consiga adquirir dados biométricos a longas distâncias (e.g., imagens da cara, íris, ou vídeos do tipo de passo) sem requerer a cooperação dos indivíduos no processo; e 2) desenvolver métodos de reconhecimento biométrico robustos aos fatores de degradação inerentes aos dados adquiridos por este tipo de sistemas.

No que diz respeito ao primeiro objetivo, a análise aos dados adquiridos pelos sistemas típicos de vídeo-vigilância mostra que, devido à distância de captura, os traços biométricos amostrados não são suficientemente discriminativos para garantir taxas de reconhecimento aceitáveis. Na literatura, vários trabalhos advogam o uso de câmaras Pan Tilt Zoom (PTZ) para adquirir imagens de alta resolução à distância, principalmente o uso destes dispositivos no modo *master-slave*. Na configuração *master-slave* um módulo de análise inteligente seleciona zonas de interesse (e.g. carros, pessoas) a partir do vídeo adquirido por uma câmara de vídeo-vigilância e a câmara PTZ é orientada para adquirir em alta resolução as regiões de interesse. Diversos métodos já mostraram que esta configuração pode ser usada para adquirir dados biométricos à distância, ainda assim estes não foram capazes de solucionar alguns problemas relacionados com esta estratégia, impedindo assim o seu uso em ambientes de vídeo-vigilância. Deste modo, esta tese propõe dois métodos para permitir a aquisição de dados biométricos em ambientes de vídeo-vigilância usando uma câmara PTZ assistida por uma câmara típica de vídeo-vigilância. O primeiro é um método de calibração capaz de mapear de forma exata as coordenadas da câmara *master* para o ângulo da câmara PTZ (*slave*) sem o auxílio de outros dispositivos óticos. O segundo método determina a ordem pela qual um conjunto de sujeitos vai ser observado pela câmara PTZ. O método proposto consegue determinar em tempo-real a sequência de observações que maximiza o número de diferentes sujeitos observados e simultaneamente minimiza o tempo total de transição entre sujeitos. De modo a atingir o primeiro objetivo desta tese, os dois métodos propostos foram combinados com os avanços alcançados na área da monitorização de humanos para assim desenvolver o primeiro sistema de vídeo-vigilância completamente automatizado e capaz de adquirir dados biométricos a longas distâncias sem requerer a cooperação dos indivíduos no processo, designado por sistema QUIS-CAMPI.

O sistema QUIS-CAMPI representa o ponto de partida para iniciar a investigação relacionada com o segundo objetivo desta tese. A análise do desempenho dos métodos de reconhecimento biométrico do estado-da-arte mostra que estes conseguem obter taxas de reconhecimento quase perfeitas em dados adquiridos sem restrições (e.g., taxas de reconhecimento maiores do que 99% no conjunto de dados LFW). Contudo, este desempenho não é corroborado

pelos resultados observados em ambientes de vídeo-vigilância, o que sugere que os conjuntos de dados atuais não contêm verdadeiramente os fatores de degradação típicos dos ambientes de vídeo-vigilância. Tendo em conta as vulnerabilidades dos conjuntos de dados biométricos atuais, esta tese introduz um novo conjunto de dados biométricos (imagens da face e vídeos do tipo de passo) adquiridos pelo sistema QUIS-CAMPI a uma distância máxima de 40m e sem a cooperação dos sujeitos no processo de aquisição. Este conjunto permite avaliar de forma objetiva o desempenho dos métodos do estado-da-arte no reconhecimento de indivíduos em imagens/vídeos capturados num ambiente real de vídeo-vigilância. Como tal, este conjunto foi utilizado para promover a primeira competição de reconhecimento biométrico em ambientes não controlados. Esta tese descreve os protocolos de avaliação usados, assim como os resultados obtidos por 9 métodos especialmente desenhados para esta competição. Para além disso, os dados adquiridos pelo sistema QUIS-CAMPI foram essenciais para o desenvolvimento de dois métodos para aumentar a robustez aos fatores de degradação observados em ambientes de vídeo-vigilância. O primeiro é um método para detetar características corruptas em assinaturas biométricas através da análise da redundância entre subconjuntos de características. O segundo é um método de reconhecimento facial baseado em caricaturas automaticamente geradas a partir de uma única foto do sujeito. As experiências realizadas mostram que ambos os métodos conseguem reduzir as taxas de erro em dados adquiridos de forma não controlada.

Palavras-chave

Sistemas de Vídeo-vigilância Automatizados, Câmaras PTZ, Configuração *Master-slave*, Calibração de Câmaras, Planeamento de Câmaras, Reconhecimento Biométrico, Conjuntos de Dados Biométricos, Reconhecimento Facial, Reconhecimento Facial à distância, Reconhecimento facial baseado em caricaturas, Monitorização de Humanos.

Resumo alargado em Português

A crescente preocupação com a segurança dos indivíduos tem justificado o crescimento do número de câmaras de vídeo-vigilância instaladas tanto em espaços privados como públicos. Contudo, ao contrário do que normalmente se pensa, estes dispositivos são, na maior parte dos casos, usados apenas para gravação, não estando ligados a nenhum tipo de software inteligente capaz de inferir em tempo real informações sobre os indivíduos observados. Assim, apesar de a vídeo-vigilância ter provado ser essencial na resolução de diversos crimes, o seu uso está ainda confinado à disponibilização de vídeos que têm que ser manualmente inspecionados para extrair informações relevantes dos sujeitos envolvidos no crime. Como tal, atualmente, o principal desafio da comunidade científica é o desenvolvimento de sistemas automatizados capazes de monitorizar e identificar indivíduos em ambientes de vídeo-vigilância.

Esta tese tem como principal objetivo estender a aplicabilidade dos sistemas de reconhecimento biométrico aos ambientes de vídeo-vigilância. De forma mais específica, pretende-se 1) conceber um sistema de vídeo-vigilância que consiga adquirir dados biométricos a longas distâncias (e.g., imagens da cara, íris, ou vídeos do tipo de passo) sem requerer a cooperação dos indivíduos no processo; e 2) desenvolver métodos de reconhecimento biométrico robustos aos fatores de degradação inerentes aos dados adquiridos por este tipo de sistemas. De forma a alcançar estes objetivos, esta tese apresenta várias contribuições descritas ao longo de seis capítulos.

O primeiro capítulo define o âmbito e o problema onde esta tese se enquadra. Para além disso, são também descritos os principais objetivos do presente trabalho de investigação, assim como as principais contribuições desta investigação no melhoramento do desempenho do reconhecimento biométrico em ambientes de vídeo-vigilância.

O segundo capítulo apresenta uma revisão da literatura nas três áreas de investigação necessárias ao desenvolvimento de um sistema automatizado de vídeo-vigilância capaz de reconhecer humanos à distância e de maneira sub-reptícia. Estas áreas são as seguintes: 1) monitorização de humanos; 2) sistemas de vídeo-vigilância; e 3) reconhecimento biométrico. No que diz respeito à monitorização de humanos, são apresentados os trabalhos mais relevantes com possível aplicação em ambientes de vídeo-vigilância. Em primeiro lugar, foi feita uma revisão dos algoritmos de subtração de fundo, dado que a deteção das zonas de movimento é utilizada habitualmente pela maioria dos métodos de deteção e *tracking*. Com esta revisão foi possível concluir que existe uma crescente preocupação em desenvolver métodos capazes de operar em ambientes de vídeo-vigilância, e que a robustez dos algoritmos estado da arte neste tipo de ambientes tem vindo a aumentar. De seguida, foi feita a revisão das duas principais estratégias utilizadas na deteção de objetos: 1) deteção holística; e 2) deteção baseada em partes. De forma semelhante à fase anterior, também existe uma preocupação crescente em estender a aplicabilidade destes métodos para cenários sem restrições, sendo a estratégia baseada em partes a mais utilizada até ao momento. Por fim, a análise aos métodos de *tracking* foi feita de acordo com o tipo de estratégia adotada e com o tipo de características utilizadas. Esta análise permitiu perceber que as características de aparência juntamente com a estratégia *tracking-by-detection* são as mais utilizadas em ambientes mais dinâmicos. No que diz respeito aos sistemas de vídeo-vigilância, foi feita uma revisão das arquiteturas utilizadas para adquirir dados biométricos à distância e sem a cooperação dos sujeitos no processo. Em primeiro lugar, as principais arquiteturas foram comparadas, de modo a evidenciar que o uso de câmaras PTZ é a estratégia mais prática e eficiente para a aquisição de dados biométricos. Nas arquiteturas baseadas em câmaras PTZ foram revistas as duas principais opções, onde foi possível perceber

as desvantagens das estratégias baseadas numa única câmara PTZ. Pelo contrário, a combinação de uma câmara PTZ com uma câmara típica de vídeo-vigilância é considerada a estratégia mais adequada para obter dados de alta resolução à distância, o que justifica o fato da maioria dos sistemas estado da arte adotarem esta configuração, designada por *master-slave*. De seguida, foram analisados com particular detalhe os sistemas estado da arte que usaram câmaras PTZ para adquirir dados biométricos à distância. Após esta revisão foi possível perceber os principais problemas deste tipo de arquitetura. No que diz respeito ao reconhecimento biométrico, foi feita uma breve revisão da evolução das formas de reconhecer indivíduos usando diferentes traços biométricos, das medidas usadas para avaliar um sistema biométrico, e dos modos de operação destes sistemas. Para além disso, os métodos estado da arte foram revistos de acordo com o traço biométrico usado. De entre os vários traços biométricos, foram apenas escolhidos a face e o tipo de passo, por serem os mais propícios a serem adquiridos à distância.

O terceiro capítulo descreve as principais contribuições desta tese na área dos sistemas de vídeo-vigilância, mais concretamente dois algoritmos que permitiram o desenvolvimento do sistema QUIS-CAMPI, o primeiro sistema automatizado de vídeo-vigilância capaz de adquirir dados biométricos em ambientes exteriores a longas distâncias (até 40m) e sem requerer a colaboração dos sujeitos no processo. Em primeiro lugar, é apresentado um algoritmo de calibração de câmaras para arquiteturas *master-slave*. Esta abordagem tem como objetivo permitir que as coordenadas da câmara *master* sejam mapeadas para os ângulos da câmara PTZ sem recorrer a dispositivos óticos adicionais ou ser necessário colocar as câmaras numa posição específica. Para isso, este método propõe usar a altura do sujeito para resolver o sistema de equações indeterminado que transforma as coordenadas da câmara *master* na orientação da câmara PTZ. De modo a permitir uma solução completamente automatizada, a altura é também inferida em tempo real usando os pontos de fuga da cena. As principais vantagens deste método são: 1) permitir a instalação da arquitetura *master-slave* no exterior sem comprometer a exatidão do mapeamento entre as duas câmaras; e 2) aumentar a distância máxima de captura em relação aos sistemas existentes na literatura. Em segundo lugar, é apresentado um algoritmo de planeamento da sequência de observações que irão ser realizadas por uma câmara PTZ. Esta abordagem tem como objetivo minimizar o tempo total de transição entre sujeitos de modo a adquirir imagens/vídeos do maior número de sujeitos. Para isso, este método propõe usar um MRF para acomodar várias métricas (e.g., número de observações já efetuadas, tempo restante até sair da cena, tempo de transição entre sujeitos) úteis na decisão da ordem pela qual os sujeitos irão ser observados pela câmara PTZ. Como principal vantagem, esta abordagem apresenta a capacidade de planear rotas quase ótimas em tempo real, o que é particularmente importante em cenários com vários indivíduos.

O quarto capítulo apresenta o conjunto de dados QUIS-CAMPI, que adquirido recorrendo ao sistema QUIS-CAMPI. Após ter sido observado que os conjuntos de dados biométricos existentes na literatura não continham os fatores de degradação típicos dos ambientes de vídeo-vigilância, procedeu-se à construção do conjunto QUIS-CAMPI, onde as imagens de prova foram adquiridas de forma totalmente automatizada num cenário de vídeo-vigilância, de forma sub-reptícia, e a distâncias entre 5 e 40 metros. Para além disso, este conjunto disponibiliza múltiplos traços biométricos tanto no conjunto de registo (imagens do corpo inteiro, vídeos do tipo de passo, modelos 3D da face) como no conjunto de prova (e.g., vídeos do tipo de passo e imagens da face). Todas estas características tornam o conjunto QUIS-CAMPI bastante importante para avaliar o desempenho dos métodos de reconhecimento biométrico estado da arte e promover o desenvolvimento de métodos capazes de operar em ambientes não controlados. Neste capítulo é também apresentado o desempenho obtido por métodos de reconhecimento estado da arte no

conjunto de dados proposto, e de forma sumária concluiu-se que os métodos atuais ainda não apresentam desempenho satisfatório neste tipo de dados.

No quinto capítulo são apresentadas três contribuições para avançar o estado da arte no reconhecimento de indivíduos em ambientes de vídeo-vigilância. A primeira contribuição diz respeito à competição ICB-RW, a primeira competição internacional para avaliar o desempenho de métodos de reconhecimento biométricos em dados adquiridos sem qualquer restrição. Esta competição foi organizada com um subconjunto de imagens faciais do conjunto de dados QUIS-CAMPI, e teve a participação de nove grupos de investigação, tendo cada grupo submetido um método para avaliação. Os resultados obtidos foram úteis para perceber quais as estratégias mais promissoras e quais os fatores de degradação que mais afetam o desempenho dos métodos. A segunda contribuição propõe um método de deteção de características corruptas e tem como objetivo melhorar o desempenho de um sistema biométrico através da exclusão das características corruptas durante a fase de *matching*. A ideia principal por detrás deste método é aproveitar a redundância da assinatura biométrica, ou seja, as correlações entre as características, para determinar na fase de teste a probabilidade de cada característica estar corrupta. A avaliação em conjuntos de imagens da íris e da face evidencia as vantagens deste método. A terceira contribuição diz respeito a um método de reconhecimento facial baseado em caricaturas. O fato de os humanos reconhecerem mais facilmente indivíduos através de caricaturas foi a base para desenvolver este método, sendo que o processo automático para a geração da caricatura 2D foi baseado na estratégia usada pelos caricaturistas. Desta maneira, foi possível obter representações onde a semelhança entre imagens de diferentes sujeitos é minimizada e a semelhança entre imagens do mesmo sujeito é maximizada.

Por último, os principais resultados deste trabalho de investigação são resumidos no capitulo seis. Para além disso, são também apontadas as principais contribuições desta tese para o desenvolvimento de um sistema de vídeo-vigilância capaz de identificar indivíduos de forma automática, assim como tópicos que carecem de trabalho adicional para atingir este objetivo.

Abstract

The concerns about individuals security have justified the increasing number of surveillance cameras deployed both in private and public spaces. However, contrary to popular belief, these devices are in most cases used solely for recording, instead of feeding intelligent analysis processes capable of extracting information about the observed individuals. Thus, even though video surveillance has already proved to be essential for solving multiple crimes, obtaining relevant details about the subjects that took part in a crime depends on the manual inspection of recordings. As such, the current goal of the research community is the development of automated surveillance systems capable of monitoring and identifying subjects in surveillance scenarios. Accordingly, the main goal of this thesis is to improve the performance of biometric recognition algorithms in data acquired from surveillance scenarios. In particular, we aim at designing a visual surveillance system capable of acquiring biometric data at a distance (e.g., face, iris or gait) without requiring human intervention in the process, as well as devising biometric recognition methods robust to the degradation factors resulting from the unconstrained acquisition process.

Regarding the first goal, the analysis of the data acquired by typical surveillance systems shows that large acquisition distances significantly decrease the resolution of biometric samples, and thus their discriminability is not sufficient for recognition purposes. In the literature, diverse works point out Pan Tilt Zoom (PTZ) cameras as the most practical way for acquiring high-resolution imagery at a distance, particularly when using a master-slave configuration. In the master-slave configuration, the video acquired by a typical surveillance camera is analyzed for obtaining regions of interest (e.g., car, person) and these regions are subsequently imaged at high-resolution by the PTZ camera. Several methods have already shown that this configuration can be used for acquiring biometric data at a distance. Nevertheless, these methods failed at providing effective solutions to the typical challenges of this strategy, restraining its use in surveillance scenarios. Accordingly, this thesis proposes two methods to support the development of a biometric data acquisition system based on the cooperation of a PTZ camera with a typical surveillance camera. The first proposal is a camera calibration method capable of accurately mapping the coordinates of the master camera to the pan/tilt angles of the PTZ camera. The second proposal is a camera scheduling method for determining - in real-time - the sequence of acquisitions that maximizes the number of different targets obtained, while minimizing the cumulative transition time. In order to achieve the first goal of this thesis, both methods were combined with state-of-the-art approaches of the human monitoring field to develop a fully automated surveillance capable of acquiring biometric data at a distance and without human cooperation, designated as QUIS-CAMPI system.

The QUIS-CAMPI system is the basis for pursuing the second goal of this thesis. The analysis of the performance of the state-of-the-art biometric recognition approaches shows that these approaches attain almost ideal recognition rates in unconstrained data. However, this performance is incongruous with the recognition rates observed in surveillance scenarios. Taking into account the drawbacks of current biometric datasets, this thesis introduces a novel dataset comprising biometric samples (face images and gait videos) acquired by the QUIS-CAMPI system at a distance ranging from 5 to 40 meters and without human intervention in the acquisition process. This set allows to objectively assess the performance of state-of-the-art biometric recognition methods in data that truly encompass the covariates of surveillance scenarios. As such, this set was exploited for promoting the first international challenge on biometric recognition in the

wild. This thesis describes the evaluation protocols adopted, along with the results obtained by the nine methods specially designed for this competition. In addition, the data acquired by the QUIS-CAMPI system were crucial for accomplishing the second goal of this thesis, i.e., the development of methods robust to the covariates of surveillance scenarios. The first proposal regards a method for detecting corrupted features in biometric signatures inferred by a redundancy analysis algorithm. The second proposal is a caricature-based face recognition approach capable of enhancing the recognition performance by automatically generating a caricature from a 2D photo. The experimental evaluation of these methods shows that both approaches contribute to improve the recognition performance in unconstrained data.

Keywords

Automated Surveillance Systems, PTZ Cameras, Master-slave Configuration, Camera Calibration, Camera Scheduling, Biometric Recognition, Non-cooperative Biometric Recognition, Biometric Datasets, Face Recognition, Face Recognition at a Distance, Caricature-based Face Recognition, Human Monitoring.

Contents

Contents

List of Figures

List of Tables

Acronyms

AFLW Annotated Facial Landmarks in the Wild

AOV Angle of View

AUC Area Under Curve

BGS Background Subtraction

CCTV Closed-circuit Television

CMC Cumulative Matching Characteristic

CNN Convolutional Neural Network

CB Calibration Patterns

ECC Error-correcting Codes

EDF Earliest Deadline First

EER Equal Error Rate

EKF Extended Kalman Filter

FAR False Acceptance Rate

FCFS First-come, First-served

FOV Field of View

FRR False Rejection Rate

FRVT Face Recognition Vendor Test

FVF Fisher Vector Faces

GPU Graphics Processing Unit

HOG Histogram of Oriented Gradients

ICB-RW International Challenge on Biometric Recognition in the Wild

KLT Kanade–Lucas–Tomasi

LBP Local Binary Patterns

LFW Labeled Faces in the Wild

MBGC Multiple Biometrics Grand Challenge

MRF Markov Random Field

NFOV Narrow Field of View

ROC Receiver Operating Curve

SVM Support Vector Machine

SOM Self-Organizing Maps

PaSC Point and Shoot Face Recognition Challenge

PTZ Pan Tilt Zoom

VLS Verilook Surveillance System

VP Vanishing Points

WFOV Wide Field of View

WCS World Coordinate System

Chapter 1

Introduction

This thesis regards the problem of recognizing individuals in surveillance scenarios without subjects cooperation, and makes two major contributions: 1) an innovative surveillance system capable of acquiring biometric samples at a distance and on the move; 2) biometric recognition methods for pushing forward the performance on data acquired in surveillance environments. The focus and scope of the thesis are further described in this chapter, together with the problem definition and objectives, the main contributions, and the document organization.

1.1 Problem Definition and Research Objectives

Surveillance is a subject undergoing intense study fostered by the increasing concerns about national security. This interest is visible in the evolution of the number of video surveillance cameras deployed worldwide (e.g., more than 4 million Closed-circuit Television (CCTV) cameras in the UK [1]). However, the availability of surveillance recordings contrasts with the limited recognition accuracy of the state-of-the-art algorithms on these data.

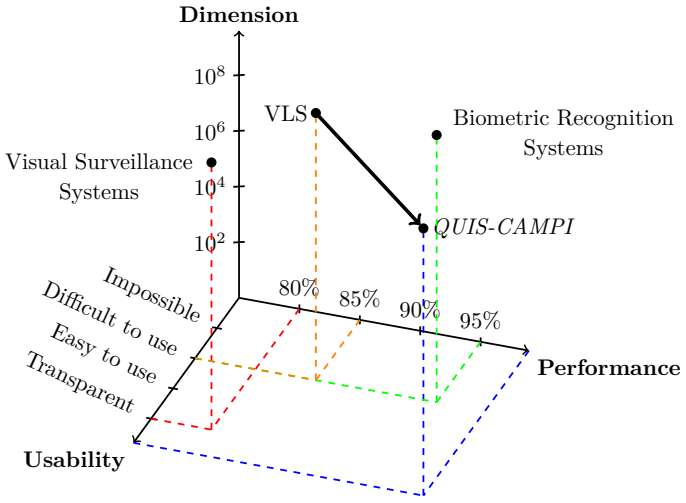


Figure 1.1: Schematic representation of the three challenges of designing a recognition system.

As depicted in figure 1.1, the complexity of designing a fully automated recognition system can be described as function of three variables. State-of-the-art biometric recognition systems can attain ideal performance at the expense of restrictive conditions during data acquisition, but their accuracy decreases dramatically when working in unconstrained scenarios. On the other hand, automated surveillance systems perform poorly when identifying individuals in totally unconstrained scenarios and require a large set of constraints to work properly in a specific scenario. To address this problem, different approaches have been proposed [2–4]. The Verilook Surveillance System (VLS) [4] is a prominent effort aiming at combining the fields of biometric recognition and visual surveillance. This system performs non-cooperative face identification from live video streams with satisfactory accuracy, but its usability is rather limited

(e.g., works only in indoor scenarios, depends on high-resolution surveillance cameras, and requires a time-consuming enrollment process). According to A. K. Jain, overcoming the trade-off between performance and usability is the ultimate barrier to the development of a biometric recognition system capable of working in surveillance scenarios, which is still regarded as the grand challenge [5].

Even though the identification of humans in surveillance scenarios is still confined to science fiction, the research community has already made significant progress towards the development of a biometric recognition system capable of working in surveillance scenarios. Currently, it is commonly accepted that the typical architecture of surveillance systems (e.g., CCTV systems) is not adequate for acquiring biometric samples with sufficient resolution for recognition purposes. To address this problem, several authors argue that Pan Tilt Zoom (PTZ) cameras are the most practical and efficient solution to acquire biometric data at a distance, particularly if configured in a master-slave architecture (refer to section 2.2 for a detailed justification).

Figure 1.2 illustrates the scope of this thesis with respect to the general processing chain of an automated master-slave surveillance system intended for biometric recognition. Even though the development of such a system depends on three distinct research areas, the scope of this thesis is restricted to the areas of surveillance systems and biometric recognition. Accordingly, this thesis aims at extending the frontiers of biometric recognition to surveillance scenarios by 1) introducing novel strategies for acquiring biometric data in these scenarios; and 2) proposing novel biometric recognition algorithms robust to the typical degradation factors of unconstrained environments. In order to achieve the primary goal of this thesis, we defined multiple objectives along the processing chain illustrated in figure 1.2:

- Evaluation of the performance of human detection and multi-target tracking algorithms in surveillance scenarios. The acquisition of biometric data in surveillance scenarios depends on the performance of human detection and tracking, which degrades significantly in unconstrained scenarios. Our goal is to assess their performance in a real surveillance scenario, and perceive the impact on the quality of the biometric data acquired.

- Proposal of a new camera calibration algorithm for master-slave surveillance systems. The existing master-slave systems either rely on rough approximations or additional constraints to estimate the mapping between image coordinates and pan-tilt parameters, and, as a consequence, the workability of these systems in outdoor environments is restrained. For this reason, we aim at creating a calibration algorithm that does not depend on stringent configurations to accurately estimate the mapping between cameras.

- Proposal of a scheduling policy to control the PTZ camera. Real-word surveillance systems should be capable of monitoring multiple individuals at the same time. As such, we aim at developing a camera scheduling approach capable of determining - in real-time - the sequence of acquisitions that maximizes the number of different targets obtained, while minimizing the cumulative transition time.

- Development of a prototype of an automated surveillance system capable of acquiring biometric data at a distance, and subsequent deployment in a real surveillance scenario. This system is hereinafter designated as the QUIS-CAMPI system.

- Creation of a dataset of biometric samples acquired in a real-world surveillance scenario. The prototype of the QUIS-CAMPI system will be used for acquiring biometric data from subjects in an unconstrained and covert manner. Also, we these data will be made publicly

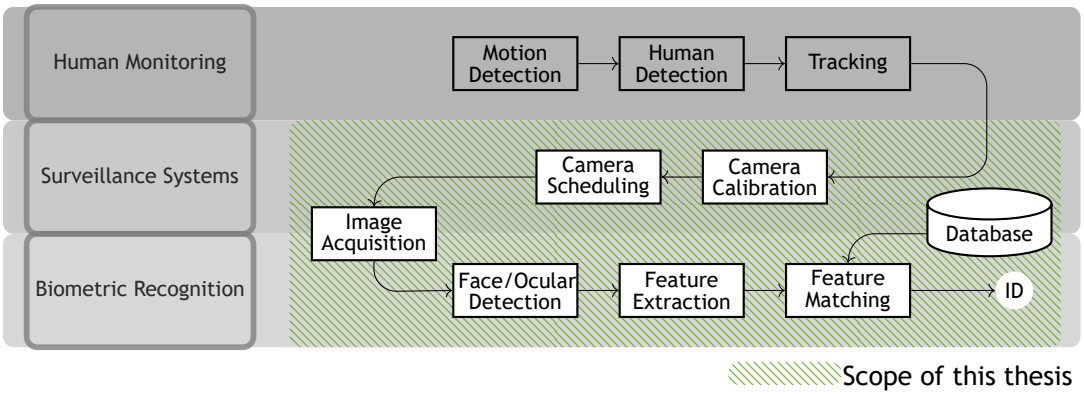


Figure 1.2: General processing chain of an automated surveillance system intended for biometric recognition. The different modules necessary for the development of an automated surveillance system capable of recognizing individuals belong to three different research areas: 1) human monitoring; 2) surveillance systems; and 3) biometric recognition. Nevertheless, the major contributions of this thesis fall in the fields of biometrics and surveillance systems.

available to the research community for pushing forward the performance of biometric recognition in the wild.

- Evaluation of state-of-the-art biometric recognition algorithms in data acquired from a fully automated surveillance system. The biometric samples acquired with the QUIS-CAMPI system are essential for determining the actual performance on data that truly encompass all the singularities of surveillance environments. We aim at using the biometric samples acquired for promoting an international competition on biometric recognition in the wild.
- Development of novel biometric recognition algorithms for improving the recognition performance in data acquired by fully automated surveillance systems. By proposing novel methods that surpass the performance of state-of-the-art biometric recognition methods on the data acquired by the QUIS-CAMPI system, we aim at contributing to the development of a fully automated biometric recognition surveillance system.

1.2 Main Contributions

This section describes the main scientific contributions of this thesis. Figure 1.3 provides a graphical description of the context of our publications in the general processing chain of an automated surveillance system intended for biometric recognition. The following paragraphs briefly describe the main proposals of this thesis to advance the state of the art in biometric recognition in the wild.

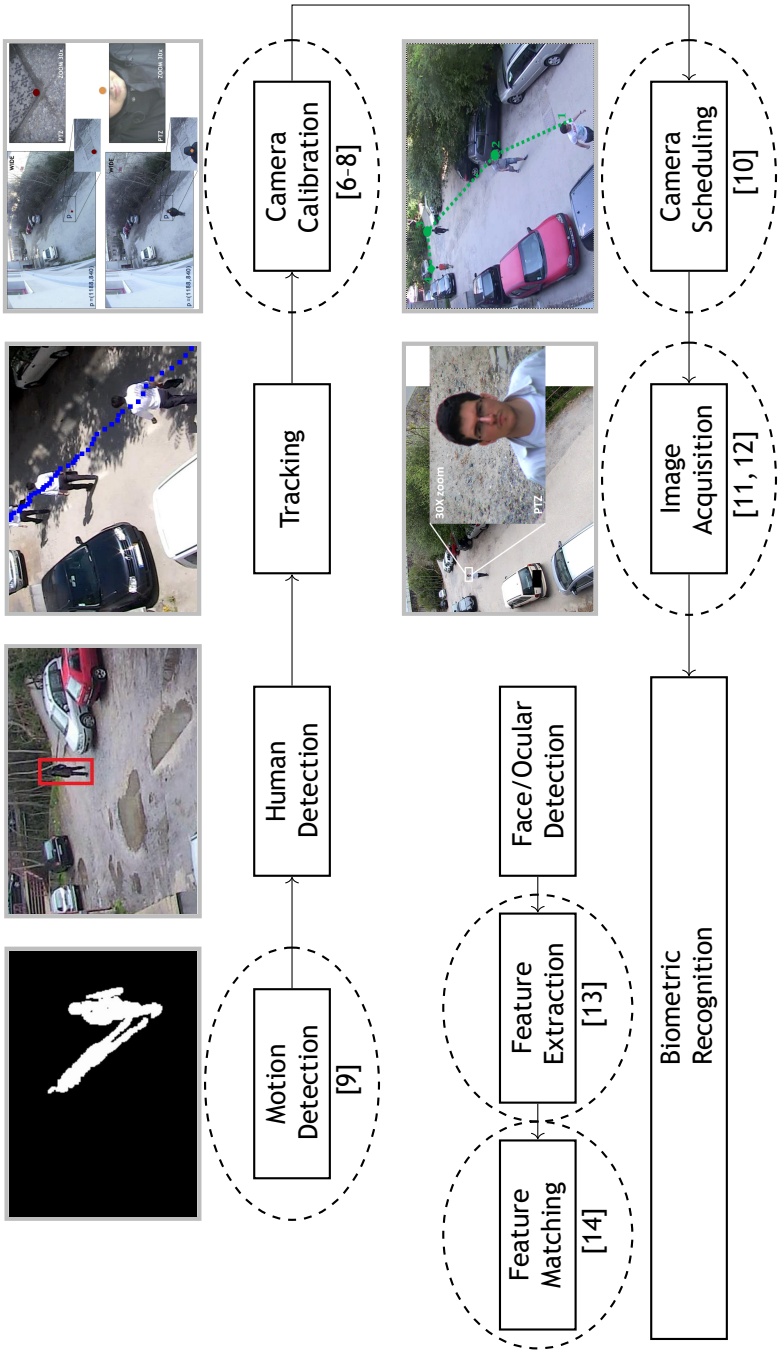


Figure 1.3: Overview of the main contributions of this thesis.

Introduction

The first contribution is a comprehensive review of the state of the art in each one of the main stages of automated surveillance systems, along with a review of the state-of-the-art biometric recognition approaches capable of working in slightly unconstrained scenarios. This work resulted in a survey published in the Springer Artificial Intelligence Review Journal [15].

The second contribution regards the evaluation of the state-of-the-art background subtraction algorithms in surveillance scenarios, resulting in a paper published in the proceedings of the IEEE International Conference on Signal and Image Processing Applications [9].

The third contribution is a novel calibration algorithm for master-slave systems, suitable for acquiring biometric data at a distance from non-cooperative subjects. Our solution provides an efficient way to estimate accurately the pan-tilt angles of the PTZ camera without relying on stringent configuration between the cameras. A preliminary version of this approach was published in the proceedings of the Iberian Conference on Pattern Recognition and Image Analysis [6], whereas the complete description of this proposal was published in the IEEE Conference on Biometrics: Theory, Applications and Systems [7].

The fourth contribution is an extension of the proposed calibration method to work with fish-eye lenses. This work was published in the Mathematical Problems in Engineering [8].

The fifth contribution regards a comprehensive review of the frameworks and protocols designed for the acquisition of unconstrained biometric data, along with a description of the state-of-the-art surveillance systems using these frameworks. This review was published as a chapter of the book Human Recognition in Unconstrained Environments [16].

The sixth contribution introduces a camera scheduling algorithm for determining the order whereby the subjects will be imaged by the PTZ camera. The proposed method is capable of determining - in real-time - the sequence of acquisitions that maximizes the number of different targets obtained, while minimizing the cumulative transition time. This method was published in the proceedings of the IEEE Conference on Advanced Visual and Signal based Surveillance [10].

The seventh contribution regards the QUIS-CAMPI surveillance system, i.e. a fully automated surveillance system for acquiring biometric data at a distance from non-cooperative subjects. This proposal was published in the proceedings of the International Workshop on Recent Advances in Digital Security: Biometrics and Forensics, which was part of the IEEE Conference on Image Analysis and Processing [17].

The eighth contribution is the QUIS-CAMPI dataset, comprising biometric samples automatically acquired by the QUIS-CAMPI system in a fully non-cooperative and covert manner. This is the first biometric set comprising data that truly represents the covariates of surveillance environments. This proposal is in the third round of the reviewing process of the IET Biometrics.

The ninth contribution is the International Challenge on Biometric Recognition in the Wild (ICB-RW) competition, the first biometric challenge carried out in data that realistically result from surveillance scenarios. The results of this competition were published in the proceedings of the International Conference on Biometrics [11].

The tenth contribution is a method for detecting degraded features in biometric signatures by exploiting feature correlation. The proposed approach was published in the proceedings of the International Workshop on Biometrics in the Wild, which was part of the IEEE Conference on Face and Gesture [14].

The eleventh contribution regards the first fully automated caricature-based face recognition system capable of working with data acquired in the wild. This work proposes a 3D-based caricature generation method for enhancing the performance of face recognition, and it was submitted for revision to the IEEE Transactions on Information Forensics and Security.

1.3 Thesis Organization

The remainder of this thesis is organized as follows: chapter 2 reviews the state-of-the-art approaches for detecting and tracking humans in surveillance scenarios. Also, a description of the architectures of automated surveillance systems is provided, along with a comparative analysis of their effectiveness for acquiring biometric data from non-cooperative subjects. Additionally, this chapter gives an overview of the concepts related to the development of biometric systems, and details the most prominent attempts for recognizing individuals in unconstrained scenarios. Chapter 3 describes the proposed QUIS-CAMPI system, with particular attention given to the proposed camera calibration and camera scheduling algorithms, which contribute significantly to the successful acquisition of biometric data from subjects at a distance and on the move. Chapter 4 introduces the QUIS-CAMPI data feed, a collection of biometric samples acquired by the QUIS-CAMPI system, along with a set of biometric data acquired under controlled conditions in the enrollment phase. Chapter 5 describes the contributions of this thesis to push forward the performance of biometric recognition in the unconstrained scenarios: 1) the ICB-RW competition, which was the first biometric challenge carried out in data that realistically result from surveillance scenarios; 2) a method capable of detecting degraded features in biometric signatures by exploiting feature correlation; and 3) the first fully automated caricature-based face recognition system capable of working with data acquired in the wild. Finally, chapter 6 presents the conclusions, summarizes our achievements and points possible directions for further work.

Chapter 2

State of the Art

In this chapter, we review the basic concepts related to the development of an automated surveillance system for recognizing individuals using their biometric traits. Section 2.1 reviews the state-of-the-art methods for detecting and tracking humans in surveillance videos, and discusses the most adequate strategies for monitoring humans in surveillance scenarios. The most relevant frameworks and protocols for acquiring biometric data in unconstrained scenarios are reviewed in section 2.2. Moreover, a comparative analysis between the use of existing architectures in surveillance scenarios is provided. Section 2.3 introduces the basic concepts related to biometrics, namely its main modes of functioning and the common metrics used for evaluating biometric recognition systems. Finally, section 2.4 summarizes the most relevant conclusions of this chapter.

2.1 Human Monitoring in Surveillance Scenarios

Automated surveillance systems usually share three main stages: pre-detection, detection, and tracking. This section provides a comprehensive review of the state of the art in these three phases with particular focus on surveillance scenarios.

2.1.1 Motion Detection

Motion information is commonly used to prune the scene in a pre-detection phase, selecting regions of interest for the detection phase. Usually, the pre-detection step relies on background subtraction to highlight the regions of interest, but some alternatives are also possible, such as optical flow.

Background subtraction methods aim at dividing the scene into foreground and background regions using the typical appearance values of static regions of the scene. Even though the detection of specific objects is not attained, the scene is pruned and the computational burden of subsequent phases is reduced. For this reason, background subtraction has been used as a pre-detection phase in different approaches such as human detection [18] and tracking [19].

Although most Background Subtraction (BGS) methods rely on a background model, they differ with respect to the strategy used to construct this model. Statistical-based approaches analyze the typical pixel intensities to distinguish between foreground and background regions. The most simple approaches infer a background model from the median of the last N frames, and obtain foreground regions by thresholding the difference between the observed image and the background model [20–22]. A more robust strategy is the use of Gaussian-based approaches where the typical values of the background are encoded by a single Gaussian [23] or a mixture of Gaussians [24,25]. Instead of using a threshold, a confidence interval is defined to perform foreground detection, ensuring the correct classification of both high and low variance background pixels. Besides, the use of a mixture of Gaussians permits the modeling of multiple sources of background.

Clustering-based approaches estimate the background by grouping pixels in K different clusters, corresponding to multiple sources of background. The codebook model [26] uses a set of codewords to represent each cluster, while color and brightness information is used to define the distance function. Different features are used to describe clusters, such as luminance [27, 28] and chrominance [29]. Recently, unsupervised neural networks models have been explored to provide BGS methods with further robustness in surveillance scenarios. Maddalena and Petrosino [30] exploited Self-Organizing Maps (SOM) to model the background by storing the RGB values of each pixel in the neuron's weights. Competitive neural networks [31] use a similar idea by adjusting the weights of output layer neurons, however, contrary to SOM, learning reinforcement is only applied to the winner neuron.

Contrary to BGS approaches, which compare the scene with a background model to detect moving regions, optical flow approaches rely on displacements between consecutive frames. By assuming small movement and brightness constancy, the displacement of each pixel can be computed [32–34].

2.1.2 Human Detection

When compared to the pre-detection phase, detection algorithms are more specific because they aim at providing the location of a specific object in the scene. In general, detection algorithms do not require a pre-detection phase, yet the majority of them rely on this phase to alleviate the computational burden and ease the detection phase. Moreover, in some cases, human detection algorithms do not use pre-detection only as an attentive filter. Instead, they rely on the shape information that is yielded from BGS methods because it has been found that it greatly improves performance when combined with appearance cues [35].

To achieve human detection, two different strategies are commonly employed: 1) holistic detection, where a whole-body search is conducted; and 2) part-based detection, where the search is oriented to locate a single body part or a combination of parts. Currently, the second approach is attracting more attention, especially in surveillance scenarios, where the head and shoulder regions are commonly used as discriminative features.

2.1.2.1 Holistic Approaches

Most holistic approaches train a discriminative classifier to exhaustively search for a specific object. Viola and Jones adapted their general object detector [36] to locate humans in surveillance scenarios using motion patterns [37]. In a similar fashion, Dalal and Triggs [38] introduced the Histogram of Oriented Gradients (HOG) features to perform human detection by training a discriminative classifier, such as a Support Vector Machine (SVM). HOG features have been explored in several approaches for the purpose of increasing robustness in surveillance scenarios [39, 40]. Local Binary Patterns (LBP) features [41] have also been widely used for human detection purposes, especially in surveillance scenarios [42, 43]. Yao and Odobez [35] improved the performance of a cascade of detectors by including shape information that was acquired in the pre-detection phase. In the work of Gurwicz et al. [44], moving objects were obtained with a background estimation method. Several features were extracted, such as image moments and horizontal and vertical projections, but only the features that were capable of the most discrimination were retained, based on the entropy gain. The selected features were provided to a SVM to distinguish between human regions and clutter in surveillance scenarios.

2.1.2.2 Part-based Approaches

Part-based approaches represent a solution to deal with the typical covariates of unconstrained scenarios (e.g., occlusion and changes in perceived 2D shape). The detection strategy is similar to the one used in holistic approaches, but in this case different body parts are detected by distinct classifiers. The final result is achieved by globally reasoning the score of all classifiers. Regarding the type of features used for classifier training, most methods rely on gradient features [18, 45–50], whereas few approaches have exploited color information [51, 52].

Mikolajczyk et al. [45] used a probabilistic assembly of parts to attain human detection. A coarse-to-fine cascade approach was used for parts detection, and a parts assembly strategy pruned incorrect detections by imposing geometric constraints. Lin et al. [46] focused on head detection to estimate the number of people in a large crowd. Subburaman et al. [47] also used head features for crowd counting, attaining state-of-the-art results in the PETS2012 dataset. Zhao and Nevatia [18] addressed human detection by analyzing the silhouette boundaries that were obtained from the foreground mask. Head detection was attained by checking local vertical peaks on the foreground contour. Detections were filtered by cross-checking silhouette information with human anthropometric data. Wu and Nevatia [49] used four different body parts (full-body, head-shoulder, torso, and legs) to detect humans in non-cooperative scenarios. Parts detectors were learned by boosting a number of weak classifiers based on edgelet features (short segments of edge pixels). The detectors' responses were combined to provide robustness to occlusions. Later, this work was extended not only to improve detection performance but also to achieve human segmentation using hierarchical body part detectors [50].

2.1.3 Human Tracking

Given an initial estimation of the object location, visual tracking approaches are expected to determine occurrences of the same object in consecutive frames. In general, tracking approaches can be distinguished by the tracking strategy adopted and the type of information used to model target objects, usually denoted as target representation.

2.1.3.1 Type of Features / Target Representation

Tracking algorithms should be provided with an object description that is usually obtained from distinctive features such as motion, shape or appearance. The model comprising all the information associated with objects of interest is denoted as the target representation.

Motion. Motion-based tracking exploits object dynamics. In the particular case of human tracking, different cues are combined to model the target (e.g., typical human velocity, articulation constraints and periodic motion). Motion models are usually related to Bayesian tracking approaches, where temporal dynamics are used to update the target state over time [48, 53, 54]. However, these models can also be independently used to exploit appearance or shape information [55, 56]. Motion information is also widely used to reduce the search space. Tracking based on optical flow estimation, namely the Kanade–Lucas–Tomasi (KLT) tracker [57], combines the assumption of small movement between frames with brightness constancy to follow a set of keypoints. Tracking-by-detection approaches have also used this strategy [58, 59]. In [58] the predicted position of the target is constrained to a predefined radius [59, 60]. In [59] the optical flow is exploited to provide further robustness to discriminative classifiers. More complex

methods have analyzed the motion relations between different regions of the scene to attain additional robustness to occlusions [60].

Appearance. Albeit different tracking techniques can use any kind of appearance descriptor, the literature evidences a relation between the technique and the type of descriptor. Kernel tracking methods use a histogram of color intensities to represent the target [61]. Different color spaces (e.g., HSV and XYZ) were also used [62–64]. Mckenna et al. [64] exploited Gaussian mixture models to parametrize the objects' color distributions in hue-saturation space. An adaptive learning algorithm was used to update these color models and ensure robustness under varying illumination. Since in different scenarios the performance is maximized by different color spaces, Stern and Efros [63] developed a method to automatically switch the color space with respect to the ambient conditions. Tracking-by-detection approaches encode appearance information to train discriminative classifiers, using multiple descriptors such as Haar wavelets [58, 59, 65], LBP [66, 67] or HOG [68]. Regarding Bayesian tracking, several approaches have exploited a large number of appearance descriptors [53, 69, 70], but, recently, a large number of approaches [71–74] adopted the use of sparse representation.

Shape. Compared to appearance-based tracking, shape modeling is invariant to illumination and appearance changes *per se*, but in turn, this cue is highly sensitive to occlusion and pose. Although some tracking methods consider shape as a key feature [75], it is often regarded as a pruning feature or as a way to take advantage of other cues. This holds particularly in surveillance scenarios, where the limited number of pixels representing the object restrains the use of complex shape models. Notwithstanding, the fusion of simple shape models with other features, such as appearance and motion, proved successful in surveillance scenarios. KaewTrakulPong et al. [19] combined shape cues with position, appearance and motion information to determine the temporal associations between a set of blobs, corresponding to human targets in an outdoor surveillance scenario. Wu and Yu [76] used a Markov field to learn a prior shape model for human edges. Pedestrian tracking was considered as a posterior density estimation according to the shape model learned, where target state is propagated using a simple motion model. Albeit edges are the most frequent shape feature used, other alternatives have been currently exploited to track objects in dynamic scenarios (e.g., the shape context descriptor [77, 78]).

2.1.3.2 Tracking Strategy

Classical approaches attempted to track an object by searching for a specific pattern in the neighborhood of the previous location (Kernel / Model Tracking) or by evolving the state of the target according to a motion and appearance model (Bayesian Tracking). Recently, a new strategy - denoted as tracking-by-detection - has gained popularity as the demand for arbitrary object tracking in unconstrained scenarios increased. The recent developments of each technique are reviewed with particular attention given to the robustness in unconstrained environments.

Bayesian Tracking. In a Bayesian framework, tracking is regarded as the estimation of the target state x_k given all the measurements $z_{1:k}$, which is equivalent to maximize the probability $p(x_k|z_{1:k})$. Bayesian filters solve this recursively using two steps: 1) *prediction* step infers the next state distribution, $p(x_k|z_{1:k-1})$, with respect to a motion model describing the target state over time; 2) *update* step uses the current observation z_k to update $p(x_k|z_{1:k-1})$, yielding

$p(x_k|z_k)$. This process permits the estimation of the latent or unobservable variable x_k through noisy measurements z_k. Regarding the type of noise, different Bayesian filters can be used. When the system is affected by Gaussian noise and the motion model is linear, the Kalman filter [79] can be employed. Despite being based on restrictive assumptions, some approaches used it in surveillance scenarios [48, 54, 80]. Zhao and Nevatia [48] used the Kalman filter with a constant velocity model to estimate the state of humans. In [54], the combined observations of multiple cameras were provided to the Kalman filter to obtain a more accurate target state. The Extended Kalman Filter (EKF) [81] was introduced to handle non-linear systems. Mittal and Davis [82] used this technique in a multi-view approach so that severe occlusion could be handled. Oliver et al. [83] combined the EKF predictions with appearance information to track persons in outdoor scenes for action-recognition purposes. In general, particle filters or sequential Monte Carlo methods are preferred in Bayesian tracking [84–88], since they can handle any kind of noise and do not require the motion model to be linear. Okuma et al. [70] used appearance cues by combining the particle filter with AdaBoost. Hu et al. [89] combined appearance, shape and motion information to track occluded people also using the particle filter. Sparse representation was also exploited by some state-of-the-art tracking methods [71–74]. Each candidate location was represented as a combination of the training templates so that the smallest projection error candidate was chosen. Mei and Ling [73] used this strategy in the L1 tracker. The target motion in consecutive frames was modeled as an affine transformation and was estimated in a particle filter framework. The importance of each transformation (i.e., the particle weights) was a function of the sparse reconstruction error. The MTT tracker [72] was later introduced as a generalization of L1 since it accounted for the dependence between transformations.

Kernel Filter. Kernel-based tracking gathers appearance information over an image patch by constructing a weighted feature histogram. The first representative kernel-based method was proposed by Comaniciu et al. [61], where the Mean Shift [90] technique was adapted to track objects based on their appearance. Target location was achieved by maximizing a similarity measure and the mean shift procedure guided the search for conditional probability maximum, avoiding a brute force search. Although this strategy provides invariance to some pose changes, the loss of spatial information is the primary drawback of kernel-based approaches. To address this issue, Kang et al. [91] divided the object according to its polar representation and modeled the typical RGB color of each part with a Gaussian distribution. Zhao and Tao [92] included spatial information in the appearance model using the correlogram technique [93], allowing to infer not only the objects' trajectory but also their orientation. Recently, distribution fields [94, 95] have also been introduced to preserve the spatial information by constructing a histogram at each pixel. Robustness to dynamic environments has also been recently proposed [96]. Chu et al. [96] used multiple kernels to improve tracking under occlusion. Zhang et al. [69] devised a head tracker using a kernel-Bayesian framework, where appearance and shape information were combined. A Gaussian mixture model was used to model the appearance and the Chamfer distance [97] was used for shape comparison. Liu et al. [98] approached human tracking using eigenshape. The arbitrarily shaped kernel allowed the tracker to adapt to the object shape avoiding background noise.

Model / Shape Tracking. Maximizing the similarity between the shape model and the contour-map of the image is the rationale of shape tracking. In general, contour information is provided by an edge-map representation and shape similarity is evaluated either with the Chamfer match-

ing [97] or with the Haussdorf distance [99]. Both shape matching techniques are computationally expensive and are not suitable to work in real-time systems. To efficiently compute the Chamfer matching or the Haussdorf distance, Gravila et al. [100, 101] proposed a solution based on the distance transform. In a later work [102], hierarchical matching was proposed to further increase the efficiency of shape matching. A set of training shapes were clustered so that a tree of shape models could be constructed with the representative model of each cluster in the first layer. Besides, a Markov transition matrix was used to encode the probabilities between shape transitions, so that, during the tracking, the most likely poses were prioritized. These approaches were combined in [103] to develop a complete pedestrian detection and tracking system, where motion and appearance cues were also exploited. The tracking module used pose clusters and a tree of pose models to efficiently search for the model that best fitted the data. In dynamic environments, shape tracking is particularly sensitive to occlusion. For this reason, Saber et al. [104] devised a matching strategy robust to partial occlusion, the partial shape matching. Husain et al. [105] used this technique to track objects in surveillance scenarios. However, even these improvements fail to produce a robust solution in surveillance scenarios, mainly due to the reduced size of objects of interest.

Tracking-by-Detection. The use of detectors in tracking has gained wide notoriety, mainly driven by the possibility of tracking arbitrary objects. Tracking-by-detection algorithms estimate the target position by searching the location that maximizes a function $F(\vec{x}) \in [-1, 1]$, where F is usually determined by a classifier and $\vec{x}$ is the feature vector of the target state. Contrary to other tracking methods, no *a priori* target representation is required, postponing the learning of this representation to the online training of the classifier. Online training allows the classifier to adapt to any kind of object and also to appearance variations. Currently, the main research line in tracking-by-detection is focused both in improving the classifier learning scheme and in exploiting multiple cues. Regarding the learning scheme, the use of online boosting classifiers was a common strategy in initial approaches [106, 107]. At each frame, the target location was sampled for positive examples while its neighborhood was sampled for negative examples. However, this strategy is highly sensitive to appearance changes, since small displacements from the ground truth location may introduce incorrect positive examples in the learning process. Babenko et al. [58] proposed a method to overcome this problem, where examples were presented as bags containing a set of instances. Bags containing at least one positive instance, corresponding to the instances sampled at the target location, were labeled as positive, otherwise they were labeled as negative. Although this strategy required the classifier to distinguish between positive and negative instances in some bags, previous results had shown that, in fact, it was more flexible and outperformed the traditional learning strategies [108]. In a similar fashion, the Struck tracker [65] used a structured output SVM [109] to perform learning. The TLD [66] and the PROST [59] methods found a different solution by combining an optic flow tracker with an online learned random forest. Negative examples were only sampled from unlikely locations of object presence based on motion constraints. Besides, new examples required an appearance confirmation to be provided to the classifier. ConTra [67] improved this strategy by taking in account distracters, i.e., objects sharing the same appearance as the target.

2.1.3.3 Multi-target Tracking

Despite multiple instances of each algorithm could be used to address multiple target tracking, this strategy would require an additional data association module. The joint probabilistic data association filter [110] and multiple hypothesis tracking [111] are two classical approaches for this purpose, but the exponential growth of computational complexity restrains their use when the number of targets is high. Greedy strategies have been used as an alternative, where correspondences are regarded as an assignment problem based on spatial distance [49, 112] or appearance similarity [113].

Offline or batch techniques methods are an alternative solution for multiple target tracking, which, in contrast to online methods, use the complete set of detections before performing trajectory estimation. This problem is usually regarded as an optimization problem, where a function describes the cost of each solution [114–116]. Linear programming was employed by several works [116–119] to solve this problem, where the possible target locations were discretized and modeled as a graph. A continuous formulation of the problem was later introduced by Andriyenko and Schindler [120–122]. The main drawback of these approaches is the high latency required to analyze a video, which is incompatible with real-time surveillance requirements. To address this issue, Benfold and Reid suggested the use of a small subset of frames [123]. In [123] the most recent six seconds of video were analyzed to track multiple pedestrians by combining information from a HOG-based detector and a KLT tracker.

2.2 Biometric Data Acquisition Frameworks

This section reviews the most relevant frameworks and protocols for acquiring biometric data in unconstrained scenarios. Section 2.2.1 provides a comparative analysis between the most common data acquisition architectures and reviews the most relevant works proposed in each category. In section 2.2.2, we discuss the challenges of PTZ-based approaches. Section 2.2.3 describes the state-of-the-art surveillance systems devised for acquiring biometric data in unconstrained scenarios.

2.2.1 Architecture of the Acquisition Systems

As illustrated in figure 2.1, surveillance systems are divided into two major categories: 1) systems using fixed-angle cameras with a Wide Field of View (WFOV); 2) systems using orientable magnification devices, such as PTZ cameras. In the former, cameras are arranged in a maximum coverage strategy to monitor multiple subjects in a surveillance area. These systems are popular for its flexibility and reduced cost, however the limited resolution of biometric data is regarded as their major drawback. The second group comprises systems using PTZ cameras for acquiring high-resolution imagery of regions of interest in the scene. In spite of the vast number of challenges, it is commonly accepted that these devices are the most efficient solution for acquiring biometric data at a distance.

2.2.1.1 Fixed-angle Surveillance Systems

Most automated surveillance systems operate with fixed CCTV cameras. The major reasons for relying on fixed-angle surveillance systems are the following: 1) the reduced cost; 2) the large number of outdoor CCTV cameras; and 3) the effectiveness of detection and tracking

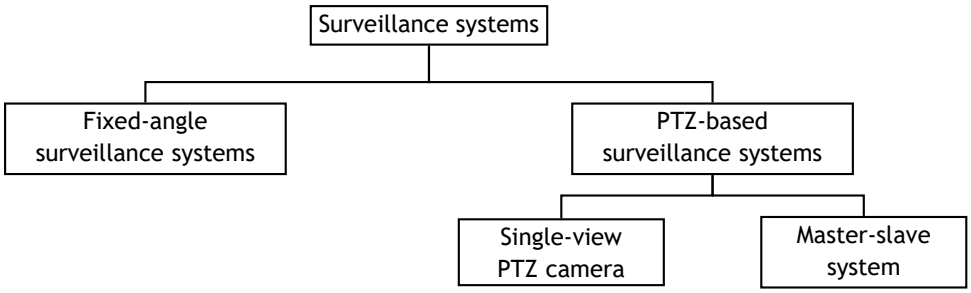


Figure 2.1: Taxonomy of the most common architectures of surveillance systems.

algorithms in the acquired data. However, in wide open scenarios, the obtained resolution is not sufficient to represent the important patterns of biometric samples, restraining the recognition of humans at a distance. With the rise of high-resolution cameras, they have been considered the substitutes of old CCTV cameras and suggested as the solution for remote human recognition. Even though high-resolution cameras can be a practical solution for mid-term distances, they still can not outperform PTZ-based systems. Figure 2.2 illustrates the relation between the interpupillary distance and the stand-off distance (the distance between the front of the lens and the subject) when using different optical devices. In this comparison, the Angle of View (AOV) of wide-view cameras was considered as $70°$, while the AOV of the PTZ camera at the maximum zoom was assumed to be $2.1°$. The comparison between the resolution of different cameras demonstrates that only PTZ cameras can acquire high-resolution face imagery at a distance, i.e., face images with an interpupillary distance greater than 60 pixels and acquired with a stand-off distance higher than 5m. For this reason, PTZ cameras are considered as the most efficient solution for acquiring high-resolution biometric data in surveillance scenarios.

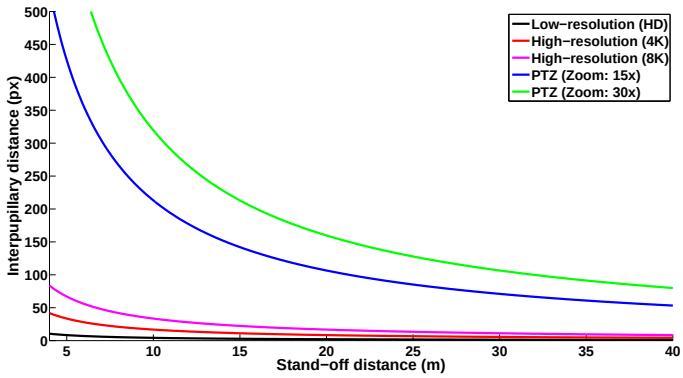


Figure 2.2: Relation between the interpupillary resolution and the stand-off distance when using different acquisition devices. The number of pixels between the eyes is determined with respect to the stand-off distance and four acquisition devices: 1) a typical surveillance camera (720p, $70°$); 2) a high-resolution camera (4K, $70°$); 3) a high-resolution camera (4K, $70°$); 4) a PTZ camera with 15x zoom (1080p, $4.2°$); and 5) PTZ camera with 30x zoom (1080p, $2.1°$). Note the evident advantages of using PTZ cameras, the resolution of face traits is more than 5 times the resolution of 8K cameras.

2.2.1.2 PTZ-based Surveillance Systems

This section provides a detailed review of PTZ-based surveillance systems. These systems can be broadly divided into two architectures: 1) master-slave; and 2) single PTZ.

Table 2.1 provides a comparison between state-of-the-art PTZ-based surveillance systems. It must be noted that the majority of these systems were not designed specifically for the ac-

Table 2.1: State-of-the-art PTZ-based video surveillance systems. Master-slave systems are organized with respect to the type of master camera used, the accuracy of pan-tilt estimation, the required camera disposal, the need for intermediate zoom states and the use of calibration marks.

System	Architecture	Master Camera	Pan-Tilt Estimation	Camera Disposal	Int. Zoom States	Calibration Marks
Kumar et al. [124]	Single PTZ	-	-	-	-	-
Varcheie and Bilodeau [125]	Single PTZ	-	-	-	-	-
Yao et al. [126]	Single PTZ	-	-	-	-	-
Varcheie and Bilodeau [127]	Single PTZ	-	-	-	-	-
Tordoff and Murray [128]	Single PTZ	-	-	-	-	-
Yao el al. [126]	Single PTZ	-	-	-	-	-
Zhou et al. [129]	Master-Slave	PTZ	Approximated	Specific	Yes	No
Liao and Chen [130]	Master-Slave	PTZ	Approximated	Specific	Yes	No
Bodor et al. [131]	Master-Slave	Wide	Approximated	Specific	No	Yes
Del Bimbo et al. [132]	Master-Slave	PTZ	Approximated	Arbitrary	Yes	No
Everts et al. [133]	Master-Slave	PTZ	Approximated	Arbitrary	No	No
Chen et al. [134]	Master-Slave	Omnidirectional	Approximated	Arbitrary	No	Yes
Tarhan and Altug [135]	Master-Slave	Catadioptric	Approximated	Specific	No	No
Xu and Song [136]	Master-Slave	Wide	Exact	Arbitrary	Yes	No
Lu and Payandeh [137]	Master-Slave	Wide	Exact	Arbitrary	Yes	Yes
Scotti et al. [138]	Master-Slave	Catadioptric	Exact	Specific	Yes	Yes
Krahnstoever et al. [139]	Master-Slave	PTZ	Exact	Arbitrary	No	No
Yang et al. [140]	Master-Slave	PTZ	Exact	Arbitrary	No	Yes

quisition of biometric data. Instead, this section is devoted to review PTZ-based approaches capable of instructing a PTZ camera to acquire high-resolution images/videos of specific parts of scene. A detailed description of PTZ-based surveillance systems specially conceived for biometric data acquisition is provided in section 2.2.3.

Single PTZ System. Single PTZ camera systems work by locating the region of interest (e.g., facial region) followed by increasing the zoom level for acquiring high-resolution biometric samples. When compared to multiple camera PTZ-based systems, this strategy is advantageous because it does not depend on inter-camera calibration to accurately determine the pan-tilt angles. However, the continuous change in the zoom level increases significantly the likelihood of tracking failure.

Kumar et al. [124], and Varcheie and Bilodeau [125, 127] used a single PTZ device in surveillance scenarios, where pan-tilt values were adjusted to keep the tracked subject in the central region of the camera view. In both approaches, zoom adjustment was not implemented, restraining the acquisition of hard biometric traits.

A common approach in single-view PTZ-based systems is the use of size preserving tracking algorithms [126, 128, 141]. In this strategy, the PTZ orientation and focal distance can be adjusted using distinct strategies: 1) region-based features [142, 143]; 2) image velocity features [141]; and 3) target depth inference [126, 128]. Region-based approaches rely on features

extracted from the object of interest to control the PTZ camera. As an example, Shah et al. [143] incorporated the zoom level variable in the particle filter tracking algorithm by adjusting the zoom level with respect to the visible percentage of the target. Methods based on image velocity rely on the motion gradient to determine the angle displacement. Fayman et al. [141] proposed a closed-loop feedback algorithm based on the optical flow and on the depth information obtained from the auto-focus camera sensor. Approaches based on target depth estimation are regarded as the most adequate, since they are capable of recovering the 3D motion of the target, which improves the zoom level adjustment accuracy. Tordoff and Murray [128] used the weak perspective projection model, i.e., a highly simplified representation of the real imaging process that ignores the influence of the center offset. This approach was improved by the work of Yao et al. [126] by using the paraperspective projection model.

Master-slave System. As illustrated in figure 2.3, the master-slave architecture regards surveillance systems where a magnification device, usually a PTZ camera, is controlled by one or more wide-angle cameras, which are responsible for monitoring a wide surveillance area. In this architecture, the WFOV cameras furnish the input data for detection and tracking modules, while a control module relies on the output of these modules to point the Narrow Field of View (NFOV) and acquire high-resolution images of the region of interest. The denomination master-slave is justified by the fact that WFOV cameras provide the data used in the decision-making process, while the NFOV camera depends on the all remaining modules being just used as a foveal sensor.

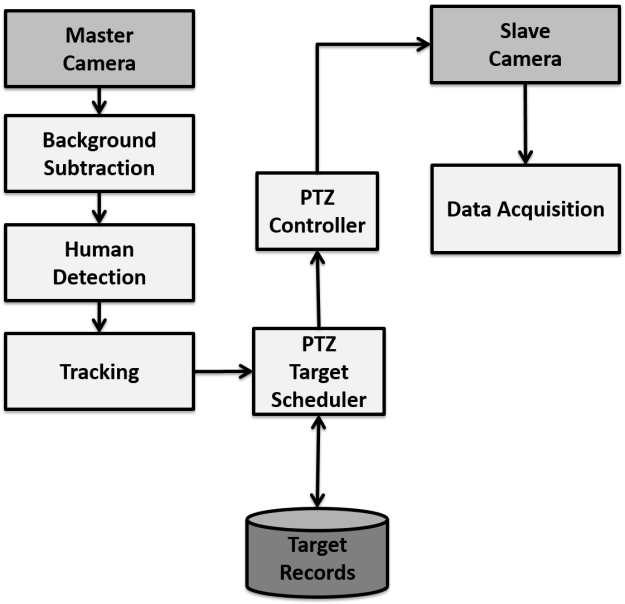


Figure 2.3: Overview of a typical master-slave video surveillance system.

As previously described in section 2.2.1.2, the limitations of single PTZ systems restrain their use in outdoor surveillance scenarios, where extreme zoom levels are required for acquiring high-resolution data. For this reason, the majority of works have focused on master-slave approaches (as evidenced by table 2.1), and is commonly accepted that they represent the most appropriate solution to address the challenges of biometric recognition in video surveillance scenarios. In spite of the multiple advantages of this strategy, the design of fully automated master-slave surveillance systems is not straightforward. Inter-camera calibration is the major bottleneck of this configuration (see section 2.2.2.4), since determining the mapping function from image coordinates to pan-tilt parameters requires depth information. Accordingly, the

existing master-slave systems mainly differ with respect to the accuracy of pan-tilt estimation parameters.

The use of 2D-based approximation is the most common approach. By constructing a mapping between the wide-view coordinates and the pan-tilt values, the inference of depth information is avoided, but, in turn it is necessary to rely on different assumptions (e.g., similar points-of-view [144], intermediate zoom states [132, 138]) to alleviate pan-tilt inaccuracies. Zhou et al. [145] relied on manually constructed look-up tables and linear interpolation to map pixel locations of the master camera to pan-tilt values. In a similar fashion, Liao and Cho [146] approximated the target position as its projection in the reference plane, to which a pixel to pan-tilt mapping had been previously constructed. To alleviate the burden of manual mapping, Liu et al. [147] presented an automatic calibration approach by estimating an approximate relation between camera images using feature point matching. Del Bimbo et al. [132] proposed a dual PTZ system, where monitoring onus is interchangeable. In the offline phase, the pan and tilt parameters of the cameras are changed in a step-by-step manner for extracting appearance-based features from the whole scene and creating a correspondence table between visual landmarks and pan-tilt angles. At run-time, features are extracted in the current master camera view and matched with the pre-built feature map, allowing to localize the camera with respect to the scene and hence estimate the position of the target. Self-calibration is regarded as the major advantage of this approach (see column 'calibration marks' in table 2.1). On the other hand, the dependency of stationary visual landmarks for calibration may be problematic in dynamic surveillance scenarios (e.g., a crowded scene, moving objects that significantly change the appearance of the scene). Zhou et al. [129] and Liao et al. [130] also used dual-PTZ systems for tracking subjects in an intermediate zoom level. An alternative approach is the use of omni-directional [134] or catadioptric cameras[1] [135, 138]. The major advantage of these systems is the possibility to observe a scene at about $360°$.

The exact inference of pan-tilt parameters is regarded as the most promising solution for the development of a realistic surveillance system. However, the accuracy of pan-tilt estimation requires the inference of subject depth. Xu and Song [136] relied on multiple consecutive frames to approximate target depth, but this strategy is time-consuming, and consequently, increases the delay between issuing the order and directing the PTZ. You et al. [148] estimated the relationship between the master and the slave camera using a homography for each image of the mosaic derived from the slave camera. An innovative solution for this problem is the use of a beam splitter[2]. This device ensures that both the master camera and the PTZ camera have the same scene view, which eases inter-camera calibration. Figure 2.4 illustrates the functioning mode of the beam splitter in the context of a master-slave system. Park et al. [149] were pioneers in exploiting this device to resolve the problem inter-camera calibration in PTZ-based systems. In this system, the cameras were installed in a dark box for obtaining sharp images. Also, the cameras and the beam splitter were disposed so that an incident beam is projected in both cameras in the same sensor position, ensuring a trivial mapping between the pan-tilt angles and the pixel position of the master camera. Regarding zoom control, a quadratic mapping function between the size of the human silhouette in the master camera and zoom values of the PTZ camera was inferred during system installment.

[1] A catadioptric optical system is one where refraction and reflection are combined in an optical system, usually via lenses (dioptrics) and curved mirrors (catoptrics).

[2] A beam splitter is an optical device that splits a beam of light into two.

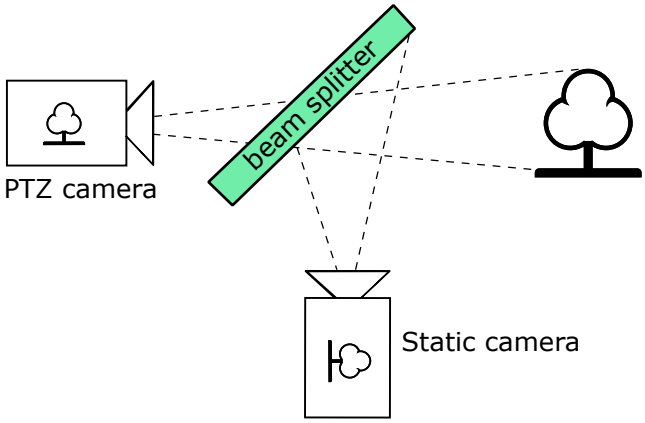


Figure 2.4: Schematic view of a multi-camera system using a beam splitter. The beamer splits the light into two so that the PTZ camera and the master camera can share the same view of the scene.

2.2.2 Typical Challenges of PTZ-based systems

As discussed in section 2.2.1.2, PTZ-based approaches are currently the best strategy for acquiring biometric data in outdoor environments. State-of-the-art PTZ cameras can achieve optical zoom magnifications up to 30x with an AOV of about 2°, ensuring the acquisition high-resolution samples at a distance. Despite these advantages, the use of a NFOV cameras also entails several challenges.

2.2.2.1 Optics Distortions

The use of high zoom levels has a tremendous impact on the quality of the acquired images, since optical magnification is achieved by increasing the focal distance of the camera (f) and reducing its AOV. As a consequence, the amount of light reaching the sensor is considerably less as the AOV decreases, which is particularly critical in outdoor scenarios where illumination is nonstandard.

To compensate for this effect, most cameras increase the aperture of the diaphragm (D) in the same proportion of f. The ratio between f and the aperture of the camera is denoted as F-number (see equation (2.1)), and is commonly used in photography to maintain image brightness along different zoom magnifications.

$$\text{F-number} = \frac{f}{D} \quad (2.1)$$

However, its side effect is the reduction of the depth of field, which in turn increases the likelihood of obtaining blurred images. As an alternative, it is possible to increase the exposure time E for balancing the impact that extreme *f* values may have on the amount of light that reaches the sensor. However, higher values of E also increase the motion-blur level in the images.

A more robust solution is to adjust simultaneously both D and E, which is, in general, the strategy adopted by PTZ devices. However, as illustrated in figure 2.5, the number of ideal configurations for (D,E) is greatly dependent on zoom magnification.

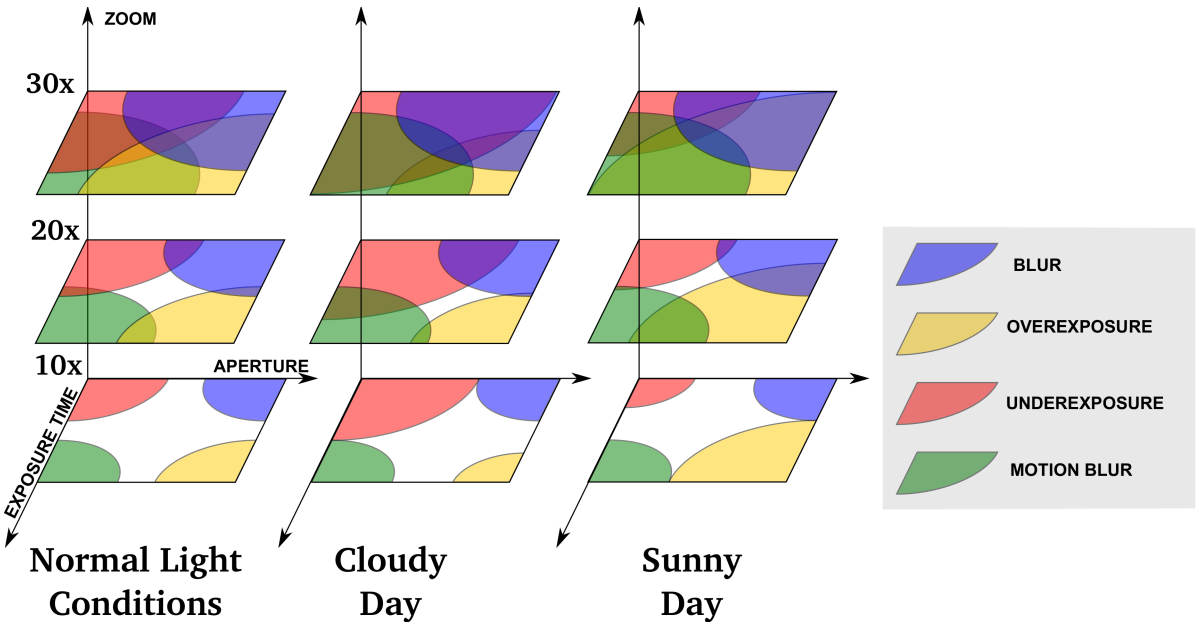


Figure 2.5: The spectrum of optical distortions with respect to exposure time, aperture and zoom level. The set of (D, E) combinations that produces non-degraded images decreases significantly as the focal distance increases. Besides, it is worth noting that the ideal set of (D, E) values (in white) varies with respect to the illumination conditions.

2.2.2.2 Non-comprehensive view of the scene

In single PTZ systems, zooming enables the close inspection of narrow regions in the scene, but it also inhibits scene monitoring. As a consequence, the detection and tracking of individuals can hardly be attained when using extreme zoom levels.

To mitigate this shortcoming, some systems alternate between different zoom levels, i.e., subject detection and tracking is performed in minimal zoom levels, and high-resolution data is obtained using maximum zoom levels. However, zoom transition is the most time-consuming task of PTZ devices, which significantly restricts the efficiency of using a single PTZ camera for biometric recognition purposes.

2.2.2.3 Out-of-Focus

As previously discussed in section 2.2.2.1, the use of extreme zoom levels reduces significantly the depth of field. To correctly adjust the focus distance to the subject position in the scene two different strategies can be exploited: 1) auto-focus; 2) manual focusing.

In the former, focus adjustment is guided by an image contrast maximization search. Even though this approach is highly effective in wide-view cameras, it fails at providing focused images of moving subjects when using extreme zoom magnifications. First, the reduced width of view perceived by the camera significantly reduces the amount of time the subject is imaged (approximately 1s), and the auto-focus mechanism is not fast enough (approximately 2s). Second, the motion blur introduced in the image compromises the contrast adjustment scheme.

As an alternative, the focus lens can be manually adjusted with respect to the distance of the subject to the camera. Given the 3D position of the subject, it is possible to infer its distance to the camera. Then, focus is dynamically adjusted using a function relating the subjects distance and the focus lens position. In this strategy, the estimation of 3D subject position is regarded as the major bottleneck, since it depends on the use of stereo reconstruction

techniques. However, this issue has been progressively addressed by state-of-the-art methods since 3D information is critical for accurately pointing the PTZ camera.

2.2.2.4 Calibration of multi-camera systems

In a multi-camera system the cameras are, in general, supposed to cooperate and share the acquisitions of the scene. Therefore, apart from calibrating each camera separately, in such systems it must be defined a mapping function between the camera views that can convert a point in the coordinate system of a camera into the one of another.

However, this is an ill-posed problem because of an important constraint of multi-camera systems that is related to epipolar geometry. The epipolar geometry [150] is used to represent the geometric relations of two points onto 2D images that come from two cameras when pointing at the same location in the world coordinates (in a 3D space). Figure 2.6 shows a typical example of epipolar geometry where a shared 3D point X is observed by both O_1 and O_2. We can see that, by changing the position of X (see dots along the view-axis of O_1) its projection X_1 remains the same but it changes in X_2. Only if the relative position of the two cameras is known it is possible to estimate the match between the two image planes and therefore obtain the exact position for both cameras. Assuming that O_1 is the wide-view camera of a master-slave system and O_2 is the PTZ camera, it is not be possible to determine the pan-tilt angle necessary to observe X by only using the information of its projection X_1.

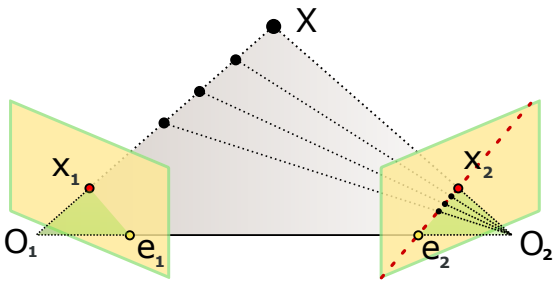


Figure 2.6: Epipolar geometry of a 3D point over two image planes. Two cameras, with their respective centers of projection points O_1 and O_2, observe the point X. The projection of X onto each of the image planes is denoted x_1 and x_2. Points e_1 and e_2 are the epipoles.

2.2.3 State-of-the-art Biometric Data Acquisition Systems

The previous sections described the principal frameworks for acquiring high-resolution images/videos at distance, along with the typical challenges of each strategy. This section is devoted to review the state-of-the-art systems particularly designed to acquire a specific biometric trait using these frameworks. Similarly to section 2.2.1.2, the most relevant systems were summarized in table 2.2 according to the type of master camera used, the accuracy of pan-tilt estimation, the required camera disposal, the need for intermediate zoom states and the use of calibration marks.

2.2.3.1 Iris

Commercial iris recognition systems can identify subjects with extremely low error rates. However, they rely on highly restrictive capture volumes, reducing their workability in less constrained scenarios. In the last years, different works have attempted to relax the constraints of iris recognition systems by exploiting innovative strategies to increase both the capture volume and the stand-off distance, i.e., the distance between the front of the lens and the subject. Successful identification of humans using iris is greatly dependent on the quality of the iris image. To be considered as acceptable quality, the standards recommend a resolution of 200 pixels across the iris (ISO/IEC 2004), and an in-focus image. Also, sufficient near infra-red (IR) illumination should be ensured (more than 2 mw/cm^2) without harming human health (less than 10 mw/cm^2 according to the international safety standard IEC-60852-1). The volume of space in front of the acquisition system where all these constraints are satisfied is denoted as the capture volume of the system. Considering all these constraints, the design of an acquisition framework capable of acquiring good quality iris images in unconstrained scenarios is extremely hard, particularly at large stand-off distances. This section reviews the most relevant works and acquisition protocols for iris and periocular recognition at a distance.

Current strategies to perform the acquisition of iris data in less constrained conditions can be divided into two families, depending of whether they use (or not) magnification devices. In terms of the approaches that make no use of magnification devices, the Iris-on-the-Move [2] system is notable for having significantly decreased the cooperation levels required for image acquisition, allowing subjects continuous movement through a portal equipped with near-infrared illuminators. Another well known commercial device is the LG IrisAccess4000, where image is acquired at-a-distance, provided that subjects' gaze point at a specific direction. Magnification devices, such as PTZ cameras, extend the system stand-off distance while providing enough resolution for reliable iris recognition. Wheeler et al. [151] introduced a system to acquire iris data at a resolution of 200 pixels from cooperative subjects at 1.5 m, using a PTZ camera assisted by two wide-view cameras. Dong et al. [152] also proposed a PTZ-based system, that images iris data up to distances of 3 m with more than 150 pixels across the iris diameter. Yoon et al. [153] relied on a light stripe to determine the 3D position, avoiding the use of an extra wide camera. The Eagle Eye system [154] uses one wide-view camera and three close view cameras, for capturing simultaneous images of both irises. This system has a stand-off distance of about 5 m with a operational range of 3mx2mx3m. This system uses a bi-ocular setup, that enables to recover the 3D world position of the subject by stereo reconstruction. Depth information cues are used both for pan/tilt angles estimation and for getting focused data. Despite being considered more reliable, the use of two wide-angle cameras significantly increases the system cost and limits its flexibility. To address this problem, various commercial solutions were introduced: Mitsubishi corporation developed a scheme where depth is estimated using the disparity between facial features [155]. Yoo et al. [156] combined the wide-view and narrow-view cameras with a beam splitter to simultaneously acquire facial and iris images. This integrated dual-sensor enables the same ray to be mapped to same position in both cameras sensors, avoiding the need for depth estimation.

2.2.3.2 Face

Face is the most popular biometric trait in surveillance scenarios when using PTZ-based systems. This can be explained by the fact that face is the most viable trait for recognition at a distance, due to its visibility and capability of being imaged in a covert manner.

Table 2.2: State-of-the-art PTZ-based systems designed for acquiring biometric data. The master-slave systems are organized with respect to the type of master camera used, the accuracy of pan-tilt estimation, the required camera disposal, the need for intermediate zoom states and the use of calibration marks.

System	Architecture	Master Camera	Pan-Tilt Estimation	Camera Disposal	Int. Zoom States	Calibration Marks
FACE						
Hampapur et al. [157]	Master-Slave	Wide	Exact	Arbitrary	No	Yes
Stillman et al. [158]	Master-Slave	Wide	Approximated	Specific	No	No
Wheeler et al. [144]	Master-Slave	Wide	Approximated	Arbitrary	No	Yes
Marchesotti et al. [159]	Master-Slave	Wide	Approximated	Arbitrary	Yes	Yes
Park et al. [149, 160]	Master-Slave	Wide	Exact	Specific	Yes	No
Amnu et al. [161]	Master-Slave	Wide	Exact	Specific	No	No
Bernardin et. al [162]	Single PTZ	-	-	-	-	-
Mian [163]	Single PTZ	-	-	-	-	-
IRIS						
Wheeler et. al [151]	Master-Slave	Wide	Exact	Specific	Yes	No
Yoon et. al [153]	Master-Slave	Wide	Approximated	Specific	Yes	Yes
Bashir et. al [154]	Master-Slave	Wide	Exact	Specific	No	No
Venug. and Savv. [164]	Single PTZ	-	-	-	-	-

Bernardin et al. [162] performed human detection using fuzzy rules to simulate the natural behavior of a human operator, which ensured a smoother camera handling. A KLT tracker [33] was used to track face keypoints over the time. Mian [163] also proposed a single PTZ-camera system to detect and track faces over the video stream by exploiting the Camshift algorithm [165]. As already discussed in previous sections, using a single camera for detection and tracking avoids the problems related to excessive calibration. However, especially when facing with biometrics, multi-camera systems become necessary to deal with the problem of off-pose or occlusions.

Regarding master-slave systems, the work of Stillman et al. [158] represents one of the first attempts where multiple cameras were combined for biometric data acquisition in surveillance scenarios. Simple skin-color segmentation and color indexing methods were used to locate multiple people in a calibrated space. Hampapur et al. [157] and Marchesotti et al. [159] used both background subtraction techniques to extract the people silhouettes from the scene and used appearance information to detect and track people's faces. Appearance-based techniques are in general computationally inexpensive but are also affected by several limitations related to illumination and occlusions. However, in surveillance scenarios these techniques remain as the most feasible solution to adopt. Amnuaykanjanasin et al. [161] used stereo-matching and triangulation between a pair of camera streams to estimate the 3D position of a person. The proposed method relies on color information of the skin to detect the faces, and the depth information from stereo-matching ensures a good estimation of the PTZ parameters to point the camera. Wheeler et al. [144] combined a WFOV camera with a NFOV PTZ camera for acquiring high-resolution face images at a maximum distance of 20m. However, in order to ease inter-camera calibration, the two cameras need to be installed side-by-side and the line defined by

the focal points of the cameras should be parallel to the ground. Park et al. [149] proposed a very similar solution, but in this case the inter-camera calibration was obtained by relying on a beam splitter.

2.3 Biometric Recognition in Surveillance Scenarios

This section overviews the evolution of the biometrics field over time and the main concepts related to biometrics. Also, it provides a comprehensive review of the state of the art in biometric recognition in the wild.

2.3.1 Historical Background

The term 'biometrics' is derived from the Greek words 'bio' meaning life and 'metric' meaning to measure. Accordingly, biometric recognition denotes the identification/authentication of individuals based on their physical or behavioral traits.

The first recognition system using biometric data was proposed in 1883 by Bertillon where specific lengths and widths of the head and body were used for identifying convicted criminals. This proposal represents an important mark in the development of objective methods for identifying individuals based on their biometric traits, but it fails to provide a unique description of an individual, as it was found that two subjects could share the same measures. The failure of the Bertillon system allowed to conclude that a biometric should not only be easy to measure and stable over time, but also be unique per individual. As a solution, Sir Francis Galton and Sir Edward Henry studied the idea introduced by Henry Faulds of using fingerprints for identification [166], and developed the first elementary fingerprint recognition system. The uniqueness of fingerprints granted the proliferation of fingerprint recognition, being currently one the most popular and used biometric traits used worldwide.

In spite of the good performance of fingerprint, the research community continued searching for different traits (e.g., iris, face, ear, gait, keystroke, palmprint, voice, hand vein). Also, the biometric recognition research evolved towards the development of identification/authentication systems capable of working in unconstrained and non-cooperative scenarios. Among the several proposed biometric traits, iris, periocular region, and gait are regarded as the most promising for being acquired at a distance and without subject cooperation. These efforts have proven to be fruitful as evidenced by the development of the Iris-On-The-Move system [2], where subjects moving at a normal walking pace through a minimally confining portal are recognized based on automatically acquired iris images. Currently, the focus is put on the development of fully automated biometric recognition systems capable of operating in uncontrolled conditions, such as the ones observed in surveillance scenarios.

2.3.2 Effectiveness Measures

The objective evaluation and comparison between biometric systems can be a hard task. Even though it is impossible to give a single value that reflects the accuracy of a recognition system, there are metrics that are commonly used for assessing the performance of these systems.

Let FA and N be the number of times an impostor was accepted by the system, and N the number of recognition processes performed by impostors, respectively. The ratio between these two metrics defines the False Acceptance Rate (FAR) of the system:

$$FAR = \frac{FA}{N}, \quad (2.2)$$

which measures the probability that the identity of valid users is denied.

Let FR and P be the number of times a registered user is rejected by the system, and P the number of recognition processes performed by a registered user, respectively. The ratio between these two metrics defines the False Rejection Rate (FRR) of the system:

$$FRR = \frac{FR}{P}, \quad (2.3)$$

which measures the probability of a registered user be confounded with an impostor.

Both FAR and FRR vary with respect to the similarity threshold t used for accepting a comparison between biometric signatures. By decreasing t, the number of accepted impostors is reduced and consequently FAR diminishes, but in turn FRR increases. The inverse effect is observed when increasing t. As such, there is an inverse correspondence between FAR and FRR, and is the obligation of the system administrator to adjust the value of t to the privilege the most important ratio according to the system requirements. A common approach is to determine the threshold t that yields an equal FAR and FRR, and the value of these ratios is denoted as the Equal Error Rate (EER). The relation between the FAR and the FRR represents the recognition performance of the biometric system and is usually represented by the Receiver Operating Curve (ROC), whose area under curve is commonly used to summarize the system performance. The described measures are the most common in the evaluation of systems working in the verification mode.

In the identification mode, the system performance is described by the relation between the probability of observing a correct identity within the list of the most similar identities in database, and the length of the list.

Let $C(k)$ be the number of times that the identity of a registered user is among the k identities retrieved by the system, and N the number of identities in the database. The identification accuracy with respect to k is given by:

$$TPIR(k) = \frac{C(k)}{N}. \quad (2.4)$$

The relation between $TPIR$ and k describes the system performance and is usually represented in the Cumulative Matching Characteristic (CMC) curve.

2.3.3 Operating Modes

Independently of the used trait, biometric recognition systems follow the processing chain illustrated in figure 2.7. First, the data acquisition module collects a biometric sample, which is passed through the feature extraction module for producing a biometric signature of a person. Next, the matching phase compares the signature to a set of biometric templates available in the database and acquired during enrollment. The number of comparisons performed distinguishes between the two modes of performing biometric recognition: 1) verification; and 2) identification.

In the verification mode, the system aims at answering the question: 'is this person who

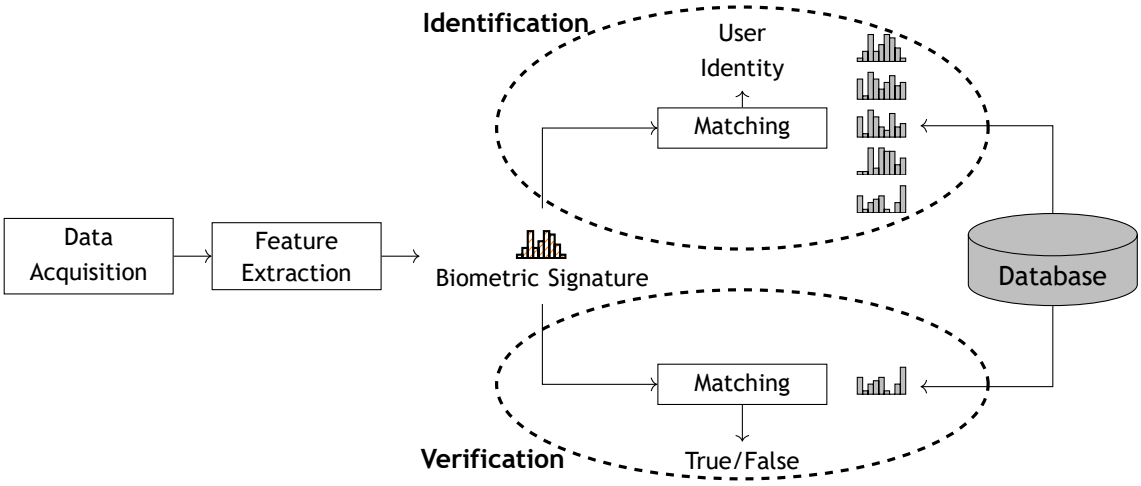


Figure 2.7: Typical stages of a biometric recognition system and the two operating modes.

he/she claims to be?' Accordingly, the user id is provided along with the biometric sample, and the system outputs a binary answer according to the similarity score between the biometric signature and the biometric template available in the database.

In the identification mode, the system tries to answer the question: 'who is this person?' For this purpose, the biometric signature is compared to the N templates of the database, and the index of the k highest similarity scores are returned as possible user identities.

2.3.4 State-of-the-art Biometric Recognition Methods

This section reviews the most relevant biometric recognition approaches based on face or gait. The rationale for confining the revision to these two traits is their feasibility to be acquired at a distance. While gait can be easily acquired by any kind of surveillance camera, the face can also be properly imaged using PTZ-based systems. However, as evidenced in section 2.2.3, the other facial traits, such as iris, require further improvements in the resolution of the existing frameworks to permit the acquisition of data with sufficient quality for recognition purposes.

2.3.4.1 Face Recognition

The search for algorithms that are capable of recognizing humans using the facial region has occurred over more than 50 years. The first attempt dates back to 1964, when Bledsoe [167] developed a facial recognition system that was based on a set of 20 distances measured from facial keypoints. During his experiments, Bledsoe stressed that the "great variability in head rotation and tilt, lighting intensity and angle, facial expression and aging" make face recognition an extremely difficult challenge. To date, these variability factors remain the primary focus of face recognition research studies.

Turk and Pentland [168] introduced the notion of eigenfaces to represent facial features in a low-dimensional space. Recognition was attained by projecting the new image, which is considered to be a point in N-dimensional space, in the face space and determining the nearest neighbor. Although the eigenfaces method is regarded as one of the first facial recognition technologies, robustness to degradation factors, such as lighting and pose, is barely attained. Later, Belhumeur et al. [169] improved this idea by using LDA instead of PCA to represent the facial features. To address the pose variation, Blanz and Vetter [170] introduced morphable models. Still images, captured at different poses, were used to build a 3D face model that

contained shape and texture information. The model was used to generate synthetic images under varying poses, with a view to enlarging the training set with representative images of all possible variations.

The use of LBP [171] to encode facial features has made a significant contribution toward increasing facial recognition performance in non-ideal scenarios. This strategy attained state-of-the-art results not only in frontal faces but also in faces that were subjected to varying illumination and expression. Again, several studies used this idea to provide further robustness to unconstrained face recognition. Li et al. [172] developed an illumination-invariant face recognition system by combining near-infrared imaging with a LBP-based face description. Tan and Triggs [173] extended the LBP to LTP to address difficult lighting conditions. Recent methods [174, 175] have found the LPQ descriptor [176] to be more robust than LBP to specific degradation factors, such as blur. Occlusions are another typical degradation factor of face recognition systems, and this factor has been addressed in several studies [177]. Nevertheless, robustness to occlusion was attained only when sparse representation techniques were introduced in facial recognition [178]. These results were subsequently improved and the processing time decreased by combining sparse coding with the ELM algorithm [179]. The advances in face recognition performance in less constrained conditions have paved the way for face recognition in real-world scenarios, whose popularity has exponentially risen with the introduction of LFW database [180]. The particularities of this set, such as the large variability in expression, pose, illumination and the objective evaluation protocol, established it as the reference benchmark for unconstrained face recognition and fostered the development of approaches robust to non-cooperative scenarios [181–183].

The improvement of the state-of-the-art performances of face recognition has been supported by the introduction of Convolutional Neural Networks (CNNs) [182, 184–187]. Rather than construct classification models over traditional hand-crafted features, the data-driven nature of deep learning has successfully enhanced the robustness of learned facial features. As such, if sufficient training data is provided, CNNs are capable of handling pose, occlusion and illumination variations of face images to a considerably high degree [188].

2.3.4.2 Gait Recognition

The way humans walk can be used for identification purposes and is usually known as gait recognition [189, 190]. This trait is advantageous for the following reasons: 1) it can be easily measured at a distance; 2) it is difficult to disguise or occlude; and 3) it is robust to low-resolution images. Moreover, a recent study about the covariate factors affecting recognition performance has found that gait is time-invariant in the short and medium term [191] thus gaining a special attention among reliable biometric traits. On the other hand, gait strongly depends on the control over clothing and footwear, which impacts negatively its feasibility in surveillance scenarios.

Notwithstanding, many methods have been introduced in the literature to optimize gait recognition systems. Ran et al. [192] created a gait signature from surveillance videos by stacking the sequence in the spatiotemporal space. The symmetries of the signature patterns enable a reliable and effective learning in the presence of imperfect gait period, self occlusion, and clutter. Venkat and De Wilde [193] addressed the problem of low-resolution videos by combining the information from sub-gaits (a part of the silhouette of a moving body) in a probabilistic approach. Moustakas et al. [194] combined the height and stride length in a probabilistic framework to improve the accuracy of a gait recognition system. Conversely, Jung et al. [195]

exploited gait to estimate the head pose in surveillance scenarios. In this approach, a 3D face model was also inferred to improve recognition performance. Choudhury and Tjahjadi [196] analyzed the human silhouettes inferred from gait sequences to attenuate the presence of noise during recognition. Considering that this strategy is highly dependent on clothing, the authors extended their approach in [197], where they introduced a strategy for handling occlusion factors caused by variations of view (e.g., subject's clothing and the presence of a carried item). Kusakunniran [198] introduced the space-time interest points for encoding gait. The interest points of the walking pattern were directly estimated from raw video sequences on the spatio-temporal feature domain, avoiding the use of pre-processing techniques (e.g., background subtraction, edge detection, human silhouettes and so on). The proposed method is robust to partial occlusion caused by carrying items or variations in hair/clothes/footwear.

2.4 Summary

This chapter presented a comprehensive review of the concepts related to the stages of automated surveillance systems intended for biometric recognition purposes. First, we reviewed the most relevant approaches in each one of the typical stages of an automated surveillance system. With regard to the pre-detection phase, it was interesting to note that an increasing number of background subtraction algorithms were especially interested in providing additional robustness to surveillance scenarios [199, 200]. This trend was confirmed by the development of benchmarks that are specifically focused in assessing the performance of background subtraction in these scenarios [201]. Similarly, in the detection phase, there was an increasing interest in extending human detection to highly challenging conditions, where a large number of subjects move freely in outdoor scenarios. In the tracking field, in spite of the majority of the approaches being not specifically focused on surveillance scenarios, a large effort has been made to benchmark state-of-the-art algorithms with the VOT challenges [202], which has consequently contributed to propelling forward the performance of tracking algorithms in complex scenes. Next, we performed a comparative analysis between the most common architectures of surveillance systems with respect to their capability of acquiring biometric samples at a distance. Besides, we reviewed the state-of-the-art optical frameworks capable of acquiring biometric data at a distance. Finally, we introduced the concepts related to biometrics, such as the modes of functioning and the effectiveness measures. Also, we reviewed the state of the art in biometric recognition in surveillance scenarios by focusing on the two most promising traits to be acquired at a distance: face and gait.

Chapter 3

The QUIS-CAMPI System

In this chapter, we describe the development of the QUIS-CAMPI surveillance system, an automated surveillance system intended for acquiring biometric data at a distance from non-cooperative subjects. As illustrated in figure 3.1, the proposed surveillance system is divided in five major modules, broadly grouped in three main phases: 1) human monitoring; 2) inter-camera calibration; 3) camera scheduling. The rationale behind the proposed system is to use the PTZ camera as a foveal sensor, i.e., the video stream obtained from the wide camera is analyzed to obtain the location of subjects' head, so that the PTZ camera can image the facial region at a high-magnification state. In the former phase, the master camera is responsible for detecting and tracking multiple subjects in the surveillance area.

The methods used in each processing module are described along this chapter. First, we detail the methods used for monitoring humans in surveillance scenarios. Then, we describe two methods for increasing the workability of the QUIS-CAMPI system in three aspects: 1) allowing the deployment in outdoor scenarios; 2) extending the maximum acquisition distance; and 3) increasing the acquisition accuracy of the system. Finally, section 3.4 summarizes the major contributions of this chapter.

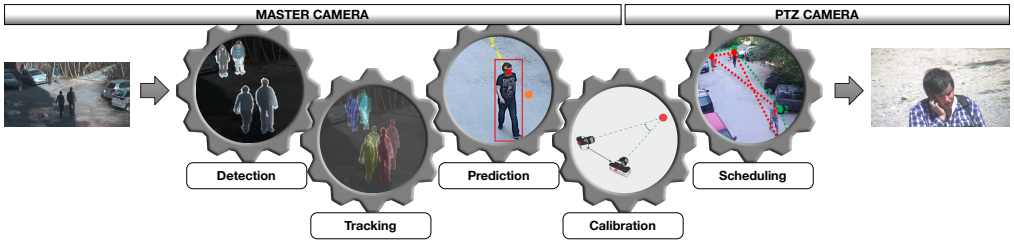


Figure 3.1: Processing chain of the QUIS-CAMPI surveillance system. A master-slave architecture is adopted for the proposed surveillance system, where the master camera is responsible for monitoring a surveillance area and providing a set of interest regions (in this case the location of subjects face) to the PTZ camera.

3.1 Human Monitoring

As described in section 1.1, human monitoring is beyond the scope of this thesis. Nevertheless, the detection and tracking modules are a necessary element in any surveillance system intended for biometric recognition purposes. Accordingly, this section describes the methods chosen for accomplishing the development of the proposed system.

3.1.1 Motion Detection

Background subtraction is typically the first phase of the processing chain of automated surveillance systems and holds the feasibility of all the subsequent phases. Hence, it is particularly important to perceive the relative effectiveness of BGS with respect to the kind of

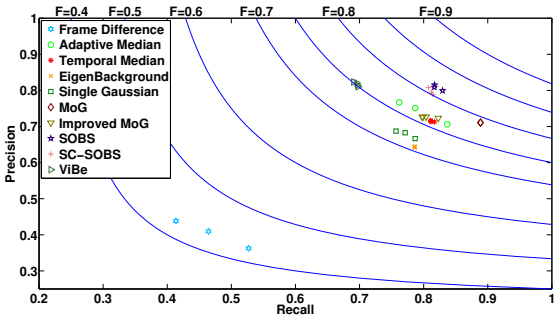


Figure 3.2: The best three configurations obtained for each BGS method using an exhaustive search of the parameter space. Blue lines denote the set of points with constant f-measure.

environment. For this reason, we performed an objective evaluation of the state-of-the-art BGS algorithms on unconstrained outdoor environments.

For this evaluation, we selected 10 state-of-the-art background subtraction algorithms, namely Frame Difference [20], Adaptive Median [21], Temporal Median [22], EigenBackground [83], Single Gaussian [23], MoG [24], improved MoG [25], SOBS [199], SC-SOBS [30] and ViBe [200]. In order to objectively distinguish between controlled and unconstrained scenarios, we introduced the wildness metric to measure the hardness of an environment. Considering that the performance of human detection is directly related to environment wildness, we defined the wildness of an environment as the miss rate of a person detector:

$$w = \frac{FN}{TP + FN}, \quad (3.1)$$

where FN and TP denote the number of false negatives and true positives yielded by the person detector.

In our experiments, we combined two detectors to achieve human detection: 1) the Viola-Jones detector [36], trained with human upper parts; and 2) the HOG-based person detector [38]. The detectors were combined at the decision level. For evaluation, we collected 15 surveillance videos acquired from a parking lot, along with 20 videos commonly used for BGS evaluation. The wildness metric was used to separate the test videos in two datasets: unconstrained scenarios ($w > 0.5$) and controlled scenarios ($w \leq 0.5$). Also, we performed an exhaustive search through the parameter space to find the optimal configuration of each method.

The average precision and recall were used to summarize the performance of each method and are provided in figure 3.2. The blue curves represent constant f-measure values and improve the visual perception of the overall performance of each method with respect to the others.

In short, the most important findings of our study were the following:

- BGS methods suffer from performance degradation in unconstrained scenarios when compared to controlled environments;
- Median-based methods adapt quickly to sudden changes in the scene, maintaining an acceptable recall rate. These methods have a good trade-off between the performance in image degradation factors and their general performance in wild scenarios;
- Although MoG has attained good performance in unconstrained scenarios, it is not adequate for highly dynamic environments containing non-periodic changes;

Figure 3.3: Illustration of the importance the human path prediction in the acquisition of high-resolution face images with a PTZ camera. The delay introduced by the mechanical operations performed by the PTZ camera yields incorrect acquisitions when pointing the PTZ device to the current head location. Instead, it is necessary to predict the head location for compensating this effect.

- ViBe has distinguished itself by its precision, but miss-detection of object parts represents its major drawback;
- By maintaining a good performance in the different image degradation factors and by attaining the best general performance, SOBS is the best method to address in-the-wild scenarios;
- In general, BGS methods are not robust to shadows. No algorithm has stood out in the dynamic shadows, whereas Gaussian-based methods attained the best performance in static shadows, mainly due to their high sensitivity to changes in the background;
- The best method (SOBS) attained an f-measure of approximately 81%, thereby we can conclude that BGS in-the-wild remains an open problem.

3.1.2 Human Detection/Tracking

After the background modeling phase, the detection search space was confined to foreground regions, ensuring real-time human detection. Considering the reduced number of pixels representing a person when entering the surveillance area, we opted for using a holistic approach based on gradient features (refer to section 2.1 for a comprehensive review of the detection strategies). Accordingly, we trained a cascade classifier using Haar-like features for detecting the human full-body. Next, the detections were used to instantiate a tracking algorithm. Multi-person tracking was achieved by using multiple instances of the Camshift algorithm [165] running simultaneously.

The tracking record of each subject was stored both for controlling the number of times that a person had been imaged by the PTZ camera and for inferring its position some seconds ahead. As illustrated in figure 3.3, the latter task is particularly important to counterbalance the time offset introduced by the mechanical delay of PTZ devices. For human position prediction, a regression neural network was trained using 10,000 paths automatically acquired from the tracking algorithm. At the end, this processing chain outputs a set of pixel locations at each frame, hereinafter designed as $S(t) = (x_i(t), y_i(t)),\ i \in \mathbb{N},\ t \in \mathbb{R}$, containing the coordinates of subjects face in the next t seconds.

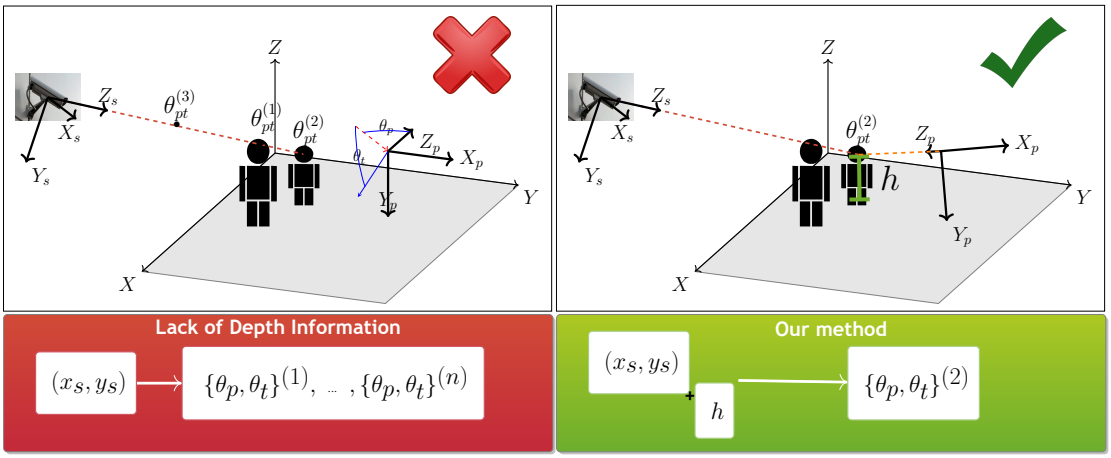


Figure 3.4: Illustration of the principal bottleneck of master-slave systems and the proposed strategy to address this problem. The same image pixel (x_s, y_s) corresponds to different 3D positions and consequently to different pan-tilt $\{\theta_p, \theta_t\}$ values. Our work is based on the premise that human height can be exploited to infer depth information and avoid that ambiguity.

3.2 Proposed Master-slave Calibration Method

As previously discussed in section 2.2, the use of a PTZ camera assisted by a wide-view camera (master-slave configuration) is the most effective strategy for acquiring biometric data at a distance in surveillance scenarios. However, as illustrated in figure 3.4, the calibration between the cameras is the major bottleneck of this approach. To address this problem, most master-slave systems use 2D-based approximations, but, in turn, they are compelled to rely on different assumptions (e.g., similar points-of-view [144], intermediate zoom states [132, 159]) to alleviate pan-tilt inaccuracies. The use of multiple optical devices has been pointed as a solution to infer depth information through triangulation [149, 160], but the highly stringent disposal of the cameras restrains its use in outdoor environments and its operating range (up to 15m). A comparative analysis between the most relevant master-slave systems is provided in section 2.2.1.2.

Considering the drawbacks of the state-of-the-art master-slave systems, we propose a calibration algorithm capable of accurately estimating pan-tilt parameters without resorting to intermediate zoom states, multiple optical devices or highly stringent configurations. Our approach exploits geometric cues, i.e., the vanishing points available in the scene, to automatically estimate subjects height and thus determine their 3D position. Furthermore, we have built on the work of Lv et al. [203] to ensure robustness against human shape variability during walking. Considering that the proposed calibration algorithm is intended to be integrated in an automated surveillance system, we have also assessed the performance of the proposed algorithm using two challenging scenarios: 1) automatic estimation of head and feet locations using a tracking algorithm; and 2) incorrect vanishing point estimation.

3.2.1 Our Method

We start by introducing the notation used in our description:

- (X, Y, Z) : the 3D world coordinates;
- (X_s, Y_s, Z_s) : the 3D coordinates in the static camera world referential;
- (X_p, Y_p, Z_p) : the 3D coordinates in the PTZ camera world referential;

- (x_s, y_s) : the 2D coordinates in the static camera image referential;
- (x_t, y_t) : the 2D coordinates of a head in the static camera image referential;
- (x_p, y_p) : the 2D coordinates in the PTZ camera image referential;
- $(\theta_p, \theta_t, \theta_z)$: the pan, tilt and zoom parameters of the PTZ camera.

In the pin-hole camera model, the projective transformation of 3D scene points onto the 2D image plane is governed by:

$$\lambda \begin{pmatrix} x_t \\ y_t \\ 1 \end{pmatrix} = \underbrace{\mathbf{K}\,[\,\mathbf{R}\,|\,\mathbf{T}\,]}_{\mathbf{P}} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \quad (3.2)$$

where λ is a scalar factor, $\mathbf{K}$ and $[\,\mathbf{R}\,|\,\mathbf{T}\,]$ represent the intrinsic and extrinsic camera matrices, which define the projection matrix $\mathbf{P}$.

Let $\mathbf{p}_t = (x_t, y_t)$. Solving equation 3.2 for (X, Y, Z) yields an under-determined system, i.e., infinite possible 3D locations for the face. As such, we propose to solve equation 3.2 by determining one of the 3D components previously.

By assuming a World Coordinate System (WCS) where the XY plane corresponds to the reference ground plane of the scene, the Z component of a subject's head corresponds to its height (h). The use of height information reduces the equation (3.2) to:

$$\lambda \begin{pmatrix} \mathbf{p}_t \\ 1 \end{pmatrix} = [\mathbf{p}_1 \quad \mathbf{p}_2 \quad h\mathbf{p}_3 + \mathbf{p}_4] \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix}, \quad (3.3)$$

where $\mathbf{p}_i$ is the set of column vectors of the projection matrix $\mathbf{P}$. As such, our algorithm works on the static camera to extract (x_t, y_t) and infer the subject position in the WCS using its height.

3.2.1.1 Height Estimation

To perform height estimation, we rely on the insight that surveillance scenarios are typically urban environments with useful geometric information that can be exploited, such as vanishing points and vanishing lines.

As in [204], three vanishing points $(\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z)$ are used for the X, Y and Z axis, in order to infer the height of a subject, which is vertical to a planar surface. $\mathbf{v}_x$ and $\mathbf{v}_y$ are determined from parallel lines contained in the reference plane, so that the line $\mathbf{l}$ defined by these points represents the plane vanishing line.

Given $\mathbf{l}$, $\mathbf{v}_z$, the head ($\mathbf{p}_t$) and feet ($\mathbf{p}_b$) points in an image, the height of a person can be obtained by:

$$h = -\frac{\|\mathbf{p}_b \times \mathbf{p}_t\|}{\alpha(\mathbf{l}.\mathbf{p}_b)\|\mathbf{v}_z \times \mathbf{p}_t\|}, \quad (3.4)$$

where $\alpha = -\|\mathbf{p}_{rb} \times \mathbf{p}_{rt}\|/(h_r(\mathbf{l}.\mathbf{p}_{rb})\|\mathbf{v}_z \times \mathbf{p}_{rt}\|)$, whereas $\mathbf{p}_{rt}$ and $\mathbf{p}_{rb}$ are the top and base points of a reference object in the image with height equal to h_r.

3.2.1.2 Pan-Tilt Angle Estimation

Considering the referential depicted in figure 3.4 and assuming $\theta_{ip} = 0$, the center of rotation of the PTZ camera is given by $C = (0, \rho \sin \theta_{it}, -\rho \cos \theta_{it})$, being θ_{it} and θ_{ip} the initial pan and tilt angles,respectively, and ρ the displacement between the mechanical rotation axis and the image plane. In general, ρ can be approximated by the camera focal distance f.

Given the 3D coordinates (X, Y, Z) of a point of interest in the WCS, the location of that point in the PTZ referential is obtained by:

$$\begin{pmatrix} X_p \\ Y_p \\ Z_p \end{pmatrix} = [\,\mathbf{R}\,|\,\mathbf{T}\,] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \quad (3.5)$$

and the corrected coordinates are given by:

$$\begin{pmatrix} X'_p \\ Y'_p \\ Z'_p \end{pmatrix} = \begin{pmatrix} X_p \\ Y_p - \rho \sin \theta_{it} \\ Z_p + \cos \theta_{it} \end{pmatrix}. \quad (3.6)$$

The corresponding pan and tilt angles are given by:

$$\theta_p = \arctan\left(\frac{X'_p}{Z'_p}\right), \quad (3.7)$$

and

$$\theta_t = \arcsin\left(\frac{Y'_p}{\sqrt{(X'_p)^2 + (Y'_p)^2 + (Z'_p)^2}}\right). \quad (3.8)$$

3.2.2 Experimental Results

To validate the proposed approach, the following procedure was adopted: given (x_s, y_s) and its corresponding (x_p, y_p) point, the algorithm error $\Delta\theta$ was determined by the angular distance between the estimated (X_p, Y_p, Z_p) and the 3D ray associated with (x_p, y_p). As compared to the typical reprojection error, this strategy was advantageous in the sense that it allowed a direct comparison with the PTZ Field of View (FOV). Additionally, the height estimation performance in surveillance scenarios was also assessed by determining the deviation Δh from the true height of the subjects.

The performance of our approach was assessed by carrying out three distinct evaluations: 1) height estimation performance; 2) independent performance analysis; 3) integration in an automated surveillance system.

In all evaluations, we used videos of ten different persons - comprising more than 1,000 frames - acquired both by the static and the PTZ camera while walking throughout an outdoor parking lot. Each pair of corresponding frames was annotated to mark the pixel location of the head and feet, in order to determine the performance of the proposed method with respect to

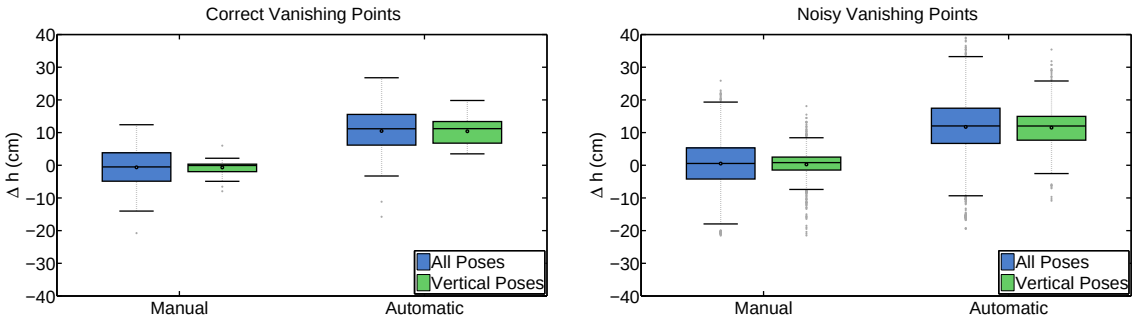


Figure 3.5: Height estimation performance in surveillance scenarios. Two distinct evaluations were carried out: an independent analysis (using manually annotated data) and the integration in an automated surveillance system (using data automatically obtained from a tracking module). Also, the accuracy of height estimation for vertical poses is presented, as well as the impact of noisy vanishing points.

Δh and $\Delta\theta$, which, in this case, corresponds to the angular distance between the estimated face location and its real position.

Besides, it is worth noting that in all evaluations a comparative analysis between inferring intrinsic and extrinsic camera parameters from Calibration Patterns (CB) and Vanishing Points (VP) was performed.

Furthermore, the inherent difficulties in accurately estimating vanishing points locations were taken into account. To assess the impact of incorrect vanishing point estimation, the previous experiments were replicated and the vanishing points location corrupted by a zero mean, Gaussian noise with standard deviation of 10px.

Finally, the feasibility of our approach in surveillance scenarios was determined by confronting $\Delta\theta$ with the PTZ FOV at different zoom magnifications (figure 3.8). The percentage of faces successfully acquired summarizes the overall performance of the proposed calibration algorithm. The attained results for the several evaluations described are presented in table 3.1 and compared with the work of Senior et al. [205].

3.2.2.1 Height Estimation Performance

The obtained results for height estimation in surveillance scenarios are presented in figure 3.5. With regard to the type of data used, the distribution of Δh evidences that, in average, automatic height estimation is accurate ($\Delta h \approx 0$ and $\sigma_{\Delta h} = 6$ cm) for manually annotated data, while, it tends to overestimate the subjects height when using automatic annotations. We believe that a more robust tracker is likely to provide closer approximations to the manual annotations. Figure 3.6 illustrates three examples of incorrect height estimation due to the output of the tracking algorithm.

Furthermore, the approximately similar distribution of the second and third quantiles for correct and noisy vanishing points suggest that in the majority of the cases the aggregated noise did not significantly affect the height estimation performance. Only strong deviations in the vanishing points - typically more than 10px (the standard deviation used in our experiments) - severely affect height estimation performance.

Finally, it is worth noting that, by building on [203], it is possible to narrow the height estimation error by relying solely on vertical poses. This result constitutes the basis of our future work, since the method accuracy may be improved by correcting height estimation on

(a) Feet and head marked manually

(b) Feet and head marked automatically

Figure 3.6: Examples of height estimation in surveillance scenarios using manually annotated data and automatic annotations obtained from a tracking algorithm. Note that the true height is 168 cm.

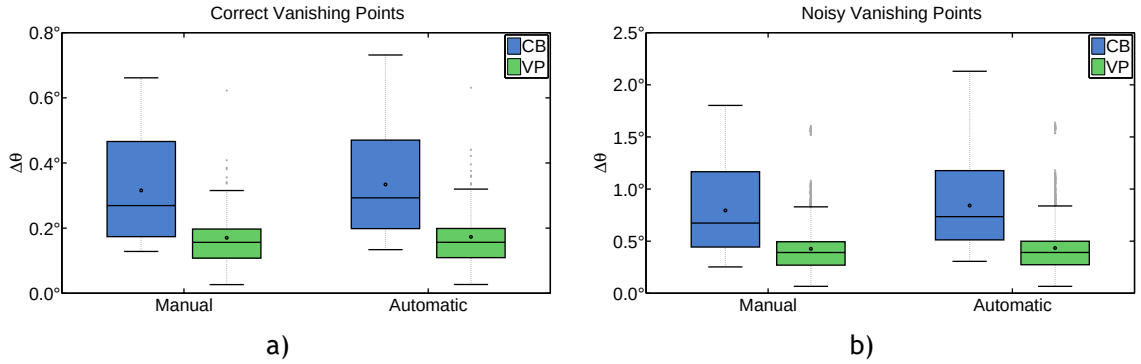


Figure 3.7: Overall performance of the proposed system. Two distinct evaluations were carried out: an independent analysis (using manually annotated data) and the integration in an automated surveillance system (using data automatically obtained from a tracking module). Additionally, two calibration strategies are compared, as well as the impact of noisy vanishing points.

non-vertical poses with information obtained from the vertical ones.

3.2.2.2 Independent Performance Evaluation

To assess the performance of the proposed calibration algorithm apart from the errors induced by the preceding phases of a surveillance system, the test videos were manually annotated, as described in section 3.2.2. The attained results are presented in figure 3.7.

Regarding the strategy used for determining the camera projection matrix, it is evident that the use of vanishing points is advantageous in surveillance scenarios. The failure of typical calibration algorithms using planar calibration patterns can be explained by the arduousness in estimating the extrinsic parameters in outdoor scenarios. Ground irregularities and the reduced size of calibration patterns, when compared to the extent of surveillance environments, are the principal factors of inaccurate estimation of the rotation and translation matrices. On the contrary, in such scenarios, vanishing points are straightforward to determine using pairs of parallel lines. Also, small inaccuracies in their estimation do not affect severely the performance of our approach (compare the differences in the average of $\Delta\theta$ when using VP), which provides additional support to the idea that a calibration based on vanishing points is preferred in surveillance scenarios.

Table 3.1: Percentage of faces successfully acquired. The performance of our method is compared for different calibration strategies (CB and VP) with [205], when using manual and automatic annotations.

Evaluation	Method	Without Noise (%)	With Noise (%)
Manual	Senior et al. [205]	30.3	-
	Our approach (CB)	58.4	57.2
	Our approach (VP)	89.4	89.1
Automatic	Senior et al. [205]	4.8	-
	Our approach (CB)	54.0	52.12
	Our approach (VP)	87.6	87.5

Finally, the overall performance of the proposed algorithm has been summarized as the percentage of faces successfully acquired. This analysis was performed by comparing $\Delta\theta$ to the PTZ FOV at a given distance and the attained results are presented in table 3.1. A comparative analysis with the results presented in [205], which also used height information to determine the 3D location of subjects, evidences a great improvement in the success rate, when considering the independent performance of the calibration module (manual data).

3.2.2.3 Integration in an Automated Surveillance System

Contrary to the previous evaluations, this experiment aims at analyzing the impact of inaccuracies yielded by a tracking algorithm. For this purpose, the test videos were provided to an adaptive background subtraction algorithm [25] to automatically obtain head and feet locations through morphological operations.

The attained results are presented in figure 3.7 and, as in section 3.2.2.2, it is clear that the use of vanishing points is advantageous in surveillance scenarios when compared to typical calibration approaches. With regard to the type of data used, it is interesting to note that the integration of the proposed algorithm in an automated surveillance system does not severely degrade the accuracy of the method (note the small differences in the average of $\Delta\theta$). Also, the same conclusion holds when comparing the use of correct and noisy vanishing points. These conclusions are also supported by the attained results presented in table 3.1. An automated surveillance system using the proposed method achieves an 87% success rate in capturing facial images at a distance using the maximum camera zoom, which outperforms the 4.8% success rate attained in [205].

In order to provide further insights about the success rate of the proposed approach with respect to the zoom magnification used, we compare the pan $\Delta\theta_p$ and tilt $\Delta\theta_t$ displacements with the PTZ FOV for different zoom magnifications in figure 3.8. Notice the extremely narrow FOV when using large zoom magnifications, and consequently, the importance of an accurate estimation of pan-tilt parameters.

3.2.3 Conclusion

In the previous sections, we described the proposed master-slave calibration algorithm to accurately estimate the pixel to pan-tilt mapping using 3D information. Our work was based on the premise that inverse projective transform was feasible if one of the 3D components was known. Accordingly, we have shown that subjects height - which can be estimated from geometric available in the scene - is a valid information to solve this problem.

The workability of the proposed system was evidenced by the following results: 1) automatic height estimation is feasible in surveillance scenarios, particularly if combined with

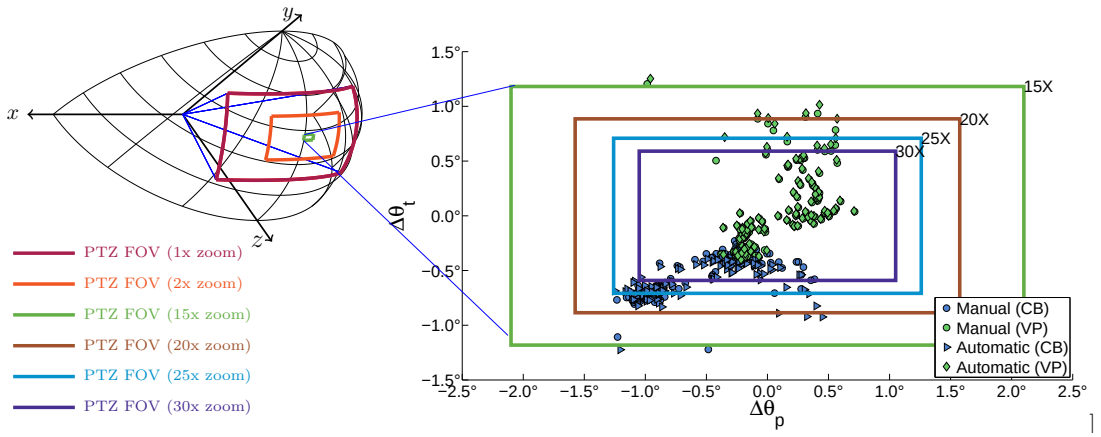


Figure 3.8: Accuracy of the proposed calibration algorithm. The pan and tilt errors are presented for two distinct calibration strategies (CB and VP). Notice the large difference between the FOV of the minimum (1x) and maximum (30x) zoom magnifications, and the importance of accurate pan-tilt estimation (at the maximum zoom the tilt error cannot exceed 0.5° to ensure a successful capture). Even so, our method has a success rate of 89% when provided with manual data.

a vertical pose filter; 2) the typical displacement between the estimated 3D position and the actual face location enables face acquisition in 89% of the cases and in 87% of the cases when integrating the calibration algorithm in automated surveillance system; 3) the system performance is not severely affected by small deviations in vanishing point locations (up to 10 px).

3.3 Proposed PTZ Scheduling Method

In real surveillance scenarios, it is quite common that the number of targets exceeds the available PTZ cameras, and the surveillance system should be capable of autonomously deciding the best sequence of observations. For this purpose, we propose a camera scheduling approach capable of determining - in real-time - the sequence of acquisitions that maximizes the number of different targets obtained, while minimizing the cumulative transition time. Our approach models the problem as an undirected graphical model Markov Random Field (MRF), which energy minimization can approximate the shortest tour to visit the maximum number of targets. A comparative analysis with the state-of-the-art camera scheduling methods evidences that our approach is able to improve the observation rate while maintaining a competitive tour time.

3.3.1 Camera Scheduling Methods

Camera scheduling in PTZ-based systems can be broadly divided in coverage and saccade approaches. In the former, the cameras are set in an intermediate zoom state so that multiple targets are observable by the same device. The goal is to maximise the number of targets seen by the complete set of cameras [136, 206, 207].

On the contrary, in a saccade approach each camera just observes one target at a time. A sequence of saccades is planned, in real-time, to maximize the number of different targets observed and minimize the cumulative transition time. Some works have presented solutions to variants of this problem [208], but Costello et al. [209] were the first to explicitly define and propose a solution to this problem. Considering the similarities with the network packet routing

problem, the authors proposed the use of the current minloss throughput optimal to schedule a set of observations. Targets weights were determined by their residual time to exit the scene and the observation sequence was constructed by minimizing the expected weighted loss, i.e the sum of targets weights not observed. Bimbo and Pernici [210] addressed the problem by modeling it as the kinetic traveling salesman problem, an extension of the classical traveling salesman problem where the cities positions change over time. However, this problem has not a known solution that runs in polynomial time, which restrains its use in real-time scenarios. To address this issue the kinetic traveling salesman problem is solved, by exhaustive search, for the six targets with the shortest deadlines. A similar strategy was used in [211], where a greedy best-first search was employed to determine the optimal plan. Qureshi and Terzopoulos [212] relied on greedy algorithms such as the shortest elapsed time first and weighted round robin. The weighted round robin is able to efficiently distribute targets to different cameras, however, at each camera, the waiting list was scheduled based on a multi-class First-come, First-served (FCFS) policy, i.e. the class was determined by the number of times the person had been imaged. In [139] the best-first heuristic was advocated as a good approximation to dynamically estimate new observation plans. Targets were modeled as graph nodes and transition costs were defined according to their distance to the camera and expected time to exit the scene. Lim et al. [213] constructed a directed acyclic graph based on the starting time of 'task visibility intervals', which were inferred by prediction. The scheduling problem was formulated as a maximal flow problem and a dynamic programming scheme was proposed to solve it. Ilie and Welch [214] relied on a greedy algorithm to determine a plan based on geometric reasoning.

3.3.2 Our Method

As figure 3.9 illustrates, the proposed model is composed of N vertices, which represent the position of each target in the sequence of saccades. Also, each vertex can be assigned to N different labels, corresponding to the N targets in the scene. This structure allows to determine the order that each target will be observed by taking into account both the temporal constraints (vertex information) and the transition costs (pairwise relations between vertices).

Let $G = (V, E)$ be a graph representing a MRF, composed of a set of t_v vertices V, linked by t_e edges E. The MRF is a representation of a discrete latent random variable $L = \{L_i\}, \forall i \in V$, where each element L_i takes one value l_u from a set of labels.

In this problem, a MRF configuration $l = \{l_1, ..., l_{t_v}\}$, determines an acquisition sequence of N targets. Besides, we define G to be a complete graph, whose edges encode the cost of assigning the target l_u to the i^{th} position and the the target l_v to the j^{th} position. The edges between consecutive vertices correspond to the transition cost of moving the camera from the target u to v, whereas the edges of non-consecutive vertices are used to avoid repetitions in the sequence of observations. The energy of a configuration l of the MRF is the sum of the unary $\theta_i(l_u)$ and pairwise $\theta_{i,j}(l_u, l_v)$ potentials:

$$E(l) = \sum_{i \in \mathcal{V}} \theta_i(l_u) + \sum_{(i,j) \in \mathcal{E}} \theta_{i,j}(l_u, l_v). \quad (3.9)$$

According to this formulation, determining the best tour is equivalent to infer the random

variables in the MRF by minimizing its energy:

$$\hat{l} = \arg\min_{l} E(l), \quad (3.10)$$

where $\hat{l_1}, ..., \hat{l_{tp}}$ are the targets index. As an example, if four targets are considered, the configuration $\{2, 3, 1, 4\}$ determines p_2 as the first subject to be visited, p_3 as the second, and so on.

The MRF was optimized according to the Loopy Belief Propagation [215] algorithm. Even though it is not guaranteed to converge to global minimums on loopy non-submodular graphs (such as our MRF), we concluded that the algorithm provides good approximations (refer to section 3.3.3).

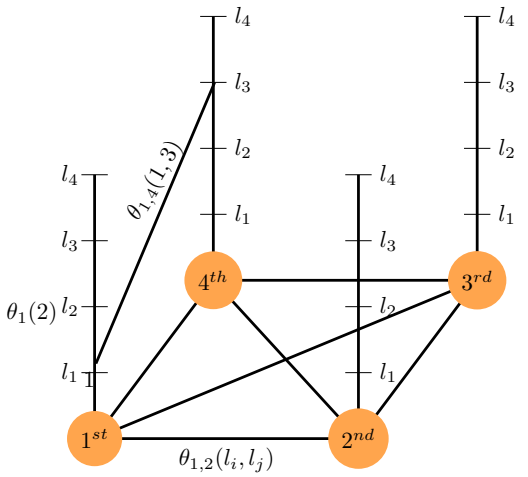


Figure 3.9: Illustrative example of the MRF used in our approach when four targets are in the scene. Labels encode the set of targets in the scene, whereas the nodes correspond to the order that they will be imaged.

3.3.2.1 Unary and Pairwise Potentials

We first define the notation used to describe the proposed approach.

- $p_u(t) = (x_u(t), y_u(t), z_u(t))$: the 3D position of the u^{th} target at time t;
- $\alpha(p_u)$: the pan angle corresponding to the cartesian coordinates of p_u;
- $\beta(p_u)$: the tilt angle corresponding to the cartesian coordinates of p_u;
- $\Lambda_u(t)$: expected time to target p_u leave the scene;
- τ : average time required to acquire a target.

In this problem the unary costs of the first vertex have been modelled as the transition cost to move the camera from the actual position to each target. Besides, targets deadline (Λ_i) is also taken into account by greatly penalizing sequences with $\Lambda_i(t) < \varepsilon$ in last vertices:

$$\theta_i(l_u) = \begin{cases} \mathcal{K}\Big(\alpha(C) - \alpha(p_u(t)), \beta(C) - \beta(p_u(t))\Big), & \text{if } \Lambda_i(t) > \varepsilon, \\ 0, & \Lambda_i(t) < \varepsilon \text{ and } i = 1, \\ \infty, & \text{otherwise}, \end{cases} \quad (3.11)$$

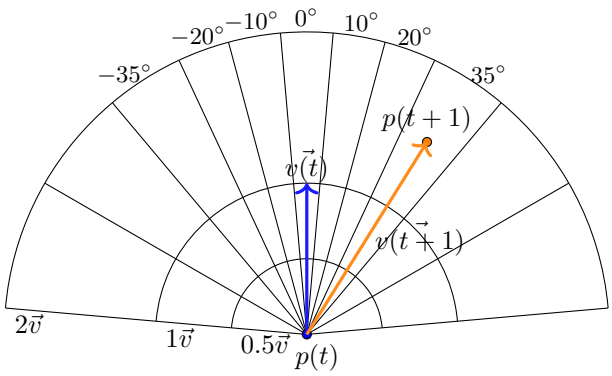


Figure 3.10: Illustration of the discrete grid used to model human transitions with respect to angular direction and velocity module. Adapted from [216].

where C is the ground cartesian coordinate to which the camera is pointing, whereas $\mathcal{K} : (\alpha, \beta) \rightarrow \Delta$ is a camera dependent function that determines the consumed time Δ to change pan and tilt values by α and β, respectively. The pairwise potential between two adjacent vertices $\theta_{i,j}(l_u, l_v)$ is defined as the time required to point the camera to p_v assuming that is pointing to p_u:

$$\theta_{i,j}(l_u, l_v) = \begin{cases} \mathcal{K}(a,b), & \text{if } u \neq v \text{ and } c(u,v) = 1, \\ 0, & \text{if } u \neq v \text{ and } c(u,v) = 0, \\ \infty, & \text{otherwise,} \end{cases} \quad (3.12)$$

where $a = \alpha(p_u(t + \tau * i)) - \alpha(p_v(t + \tau * j))$ and $b = \beta(p_u(t + \tau * i)) - \beta(p_v(t + \tau * j))$. The logical function c determines if u and v are two consecutive vertices. Besides, the estimation of $p_u(t + \tau * i)$ is attained by predicting targets position using a constant velocity model.

3.3.2.2 Virtual Path Generation

The assessment of camera scheduling performance can be carried out using two distinct strategies: 1) integration in a running automated surveillance system; 2) performing an independent evaluation using pre-acquired human walks from the tracking module of a calibrated camera. In the former case, the results may be misleading, since it is difficult to separate the performance of the control module from the overall system. On the other hand, relying on pre-acquired human walks greatly limits the number of available paths. The use of randomly generated walks can overcome dataset size limitations, but it is highly prone to generate non-plausible paths.

Consequently, we used a virtual human walk generator to perform an independent evaluation of camera scheduling algorithms. Rather than assume constant direction and velocity model [210] - which restricts data variability - we built on the works [216, 217] to generate a virtually unlimited number of synthetic human walks.

Let $(x(t), y(t), z(t))$ be the position of a target p at the time t and $v(t)$ the velocity vector, targets movement is discretized into a grid with eleven possibilities in angular direction and three possibilities in acceleration ratio, as illustrated in figure 3.10. Path generation is performed by iteratively sampling the $P(\theta, r)$ distribution to determine $p(t + 1)$. In order to capture the typical behavior of humans in surveillance scenarios, the distribution P is inferred

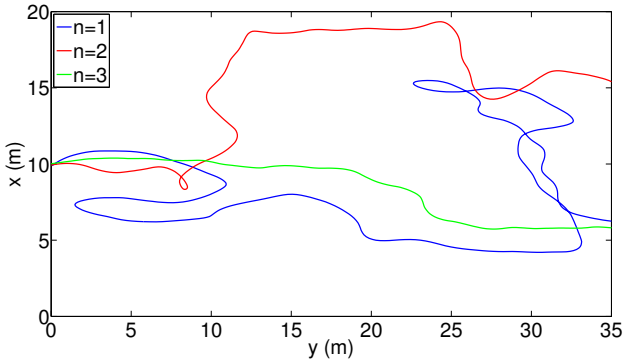


Figure 3.11: Examples of three virtual paths generated using the conditional distribution $p(\{\theta_t, r_t\}|(\{\theta_{t-1}, r_{t-1}\}, ..., \{\theta_{t-n}, r_{t-n}\})$ for different values of n. Note the increasingly linear shape of paths with respect to n.

from a set pre-acquired human walks. Additionally, we adopt the 'toward destination' behavior - described in [217] - by dynamically re-weighting P with respect to the desired destination.

However, this strategy is memoryless, i.e., it does take in account the previous (θ, r) transitions to decide the next state, which, again, may yield non-plausible paths. To address this issue we rely on the conditional distribution $p(\{\theta_t, r_t\}|(\{\theta_{t-1}, r_{t-1}\}, ..., \{\theta_{t-n}, r_{t-n}\})$.

In our experiments, we have acquired ten paths from ten persons walking through a parking lot of 20 m by 40 m at ten frames per second - corresponding to more than 30,000 human path positions - to infer $P(\theta, r)$.

Figure 3.11 illustrates the effect of n on path irregularities, such as small loops. Even though higher values could improve path reality, it would also require a higher number of training data to accurately infer the distribution p. As such, we use $n = 3$ in the evaluation of the camera scheduling algorithms.

3.3.3 Experimental Results

In this section, we evaluate the proposed approach using a virtual simulation. To replicate the conditions of a common surveillance scenario, the scene size used was similar to a typical parking lot (20 x 40 m) and the camera was assumed to be located at (0,0,5m). Also, targets paths were generated using the method described in section 3.3.2.2 and their initial positions were randomly selected. A Hikvision DS-2DF PTZ camera was used to estimate the function $\mathcal{K}$ in a similar fashion as in [210]. All experiments were performed in an Intel Core i7-2700K @ 3.50GHz.

Considering that we were interested in evaluating the time required to observe all the targets and the number of targets successfully acquired, the simulation S was defined as $S\colon P \rightarrow \{\Gamma, \Theta\}$, where $P = \{p_1(t), p_2(t), ..., p_n(t)\}$ defines the targets positions with respect to time, $\Gamma = \{t_1, t_2, ..., t_k\}$ defines the consumed time during the k^{th} acquisition tour when all the targets were in the scene, and $\Theta = \{c_1, c_2, ..., c_k\}$ the number of targets successfully acquired in each acquisition tour. To avoid disparate values of k in the same simulation for different algorithms, we opted to restrain the simulation to a single tour, i.e. $k = 1$. Algorithm 1 presents the pseudocode of the proposed simulation.

Our approach was compared to typical schedule routines adopted in [209, 212, 218], namely the FCFS and the Earliest Deadline First (EDF). Moreover, a comparison with the works [210] and [139], hereinafter designated as TDO and Krahn et al., was also performed.

Data: P, camera schedule algorithm F, service time s
Result: t_1,c_1
t=0;
t_1=0,c_1=0;
waitList=$\{1, 2, ..., \#P\}$;
while *!isempty(waitList)* **do**
  select next target $p = F(P)$;
  compute transition cost Δ;
  remove p from waitList;
  if $\Lambda_a(t) < \varepsilon$ **then**
    $t_1 : t_1 + \Delta + s$;
    $c_1 : c_1 + 1$;
  end
end

Algorithm 1: Pseudocode for the simulation used to evaluate the performance of camera scheduling approaches.

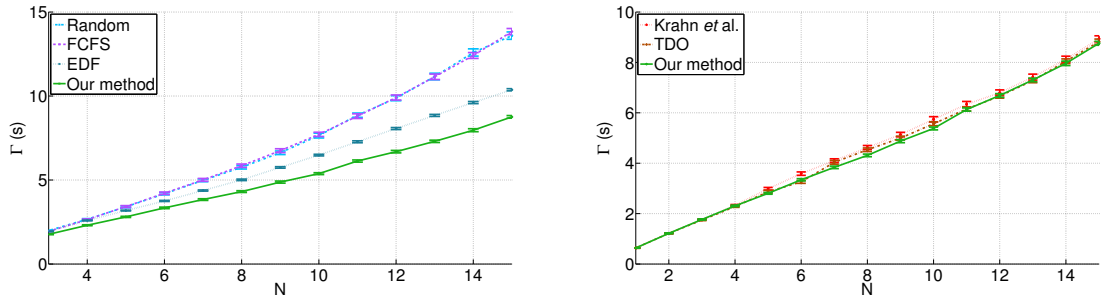


Figure 3.12: Comparative analysis of the consumed time (Γ) required to observe N persons in the scene. a) Our approach is compared with common scheduling routines previously used in PTZ-based systems [209, 212, 218]. b) The comparison with the most competitive state-of-the-art methods is presented separately for visualization purposes.

3.3.3.1 Cost Time

The analysis of Γ with respect to N furnishes insight about the algorithms efficiency to acquire a set of N targets. Figure 3.12 depicts the results attained using 100 simulations for up to 15 targets. Regarding the comparison with naive schedule approaches - figure 3.12a) - it is evident that the MRF-based algorithm can acquire a set of N persons faster, allowing the camera to repeat the acquisition sooner and thus collect more pictures. When considering the comparison with the work of Bimbo and Pernici [210] - figure 3.12b) - it is worth noting that our approach is unable to improve TDO results up to $N = 6$. This is explained by the six element queue used to prioritize targets with the shortest deadlines and the use of an exhaustive search to determine the best solution for this subset. As compared to the algorithm of Krahnstoever et al. [139], the improvements can be explained by the assumption that the best target is the one with the lowest transition cost. Even though this solution can provide good approximations, it can be improved by taking into account the positions of the remaining targets as performed in our MRF model.

3.3.3.2 Observation Rate

Additionally, we have also evaluated the average observation rate ($\frac{\Theta}{N}$) with respect to the number N of targets in the scene. The results presented in figure 3.13 clearly evidence an

improvement in the number of successfully observed targets as compared to the most competitive alternatives regarding the Γ performance. This difference can be explained by the fact that the remaining approaches are mainly concerned with the minimization of tour cost.

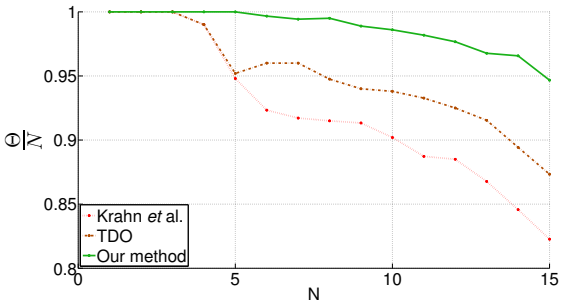


Figure 3.13: Comparative analysis of the average observation rate of the proposed algorithm with the most competitive alternatives regarding the Γ performance.

3.3.3.3 Run-time Analysis

Considering the real-time requirement of the camera scheduling problem, we have estimated the average speed of the proposed algorithm with respect to the number of targets in the scene. For this purpose, 100 simulations were used to estimate the average running time for up to twenty targets, as illustrated in figure 3.14. Our approach is capable of planning a sequence of saccades in less than 30 ms for up to 15 targets, which is residual when compared to the average time (600 ms) that the PTZ camera takes to move and acquire a shot of a target. Moreover, it is worth noting that for $N = 15$, the proposed algorithm is 10^8 times faster than an exhaustive search.

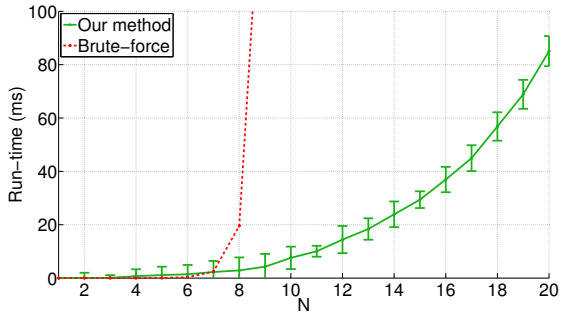


Figure 3.14: Average running time.

3.3.4 Conclusion

In the latter sections, we were concerned about the time costs of dynamic camera scheduling algorithms, which are prohibitive for crowded scenes, i.e., with over 15 subjects in the scene. Accordingly, we modeled the dynamic camera scheduling problem using a MRF model. By denoting each vertex as the position of a target in the sequence plan, our approach can take into account temporal constraints (targets deadlines) and transitions costs between consecutive vertices. The energy minimization of the MRF model yields - in real-time - a tour to acquire the maximum number of different targets while minimizing the total travel time.

Additionally, a realistic virtual simulation was proposed to assess the performance of camera scheduling algorithms. The use of realistic human walk generator - trained from real human paths - permitted to overcome dataset size constraints while maintaining the plausibility of human walks. A comparative performance analysis with state-of-the-art approaches evidences that the proposed model is able to improve the observation rate while maintaining a competitive tour time.

3.4 Summary

In this chapter, we described the necessary modules for the development of an automated acquisition system capable of acquiring high-resolution biometric data at a distance from non-cooperative subjects, the QUIS-CAMPI system.

Having concluded that the master-slave architecture is the most practical configuration for ensuring that the acquired biometric samples have a sufficient resolution to be used for recognition purposes, we focused our research on developing new strategies to address the drawbacks of this architecture.

Accordingly, we proposed a master-slave calibration algorithm capable of accurately estimating the correspondence between the master camera coordinates and the orientation angle of the PTZ camera. When compared to the state-of-the-art master-slave systems, this proposal is advantageous because it avoids the use of extra optical devices for ensuring an accurate mapping. Besides, our approach does not depend on stringent configurations, which is beneficial for increasing the working distance of the acquisition system.

The other proposal concerns the time costs of PTZ camera scheduling algorithms, which are prohibitive for crowded scenes, i.e., with over 15 subjects in the scene. We described a method to determine - in real-time - the tour that maximizes the number of different targets observed and minimizes the total travel time.

In short, it is our belief that the proposals described in this chapter contribute to improve the workability of biometric data acquisition systems, not only by extending the stand-off distance of the system, but also by increasing the number of biometric samples acquired.

Chapter 4

The QUIS-CAMPI Data Feed

Biometric datasets are an important asset to push forward the state-of-the-art recognition performance. As an example, we highlight the evolution of face recognition datasets, which have moved towards more challenging conditions (e.g., the Labeled Faces in the Wild (LFW) [180]), as novel algorithms surpass the challenges of the hardest sets. The LFW dataset has paved the way for biometric recognition in the wild, and fostered the development of even more challenging datasets. Nonetheless, as noted by Klare *et* al. [219], one explanation for unconstrained face recognition being still far from solved is that the LFW and similar datasets are not fully unconstrained. To close this gap, Klare *et* al. [219] introduced the IJB-A dataset, which follows the spirit of LFW but includes high variability in pose. However, even this challenging dataset does not encompass the complete set of covariate factors present in real surveillance scenarios, as the majority of the images were not acquired on the move and in an automated manner. For this reason, the levels of blur caused either by motion or incorrect focusing are reduced.

Considering that none of the existing datasets is suitable for the evaluation of biometric recognition in surveillance scenarios, we propose the QUIS-CAMPI data feed, whose acronym derives from Latin and summarizes its goals: 'Quis' stands for 'Who is' and 'Campi' refers to a delimited space. Hence, this set aims at fostering the development of biometric recognition systems that work outdoors, in fully unconstrained and covert conditions. To this end, we relied on the QUIS-CAMPI system (described in section 3) to capture both full body video sequences and high-resolution face images of subjects in a parking lot. The particularities of the surveillance system permit the continuous acquisition of novel biometric samples that are supplied to the dataset after being manually screened and associated to the corresponding gallery subjects.

The remainder of this chapter is organized as follows: section 4.1 overviews the datasets for assessing the recognition performance in these environments. A detailed description of the proposed dataset is given in section 4.2. Section 4.3 describes the evaluation protocol and compares the results attained by state-of-the-art face recognition algorithms in the QUIS-CAMPI and LFW datasets, and the major conclusions are summarized in section 4.4.

4.1 Biometric Datasets

About 25 years ago, biometric recognition emerged as an interesting topic, leading to the development of many novel algorithms, usually validated in small, non-representative and proprietary databases, according to distinct evaluation protocols. To meet the growing demands for objective evaluation tools, sets of biometric samples comprising different covariate factors were introduced as a solution. The ORL database of faces [233], the AR face database [234] and the Yale face database were pioneer sets on face recognition, while FERET [235] was the first benchmark on this topic. Despite their valuable contribution in providing objective and trustworthy tools for assessing recognition performance, these sets soon became outdated as novel algorithms reported almost ideal accuracy on these data.

Table 4.1: Comparative analysis between the datasets particularly devised for studying unconstrained biometric recognition. Datasets are compared with respect to the number of subjects available and the key covariate factors of recognition in the wild: expression (E), occlusion (O), illumination (I), pose (P), motion-blur (M) and out-of-focus (F). Also, the key aspects to ensure that the data realistic result from real-world scenarios are also included. The abbreviations of these aspects refer to non-cooperative (NC), on the move (OM), at a distance (AD), outdoor (OU) and automated image acquisition (AA).

	Number of subjects	Covariate Factors	NC	OM	AD	OU	AA	Observations
XM2VTS [220]								
BANCA [221]	26	E, I, P		✓				A database of face videos comprising twelve recordings per subject were acquired under controlled, uncontrolled and adverse conditions.
FRGC [222]								
FRVT 2006 [223]	$>$ 35 000	E, I				✓		The first independent performance benchmark for 3D face recognition technology. Also, it comprises still frontal face images acquired under controlled and uncontrolled illumination.
GBU [224]								
MBGC [225]	570 (still) 147 (video)	E, I				✓		This set was used to promote two distinct face recognition challenges: 1) the still face challenge problem, comprising frontal and off angle still face images taken under uncontrolled indoor and outdoor lighting; 2) the video challenge problem containing videos acquired in unconstrained environments.
SC-FACE [226]								
LDHF-DB [227]	100	I, F			✓	✓		This set comprises both visible and near-infrared face images at distances of 60m, 100m, and 150m acquired outdoors.
LFW [180]								
PubFig [228]	200	E, O, I, P	✓			✓		Similar in spirit to the LFW dataset, but the number of images per person is higher.
FaceScrub [229]								
IJB-A [219]	500	E, O, I, P,M	✓			✓		Similar in spirit to the LFW dataset, but containing high-variability in pose.
YouTube Faces [230]								
Choke Point [231]	25	E, O, I, P, M	✓	✓			✓	A database of face videos acquired indoors in a non-cooperative manner.
PaSC [232]								
QUIS-CAMPI	268 (v1) 320 (v2)	E, O, I, P, M, F	✓	✓	✓	✓	✓	The first data feed of biometric samples automatically acquired by an outdoor surveillance system, with subjects on the move and at a distance.

These improvements fostered the development of more challenging datasets, such as the CMU PIE [236], the Multi-PIE [237], the XM2VTS [220] and the BANCA [221] databases, comprising biometric samples with significant variations in illumination, pose and expression. Also, different challenges were introduced for assessing the accuracy of state-of-the-art face recognition methods in less constrained scenarios (e.g., the Face Recognition Grand Challenge [222] and the Face Recognition Vendor Test (FRVT) 2006)).

Aiming at providing more realistic data, the research has advanced towards the acquisition of unconstrained samples along the diverse biometric traits, such as iris [238], periocular [239, 240] and face [241]. Regarding face recognition, LFW was the first database particularly devised for studying face verification in the wild and was, therefore, responsible for promoting the development of more robust algorithms (an increment of 10% in the recognition accuracy in last years), as well as for fostering the emergence of more challenging collections of data (e.g., PubFig [228], FaceScrub [229], IJB-A [219] and Disguise and Makeup Faces Database [242]).

Simultaneously, still to video modality has also gained increasing attention leading to the development of novel datasets and biometric challenges on this topic. The video challenge portion of the Multiple Biometrics Grand Challenge (MBGC) contained subjects walking towards the camera and non-frontal footage of subjects performing an activity. Later, the Point and Shoot Face Recognition Challenge (PaSC) was introduced, comprising unconstrained video sequences of subjects performing multiple activities outdoors. The YouTube Faces database [230] contains unconstrained recordings obtained from the internet, and it was particularly designed for studying the problem of unconstrained face recognition in videos. On contrast, the SC-FACE [226] and the ChokePoint [231] datasets were originally intended to provide data acquired in realistic indoor surveillance scenarios. Regarding surveillance scenarios, the PETS [243], i-LIDS [244], CAVIAR [245] datasets and the VISOR [246] repository comprise video sequences of pedestrians in realistic surveillance scenarios. Even though the low resolution of data inhibits its use for face recognition purposes, it has been showed that the fusion of face with gait information can significantly increase the recognition performance [247,248]. However, these results have been obtained in gait datatsets where subjects walk in a cooperative and predefined pose [249,250].

A comparative analysis between state-of-the-art databases concerning unconstrained face recognition is given in table 5.1. It is interesting to note that despite these sets comprise highly challenging biometric data, most of them were manually captured by human operators and they still lack several crucial covariate factors of surveillance environments, such as motion-blur.

4.2 Description of the QUIS-CAMPI Dataset

When planning the QUIS-CAMPI dataset, we had two main concerns: 1) to acquire biometric data of subjects in a real surveillance scenario, covertly, on the move and at a distance; and 2) to provide multiple biometric enrollment data to perceive the advantages of using media collection [251] for identifying humans in the wild. For that purpose, we collected multiple biometric samples that where organized into two distinct groups: 1) enrollment data; 2) probe data. In the former, we have enrolled volunteers who have provided written authorization for image acquisition and distribution. The enrollment was conducted in an indoor controlled scenario and the following samples were collected: soft biometrics, full-body imagery, gait sequences and a 3D face model. Before the enrollment, each subject signed the informed consent provided in appendix A to provide written authorization for the acquisition of enrollment and probe data. In the latter, the data were acquired by the system described in chapter 3,

Figure 4.1: Illustrative example of the biometric data available in QUIS-CAMPI. For each subject of the database distinct biometric traits are acquired during enrollment, comprising soft biometrics, full body imagery (a), gait sequences (b) and a 3D model (c). Subsequently, non-cooperative biometric data are automatically collected each time a subject enters the surveillance area, comprising high-resolution face images (d), gait sequences and their corresponding foreground (e). Note that these data are acquired under varying lighting and weather conditions, at different times of the day, while subjects are on the move and at a distance.

Table 4.2: List of the soft biometric traits collected during enrollment.

Trait	Labels
Age	$\mathbb{N}$
Height	$\mathbb{N}$
Weight	$\mathbb{N}$
Sex	Male, Female
Ethnicity	Caucasian, African, Hispanic, Asian, Indian
Skin Color	White, Tanned, Oriental, Black
Hair Color	None, Black, Brown, Red, Blond, Grey, Dyed
Hair Length	None, Shaven, Short, Medium, Long
Facial Hair Color	None, Black, Brown, Red, Blond, Grey
Facial Hair Length	None, Stubble, Mustache, Goatee, Full beard
Hair Style	None, Straight, Curly, Wavy, Frizzy

while subjects walked throughout a surveillance area. Probe samples comprise high-resolution face images automatically captured by the PTZ camera and the corresponding gait sequences recorded by the master camera. It is important to note that the large majority of subjects use this area in their normal routine, which ensures a faithful representation of surveillance covariates.

Figure 4.1 illustrates the biometric data available for each subject.

4.2.1 Enrollment Data

Enrollment data provide good quality samples acquired indoor: soft biometrics, full-body imagery, gait sequences and a 3D face model.

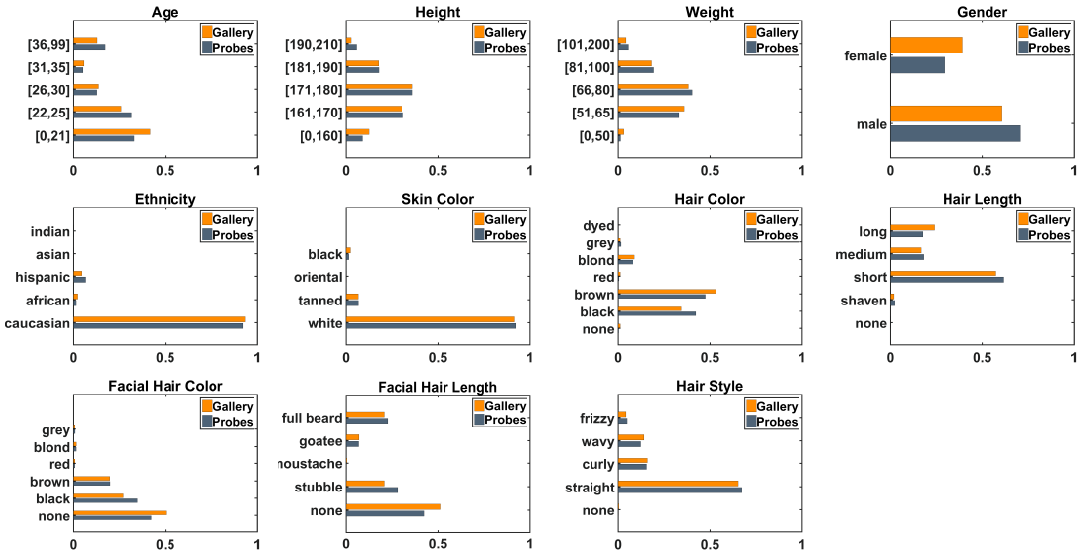


Figure 4.2: QUIS-CAMPI statistics. A set of statistics was collected for distinct types of biometric samples: 1) distribution of the soft traits along the enrollment data (denoted as gallery) and probe data; 2) distribution of the interpupillary distance in the probe images; 3) distribution of the tracking sequences width collected outdoors; and 4) distribution of the number of days elapsed between the acquisition of probe data and the enrollment process.

Soft biometrics. Eleven types of soft biometric labels were registered for each subject. The full list is presented in table 4.2 and the rationale behind the choice of these features was their discernibility at a distance and the discrimination power reported in the study of Tome et al. [252]. The distribution of each trait with respect to the labels adopted is depicted in figure 4.2.

Full-body shots. A high-resolution image of the person body was acquired at three different angles (frontal, left-side and right-side). Also, the intrinsic and extrinsic parameters of the camera were registered, along with five keypoints of the body in the frontal view. These data can be used to infer real-world measurements of body components (e.g., height and face metrology).

Gait sequences. Persons were asked to walk naturally during 5m while being filmed at three different angles. This was performed twice to obtain six different viewing angles of the gait sequence.

3D face model. A set of images acquired at different viewing angles was used to construct a textured 3D model of face using Visual-SFM [253].

4.2.2 Probe Data

Fully unconstrained biometric samples are the key novelty of the QUIS-CAMPI dataset and comprehend two main components: 1) face images automatically acquired by the PTZ camera and 2) gait sequences. Additionally, the foreground regions, obtained from a background subtraction algorithm, are also recorded.

High-resolution facial shots. The master-slave surveillance system described in chapter 3 is used to automatically acquire high-resolution face images of the enrolled subjects, while they walk throughout the surveillance area. Considering that not all the acquired data contain the facial region (e.g., incorrect human detection or tracking) and that the surveillance rig is not able to distinguish between enrolled and non-enrolled subjects, the data are manually screened before being supplied to the database. Also, the face location of the subject of interest in the image is provided as metadata. These annotations are determined by a state-of-the-art face detection algorithm [254] and cross-verified manually. On average, the interpupillary distance of a face image is 116 px with a standard deviation of 35 px, and about 99% of the images have an interpupillary distance higher than 60 px (the minimum resolution required for commercial face recognition engines).

Tracking Sequences. A set of videos automatically acquired by the master camera while the person is passing in the surveillance area. On average, each sequence has a resolution of 73 x 114 px with a standard deviation of 43 px for both dimensions.

Background Subtraction Sequences. The output of the background subtraction algorithm for each tracking sequence.

4.2.3 Database Versioning

The automated acquisition of biometric samples and their regular deployment to the dataset is the reason for denoting QUIS-CAMPI as a data-feed and at the same it is one of the key novelties of this tool. Moreover, this singularity is the rationale to argue that QUIS-CAMPI is the first *open* dataset, which is particularly advantageous to avoid inappropriately fitting classifiers to the final test data. Despite the advantages of this choice, it also introduces significant challenges that have to be carefully addressed to ensure that the performance reported in this dataset can be compared in a practical and fair manner.

To this end, we relied on *git* - one of the most commonly used version control systems - to organize the QUIS-CAMPI data feed in two distinct types of branches: 1) the master branch comprises the most updated version of the entire biometric data; 2) the evaluation branches encompass a former snapshot of the master branch plus the evaluation files defined according to the evaluation protocol of QUIS-CAMPI (refer to section 4.3.1). This structure is depicted in figure 4.3, where the advantages of this strategy can be easily perceived. First, the version control capabilities allow users to navigate through any state of the QUIS-CAMPI data feed using the master branch, which is useful to obtain new biometric samples without the burden of redownloading the entire set. Second, the evaluation branches are static and independent of any updates on master branch, allowing researches to compare their approaches by referring to a specific evaluation branch.

4.2.4 Database Availability

Regarding the dataset structure, the file names correspond to the acquisition date in the following format: Y<*a*>M<*b*>D<*c*>h<*d*>m<*e*>s<*f*>, where *a*, *b*, *c*, *d*, *e*, and *f* denote the acquisition year, month, day, hour, minute and second, respectively. The correspondences between the files and enrolled subjects, as well as the soft biometric traits, are provided in a relational database, which is deployed as a backup SQL file. For convenience, we include a view in the database

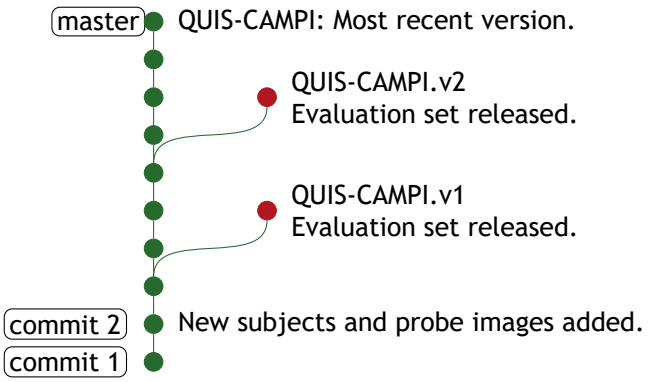


Figure 4.3: History graph of the QUIS-CAMPI data feed using a version control software. A *git* repository was used to deploy new samples acquired by the data feed (represented by the master branch), while maintaining static evaluation sets released at much lower rate (represented by branches). This strategy permits researchers to access any state of the QUIS-CAMPI data feed for development purposes, while it also ensures that algorithms can be compared by reporting performance on the different evaluation set versions.

that eases the access to biometric data using simple SQL queries. For additional information on how to get and use the dataset, please refer to the QUIS-CAMPI web site [1].

4.3 Experimental Evaluation

In this section, we introduce the evaluation protocols that should be adopted for reporting the algorithms performance in the QUIS-CAMPI set. We believe that the proposed guidelines for the different recognition modalities are adequate for the majority of biometric recognition algorithms. However, as in the recent case of the updated guidelines of LFW [241], additional protocols may be included in the future to meet novel requirements.

Table 4.3: Description of the performance metrics adopted for the different evaluation paradigms.

Paradigm	Setting	Performance Metric
Verification	Image-restricted	ROC plot AUC ACC
	Unrestricted	ROC plot AUC ACC
Identification	Closed-set	CMC plot AUC Rank-1 accuracy
	Open-set	ROC plot AUC ACC

4.3.1 Evaluation Protocol

Having in mind the main purpose of QUIS-CAMPI, i.e., to provide an objective tool for assessing the performance of biometric recognition algorithms in surveillance scenarios, we

[1] http://quiscampi.di.ubi.pt

Table 4.4: Description of the QUIS-CAMPI evaluation protocols under the verification paradigm. As in LFW [241], the protocols are defined according to the use of the image-restricted or unrestricted setting, and to the use of additional training data.

	Unrestricted, No Outside Data	Image-restricted, Label Free Outside Data	Unrestricted, Label Free Outside Data	Unrestricted, With QUIS-CAMPI Biometric Data	Unrestricted, With Labelled Outside Data
Identity information in training images					
Annotations in training images					
External images					
Binary label for external image pairs					
Identity label for external images					

introduce two evaluation protocols for the two recognition modalities: 1) verification and 2) identification.

4.3.1.1 Verification

Regarding the verification paradigm, we adopt the protocol defined in LFW [180, 241], which is an objective, simple and well established way of assessing face verification algorithms. Accordingly, the PTZ face images are used to form pairs of matched images (positive pairs) and mismatched images (negative pairs) organized in two groups: 1) model selection and algorithm development; and 2) performance reporting. In the first group, random pairs are used as training (2200 pairs) and test (1000 pairs) sets. This group is particularly intended for tuning algorithm parameters, and thus, preventing the bias introduced by adjusting the method to the final evaluation set. In the second group, 10 splits - containing 300 positive and negative pairs of PTZ face images - are built for evaluating algorithms performance using leave-one out validation. In the training phase, two distinct paradigms are available: image-restricted, where only the training split pairs can be used, and unrestricted, where identity information is provided, which allows forming additional training pairs. In addition, one can increase the algorithms robustness by exploiting metadata from outside of QUIS-CAMPI or external training images. The ensemble of training paradigms for face verification in the QUIS-CAMPI dataset is listed in table 4.4. During the test phase, algorithms must be evaluated using mean classification accuracy (ACC) over the 10 splits. If possible, the ROC curve and its corresponding Area Under Curve (AUC) should be also reported. The metrics that shall be adopted under the different recognition paradigms are described in table 4.3.

4.3.1.2 Identification

Under the identification paradigm, a probe image must be matched against all gallery subjects. This task can be approached in two distinct manners: 1) assuming that all probe images correspond to one subject in the gallery (closed-set recognition); and 2) assuming that not all the probe identities are represented in the gallery (open-set recognition).

Probe data comprise the PTZ face images, while the frontal mugshots acquired during enrollment are used as gallery. Probes are divided in two mutually exclusive sets: 1) training (containing 70% of the subjects); and 2) test (containing 30% of the subjects). Besides, two distinct sets of gallery subjects are provided. One comprising all the available subjects of the dataset (closed-set) and the other containing just 80% of the subjects. At last, in order to avoid

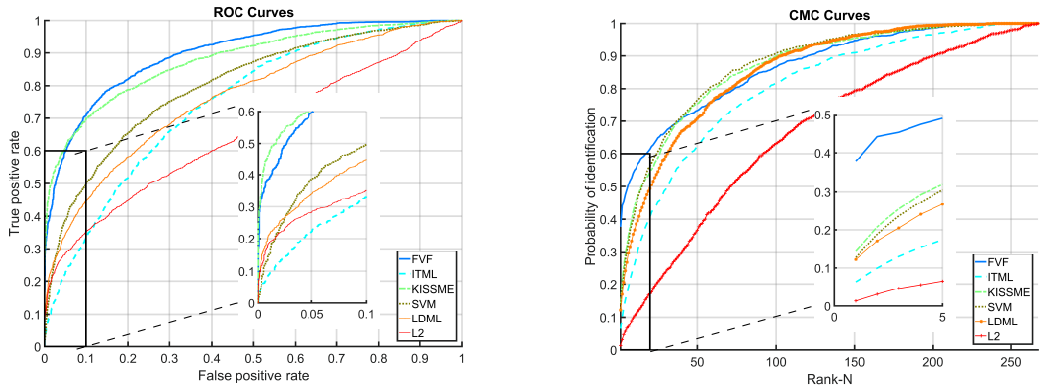


Figure 4.4: Recognition performance in the QUIS-CAMPI dataset. At left: verification performance attained in the QUIS-CAMPI dataset under the unrestricted setting with no outside data. At right: identification performance attained in the QUIS-CAMPI dataset under the closed-set setting with no outside data.

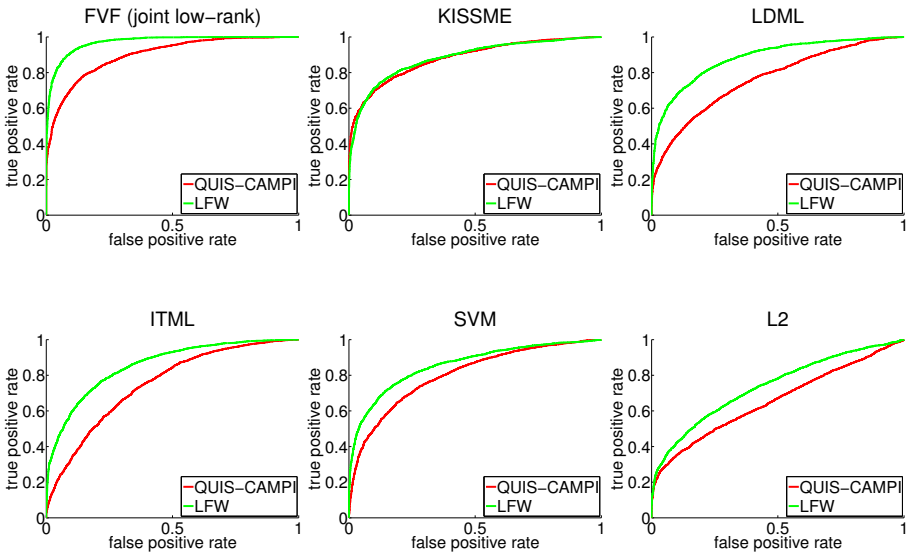


Figure 4.5: Comparison between the recognition performance observed per algorithm in the QUIS-CAMPI and LFW datasets, under the unrestricted setting.

bias in the subject separation, this process is repeated 10 times defining 10 independent evaluation sets. At the training phase three distinct paradigms can be exploited: 1) no outside data; 2) QUIS-CAMPI biometric data; and 3) biometric data from external sources. At the test phase, algorithms must be evaluated using leave-one out validation and the average performance should be reported according to the metrics listed in table 4.3.

4.3.2 Results and Discussion

In order to perceive the robustness of state-of-the-art face recognition algorithms to the QUIS-CAMPI degradation factors, we report results for six face recognition algorithms under the unrestricted setting with no outside data (for verification) and under the closed-set setting with no outside data (for identification) using QUIS-CAMPI.v1 . The tests are conducted using three metric learning based approaches (ITML [255], LDML [256] and KISSME [257]), the Fisher Vector Faces (FVF) [258], a SVM and the L2 distance between SIFT features. The justification

for choosing these methods is twofold: 1) they are well established face recognition methods with competitive performance reported on LFW; and 2) the source code is freely available, which ensures a reliable assessment of their performance. In these experiments, the QUIS-CAMPI metadata are used to provide the algorithms with 256x256 cropped images of the facial region. Considering that, apart from FVF, all the evaluated algorithms expect pre-processed face descriptors, we adopt the widely used strategy of Guillaumin et al. [256], which exploits the SIFT descriptors [259] of nine automatically detected face landmarks [260].

In accordance with the QUIS-CAMPI evaluation protocol, the methods are optimized using the first group of pairs, i.e., the parameters are determined based on maximum recognition accuracy obtained. Subsequently, the performance of each method is determined using leave one-out validation in the 10 splits of the second group. Aiming at providing a comparative performance analysis between QUIS-CAMPI and state-of-the-art benchmarks, the performance of these algorithms is also assessed in LFW.

The results are reported using the ROC curves (for verification) and CMC curves (for identification), as well as its corresponding AUC. Figure 4.4 depicts the ROC curves and CMC curves obtained in the QUIS-CAMPI dataset under the unrestricted setting and the closed-set setting, respectively. The comparison between the ROC curves in the QUIS-CAMPI and LFW sets is given in figure 4.5. Additionally, these results are summarized in table 4.5 with respect to the AUC values.

Regarding the algorithms performance, it is important to note that FVF clearly outperforms the remaining approaches, whereas the use of L2 norm on the SIFT descriptors - without using any additional learning metric - is clearly behind the remaining methods. While the latter is not surprising, the performance gap of the former can be explained by the use of automatic keypoint detection. Due to the large variation in pose, it is likely that the use of such strategy in completely unconstrained environments fails, yielding thus, incorrect face descriptors. On contrast, holistic approaches, such as FVF, may be a more adequate solution for addressing these challenging conditions.

Regarding the comparison of the recognition performance per dataset, the results obtained sustain our claim that QUIS-CAMPI is much more challenging than LFW. This is particularly evident in figure 4.5, since none of the state-of-the-art algorithms was able to improve the results obtained in the LFW when addressing the QUIS-CAMPI dataset.

Besides, the verification accuracy achieved for QUIS-CAMPI justifies the need for such a dataset, since much more has to be done to close the gap between surveillance and biometric recognition.

4.4 Summary

In this chapter, we described the QUIS-CAMPI data feed, comprising biometric samples automatically acquired outdoors, at a distance, on the move and in a fully non-cooperative manner. A key difference between QUIS-CAMPI and related sets is that samples are automatically acquired in a real surveillance scenario, i.e., in a covert way and without any human intervention in the process. This assures that the collected data completely encompass the set of covariate factors of real-world scenarios. Additionally, to the best of our knowledge, QUIS-CAMPI is the first *open* biometric data feed, i.e., new probes and subjects are continuously being added as the system automatically acquires more data. This fact is beneficial for inhibiting bias introduced by the complete knowledge of the entire dataset, particularly in the case of

Table 4.5: Recognition performance on the QUIS-CAMPI and LFW datasets under the verification and identification modalities. Algorithms performance was determined on both datasets according to the QUIS-CAMPI evaluation protocol under the unrestricted setting (verification) and the closed-set setting (identification). The comparative analysis between the results of QUIS-CAMPI and LFW confirms the additional challenges of the proposed dataset. Performance was assessed using AUC of the ROC curve and its corresponding standard deviation.

	Verification				Identification			
	QUIS-CAMPI		LFW		QUIS-CAMPI		LFW	
Algorithm	AUC (%)	ACC (%)	AUC (%)	ACC (%)	AUC (%)	Rank-1 (%)	AUC (%)	Rank-1 (%)
FVF [258]	89.9 $\pm$ 2.8	81.7 $\pm$ 3.0	97.1 $\pm$ 0.8	90.2 $\pm$ 1.6	88.5	26.4	86.8	37.7
ITML [255]	74.9 $\pm$ 4.0	68.4 $\pm$ 2.9	85.7 $\pm$ 1.8	77.5 $\pm$ 2.0	93.9	4.9	80.3	6.4
KISSME [257]	88.0 $\pm$ 3.9	79.2 $\pm$ 4.0	88.2 $\pm$ 1.7	80.9 $\pm$ 2.0	95.8	10.7	86.8	14.3
LDML [256]	76.4 $\pm$ 3.1	69.1 $\pm$ 2.2	88.1 $\pm$ 1.4	79.7 $\pm$ 2.0	96.9	3.9	85.6	12.2
SVM [261]	80.2 $\pm$ 2.8	72.8 $\pm$ 2.6	85.8 $\pm$ 1.3	78.1 $\pm$ 1.0	94.9	9.4	87.3	12.7
L2	65.6 $\pm$ 4.3	62.5 $\pm$ 3.6	73.7 $\pm$ 1.5	67.3 $\pm$ 1.8	70.5	3.0	66.7	1.3

open-set recognition. Moreover, QUIS-CAMPI comprises multi-biometric traits, permitting the exploitation of multi-modal recognition strategies. To objectively justify the need for this set, six face verification algorithms were evaluated both in QUIS-CAMPI and LFW under the image-restricted and unrestricted settings. The conclusions were twofold: 1) the algorithms accuracy in the QUIS-CAMPI set is much lower than in LFW, which confirms that the proposed dataset is more challenging; 2) the state-of-the-art algorithms are still far from optimal recognition rates, and substantial improvements in the recognition technology should be made before saturating the announced set.

Chapter 5

Biometric Recognition in Surveillance Scenarios

This chapter regards the recognition of individuals from biometric data acquired without restrictions, i.e., in an unconstrained and non-cooperative manner. The contributions of this thesis to advance the recognition performance in this kind of data are described in the following sections. Section 5.1 introduces the ICB-RW competition and reports the performance achieved by the nine methods specially designed for this challenge. In section 5.2, we describe our proposal for detecting corrupted features in biometric signatures by relying on the correlation between subsets of features. Section 5.3 describes a novel face recognition approach based on caricatures, and section 5.4 summarizes the major conclusions of the chapter.

5.1 Performance Evaluation of Biometric Recognition in Surveillance Scenarios

The ICB-RW competition was promoted to support this endeavor, being the first biometric challenge carried out in data that realistically result from surveillance scenarios. The competition relied on an innovative master-slave surveillance system for the acquisition of face imagery at a distance and on the move. This section describes the competition details and reports the performance achieved by the participants algorithms.

5.1.1 ICB-RW Competition

The ICB-RW competition took place from September to December, 2015. The website had more than 700 visitors from 24 countries. There was a total of 19 registrations in the competition. Most of the registered users were members of academic institutions, whereas a smaller number was from and private companies. At the end, nine participants submitted their executable to be evaluated in the ICB-RW sequestered data.

5.1.2 ICB-RW Dataset

5.1.2.1 Data Summary

The ICB-RW evaluation set comprises biometric data from 90 volunteers who have provided written authorization for image acquisition and distribution. The data were organized into two sets: 1) gallery data; and 2) probe data. Gallery data comprise 3 images of the subject's head acquired in a controlled scenario: 1) frontal image; 2) left-side image; and 3) right-side image. Probe data were automatically acquired by the QUIS-CAMPI surveillance in a covert and non-cooperative manner while subjects walk throughout the surveillance area. For each subject, ten probe images were selected aiming at preserving the following degradation factors throughout the dataset: 1) variations in pose; 2) variations in expression; 3) varying illumination; 4) occlusions; and 5) blur.

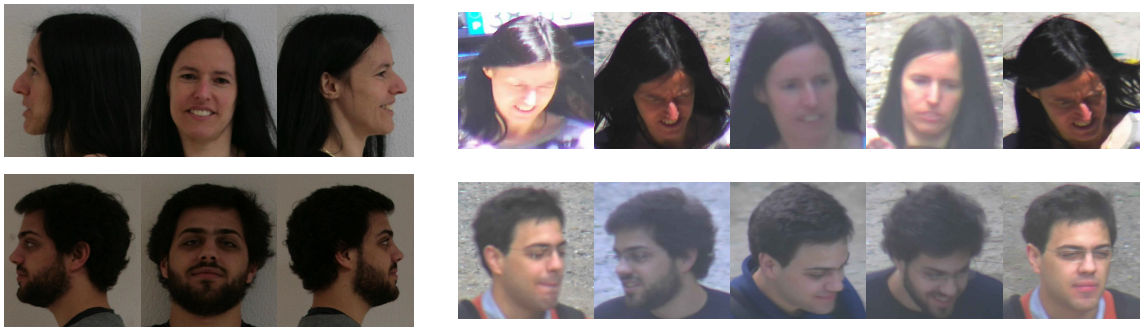


Figure 5.1: Example of the gallery and probe data of two subjects in the ICB-RW dataset. Gallery data comprise one frontal view and two side view images of the subject (left), while five probe images were provided to the participants for algorithm training (right).

Moreover, the face location of the subject of interest in both probe and gallery images was provided as a bounding box. These data were automatically inferred from a state-of-the-art head-landmark localization method [254] and corrected manually. The annotations were provided in the dataset as metadata. Figure 5.1 illustrates the gallery and probe images of two subjects in the dataset.

5.1.3 ICB-RW Protocol

For evaluation purposes, the probe images were randomly divided into two subsets comprising five images each. Participants were provided with gallery data and one probe data subset, while the other subset was kept as sequestered data. During the competition period, participants were expected to build and train their algorithms using the publicly available data, while the final evaluation would be performed by the competition organizers upon algorithm submission.

5.1.3.1 Evaluation Metrics

The algorithm performance was determined by the AUC of the CMC curve. For each probe image P_i , a rank-K list was constructed by selecting the K most similar gallery subjects according to the algorithm scores. The CMC curve relates the percentage of correct identification for all probe images with the size K of the rank-K list.

It is worth noting that the evaluation was conducted using the sequestered subset of probe images, which are disjoint from the ones used by the participants and ensure a non-biased evaluation.

5.1.4 Results and Discussion

In this section, we report the results attained in the ICB-RW challenge. The performance of the nine algorithms submitted to ICB-RW is presented in figure 5.2 and summarized in table 5.1 with respect to the rank-1 and rank-5 identification rate and the AUC of the CMC curve. Also, a brief description provided by the authors is included.

The competition results are useful for two major reasons: 1) perceive how far research has come in fully automated human recognition; 2) provide insight into which strategies are the most adequate for extracting discriminant information from extremely degraded images.

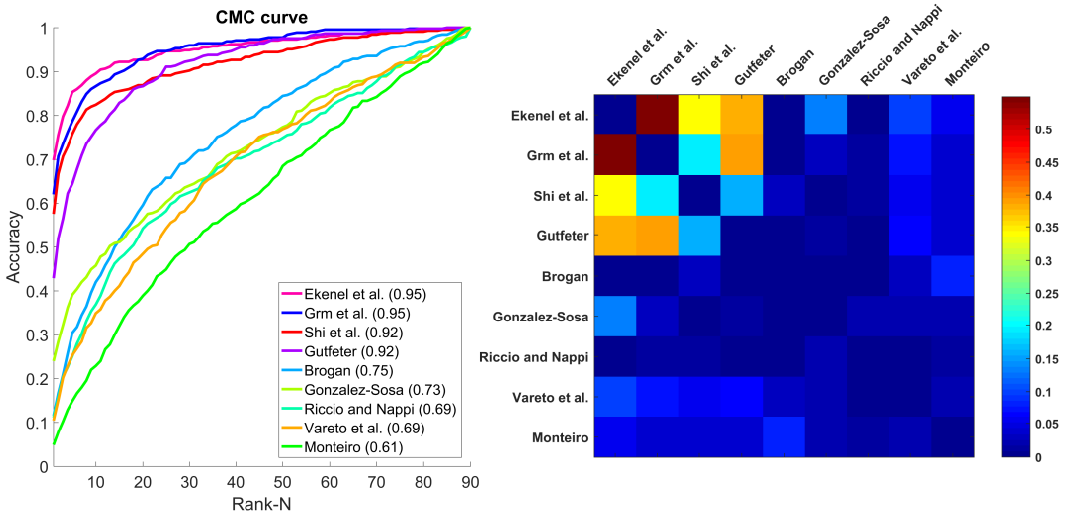


Figure 5.2: Left: Algorithms performance in the ICB-RW competition. Algorithms were evaluated in the sequestered probe data and the CMC curves were used to assess the identification performance. Right: Correlation between the nine algorithms submitted to ICB-RW. The similarity scores obtained during the evaluation phase were used to determine the correlation between the different algorithms submitted to ICB-RW.

In the former, it is interesting to note that the best performing approaches reach relatively high levels of accuracy, particularly if more than one match is considered (85.3% accuracy by retrieving 5% of the database). Despite these results suggest that biometrics has come quite close to recognizing humans in completely unconstrained scenarios in an immediate future, there exist some issues that should be taken into account. First, it would be important to study the impact of gallery size on the algorithms performance, as the number of gallery subjects of ICB-RW was rather limited when compared with real-world scenarios. Second, it is worth noting that in this challenge, a closed-set identification was adopted. Again, it is important to study the impact on the performance when considering open-set identification.

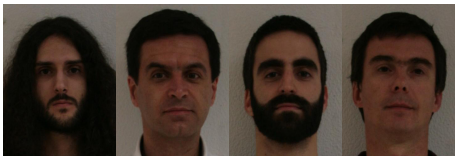
In the latter, we use the general description of the methods, their performance in the ICB-RW competition and the correlation between the similarity scores (figure 5.2) to conclude about the most valuable strategies to address highly degraded biometric data. Regarding the best performing approaches, it is interesting to note that all of them rely on deep learning, in particular they use deep convolutional neural networks. The similarity between them is also corroborated by high correlation values presented in figure 5.2. On the other hand, the approaches using typical descriptor-based features such as SIFT [259] and LBP [41] have attained poor identification rates. Even though these approaches compensated for illumination variations, it is likely that these descriptors are not invariant to the extreme variations in blur, occlusion, and expression of unconstrained scenarios. Regarding the high variability in pose, this factor is addressed in most approaches with the use of face landmark localization and frontalization techniques, whose performance in unconstrained scenarios has been significantly improved [254, 262].

In addition, we analyzed the most easy-to-identify and hard-to-identify probes and subjects to provide additional insight into the most discriminant and confounding factors of unconstrained biometric recognition. For this purpose, we relied on the similarity scores of the three best performing approaches to determine the percentage of gallery subjects that have to be retrieved for correctly identifying a probe image. The percentages of methods were fused using the maximum value and some of the best and worst performing images are depicted in

a) Easy-to-identify probe images.

b) Hard-to-identify probe images

Figure 5.3: The most easily and hardly identifiable probe images according to the methods performance. The percentage of gallery subjects necessary to correctly identify a probe image was used to rank probe images. The first and last four images are depicted in a) and b), respectively. Note that in a) the subjects are almost frontal, the expression is neutral and the face is not occluded, whereas in b) there exists extreme variation in pose, the expressions are not standard, and the facial region is degraded by shadows or self-occlusion.



a) Easy-to-identify subjects.

b) Hard-to-identify subjects

Figure 5.4: The most easily and hardly identifiable subjects according to the methods performance. The percentage of gallery subjects necessary to correctly recognize the probe images of each subject was used to rank subjects. The first and last four subjects are depicted in a) and b), respectively. Note that in a) the subjects are significantly different regarding hair, age and face structure, whereas in b) subjects share age and facial structure.

figure 5.3a) and figure 5.3b), respectively. By averaging the results per subject we also determined the most easy-to-identify and hard-to-identify subjects, which are depicted in figure 5.4. It is interesting to note that high variability in pose greatly affects the methods performance, whereas almost frontal poses with neutral expression provided the best results. On the other hand, it is worth noting that the subjects sharing the same age group and facial features, such as facial hair style, were the most likely to be incorrectly classified.

5.1.5 Conclusion

The limitations of state-of-the-art unconstrained biometric datasets and the lack of benchmark tests on biometric recognition in surveillance scenarios were the rationale behind the ICB-RW challenge. The use of a fully automated surveillance system capable of covertly imaging subjects at a distance and on the move allowed to assess how far research has come in human recognition in totally unconstrained scenarios. The results were relatively positive, however much more has to be done to consider biometric recognition in the wild as solved, particularly when considering scenarios where the number of gallery subjects is much higher. We hope that

Table 5.1: Final results for the ICB-RW competition. The nine valid submissions of ICB-RW are listed according to their rank in the challenge. Methods are ranked according to the AUC of the CMC curve. Also, both rank-1 and rank-5 identification rate (IR) are presented, along with a brief description provided by the authors.

Method	Description	Rank-1 IR (%)	Rank-5 IR (%)	AUC (CMC curve)
H. Ekenel, G. Özbulak, E. Ghaleb Istanbul Technical University	The probe and gallery face images are aligned with respect to eye centers. Only the frontal images are used as gallery. Face representation is extracted from a CNN with a VGG face model [185]. In the test phase, the nearest neighbour classifier is used with the correlation distance as the similarity score.	69.8	85.3	0.954
K. Grm, S. Dobrisek, V. Struc University of Ljubljana	An augmented dataset was generated through oversampling the training images via bounding box noise and horizontal flipping. The pre-trained VGG face deep convolutional network [185] was used to extract features from the images. Then, a softmax classifier was trained on the features.	62.0	78.7	0.952
H. Shi, X. Zhu, S. Liao, Z. Lei, S. Li Institute of Automation, Chinese Academy of Sciences	Features are extracted from a deep convolutional network model trained on the CASIA-Webface database and the cosine similarity is used as score. Ten models learned from different facial parts are fused, and the gallery images of different poses are synthesized to ease the matching phase.	57.6	75.8	0.921
W. Gutfeter NASK, Warsaw University of Technology	The algorithm builds similarity scores by merging results obtained from a set of convolutional neural networks trained for recognizing faces from different angles.	42.9	64.4	0.918
J. Brogan University of Notre Dame	Gallery and probe images are frontalized using a modified version of [262]. Data features are extracted from a SLMSimple Neural Network [263] and four bins are created containing different versions of the gallery images. Probe descriptors are matched with one of the four bins according to yaw angle of the head, and the resulting pairs of feature vectors are input into an SVM trained with LFW [241] data.	11.6	30.4	0.755
E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez University Autonoma de Madrid	The LBP [41] of nine facial regions are extracted from a frontalized image [254] followed by illumination normalization. A fused distance score is determined by only considering the five best individual facial regions at each trial.	24.0	39.1	0.725
D. Riccio, M. Nappi University of Salerno	The algorithm locates facial points through an extended Active Shape Model and remaps the face region to a 64x100 image. It applies a local normalization process to correct illumination variations, and the matching is performed with an optimized localized version of the spatial correlation index.	11.1	25.1	0.694
R. Vareto, R. Prates, W. Schwartz University Federal de Minas Gerais	A set of facial components are obtained by performing face landmark localization, and these components are described by a variant of LBP. In the learning phase, a binary classifier based on the partial least squares model is inferred for each gallery subject. During the test phase, the identity of the probe samples is determined by the classifier with the highest score.	10.4	25.3	0.688
J. Monteiro University of Porto	During enrolment, an universal background model is used to infer a model per individual describing the statistical distribution of each feature (SIFT/GIST). In the recognition phase, features are projected onto both the UBM and the individual specific models. A likelihood-ratio between both projections outputs the final recognition score.	4.9	14.9	0.613

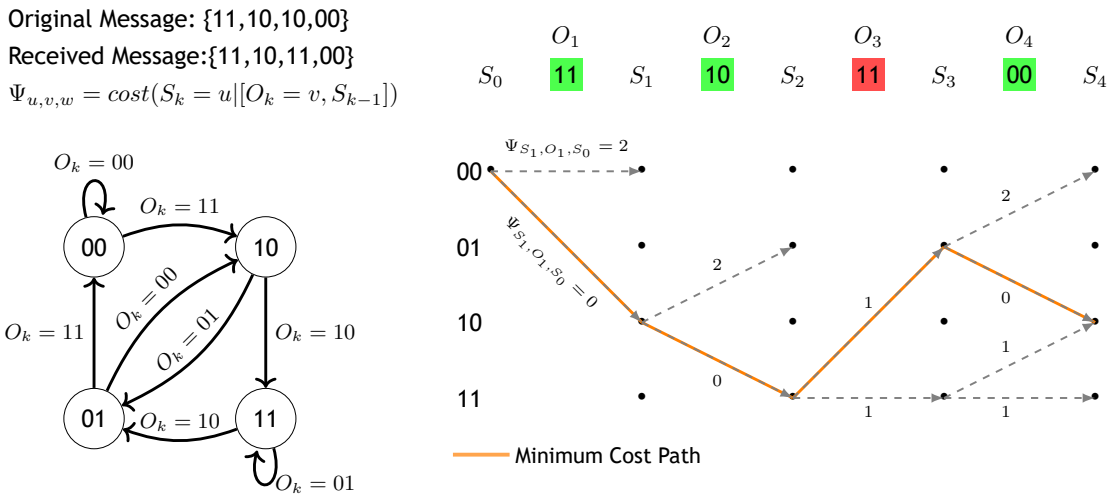
the results obtained from this contest can contribute to the understanding of the challenges of biometric recognition in the wild, and further advance the research in this field.

5.2 Proposed Feature Quality Assessment Method

Biometric systems work by extracting distinctive sets of features from body traits, which subsequently fed a classifier to determine the subject identity. Usually, the feature encoding process is performed using handcrafted descriptors yielding a compact representation, but still highly redundant [264, 265].

The use of data redundancy is a key assumption in Error-correcting Codes (ECC). In these approaches, redundant data are generated and added to the encoded representation in a deterministic manner using pre-built models. As an illustration, figure 5.5 depicts the operation mode of convolutional codes, where it can be observed the main insight of ECC: an observed variable o_t can be determined to be noisy by analyzing the previous observations $o_1, o_2, ..., o_{t-1}$ to which the o_t is dependent. In this context, an error is defined as an impossible observed value given a set of past observed values.

Motivated by the fact that most biometric methods rely on redundant descriptors to perform classification, we follow the principles of ECC to develop a method capable of detecting corrupted features in biometric signatures by relying on the correlation between subsets of features.



a) State Transition Diagram **b) Trellis Diagram**

Figure 5.5: Graphical representation of the decoding process used in convolutional codes. a) The state transition diagram defines the valid transitions according to the observed encoded pair of bits. b) The trellis diagram illustrates the cost of all valid $S_k \rightarrow S_{k+1}$ transitions for a given observation O_k (for displaying purposes not all the valid transitions are represented). The Viterbi algorithm is used to determine the most likely sequence of states visited by the transmitter, allowing to recover the original encoded message. Note that this model can detect that an error has occurred in the transition to S_3 by exploiting the dependence between S_3 and (S_2, O_3).

Unlike the existing biometric cryptosystems, the proposed method works directly on the biometric signatures. This is advantageous for highly-degraded signatures but introduces different challenges. First, feature vectors obtained from different biometric traits do not follow deterministic relations between their components as in the case of ECC methods. Second, fea-

tures dependencies greatly vary with respect to the type of data. To address these problems, our approach infers feature correlations from training data and a relaxed error definition is introduced: the soft assignment of the i^{th} to a corrupted state is modeled by the likelihood of observing o_i given $\{o_j, j \in D\}$, being D the indices of the features to which i is dependent. The proposed model operates in two phases: 1) redundancy analysis; and 2) state inference. While the former determines the dependent pairs of features, the latter analyzes these relations to decide - at test time - the most likely set of corrupted components given an observed probe descriptor. In this phase, inference is performed with a MRF because it is straightforward to encode both unary costs and pairwise constraints between features.

To illustrate the usefulness of the proposed method, we assess the performance improvement of different biometric recognition methods when coupled with the proposed error detection method. Also, we report the performance in an image classification dataset using CNNs descriptors, in order to show the flexibility of the proposed approach.

5.2.1 Error Detection in Biometric Signatures

Error control techniques operate either by adding check bits to the original message (systematic scheme) or encoding the message in a specific representation (non-systematic). The rationale behind both approaches is the augmentation of the transmitted message size by adding redundant data. The deterministic relations between the new representation and the original data allow the decoder to check if an encoded message is valid, locate the corrupted components, and in some techniques recover the original data.

Linear ECC are one kind of systematic scheme widely used in biometric cryptosystems [266–268]. Let H be a N-dimensional Hamming space, linear ECC methods work by mapping input data to elements of H, denoted as codewords. The set of valid codewords, usually denoted as C, are chosen such that they are separated at least by a Hamming distance of d. Correction is performed by transforming an encoded sequence to the its nearest codeword in C. This strategy ensures a correct assignment if no more than $\lfloor \frac{d-1}{2} \rfloor$ bit errors occur [269].

In [267, 270–274], different linear ECC methods, such as the Hadamard codes, Reed-Solomon codes and the low-density parity-check codes, are used to correct errors in iris codes. However, these methods are highly limited by the number of errors than can be corrected in the original feature vector, restraining their applicability to data with large intra-class variations.

As an example, the work of Kanade et al. [272] introduces a novel way to use ECC to reduce the variabilities in biometric data by using Hadamard codes in a block-wise manner. Even though these codes ensure successful regeneration up to 25% of bit corruptions in the encoded message, this percentage is reduced exponentially in the original feature vector.

Aiming at extending the applicability of ECC methods, the Error-Correcting Output Codes were introduced as a machine learning ensemble method [275]. This approach is believed to improve performance both by decomposing the multi-class problem into a number of two-class problems, as well as by correcting errors in the decision-making stage. Different biometric recognition methods exploit variations of this technique [276, 277]. A prominent example is the face verification method of Kittler et al. [277], where the authors assume that multiple images of the same client identity are available, and the verification score is determined by a statistical test using first order Minkowski distance.

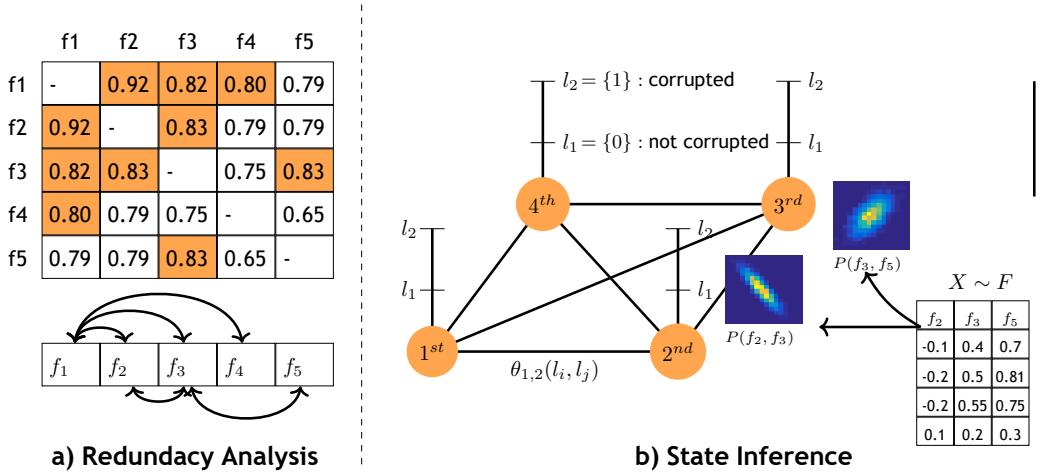


Figure 5.6: Schematic representation of the phases involved in the proposed method. a) In the redundancy analysis phase, redundant pairs of features are determined by measuring linear correlations. b) The likelihood of each feature component being corrupt is encoded by a MRF, whose energy minimization yields the maximum-likelihood sequence of states. Note that only the redundant features are connected in the MRF, allowing to discard irrelevant evidence when determining the state of a component.

5.2.2 Our Method

We first define the notation used to describe the proposed approach.

- F_c : the set of feature vectors obtained from non-degraded data;
- F_d : the set of feature vectors obtained from degraded data;
- $F = F_c \cup F_d$;
- f_i : the i^{th} component of a feature vector;
- $c_i = \{0, 1\}$: the state of f_i determining if the component is corrupted (1) or not (0).

As illustrated in figure 5.6, the proposed model is composed of two phases: 1) redundancy analysis and 2) state inference. The first aims at determining the pairs of features that are correlated, allowing the latter phase to exploit this information to infer the sequence of states that best explains the observed data.

5.2.2.1 Redundancy Analysis

Contrary to ECC methods, where bit dependence is pre-determined using handcrafted state transition models, the proposed method requires the inference of data redundancies from training data. For this purpose, we rely on the Pearson correlation to determine the pairwise dependence between all feature pairs. This produces the correlation matrix M that is used to determine the indices to which the i^{th} is connected:

$$L(i) = \{j | M(i, j) \geq \nu\}, \quad (5.1)$$

where $\nu \in [0, 1]$ denotes the minimum confidence degree for considering two components as dependent.

Subsequently, L is used to determine the pairs that influence each other during the inference phase, reducing the inference complexity.

5.2.2.2 State Inference

The proposed model is depicted in the figure 5.6 and is composed of n_v vertices, representing the components of the feature vector. Also, to each vertex can be assigned one label $\{0, 1\}$, denoting whether a component of the feature vector is corrupted or not. The rationale behind this structure is threefold: 1) the dependence between features can be modeled with the existence of edges between vertices; 2) the probability of a feature being corrupted given a subset of observed features can be modeled by pairwise costs; 3) the probability of a feature being corrupted given its observed value can be represented by the unary potentials.

Let $G = (V, E)$ be a graph representing a MRF, composed of a set of n_v vertices V, linked by n_e edges E. The MRF is a representation of a discrete latent random variable $L = \{L_i\}, \forall i \in V$, where each element L_i takes one value l_u from a set of labels.

In this problem, a MRF configuration $l = \{l_1, ..., l_{n_v}\}$, determines the set of corrupted components of the feature vector. The number of edges in G is determined by the pairwise dependence between features, i.e., the vertices are connected if and only if the feature pair (i, j) is redundant (discussed in section 5.2.2.1). Each edge encodes the cost of assigning the class l_u to the i^{th} feature vector component and the class l_v to the j^{th} feature vector component.

The energy of a configuration l of the MRF is the sum of the unary $\theta_i(l_u)$ and pairwise $\theta_{i,j}(l_u, l_v)$ potentials:

$$E(l) = \sum_{i \in \mathcal{V}} \theta_i(l_u) + \sum_{(i,j) \in \mathcal{E}} \theta_{i,j}(l_u, l_v). \quad (5.2)$$

According to this formulation, determining the corrupted components of the feature vector is equivalent to infer the random variables in the MRF by minimizing its energy:

$$\hat{l} = \arg\min_{l} E(l), \quad (5.3)$$

where $\hat{l_1}, ..., \hat{l_{tp}}$ are the labels of the n_v feature vector components. As an example, if a five length feature vector is considered, the configuration $\{0, 1, 0, 0, 1\}$ determines f_2 and f_5 as being corrupted.

The loopy belief propagation [215] algorithm was used to optimize the MRF. Even though it is not guaranteed to converge to a global minimum on loopy non-submodular graphs (such as our MRF), we concluded that the algorithm provides good approximations (refer to section 5.2.3).

5.2.2.3 Unary and Pairwise Potentials

Let $X_c \in F_c$ and $X \in F$ be samples acquired during the training phase. These data can be used to determine the posterior probability of f_i being corrupt given an observed value of this component, which according to the Bayes theorem is defined as:

$$P(c_i \mid f_i) = \frac{P(f_i \mid c_i).P(c_i)}{P(f_i)}. \quad (5.4)$$

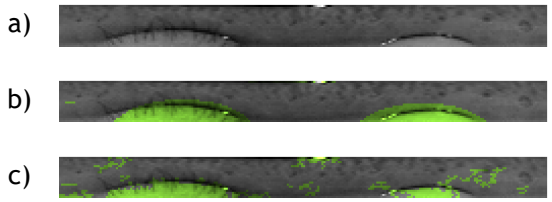


Figure 5.7: Comparison between the iris noise mask obtained from [278] and the degraded regions inferred from our approach. a) Original iris code. b) Iris noise mask of [278]. c) Degraded components identified by our method reshaped to the original iris code size. Note that our approach was able to identify non-iris regions, whereas the geometric-based approach used in [278] overestimated the boundaries of non-iris regions.

Considering that during the training phase it may be cumbersome to obtain a sample from F_d, it is not possible to obtain $P(f_i \mid c_i = 1)$ in a straightforward way. Consequently, we use X to estimate $P(f_i)$ and subsequently derive $P(c_i = 1 \mid f_i)$ from its complement. The unary costs are thus defined as $\theta_i(l_u) = 1 - P(c_i = l_u \mid f_i)$.

While the unary potentials disregard any information from neighbors and are meant to emphasize the necessity of each observation being coherent with the typical distribution of the i^{th} feature vector component, the pairwise potentials model if the observation of f_i is coherent with the observed value of f_j, to which f_i depends. Again, this can be measured using the posterior probability of f_i being corrupt given an observed value of f_i and f_j, which according to the Bayes theorem is defined as:

$$P(c_i \mid [f_i, f_j]) = \frac{P([f_i, f_j] \mid c_i).P(c_i)}{P(f_i, f_j)}. \quad (5.5)$$

The pairwise costs between two adjacent vertices $\theta_{i,j}(l_u, l_v)$ are then defined as:

$$\theta_{i,j}(l_u, l_v) = \begin{cases} P(l_u, l_v \mid [f_i, f_j]), & \text{if } l_u = 0 \text{ and } l_v = 0, \\ 0.5, & \text{if } l_u \neq l_v, \\ 1 - P(l_u, l_v \mid [f_i, f_j]), & \text{otherwise.} \end{cases} \quad (5.6)$$

Similarly to the unary costs, $P([f_i, f_j])$ is directly estimated from X.

5.2.2.4 Feature Matching

After obtaining the corrupted component mask m^y of the probe descriptor y, feature matching should be modified to allow disregarding degraded components when determining the scores between y and training samples x using a classifier Φ. Accordingly, the score is determined by:

$$s(x_i, y) = \Phi(m^y.x_i \ , \ m^y.y). \quad (5.7)$$

5.2.3 Results and Discussion

The proposed method was validated in three distinct datasets, namely, the CASIA-Thousand [279], the AR-Database [234] and the ILSVRC [280]. While the first and the second regard biometric traits (iris and face, respectively), the latter was designed for use in visual object recognition. The rationale to include this set was to evidence the flexibility of the proposed approach. It should be stressed that no particular concerns were taken in optimizing the recognition methods for the used datasets, meaning that the focus was put in the performance gap between both recognition schemes than in the recognition errors in absolute values, which are out of the scope of this contribution.

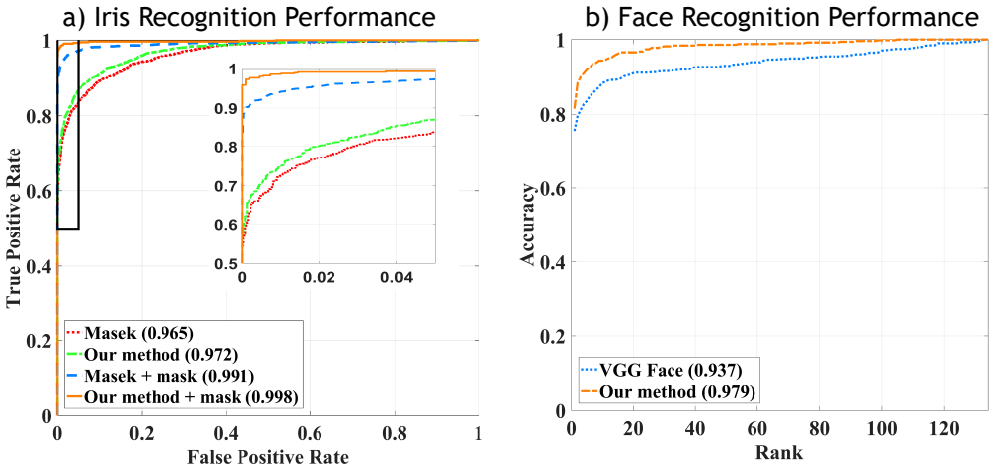


Figure 5.8: Comparison between the original performance of the recognition methods and the performance obtained by disregarding degraded components of the image descriptor during the matching phase. a) ROC curves obtained when using the iris code from Masek's algorithm, the error mask inferred from our approach, the error mask from the original method, and the combination of both masks. The CMC curves reporting the results for face recognition and image classification obtained with a CNN descriptor are depicted in b). The AUC (in parentheses) is also provided for each approach.

5.2.3.1 Iris Recognition Performance

To exemplify the usefulness of the proposed method for iris recognition, we compared the performance of Masek's algorithm [278] using four strategies: 1) the original iris code (baseline); the masked iris code obtained from our error detection approach; 3) the noise free iris code obtained from the original recognition method; and 4) the masked noise free iris code obtained by combining the second and third strategies.

The experiments were carried out in a subset of the CASIA-Thousand database that was obtained by manually discarding images segmented incorrectly. These data were subsequently separated into training and test sets, comprising 600 and 400 different eyes, with about 6000 and 4000 images, respectively. Additionally, we screened, through visual inspection, the training images to obtain a sample of F_c, i.e., a set of images where iris is not heavily occluded by eyelids, eyelashes, shadows, or specular reflections. At the end, 244 images were kept as a sample of F_c, while the 6000 training images were considered as a sample of F (figure 5.9 show exemplars of these two sets).

The ROC curves and the corresponding AUC for the described variants are compared in figure 5.8 and evidence the benefits of withdrawing corrupted features from the matching phase

(0.97 vs 0.96 regarding AUC). However, the proposed method was not able to overcome the original method when coupled with its noise iris mask.

The comparison between the corrupted bit locations obtained from our method and the noise iris mask inferred from the Masek's method is depicted in figure 5.7. It can be seen that our approach failed to identify some corrupt components. This can be explained by the fact that the iris noise mask was derived using mainly geometric-based information, which is not available to our method. Nevertheless, it should be noted that combining both masks - using bit intersection - outperforms the remaining variants, suggesting that noise regions identified by the method of Masek were overestimated.

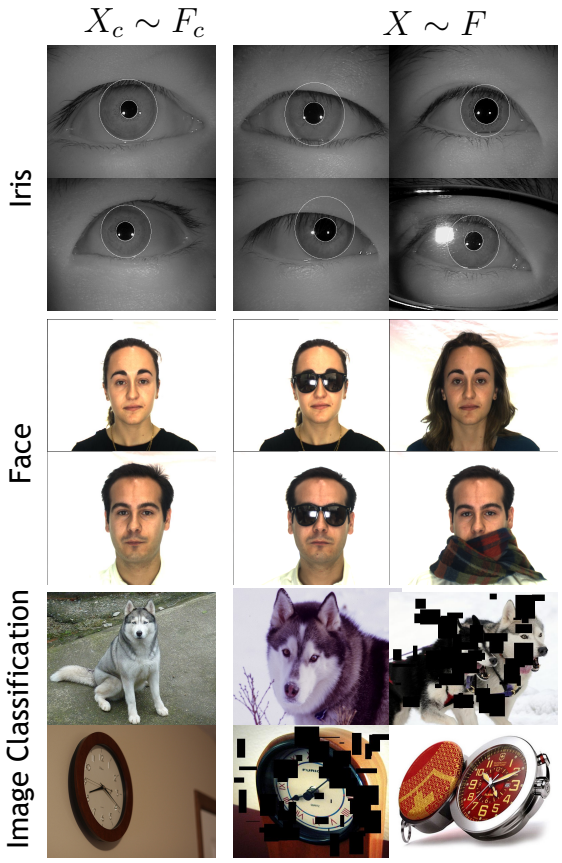


Figure 5.9: Representative examples of the data used in the experiments. The first column illustrates images assigned to F_c for being considered as non-degraded, while the remaining columns comprises both degraded and non-degraded images available in F.

5.2.3.2 Face Recognition Performance

Regarding face recognition, we used the AR database to determine the usefulness of the proposed method, since the design of this set ensured a clear distinction between the degradation factors of each image. In the experiments, non-degraded frontal images were used as training data, while images corrupted by varying illumination, occlusion and facial expression served as test data. For feature extraction, we relied on the VGG Face Descriptor [185], a CNN implementation based on the VGG-Very-Deep-16 architecture adapted for the task of face verification. In accordance to [185], data was fed to a pre-trained CNN and the 4,096-dimensional descriptor of the final fully connected layer was used as image descriptor. Figure 5.8 depicts the CMC curves for baseline and the proposed method.

The explanation for this improvement lies in the fact that the state of a component is inferred in a consensual manner. For example, degradation factors affecting a particular region of the image (e.g., occlusions) may induce large deviations in some components of the image descriptor. In an unary-based approach, each component would only be classified as degraded in case of an extreme variation. On contrast, in our method the value of a degraded component is considered incongruent with a subset of non-degraded observations, whose pairwise potentials force the component state to be corrupt in the final MRF configuration.

5.2.3.3 Image Classification Performance

In order to demonstrate the flexibility of our approach, we also assessed its performance in the field of image classification. For this purpose, we used the data available from the ILSVRC [280] - a state-of-the-art image classification challenge - containing 150,000 validation and test images of 1000 object categories. For each category, 75 images were corrupted to serve as test set, while the remaining images were used for training. Corruption was performed using synthetic occlusions (80 random patches of variable size), as illustrated in figure 5.9.

The image descriptors were extracted using one of the best performing methods on the ILSVRC 2014 challenge [281]. Again, we used the last fully connected layer of the pre-trained CNN, and the descriptors were compared with the L2 distance. The comparison between the CMC curves of figure 5.10 shows a performance improvement when correcting descriptors of images degraded by synthetic occlusions. Even though these occlusions are not realistic, it should be noted that the experiments on the ILSVRC set aimed only at evidencing the flexibility of the proposed approach.

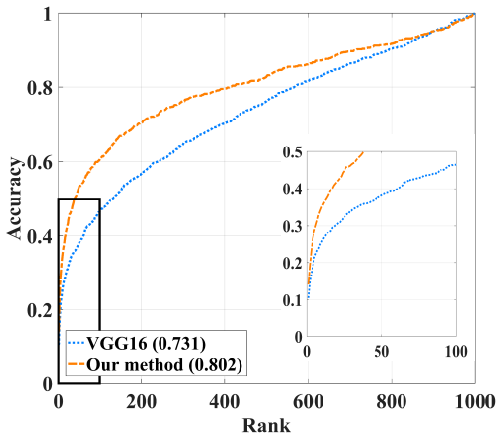


Figure 5.10: Comparison between the original performance of the recognition methods and the performance obtained by disregarding degraded components of the image descriptor during the matching phase. The image classification performance obtained with a CNN descriptor was assessed using CMC curves and the corresponding AUC (in parentheses).

5.2.4 Conclusion

In this section, a method for determining degraded components of biometric signatures was introduced. Unlike ECC-based biometric cryptosystems, our approach works directly on the visual descriptor, providing additional robustness to high-magnitude errors and highly degraded feature vectors. The proposed method assumes that data redundancy in biometric signatures

resulting from unconstrained scenarios can be used for the detection of degraded components. Though it seems a limiting assumption, the experiments show an improvement in the recognition performance when disregarding the degraded components during the matching phase. These results not only evidence the feasibility of the proposed method, but also suggest that visual descriptors actually contain redundant and low-entropy features.

As further directions for this work, we are currently investigating ways to simultaneously perform detection and correction of the degraded components.

5.3 Proposed Face Recognition Method

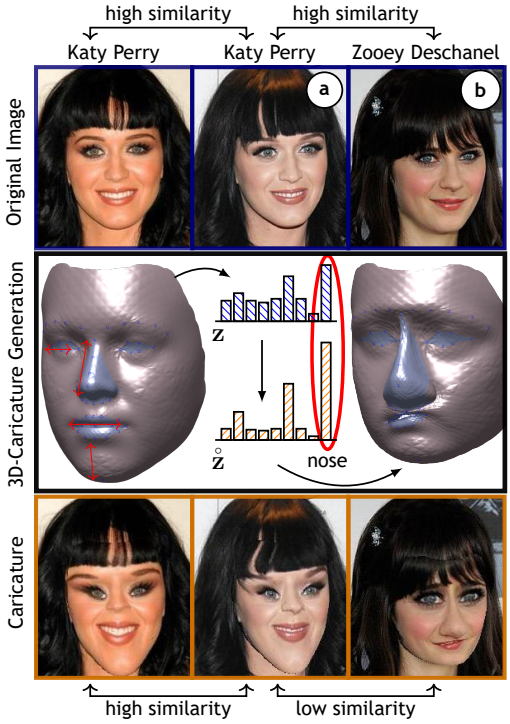


Figure 5.11: Advantages of using caricatures for face recognition. Both humans and automated systems find difficult to distinguish between visually similar subjects (e.g., Katy Perry and Zooey Deschanel). Familiarized observers overcome this problem by focusing on the most distinguishable features of each face, and several studies suggest that this task is carried out in the brain by creating a caricatured representation of the original image. Our method aims at mimicking this process by analyzing face proportions and exaggerating the most salient ones. The proposed approach is capable of producing 2D caricatures where inter-subject similarity is minimized and intra-subject similarity is preserved.

Humans have an astonishing capability of recognizing familiar faces in totally unconstrained scenarios. However, this performance decreases significantly in case of unfamiliar faces [282]. The question of how an unfamiliar face becomes a familiar face is not consensual, but there is evidence that this process is carried out in a caricatured manner [283, 284]. According to this theory, familiarization works by analyzing the most significant physical deviations of a face with respect to a mental representation of the average face, followed by the creation of a modified description of the face, where the most distinctive features are exaggerated and average features are oversimplified (similar to drawing a caricature). Moreover, different studies concluded that humans perform better at recognizing individuals from caricatures [285–288] than veridical faces, supporting the idea that the human brain encodes familiar

faces as a caricatured version of the original face.

Inspired by the idea that distinctive feature exaggeration may be the key for the incredible performance of humans on recognizing familiar faces, we introduce a fully automated face recognition approach based on a 3D caricature generation method capable of creating 2D face representations, where likeness is preserved and the inter-class separation is enlarged. The rationale behind our idea is illustrated in figure 5.11, where an unfamiliar observer perceives incorrectly figure 5.11a) and figure 5.11b) as photos from the same identity. On the contrary, it is straightforward to discern between Katy Perry and Zooey Deschanel when observing their caricatures.

For automated caricature generation, the proposed method attempts to mimic the three main stages of the caricature drawing process:

1. **Caricaturists infer 3D face structure from either a single or multiple views of the face.** This phase is replicated by estimating a 3D morphable model from an input image and a set of facial landmarks. The accuracy of the landmarks decreases significantly in unconstrained data, and for that reason, we combine multiple state-of-the-art face alignment algorithms in an ensemble learning strategy. In addition, we use a model with a reduced number of vertices to account for model stability while maintaining the dominant features of the face.

2. **The caricaturist analyzes facial features for determining the deformation applied to each one.** After inferring the 3D structure, our method compares a set of face regions with a reference 3D model regarding translation, scale and orientation. The region deviations are then normalized and exaggerated using a 'measure locally, weight globally' strategy.

3. **The artist redraws the original face using the deformed proportions.** After determining the positions of the deformed vertices, the mesh is warped with a Laplacian mesh editing technique for preserving local detail and guaranteeing smooth transitions between vertices. The final 2D caricature is obtained by projecting the 3D model in the original camera-view.

In the learning and classification phase, we replicate the strategy introduced in [185] but using caricatures rather than veridical face images. Accordingly, the VGG-Face architecture is trained from scratch on caricatures automatically generated from the VGG dataset, whereas the features produced by the 'fc6' layer are used as face descriptor.

The performance of the proposed face recognition approach is assessed on three state-of-the-art face recognition datasets (LFW [180, 289], IJB-A [219], and MegaFace [290]). To demonstrate the improvements due to the use of caricatures, we measure the relative performance between using caricatures and using original images for network training.

In summary, this method makes two major contributions: 1) a 3D-based caricature generation method for producing 2D caricatures that enhance the performance of face recognition; and 2) the first fully automated caricature-based face recognition system capable of working in real-time with data acquired in the wild.

5.3.1 Caricature-based Face Recognition

The internal process behind recognizing faces has been studied extensively during the last decades [285] and several studies suggest that the brain encodes faces with respect to

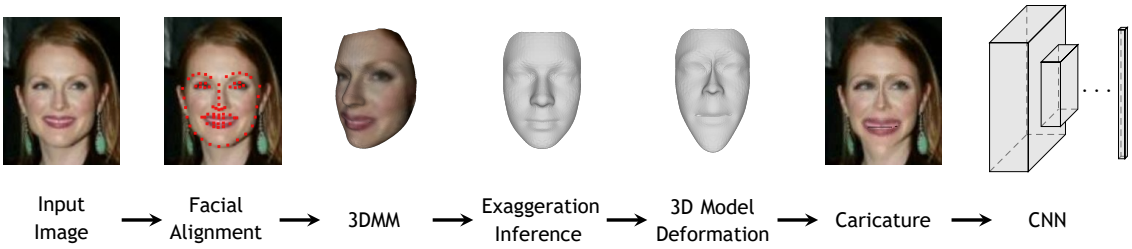


Figure 5.12: Overview of the processing chain of the proposed method. The 3D face structure of probe images is inferred by a 3DMM method coupled with a set of automatically detected facial landmarks. This three-dimensional model permits the replication of the caricature drawing process by: 1) measuring the deviation of face regions to a reference prototype; and 2) using a 'measure locally, weight globally' strategy for inferring the exaggeration of each region. Using the modified regions as constraints, the original mesh is deformed with the Laplacian mesh editing algorithm, and the 2D caricature is obtained by projecting the deformed model in the original camera-view. At the end, the caricature is passed through a CNN to obtain a caricature-based face descriptor.

a general face prototype [291]. Also, for encoding, the brain emphasizes the most deviated physical traits and disregards average features, contributing to increase the inter-class separation while retaining the stability of intra-class separation. These results explain why humans can recognize better caricatures than veridical faces [285, 288, 292, 293] and indicate that, in fact, the brain encodes faces in a caricatured manner [294]. These findings evince that automated face recognition may also benefit from the use of caricatures. However, few works have exploited this idea, and to the best of our knowledge, our work introduces the first fully automated caricature-based face recognition system. Below, we review the existing approaches for generating caricatures from 2D images, and caricature-based face recognition methods.

5.3.1.1 Caricature Generation

Creating face models in a caricatured style is a popular topic in computer graphics and can be broadly divided in two families: 1) rule-based approaches; and 2) example-based approaches.

Rule-based approaches amplify the divergence between a probe face and a reference face by modifying the point-to-point distance of a set of fiducial points marked on both images. The first representative work of this family used 165 feature points to control deformations [295]. Liao et al. [296] introduced an automated caricature generation method that detects and analyzes facial features without human assistance. In [297], the normalized deviation from the average model was used to exaggerate the distinctive features, while Tseng et al. [298] used the inter and intra correlations of size, shape and position features for exaggeration. These works are 2D-based and most of them provide semi-automated systems that depend on user input to define the regions to be deformed. With the advent of 3D face databases, 3D-based caricature generation became the most popular approach. Lewiner [299] introduced an innovative 3D caricature generation tool by measuring the face deviations in the harmonic space. Clarke et al. [300] proposed an automatic 3D caricature generator based on a pseudo stress-strain model for representing the deformation characteristics at each feature point. Sela et al. [301] introduced a general approach for deforming surfaces based on the local curvature. However, by disregarding the use of a reference model for guiding region deformation, this method decreases likeness when applied to faces.

Data-driven approaches learn a mapping between the features of original face images to its corresponding caricature [302]. Liu et al. [303] proposed a machine learning method to map 2D feature points detected in face images to the coefficients of a PCA model learned

from a dataset of 200 3D caricature models. An interactive technique was proposed in [304], where each mouse operation on vertices caused the inference of the PCA coefficients of the individual face components. Han et al. [305] introduced the first system capable of creating 3D face caricatures from 2D sketches by training a CNN with 2D sketches and the corresponding subspace of 3D shape and expression variations. For a detailed description of automated caricature generation refer to the survey of Sadimon et al. [306].

5.3.1.2 Face Recognition

The performance of face recognition in the wild has significantly increased, mainly due to the advent of deep learning [182, 184, 307]. Nevertheless, the majority of face recognition approaches focused on improving performance via new learning strategies, augmenting training data or learning an embedding in the descriptors space, instead of adjusting the input data to a more suitable representation to address this problem (e.g., using face caricatures). Regarding caricature-based face recognition, there is limited work in the literature. Klare et al. [308] used qualitative features from face images and the corresponding caricatures to train a logistic regression model that predicted the similarity score between a caricature and a photo. However, these features were manually annotated via Amazon's Mechanical Turk, restraining the usability of this approach in a real-world scenario. Abaci and Akgul [309] proposed a method to automatically extract facial attributes from photos, but the attributes of caricatures were manually labeled. On contrary, Ouyang et al. [310] introduced a completely automated approach to match photos with caricatures by using a classifier ensemble for estimating facial attributes in both domains.

5.3.2 Our Method

For comprehensibility, we use the following notation: matrices are represented by capitalized bold fonts, vectors appear in bold, and subscripts denote indexes. The proposed method is divided in six main phases, which are depicted in figure 5.12 and define the structure of this section.

5.3.2.1 Facial Alignment

The localization of facial landmarks, also known as facial alignment, is a key step in the 3DMM phase of our approach. Besides, spurious landmarks affect significantly the likeness of the caricature, as the inferred 3D face structure does not portray correctly the facial features of the subject. Despite the astonishing increase in performance of face alignment algorithms, the localization of landmarks in totally unconstrained data remains an open problem. For this reason, we combine k state-of-the-art facial alignment algorithms [254, 311–313] in an ensemble strategy for predicting the most accurate set of landmarks obtained from these methods. Let $\mathbf{q} = [x_1, y_1, \cdots, x_n, y_n]^T$ be a vector with the locations of n face landmarks in a 2D image, and $\mathbf{Q} = [\mathbf{q}^{(1)}, \cdots, \mathbf{q}^{(k)}]$ the matrix with the locations of the facial landmarks of k distinct face alignment methods. Assuming that the k face alignment methods produce uncorrelated outputs, the way they correlate in a particular image may provide insight about the correct set of landmarks, i.e., methods producing landmarks in very close locations are more likely to be correct. Accordingly, the output of the facial alignment algorithms is used to obtain $\mathbf{Q}^{(i)}, i \in \{1, \cdots, N\}$ for N annotated images, and for each image, the vector $\hat{\mathbf{y}}^{(i)} \in \{0, 1\}^k$ is determined by:

$$\hat{\mathbf{y}}_j^{(i)} = \begin{cases} 1 & \text{if } \frac{|\mathbf{q}^{(j)} - \mathbf{g}^{(i)}|}{d^{(i)}} < \varepsilon \\ 0 & \text{otherwise,} \end{cases} \quad (5.8)$$

where $\mathbf{g}^{(i)}$ and $d^{(i)}$ are the manually annotated landmarks, and the inter-ocular distance for the i^{th} image, respectively, whereas ε is a hard threshold controlling the maximum amount of inter-ocular distance that a set of landmarks $\mathbf{q}$ can differ from the ground truth. Each binary vector $\hat{\mathbf{y}}^{(i)}$ denotes the methods that produced the correct landmarks for the i^{th} image, and the vectors of all training images are used to infer the function $\Psi : \mathbb{N}^{k \times n} \mapsto \{0,1\}^k$ by minimizing the following loss function:

$$\sum_{i=1}^{N} \left\| \Psi(\mathbf{Q}^{(i)}; \mathbf{W}) - \hat{\mathbf{y}}^{(i)} \right\|_2, \quad (5.9)$$

where $\mathbf{W}$ are the weights of the neural network used for inference. Given a probe image and the respective landmarks of the k facial alignment methods, $\mathbf{y} = \Psi(\mathbf{Q}; \mathbf{W})$ provides the likelihood of each method being correct, and we choose the landmarks of the method with maximum likelihood.

5.3.2.2 3D Morphable Model

Blanz and Vetter introduced the 3D morphable models for the synthesis of 3D faces [314]. The main insight behind this approach is assuming that any face can be constructed using a linear combination of M registered face models. A face is represented by a vector $\mathbf{s}^{(o)} \in \mathbb{R}^{3N}$ and a vector $\mathbf{t}^o \in \mathbb{R}^{3N}$, containing the x, y and z components of the shape, and the RGB color information, respectively. N is the number of mesh vertices. Considering the correlation between the components of $\mathbf{s}^o$, each face is actually represented in a more compact version using the principal components (PC) of the shape and texture space, denoted by $\mathbf{s}$ and $\mathbf{t}$, respectively. Given a set of shape exemplars $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_M\}$ and texture exemplars $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_M\}$, a new fitted model ($\mathbf{s}_f$,$\mathbf{t}_f$) is expressed as:

$$\mathbf{s}_f = \sum_{i=1}^{M} \alpha_i.\mathbf{s}_i + \sum_{j=1}^{K} \lambda_j.\mathbf{b}_j \qquad \mathbf{t}_f = \sum_{i=1}^{M} \beta_i.\mathbf{t}_i, \quad (5.10)$$

where α and β are vectors with the weights assigned to each exemplar, whereas $\mathbf{B} = \{\mathbf{b}_1, ..., \mathbf{b}_K\}$ is a set of deviations components of K different facial expressions and λ is the vector with the weight of each expression. Given an input image $\mathbf{I}$ and a set of landmarks $\mathbf{q}$, the variables α, β and λ are estimated by minimizing the following energy function:

$$E = \sum_{(x,y) \in \mathbb{N}^2} |\mathbf{I}(x,y) - \mathbf{I}_f(x,y)| + \sum_{k=1} |\mathbf{q}_k - \mathbf{p}_k|, \quad (5.11)$$

where $\mathbf{I}_f$ is the image obtained by projecting the fitted model, and $\mathbf{p}_k$ is the projected position of the vertex corresponding to the k^{th} landmark.

Inferring a 3D surface from a face image is an ill-posed problem, and thus, different

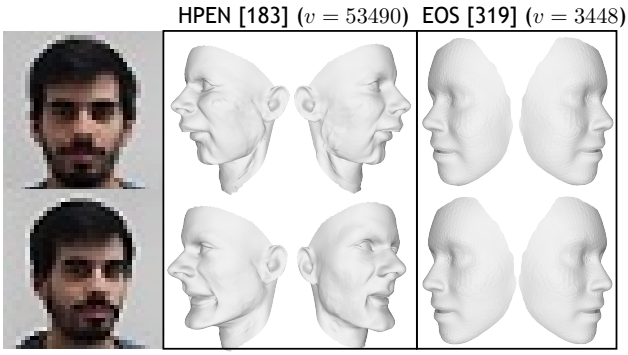


Figure 5.13: Examples of 3D models obtained by different 3DMM methods in low-resolution data. The use of low-resolution data hinders the process of recovering the latent parameters of the 3D model, particularly when using dense models. The comparison between HPEN (a common 3DMM method coupled with a dense model) and EOS (3DMM particularly adapted for low-resolution data coupled with a sparse model) evidences two major drawbacks of the first approach: 1) the models do not correspond to the face structure of the subject; and 2) they are not consistent in data of the same individual.

constraints have been proposed to ease the energy minimization (e.g., features extracted from the input image [315] and facial symmetry [316]). Despite these modifications, the energy function E remains highly non-convex in low-resolution images, and thus, it is unfeasible to build a statistical model of the facial texture that generalizes well in wild data and is, at the same time, in correspondence with the shape model. Accordingly, we opt for using a 3D face reconstruction methodology that relies solely on fitting a statistical 3D facial shape prior on a sparse set of landmarks [317–319]. Moreover, we use the Surrey Face Model [319] (a sparse 3D model with 3,448 vertices) instead of the commonly used Basel Face Model [320] with 53,490 vertices. The rationale for this choice is twofold: 1) images acquired in totally unconstrained scenarios increase the likelihood of spurious landmarks, which in turn may induce aberrations in isolated parts of models with many degrees of freedom (as illustrated in figure 5.13); and 2) the computational cost of the minimization algorithm increases significantly with number of vertices, and the proposed method is intended for real-time applications.

5.3.2.3 Exaggeration Inference

The correct assessment of which facial features should be exaggerated is the key for drawing recognizable caricatures. This process occurs internally in the human brain and is commonly accepted that is guided by the comparison to a reference model [321] or average model [322], which in our case is the Surrey Face Model.

Let $\boldsymbol{\pi} = [x, y, \theta, s]$ be the attributes of a face region, where (x, y) is the mass center in the frontal model version, θ is the region orientation in the xy-plane and s the region size. $\mathbf{r} = [\boldsymbol{\pi}^{(1)}, ..., \boldsymbol{\pi}^{(n)}]$ is the vector obtained by concatenating the attributes of n face regions, whereas $\mathbf{r}' = \mathbf{r} - \mathbf{r}^{(f)}$ is the element-wise difference between $\mathbf{r}$ and the regions of the reference model $\mathbf{r}^{(f)}$. The difference operator $\ominus$ between regions is defined as:

$$\boldsymbol{\pi}^{(1)} \ominus \boldsymbol{\pi}^{(2)} = \{x^{(1)} - x^{(2)}, y^{(1)} - y^{(2)}, \theta^{(1)} - \theta^{(2)}, \frac{s^{(1)}}{s^{(2)}}\}. \quad (5.12)$$

The vector $\mathbf{r}'$ is a compact description of the differences between a face and a reference model. However, each component is derived from attributes with distinct scales and variances.

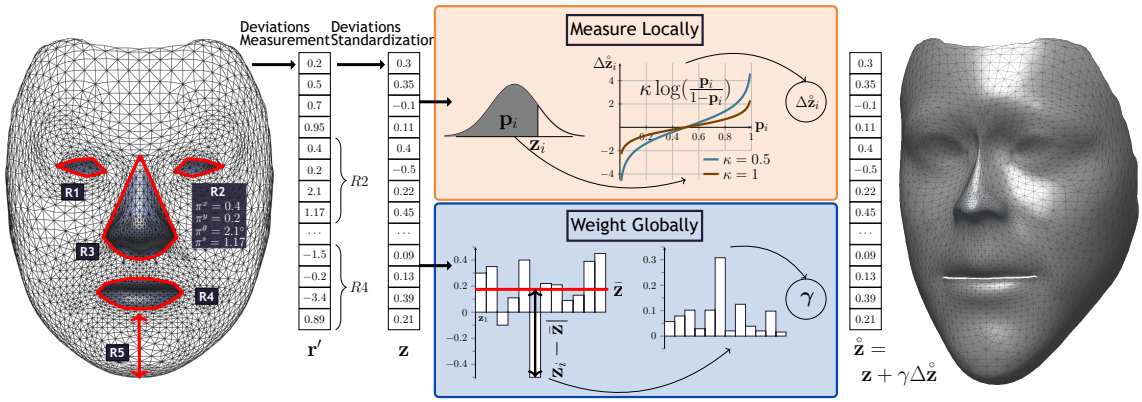


Figure 5.14: Schematic representation of the exaggeration inference phase. The key for drawing recognizable caricatures is the correct assessment of the exaggeration degree that should be applied to each facial feature. Aiming at replicating the internal brain process that guides caricature drawing, we proceed by measuring the differences between the attributes of the inferred model and a reference model, followed by standardizing these deviations using z-score normalization. The normalized deviations are subsequently deformed using a 'measure locally, weight globally' strategy, allowing to determine the exaggeration degree of each attribute not only by its the individual deviation but also from its global importance in the face context.

As such, we normalize $\mathbf{r}'$ using the standard score:

$$\mathbf{z}_i = \frac{\mathbf{r'}_i - \boldsymbol{\mu}_i}{\boldsymbol{\sigma}_i}, \quad (5.13)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are the sample mean and sample standard deviation of the i^{th} attribute (estimated from the training data). This normalization provides a comparable description of how each attribute deviates from the mean.

We believe that a very similar representation is inferred internally by caricaturists, and that they exploit it for emphasizing the most distinguishable features of the whole face in a holistic manner, i.e., determine the exaggeration degree of each feature not only by its the individual deviation but also from its global importance in the face context. Inspired by this observation, we introduce a two-step process for inferring the exaggeration degree of the normalized deviation of each attribute.

The proposed inference strategy works in a 'measure locally, weight globally' manner. In the first step, the maximum displacement in the normalized space $\Delta\mathring{\mathbf{z}}$ is individually determined by applying a transfer function to the cumulative probability of $\mathbf{z}_i$ (denoted by $\Phi_{\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i}(\mathbf{z}_i)$):

$$\Delta\mathring{\mathbf{z}}_i = \kappa \log\left(\frac{\Phi_{\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i}(\mathbf{z}_i)}{1 - \Phi_{\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i}(\mathbf{z}_i)}\right) - \mathbf{z}_i, \quad (5.14)$$

where κ is a parameter controlling the level of exaggeration applied to each attribute.

In the second step, the relative importance of each attribute is determined by measuring the absolute distance of $\mathbf{z}_i$ to the mean of the observed attributes ($\bar{\mathbf{z}_i}$), and the weight of each attribute is given by:

$$\gamma_i = \frac{|\mathbf{z}_i - \bar{\mathbf{z}_i}|}{\sum_{i=1} |\mathbf{z}_i - \bar{\mathbf{z}_i}|}. \quad (5.15)$$

Both steps are then combined to produce the deformed deviation in the normalized space:

$$\mathring{\mathbf{z}}_i = \mathbf{z}_i + \gamma_i \Delta \mathring{\mathbf{z}}_i. \quad (5.16)$$

Figure 5.14 provides a summary description of the proposed deformation inference process.

Recovering the regions attributes from the deformed deviations $\mathring{\mathbf{z}}$ is attained by reversing the normalization process:

$$\mathring{\mathbf{r}}_i = (\mathring{\mathbf{z}}_i . \boldsymbol{\sigma}_i + \boldsymbol{\mu}_i) \oplus \mathbf{r}_i^{(f)}, \quad (5.17)$$

where $\oplus$ is the sum operator between regions. At the end, the 3D position of the region vertices is adjusted to comply with new region properties, i.e., regarding $\pi^{(x)}$, $\pi^{(y)}$, and $\pi^{(\theta)}$ the vertices are simply translated or rotated, whereas for $\pi^{(s)}$ the position of each vertex is adjusted by the vector $\pi^s(\mathbf{v}_i - \mathbf{v}_i^{(f)})$, being $\mathbf{v}_i$ and $\mathbf{v}_i^{(f)}$ the vertices of the i^{th} region in the observed and reference model, respectively.

5.3.2.4 3D Model Deformation

Given the updated positions of the face regions vertices, it is necessary to deform the mesh to satisfy these constraints. The deformation applied should, at the same time, comply with the constraints and preserve local details, i.e., produce smooth deformations by varying the position of each vertex with respect to its neighbors. Laplacian mesh editing [323, 324] is a classical algorithm to address this problem. It represents vertices with respect to its neighbors using differential coordinates, denoted as Laplacian coordinates. The Laplacian coordinates are defined as $\mathscr{L}(\mathbf{v}_i) = \mathbf{v}_i - \frac{1}{d_i}\sum_{j \in N_i} \mathbf{v}_j$, where d_i is the degree of the i^{th} vertex and N_i is the set of neighbors of the i^{th} vertex. Given a subset of vertices C and their updated positions $\mathbf{u}$ (user constraints), the Laplacian mesh editing algorithm determines the deformed vertex positions $\hat{\mathbf{v}}$ by minimizing the following function:

$$\hat{\mathbf{v}} = \arg\min_{\mathbf{v}'} \sum_{i=1}^{n} ||\mathscr{L}(\mathbf{v}_i) - \mathscr{L}(\mathbf{v}'_i)||^2 + \sum_{i \in C} ||\mathbf{v}'_i - \mathbf{u}_i||. \quad (5.18)$$

The model given in the right side of figure 5.14 depicts a deformed mesh obtained with the Laplacian mesh editing algorithm, where can be observed the smoothness of the deformation.

5.3.2.5 Caricature Synthesis

The synthesis of the 2D caricature in the original pose is achieved by projecting each vertex with the camera parameters previously determined in the 3DMM phase. For maintaining the original image resolution, we adapt the number of vertices using interpolation.

5.3.2.6 Feature Encoding and Matching

After generating the caricature from a veridical photo, the VGG-Face architecture is trained to identify individuals from caricatures (refer to section 5.3.3.2 for further details). Feature encoding is attained using the learned filters, and caricatures are described by the

Figure 5.15: Examples of the data sets used in the empirical validation of the proposed face recognition method. The upper row regards the LFW data set, whereas the bottom rows are from the IJB-A and MegaFace sets, respectively.

4096-dimensional features produced in the 'fc6' layer of the VGG-Face architecture. During the matching phase, the L2 distance between the descriptors is used as dissimilarity score.

5.3.3 Results and Discussion

Five well-known data sets were selected for our experimental evaluation. The Annotated Facial Landmarks in the Wild (AFLW) [325] set was used to evaluate the results of the face alignment phase. The VGG dataset [185] was chosen for its large quantity and diversity of face images (more than 2M images from 2622 celebrities), providing an excellent tool for tuning a CNN to the task of face recognition. Finally, the LFW [289], IJB-A [219] and MegaFace datasets were used for assessing the performance of our approach in data acquired in the wild. All these sets comprise images of celebrities, except for AFLW and Megaface, which contain images of Flickr users. Figure 5.15 shows some images from the data sets considered for performance evaluation.

In addition, this section describes the modifications performed in the proposed approach for generating the caricature version of the VGG dataset in less than 4 days, and details the parameters used for training the VGG-Face network.

5.3.3.1 Facial Alignment

The AFLW dataset has 25,993 color images, each one annotated with a 21-point markup on visibility. This set was used to compare the performance of the ensemble learning strategy introduced in section 5.3.2.1 with the individual performance of four state-of-the-art face alignment methods [254, 311–313]. These methods are compliant with the popular 68 landmark format [326], while AFLW only provides a maximum of 21 landmarks depending on visibility. For evaluation, we selected a subset of 11 landmarks that share the same semantic positions in the two formats. Also, we considered exclusively samples with pose angles in the intervals yaw $\pm\frac{\pi}{4}$, pitch $\pm\frac{\pi}{2}$ and roll $\pm\frac{\pi}{5}$, according to the plausibility of observing such poses in visual surveillance scenarios. In accordance with the standard evaluation protocol [327], the average point-to-point Euclidean distance normalized by the inter-ocular distance was used as error metric, and the overall accuracy is reported by the cumulative errors distribution curve in figure 5.16. The comparative performance between the ensemble strategy and the best performing approach shows an increase of 5% in the proportion of images with an inter-ocular normalized error less

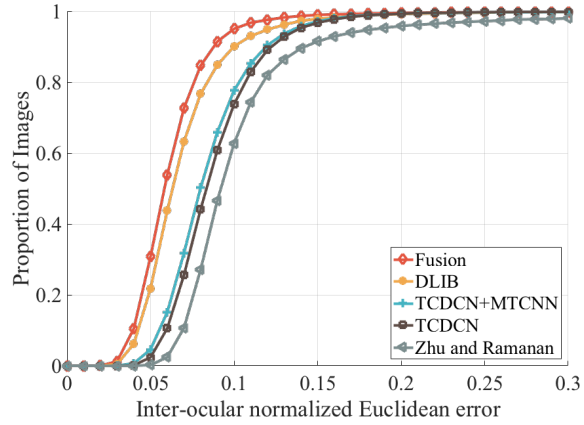


Figure 5.16: Cumulative error distribution curves for a subset of the AFLW dataset. Four state-of-the-art facial alignment methods (DLIB [311],TCDCN [313],TCDCN+MTCNN [312] and Zhu and Ramanan [254]) and their fusion were evaluated in AFLW for evidencing the advantages of combining their results with an ensemble learning strategy.

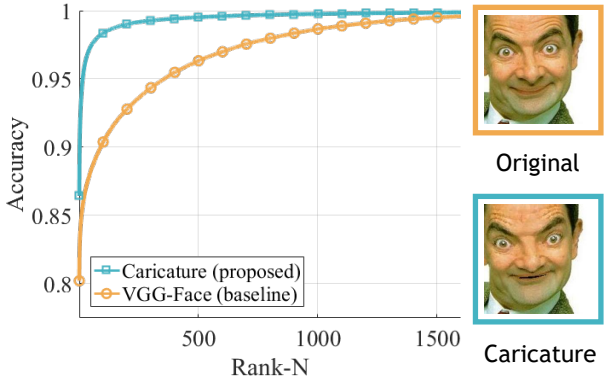


Figure 5.17: Comparison between the performance of the VGG-Face network trained on veridical images and on caricatures. The improvements in the CMC curve of the network trained on caricatures evidence the benefits of using this representation for automated face recognition. This improvement is justified by the fact that caricatures enhance the distinctive features of subjects, easing the recognition task.

than 10%. Even though these improvements seem irrelevant, they represent a significant decrease in the number of distorted caricatures caused by spurious landmarks, which in turn ease CNN training and improve the performance of the network.

5.3.3.2 CNN Training

This section details the architecture, the parameters and the training data chosen for optimizing the network to the caricature recognition task. Our goal is to show that, similarly to humans, a CNN attains higher face recognition rates if trained with caricatures than with the original face images. The approach of Parkhi et al. [185] has been found particularly useful for this endeavor because of two major reasons: 1) the authors show that is possible to obtain state-of-the-art face recognition results on different datasets solely by training from scratch a CNN with millions of images automatically retrieved from the web; and 2) the network architecture (the VGG-Face) and the set of images used for training (the VGG dataset) are publicly available, allowing to replicate the experiments of [185], and measure the performance gap between the use of veridical photos and caricatures.

Accordingly, we trained the VGG-Face architecture from scratch on two distinct types of

Table 5.2: Training configuration used for adjusting the weights of the CNN from scratch.

• Batch-size	64
• Momentum	0.9
• Weight-decay	5×10^{-4}
• Dropout Rate	0.5
• Learning Rate	10^{-2}
• Weight Initialization	$X \sim \mathcal{N}(0,\, 10^{-4})$

data: 1) original images of the VGG dataset (baseline); and 2) caricature images of the VGG dataset. The original VGG set contained 2.6M images (2622 identities with 1000 images), but at the time of our experiments, only 2.1M images were available on the web. Next, 90% of the images of each subject were randomly selected for training and the remaining were kept aside for performance evaluation. The configuration used and the regularization parameters for model optimization are described in table 5.2. For augmenting training data, a 224×224 pixel patch was randomly cropped from the image and horizontal flipping was applied with 50% probability. The model was implemented in the MATLAB toolbox MatConvNet and linked against the NVIDIA CuDNN libraries to accelerate training. All the experiments were carried on a NVIDIA Titan X Graphics Processing Unit (GPU) with 12GB of onboard memory, and each epoch took about 13h to run.

The comparative performance obtained by evaluating the trained models in 10% of the VGG set is depicted in figure 5.17. The results evidence the benefits of using caricatures for automated face recognition, and we argue that this improvement is justified by the fact that caricatures enhance the distinctive features of the subject, easing the recognition task. As an example, figure 5.17 also provides the two representations of an identity of the VGG set, where it is easier to identify the well-known actor Rowan Atkinson by its caricature than by its veridical image.

5.3.3.3 Running Time

The average running time of the caricature generation phase is a crucial variable for two major reasons: 1) evaluating the applicability of the proposed method in a real-time system; and 2) determining the time required for generating the caricatures of the training set, which can be prohibitive in the case of VGG dataset (2.1M images). The extensive processing chain and the use of off-the-shelf implementations affect substantially the processing time, and, as such, some phases of the proposed approach were modified either by using approximations or memoization.

In the 3DMM phase, the maximum number of iterations for inferring the 3D model was changed from 50 to 5, as we noticed marginal differences in the obtained models. Regarding model deformation, the off-the-shelf implementation of the Laplacian mesh editing algorithm was optimized with memoization, while caricature synthesis was speedup by using triangulation hierarchy during texture rendering. Table 5.3 provides a comparison between the original and optimized average running time of each phase on a single core of an i7-4790 CPU, as well as, the total running time per image and the total time required for processing the whole VGG set. The results show that, after the optimization, training images can be generated in few days by distributing the data into multiple computers and exploit all the cores of the CPU.

Table 5.3: Comparison between the original and optimized running time of the phases of the proposed caricature generation method.

Phase	Running Time (ms/img)	
	Original	Optimized
Face Alignment	160 ± 120	160 ± 120
3DMM	800 ± 93	200 ± 52
Face Analysis	95 ± 26	95 ± 26
3D Model Deformation	2 500 ± 155	500 ± 120
Image Synthesis	740 ± 37	330 ± 16
Total	4295 ± 431	1290 ± 334
VGG	≈ 114 days	≈ 31 days

5.3.3.4 Face Recognition Performance: Improvements due to Caricatures

To assess the performance of the proposed approach in highly unconstrained data, three state-of-the-art face recognition datasets were used, namely the LFW [289], IJB-A [219], and MegaFace [290]. The rationale for using multiple sets was twofold: 1) ensuring a non-biased evaluation of face recognition in the wild (the particularities of a single set could inadvertently overestimate the recognition rate of the proposed method); and 2) showing that the proposed approach can cope with large variations in data.

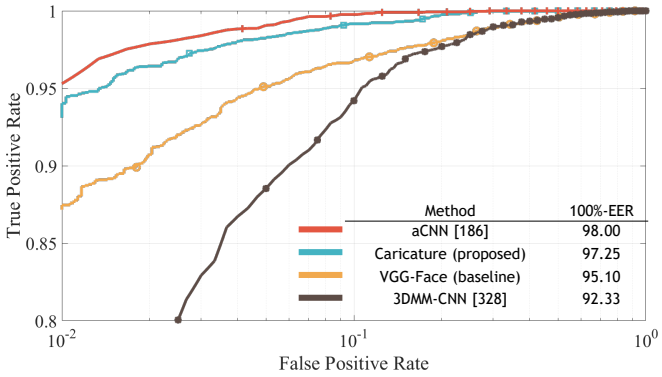


Figure 5.18: Face verification performance for the LFW dataset. The ROC curves of caricature-based face recognition (our method) and original image face recognition (baseline) show the advantages of using this representation. Also, the results show that the obtained performance is competitive with state-of-the-art algorithms.

During encoding, the metadata of the evaluation sets were used to crop each probe image to a 256x256 sub-image containing the facial region and maintaining aspect ratio. Then, five patches of 224x224 pixels were sampled from each face image (from the four corners and center), and each region was duplicated with horizontal flipping. The ten resulting patches were subsequently input to the network and the obtained descriptors were averaged to produce the face descriptor of the probe image.

Experiments on the LFW dataset. LFW is a de facto benchmark for evaluating face recognition in the wild, comprising 13,233 images from 5,749 subjects. The evaluation protocol provides 3000 pairs of images organized in 10 splits for assessing the verification performance of face recognition algorithms. Also, each method should report the results under a specific setting with respect to the type of training data used. In our case, even though the descriptor obtained

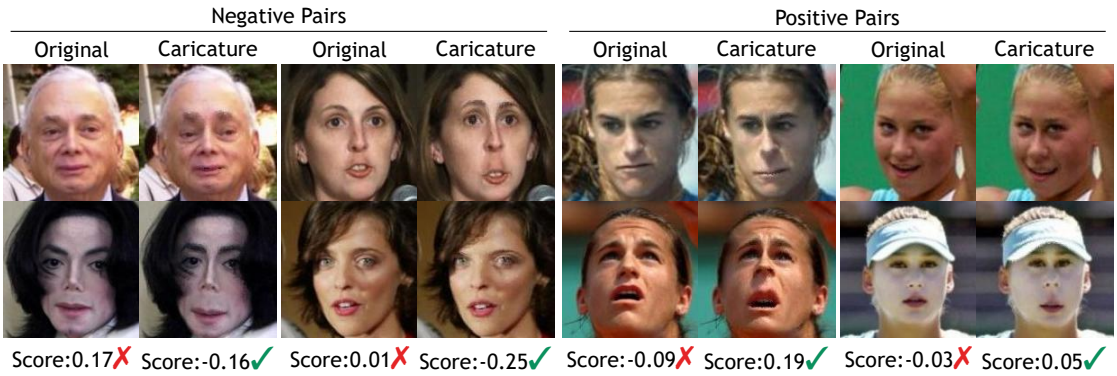


Figure 5.19: Successful cases of the proposed approach. The advantages of using caricatures for face recognition are represented by four pairs of the LFW and IJB-A sets where our approach produced a correct score, while the use of veridical photos produced an incorrect output.

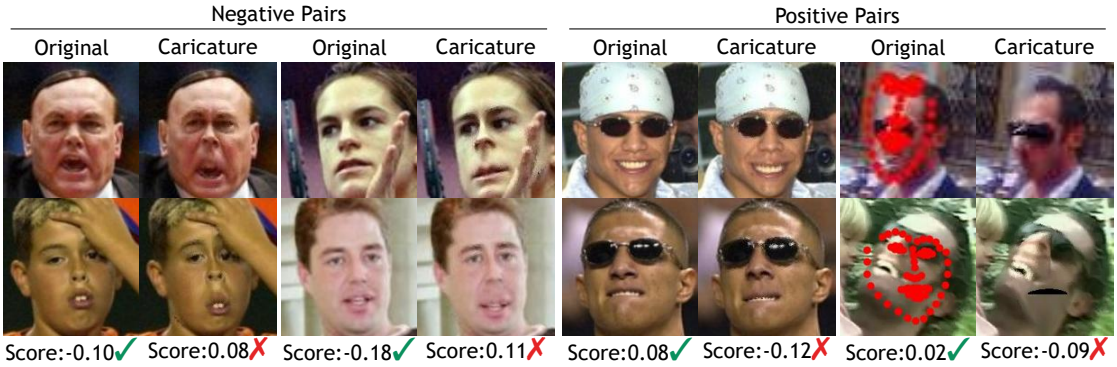


Figure 5.20: Failure cases of the proposed approach. The major causes of failure in the LFW and IJB-A sets are represented by four pairs where our approach produced an incorrect score, while the use of veridical photos produced a correct output. Occlusions, facial expressions and the failure of facial alignment are the major reasons for incorrect caricature generation.

is not tuned in the LFW training data, we should report our results under the 'unrestricted with labeled outside data' setting due to the use of the VGG dataset during model training.

Regarding the comparison with state-of-the-art methods, the works of Tran et al. [328] and Masi et al. [186] were selected for sharing similarities with our approach. In [186], the 3D face structure was inferred from a single 2D image to augment the number of training samples by rendering the original face in a distinct pose, shape and expression. In [328], the authors introduced a regression network for estimating the 3D structure from a single 2D image and used this representation for face recognition.

Results are summarized in the ROC curves of figure 5.18. When compared to the baseline, our approach achieved a significant decrease in the EER, supporting the claim that automated face recognition benefits from the use of caricatures (refer to figure 5.19 for some examples). Regarding the comparison to similar approaches, our approach performed significantly better than 3DMM-CNN [328], while it produced competing results with respect to aCNN [186]. The results of [328] suggest that texture plays a decisive role in the face recognition task, while the comparison between the performance of our method with [186] indicates that the emphasis of distinctive facial features is as effective as generating multiple views of the original image.

Table 5.4: Summary of the face recognition performance on IJB-A.

Method	Trained on IJB-A	Verification		Identification	
		FAR 0.1	FAR 0.01	Rank-1	Rank-5
GOTS [219]	Yes	62.7 ±1.2	40.6 ±1.4	44.3 ±2.1	59.5 ±2.0
OpenBR [329]	Yes	43.3 ±0.6	23.6 ±0.9	24.6 ±1.1	37.5 ±0.8
Wang et al. [330]	Yes	89.5 ±1.3	73.3 ±3.4	82.0 ±2.4	92.9 ±1.3
Chen et al. [331]	Yes	96.7 ±0.9	83.8 ±4.2	90.3 ±1.2	96.5 ±0.8
aCNN [186]	Yes	88.6	72.5	90.6	96.2
VGG-Face [185]	No	85.4 ±1.2	61.1 ±2.3	87.6 ±1.6	92.8 ±0.9
Caricature	**No**	**86.0** ±1.4	**63.5** ±2.7	**88.9** ±1.1	**94.1** ±0.7

Experiments on the IJB-A dataset. The IJB-A dataset represents an advance over LFW, by comprising data with a wider range of variations, particularly in pose. It contains 500 subjects with 5,397 images and 2,042 videos split into 20,412 frames, 11.4 images and 4.2 videos per subject. Regarding the evaluation protocol, it differs from LFW by considering template-to-template comparisons rather than image-to-image comparisons, where each template contains a combination of images or frames sampled from multiple image sets or videos of a subject. Algorithms can be evaluated in the verification (1:1 matching) or identification (1:N search) protocol over 10 splits. In the verification protocol, each split contains around 11,700 pairs of templates (15% positive and 85% negative pairs) on average, whereas the identification protocol also consists of 10 splits, each containing about 112 gallery templates and 1763 probe templates. During evaluation, each template is described by the average of image descriptors. Table 5.4 reports the performance of the baseline, the proposed approach, and competing approaches with respect to the standard accuracy metrics of IJB-A.

Regarding the comparison with the baseline, the improvements of our approach were not statistically significant, contrasting with the performance increase attained in LFW. In our view, the principal cause for this outcome was the failure of facial alignment, rather than the ineffectiveness of caricatures. The particularities of IJB-A (extreme variations in pose, face resolution, and illumination) affect significantly the accuracy of the landmark detector, which in turn distorts the generated caricature (figure 5.20 depicts some examples).

With respect to the comparison to other approaches, our method outperformed the baselines of IJB-A (GOTS and OpenBR), but it fell behind the remaining state-of-the-art face recognition methods. However, it should be stressed that, unlike the other approaches, no particular effort was made to optimize our method for this dataset (e.g., fine-tuning with IJB-A training data), since our major concern is measuring the relative performance between caricatures and original images.

Experiments on the MegaFace dataset. MegaFace [290] is a recent and very challenging dataset for evaluating face recognition at scale. The gallery set comprises more than 1 million images from 690K different individuals, while the probe set was sampled from the FaceScrub dataset. The evaluation protocol provides code for testing algorithms in the verification and identification scenarios.

Figure 5.21 compares the verification and identification performance of our approach with the baseline, and with commercial systems that took part of the MegaFace challenge, whereas table 5.5 summarizes the methods performance according to the standard metrics of

Table 5.5: Summary of the face recognition performance on MegaFace with 1M distractors.

Method	Rank-1	TAR@FAR=10^{-6}
Vocord-DeepVo1	75.13	67.32
NTechLAB-facenx	73.30	85.08
Shanghai Tech	74.05	86.34
Google-FaceNet v8	70.50	86.47
Beijing FaceAll-Norm-1600	64.80	67.12
SIAT-MMLAB	65.23	76.72
Barebones FR	59.36	59.04
VGG-Face (baseline)	75.08	74.78
Caricature (proposed)	**75.10**	**76.49**

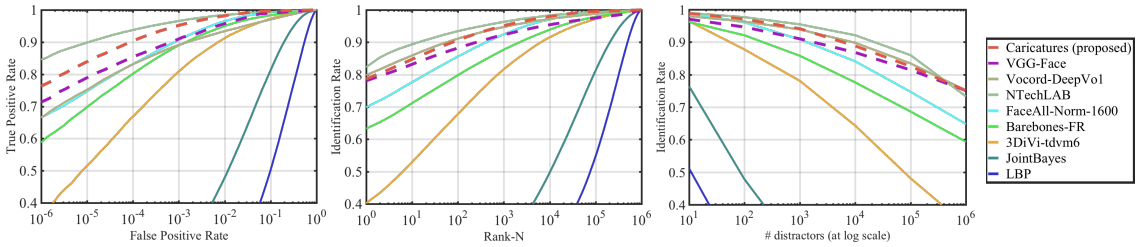


Figure 5.21: Face recognition performance on MegaFace. Left: Verification performance on MegaFace with 1M distractors. Middle: Identification performance on MegaFace with 1M distractors. Right: Rank-1 performance as a function of the number of distractors on the probe set.

the dataset. The results show that the proposed approach outperformed the baseline on both recognition settings when using 1M distractors. Improvements in performance are also observed with a variable number of distractors, as evidenced in the right plot of figure 5.21. Regarding the comparison with commercial approaches and the baseline methods of MegaFace (LBP [171] and Joint Bayes), it is interesting to note that, in the majority of the cases, caricature-based face recognition attained better performance than these systems.

5.3.4 Conclusion

In this section, we introduced the first fully automated caricature-based face recognition system capable of working in real-time with data acquired in the wild. A 3DMM method coupled with a set of automatically detected facial landmarks was used for inferring the 3D face structure of probe images. Next, the inferred model was compared to a reference prototype for determining the divergence between facial regions, and the exaggeration applied to each region was determined by a 'measure locally, weight globally' strategy. The modified regions were given as constraints to a Laplacian mesh editing algorithm for deforming the original mesh, and the 2D caricature was obtained by projecting the deformed model in the original camera-view. During the learning phase, the VGG-Face architecture was trained from scratch on 2.1M caricatures automatically generated from the VGG dataset, whereas classification was performed with the features from the 'fc6' layer.

To assess the advantages of using caricatures for automated face recognition, we used three state-of-the-art face recognition datasets for measuring the relative performance between our approach and the VGG-Face trained on the original images of the VGG dataset. The results revealed significant improvements in the recognition performance when using caricatures

rather than veridical images, confirming the usefulness of using caricature-based face recognition. Regarding the comparison with state-of-the-art methods, our approach was capable of obtaining competitive results even without being particularly tuned for any of the evaluation sets. Nevertheless, it should be noted that our goal was not to attain the best performing results on these sets, but measure the performance gap between the use of caricatures and veridical images.

5.4 Summary

This chapter was devoted to the description and experimental evaluation of the contributions made for improving the performance biometric recognition in surveillance scenarios.

Before attempting the development of novel recognition algorithms to deal with the degradation factors of the data, we organized the ICB-RW competition in order to gauge the performance of state-of-the-art approaches and perceive the major causes for the algorithms failure. Also, this competition provided important insights for the development of our biometric recognition approaches.

Next, we described our two proposals. The first approach regards an algorithm for detecting degraded features in biometric signatures, which later constraints the features that are taken into account in the matching phase. The second proposal is a caricature-based face recognition approach for mimicking the caricature drawing process of caricaturists and producing 2D representations where inter-subject similarity is minimized and intra-subject similarity is preserved. We concluded that both proposals increase the robustness of biometric recognition to the degradation factors inherent to surveillance environments. However, it should be noted that these results are dependent on the successful acquisition of biometric samples, and, even though this issue has been addressed in the previous chapters, there is room for improvement in the acquisition rate of the proposed surveillance system.

In short, we are convinced that the proposals described in this chapter contribute to extend the range of biometric recognition to more challenging data, and thus they constitute a step towards biometric recognition in the wild.

Chapter 6

Conclusion and Future Work

Throughout this thesis, several contributions were made to enable the acquisition of biometric samples in unconstrained scenarios and to push forward the performance of biometric recognition in these data. This chapter presents the main conclusions that resulted from the research work performed during the 4 years of this doctoral program. Furthermore, it provides directions for further research topics.

6.1 Conclusion

The major goal of this thesis was the development of an automated surveillance system capable of recognizing subjects at a distance and without their cooperation in the acquisition process. As described along this thesis, the successful development of such a system requires the knowledge of three distinct research domains, but the scope of this thesis was limited to two of these areas: surveillance systems and biometric recognition. Nevertheless, we believe that the contributions presented represent a step-forward towards biometric recognition in surveillance scenarios.

Regarding the surveillance area, we introduced a novel calibration algorithm for master-slave surveillance systems that avoids the use of extra optical devices and stringent configurations between cameras. As a consequence, the acquisition of high-resolution biometric data can be performed outdoors and at a wider range of distances. The second significant contribution was a camera scheduling approach that minimizes the cumulative transition time when imaging multiple subjects. This method is of particular importance for ensuring that the biometric data of each subject are acquired at least once. These contributions were then used in combination with the state-of-the-art approaches of the human monitoring field to develop a fully automated surveillance capable of acquiring biometric data at a distance and without human cooperation, designated as QUIS-CAMPI system.

The analysis of the state-of-the-art biometric databases showed that these sets do not faithfully represent the complete set of degradation factors of surveillance scenarios. For this reason, we introduced the QUIS-CAMPI dataset, comprising both full body video sequences and high-resolution face images acquired by the QUIS-CAMPI system from non-cooperative subjects in a real surveillance scenario. The acquisition process ensures that data truly encompass the covariates of surveillance environments. Moreover, this set provides multiple biometric traits in the enrollment set (full-body images, gait sequences, 3D model of the face), which makes it particularly useful for assessing the performance of biometric recognition methods in completely unconstrained environments.

In the biometrics field, we organized the ICB-RW competition, which was particularly important to assess the performance of the state of the art in biometric recognition in totally unconstrained scenarios, as well as also to provide insight about the most adequate strategies for addressing the typical degradation factors of these scenarios. Second, we devised a method for detecting corrupted features in biometric signatures. This method works at the final stage

of the biometric systems processing chain, i.e. the feature matching phase, and for that reason this approach is not specific to a particular trait, which is regarded as a strong point. The last contribution regards a caricature-based face recognition method capable of working with data acquired in the wild. Our approach is capable of creating 2D face representations, where likeness is preserved and the inter-class separation is enlarged. The advantages of our method are confirmed by the results obtained in three state-of-the-art datasets, where the recognition performance increased when using caricatures rather than veridical images.

To conclude, we believe that the contributions presented in the thesis contribute to bridge the gap between biometric recognition and surveillance. However, it should be stressed that, in spite of these achievements, the automated recognition of humans in surveillance scenarios is still to be accomplished. This is justified by the fact that the recognition rate is highly sensitive to the failure of any of the stages of the processing chain, and according to our experiments there is room for improvement in modules related to human monitoring in surveillance scenarios.

6.2 Future Work

We are currently attempting to increase the success rate of the biometric data acquisition process by studying novel strategies for detecting and tracking humans in surveillance scenarios. In spite of the successful acquisition of the QUIS-CAMPI dataset, we observed that the performance of state-of-the-art human detectors and tracking algorithms decreases significantly in surveillance scenarios, particularly in crowded scenes.

Also, we are interested in finding innovative solutions to address some particularities of the implementation of the QUIS-CAMPI prototype, which affect significantly the quality of biometric data acquired, and in turn the recognition accuracy. As an example, we aim at investigating alternative solutions for obtaining focused images from the PTZ camera, and improving the accuracy of human path prediction by automatically analyzing the constraints of the scene, i.e., perceive the locations where is unlikely to observe subjects. These improvements would significantly improve the rate of acquisitions and the data quality. This avoids the development of recognition methods capable of addressing these issues, which is a more challenging problem than these improvements.

Appendix A

Informed consent for obtaining the subjects permission to acquire biometric samples

UNIVERSIDADE da BEIRA INTERIOR
Departamento de Informática

Consentimento para recolha de dados biométricos

No âmbito do projeto de investigação ***QUIS-CAMPI*** solicitamos a vossa participação para a recolha de um conjunto de dados biométricos, utilizados no desenvolvimento de métodos de reconhecimento biométrico em ambientes não cooperativos.

A recolha de dados consiste em duas fases:

FASE 1: O **registo** é efectuado no laboratório SOCIA (Sala 6.12) do Pólo I da Universidade da Beira Interior, sendo guardados os seguintes dados:

- Informação biométrica auxiliar (exemplo: idade, sexo, altura, peso, cor da pele, características faciais, características do cabelo);
- Foto corpo completo;
- Imagens da face sob diferentes poses.
- Vídeos de sequência de passo.

FASE 2: O **reconhecimento** será feito com base em dados capturados a partir de câmaras de videovigilância instaladas no parque de estacionamento adjacente ao laboratório SOCIA.

A participação considera-se válida após recolha de dados em ambas as fases. Todos os dados recolhidos serão utilizados exclusivamente para efeitos de investigação, bem como serão partilhados com a comunidade científica para efeitos de comparação objectiva de resultados.

A equipa de investigação agradece a sua participação neste estudo. Obrigado!

Em caso de dúvidas contactar o responsável do estudo, Prof. Hugo Proença, através do e-mail hugomcp@di.ubi.pt. A recolha será realizada pelo Doutorando João Neves (jcneves@penhas.di.ubi.pt).

- -

Eu, _____, cartão de cidadão n.º _____ consinto a recolha de dados biométricos no âmbito do projecto QUIS-CAMPI, de acordo com o acima enunciado.

Universidade da Beira Interior, Covilhã, ____ de _____ de 20 ____

Assinatura

REF: ☐☐☐ _____

Other publications resulting from this doctoral research program not included in the thesis

Appendix B

Other publications resulting from this doctoral research program not included in the thesis

Segmenting the Periocular Region using a Hierarchical Graphical Model Fed by Texture / Shape Information and Geometrical Constraints*

Hugo Proença, João C. Neves and Gil Santos
IT - Instituto de Telecomunicações
University of Beira Interior, Portugal
{hugomcp,jcneves, gsantos}@di.ubi.pt

Abstract

Using the periocular region for biometric recognition is an interesting possibility: this area of the human body is highly discriminative among subjects and relatively stable in appearance. In this paper, the main idea is that improved solutions for defining the periocular region-of-interest and better pose / gaze estimates can be obtained by segmenting (labelling) all the components in the periocular vicinity. Accordingly, we describe an integrated algorithm for labelling the periocular region, that uses a unique model to discriminate between seven components in a single-shot*: iris, sclera, eyelashes, eyebrows, hair, skin and glasses. Our solution fuses texture / shape descriptors and geometrical constraints to feed a two-layered graphical model (Markov Random Field), which energy minimization provides a robust solution against uncontrolled lighting conditions and variations in subjects pose and gaze.*

1. Introduction

Motivated by the pioneering work of Park *et* al. [14], the concept of periocular recognition has been gaining relevance in the biometrics literature, particularly for uncontrolled data acquisition setups. For such cases, the idea is that - apart the iris - additional discriminating information can be obtained from the skin and sclera textures, and the shape of eyelids, eyelashes and eyebrows.

Most of the relevant periocular recognition algorithms work in a *holistic* way, i.e., they define a region-of-interest (ROI) around the eye and apply a feature encoding strategy independently of the biological component at each position. The exceptions (e.g., [17] and [6]) regard the iris and the sclera components, for which specific feature encoding / matching algorithms are used. This observation leads that some components (e.g., hair or glasses) might be erroneously taken into account and bias the recognition process.

The automatic labelling (segmentation) of the components in the periocular region has - at least - two obvious advantages: it enables to define better ROIs and conducts to more accurate estimates of subjects' pose and gaze. Hence, this paper describes an image labelling algorithm for the periocular region that discriminates between seven components (iris, sclera, eyelashes, eyebrows, hair, skin and glasses), according to a model composed of two phases:

1. seven non-linear classifiers running at the pixel level are inferred from a training set, and provide the posterior probabilities for each image position and class of interest. Each classifier (neural network) is specialized in detecting one component and receives local statistics (texture and shape descriptors) from the input data;

2. the posteriors based on data local *appearance* are combined with geometric constraints and components' adjacency priors, to feed a hierarchical Markov Random Field (MRF), composed of a *pixel* and a *component* layer. MRFs are a classical tool for various computer vision problems, from image segmentation (e.g., [10]), image registration (e.g., [8]) to object recognition (e.g., [5]). Among other advantages, they provide non-causal models with isotropic behavior and faithfully model a broad range of local dependencies. The model proposed in this paper inherits some insights from previous works that used shape priors to constraint the final model (e.g., [3]) and multiple layered MRFs (e.g., [19]).

To illustrate the usefulness of the proposed algorithm, we compare the effectiveness of the Park *et* al.'s [14] recognition method, when using the ROI as originally described and according to an improved version, that considers the center of mass of the cornea as reference point (less sensitive to gaze) and avoids that hair and glasses inside the ROI are considered in feature encoding / matching. The

*This work was supported by FCT project PEst-OE/EEI/LA0008/2013

observed improvements in performance anticipate other benefits that can be attained by labelling the periocular region before recognition: pose / gaze estimates based in the labelled data and development of component-specific feature encoding / matching strategies.

The remainder of this paper is organized as follows: Section 2 summarizes the most relevant periocular recognition algorithms. Section 3 provides a description of the proposed model. Section 4 regards the empirical evaluation and the corresponding results. Finally, the conclusions are given in Section 5.

2. Periocular Recognition: Literature Review

The first work in this field was published in 2009, due to Park *et* al. [14]. They characterised the periocular region by local binary patterns (LBP), histograms of oriented gradients (HOG) and scale-invariant feature transforms (SIFT), fused at the score level. Subsequently, the same authors [13] described additional factors that affect performance, including segmentation inaccuracies, partial occlusions and pose. Woodard *et* al. [20] observed that fusing the responses from periocular and iris recognition modules improves performance with respect to each system considered individually. Bharadwaj *et* al. [4] fused a global descriptor based on five perceptual dimensions (image naturalness, openness, roughness, expansion and ruggedness) to circular LBPs. The Chi-square distances from both types of features were finally fused at the score level. Ross *et* al. [16] handled challenging deformed samples, using probabilistic deformation models and maximum-a-posteriori estimation filters. Also concerned about robustness, Woodard *et* al. [21] represented the skin texture and color using separate features, that were fused in the final stage of the processing chain. Tan *et* al. [18] proposed a method that got the best performance in the *NICE: Noisy Iris Challenge Evaluation*[1]. contest. This method is actually a periocular recognition algorithm: texton histograms and semantic rules encode information from the surroundings of the eye, while ordinal measures and color histograms encode the iris data. Oh *et* al. [9] combined sclera and periocular features: directional periocular features were extracted by structured random projections, complemented by a binary representation of the sclera. Tan and Kumar [17] fused iris information (encoded by Log-Gabor filters) to an over-complete representation of the periocular region (LBP, GIST, HOG and Leung-Malik Filters). Both representations were matched independently and fused at the score level.

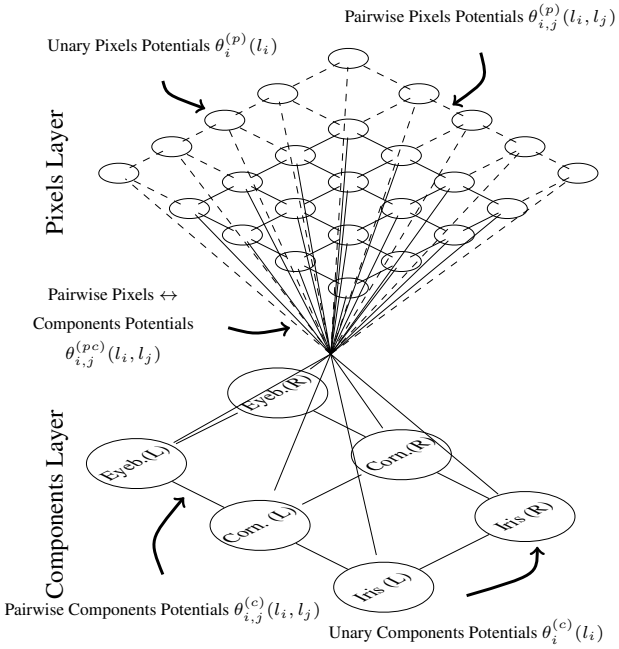


Figure 1. Structure of the MRF that segments the periocular region.

3. Proposed Method

As Fig. 1 illustrates, the proposed MRF is composed of two layers: one works at the *pixel* level, with a bijection between each image pixel and a vertex in the MRF. The second layer regards the major *components* in the periocular vicinity, with six vertices representing the eyebrows, irises and corneas from both sides of the face. The insight behind this structure is that the pixels layer mainly regards the data appearance, while the components layer represents the geometrical constraints in the problem and assures that the generated solutions are biologically plausible.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph representing a MRF, composed of a set of t_v vertices $\mathcal{V}$, linked by t_e edges $\mathcal{E}$. Let t_p be the number of vertices in the *pixels* layer and let t_c be the number of vertices in the *components* layer, such that $t_v = t_p + t_c$. Let $\mathfrak{C}(x, y)$ denote the biological component at position (x, y) of an image and $\mathfrak{T}_j$ be the component's *type* of the j^{th} component node: either 'iris', 'cornea' or 'eyebrow'.

The MRF is a representation of a discrete latent random variable $\boldsymbol{L} = \{L_i\}, \forall i \in \mathcal{V}$, where each element L_i takes one value l_i from a set of labels. Let $\boldsymbol{l} = \{l_1, \ldots, l_{t_p}, l_{t_p+1}, \ldots, l_{t_p+t_c}\}$ be one configuration of the MRF. In our model, every component node is directly con-

[1] http://nice2.di.ubi.pt/

nected to each pixel node and the pixel nodes are connected to their horizontal / vertical neighbors (4-connections). Also, the edges between component nodes correspond to geometrical / biological constraints in the periocular region: the nodes representing both irises, corneas and eyebrows are connected, as do the iris, cornea and eyebrow nodes of the same side of the face. Note that the proposed model does not use high-order potentials. Even though there is a point in Fig. 1 that joins multiple edges, it actually represents overlapped pairwise connections between one component and one pixel vertex.

The energy of a configuration $\boldsymbol{l}$ of the MRF is the sum of the unary $\theta_i(l_i)$ and pairwise $\theta_{i,j}(l_i, l_j)$ potentials:

$$E(\boldsymbol{l}) = \sum_{i \in \mathcal{V}} \theta_i(l_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{i,j}(l_i, l_j). \quad (1)$$

According to this formulation, labelling an image is equivalent to infer the random variables in the MRF by minimizing its energy:

$$\hat{\boldsymbol{l}} = \arg\min_{\boldsymbol{l}} E(\boldsymbol{l}), \quad (2)$$

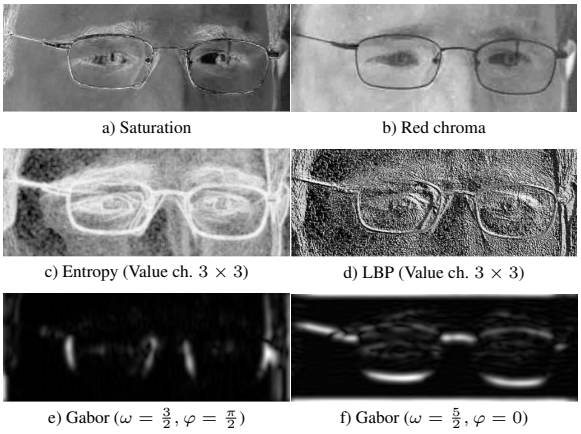
where $\{\hat{l}_1, \ldots, \hat{l}_{t_p}\}$ are the labels of the pixels and $\{\hat{l}_{t_p+1}, \ldots, \hat{l}_{t_p+t_c}\}$ specify the components' parameterizations. In this paper, the MRF was optimized according to the Loopy Belief Propagation [7] algorithm. Even though it is not guaranteed to converge to global minimums on loopy non-submodular graphs (such as our MRF), we concluded that the algorithm provides visually pleasant solutions most of the times. As future work, we plan to evaluate the effectiveness of our model according to more sophisticated energy minimization algorithms (e.g., sequential tree-reweighed message passing [11]).

3.1. Feature Extraction

Previous works reported that the hue and saturation channels of the HSV color space are particularly powerful to detect the sclera [15], whereas the red / blue chroma values provide good separability between the skin and non-skin pixels [1]. Also, the iris color triplets are typically distant from the remaining periocular components and there is a higher amount of information in patches of the eyebrows and hair regions than in the remaining components. Accordingly, a feature set at the pixel level is extracted, composed of 34 elements (Fig. 2): {red, green and blue channels (RGB); hue, saturation and value channels (HSV); red and blue chroma (yCbCr); LBP and entropy in the value channel}, all averaged in square patches of side $\{3, 5, 7\}$ around the central pixel. Also, the convolution between the value channel and a set of Gabor kernels $\boldsymbol{G}$ complements the feature set:

$$\boldsymbol{G}[x, y, \omega, \varphi, \sigma] = \exp\left[\frac{-x^2 - y^2}{\sigma^2}\right] \exp[2\pi\omega i \Phi] \quad (3)$$

being $\Phi = x\cos(\varphi) + y\sin(\varphi)$, ω the spatial frequency, φ the orientation and σ the standard deviation of an isotropic Gaussian kernel ($\omega \in \{\frac{3}{2}, \frac{5}{2}\}, \varphi \in \{0, \frac{\pi}{2}\}, \sigma = 0.65\omega$).



a) Saturation b) Red chroma

c) Entropy (Value ch. 3×3) d) LBP (Value ch. 3×3)

e) Gabor ($\omega = \frac{3}{2}, \varphi = \frac{\pi}{2}$) f) Gabor ($\omega = \frac{5}{2}, \varphi = 0$)

Figure 2. Illustration of the discriminating power of the features extracted, for the seven classes considered in this paper.

3.2. Unary Potentials

Let $\gamma : \mathbb{N}^2 \rightarrow \mathbb{R}^{34}$ be the feature extraction function, that for each image pixel (x, y) returns a feature vector $\gamma(x, y) \in \mathbb{R}^{34}$. Let $\Gamma = [\gamma(x_1, y_1), \ldots, \gamma(x_n, y_n)]^T$ be a $n \times 34$ matrix extracted from a training set, that is used to learn seven non-linear binary classification models, each one specialized in detecting a component (class) $\omega_i \in$ {Iris, Sclera, Eyebrows, Eyelashes, Hair, Skin, Glasses}. Let $\eta_i : \mathbb{R}^{34} \rightarrow [0, 1]$ be the response of the ith non-linear model, used to obtain the likelihood of class ω_i: $p\Big(\eta_i\big(\gamma(x, y)\big)|\omega_i\Big)$. According to the Bayes rule, assuming equal priors, the posterior probability functions are given by:

$$P\Big(\omega_i|\eta_i\big(\gamma(x, y)\big)\Big) = \frac{P\Big(\eta_i\big(\gamma(x, y)\big)|\omega_i\Big)}{\sum_{j=1}^{7} P\Big(\eta_j\big(\gamma(x, y)\big)|\omega_j\Big)}. \quad (4)$$

The unary potentials of each vertex in the pixels layer are defined as $\theta_i^{(p)}(l_i) = 1 - p\Big(\omega_i|\eta_i\big(\gamma(x, y)\big)\Big)$.

Each label in the components layer represents a parameterisation of an ellipse (found by the Random Elliptical Hough Transform (REHT)) [2] that roughly models the eyebrows, corneal or iris regions. Starting from images labelled by the index of the maximum posterior probability

$I_m(x,y) = \arg\max_j p\Big(\omega_j | \eta_j\big(\gamma(x,y)\big)\Big)$ (upper image in Fig. 3), a binary version per component can be obtained (bottom images in Fig. 3):

$$I_{m_i}(x,y) = \begin{cases} 1 & \text{, if } I_m(x,y) = i \\ 0 & \text{, otherwise} \end{cases} \quad (5)$$

The output of the REHT algorithm in $I_{m_i}(x,y)$ gives the unary potential of the component vertices: $\theta_i^{(c)}(l_i) = -\log\big(\kappa(i)\big), \forall i \in t_{p+1}, \ldots, t_{p+c}$, being $\kappa(i)$ the votes returned by the REHT for the ith ellipse parameterisation.

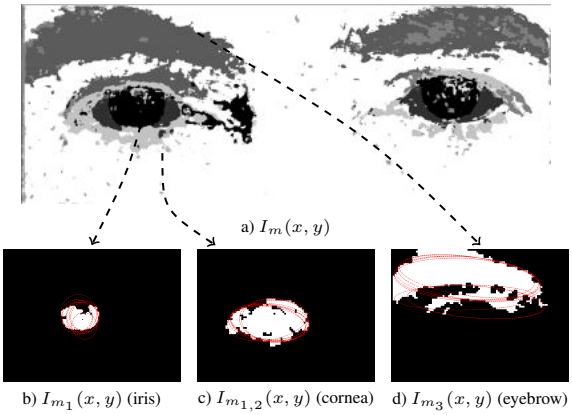


Figure 3. (Upper row) Example of an image labelled by the maximum of the posteriors given by the classification models $\eta_i\big(\gamma(x,y)\big)$. The red ellipses in the bottom images represent the parameterisations returned by the REHT algorithm for the left iris, cornea and eyebrow.

3.3. Pairwise Potentials

There are three types of pairwise potentials in our model: 1) between two pixel nodes; 2) between two component nodes; and 3) between a pixel and a component. The pairwise potential between pixel nodes spatially adjacent $\theta_{i,j}^{(p)}(l_i, l_j)$ is defined as the prior probability of observing labels l_i, l_j in adjacent positions of a training set (e.g., it is much more probable that an "eyebrow" pixel is adjacent to a "skin" pixel than to an "iris" one):

$$\theta_{i,j}^{(p)}(l_i, l_j) = \frac{1}{\alpha_0 + P(\mathfrak{C}(x',y') = \omega_i, \mathfrak{C}(x,y) = \omega_j)}, \quad (6)$$

where $P(.,.)$ is the joint probability, (x', y') and (x, y) are 4adjacent positions and $\alpha_0 \in \mathbb{R}^+$ avoids infinite costs (likewise, all α_i terms below are regularization terms).

The pairwise potentials between component nodes consider the geometrical constraints in the periocular area, i.e., enforce that the irises are inside the cornea, and below the eyebrows. Also, both irises, corneas and eyebrows should have similar vertical coordinate and similar size. Let $(x_i, y_i, a_i, b_i, \varphi_i)$ be the ith parameterisation of an ellipse, being (x_i, y_i) the ellipse centre, (a_i, b_i) its major / minor axes and φ_i the rotation. For pairs of nodes of the same type ($\mathfrak{T}_i = \mathfrak{T}_j$), similar vertical coordinates and similar sizes are privileged:

$$\theta_{i,j}^{(c1)}(l_i, l_j) = \alpha_1 |y_i - y_j| + \alpha_2 |a_i + b_i - a_j - b_j|. \quad (8)$$

For edges connecting the cornea (ith node) and the eyebrow (jth node) we privilege similar horizontal coordinates and locations having the eyebrow above the cornea:

$$\theta_{i,j}^{(c2)}(l_i, l_j) = \alpha_3 |x_i - x_j| + \alpha_4 \max\{0, y_i - y_j\}. \quad (9)$$

Regarding the iris / cornea pairwise potentials, we penalize parameterizations with portions of the iris outside the cornea:

$$\theta_{i,j}^{(c3)}(l_i, l_j) = \alpha_5\Big(1 - \frac{\sum_{x_i}\sum_{y_i} \psi(x_i, y_i, x_j, y_j, a_j, b_j, \varphi_j)}{\sum_{x_i}\sum_{y_i} 1}\Big), \quad (10)$$

being (x_i, y_i) a pixel labelled as iris and $\psi(x_i, y_i, x_j, y_j, a_j, b_j, \varphi_j)$ an indicator function that verifies if that position is inside the ellipse defined by the jth parameterisation (7). Overall, the pairwise potentials in the components layer are defined as:

$$\theta_{i,j}^{(c)}(l_i, l_j) = \sum_{k=1}^{3} \theta_{i,j}^{(ck)}(l_i, l_j). \quad (11)$$

Lastly, the pairwise potentials between pixels and components enforce that pixels inside a component parameterisation are predominantly labelled by the value that corresponds to that type of node, whereas pixels outside that parameterisation should have label different of the component's type. Let (x_{jk}, y_{jk}) be the coordinates of the ellipse defined by the jth parameterization. The pairwise cost between the ith pixel node and the jth component node is given by:

$$\theta_{i,j}^{(pc)}(l_i, l_j) = \begin{cases} \min_k ||(x_i, y_i) - (x_{jk}, y_{jk})||_2, \text{if } l_i \in \mathfrak{T}_j \\ \quad \text{and } \psi(x_i, y_i, x_j, y_j, a_j, b_j, \varphi_j) = 0 \\ 0, \text{if } l_i \notin \mathfrak{T}_j \\ \quad \text{and } \psi(x_i, y_i, x_j, y_j, a_j, b_j, \varphi_j) = 0 \\ 0, \text{if } l_i \in \mathfrak{T}_j \\ \quad \text{and } \psi(x_i, y_i, x_j, y_j, a_j, b_j, \varphi_j) = 1 \\ \max_k ||(x_i, y_i) - (x_{jk}, y_{jk})||_2, \text{if } l_i \notin \mathfrak{T}_j \\ \quad \text{and } \psi(x_i, y_i, x_j, y_j, a_j, b_j, \varphi_j) = 1 \end{cases}, \quad (12)$$

where $||.||$ is the Euclidean distance.

$$\psi(x,y,x_i,y_i,a_i,b_i,\varphi_i) = \begin{cases} 1 & \text{, if } \frac{\left(\cos(\varphi_i)(x-x_i)+\sin(\varphi_i)(y-y_i)\right)^2}{a_i^2} + \frac{\left(\sin(\varphi_i)(x-x_i)+\cos(\varphi_i)(y-y_i)\right)^2}{b_i^2} \leq 1 \\ 0 & \text{, otherwise} \end{cases} \quad (7)$$

4. Experiments

Our experiments were carried out in a data set composed of 5,551 visible-light images (with resolution 800 $\times$ 300) containing the periocular regions from both sides of the face. These images were the source for the UBIRIS.v2 dataset: they were collected in indoor unconstrained lighting environments and feature significant variations in scale, subjects' pose and gaze. For learning / evaluation purposes, 200 images were manually labelled, covering the seven classes we aim to deal with. This set was divided into two disjoint parts: 1) one used to learn the classification models and to estimate the prior unary / pairwise costs of the MRF; and 2) the complementary part served for quantitative performance evaluation.

To obtain the seven classification models, we used feed-forward neural networks with three layers and $\{34:17:1\}$ topology, with *tan-sigmoid* transfer functions in the input and hidden layers and linear transfer functions in the output layer. The learning sets were always balanced (random sampling) and the Resilient Back-propagation algorithm used to learn the classifiers. Regarding the MRF optimization, every image was resized to 200 $\times$ 75 pixels, i.e., $t_p = 15,000$ in our MRFs. Also, $\boldsymbol{\alpha} = \{0.01, 1, 2, 10, 10\}$.

4.1. Segmentation Performance

Fig. 4 illustrates the results typically attained by the proposed model. Their visual coherence is evident, where regions labelled as hair appear in pink, eyebrows in yellow, irises in green, eyelashes in black, sclera in blue and glasses in blueberry color. Also, solutions were biologically plausible in the large majority of the cases, for various hairstyles, and different subjects poses / gazes. A particularly interesting performance was observed for glasses, where the algorithm attained remarkable results for various types of frames. This was probably due to the fact that glasses were the unique non-biological component among the classes considered, which might had increased their dissimilarity with respect to the remaining components.

In opposition, the most concerning cases happened when the eyebrows and the hair were overlapped (bottom-right image in Fig. 4) . Also, for heavily deviated gazes, the sclera was sometimes under-segmented (typically, by non-detecting the less visible side). In opposition, eyelashes tended to be over-segmented, with isolated eyelashes being grouped in large eyelash regions, which might be due to excessive pairwise cost for observing different labels in adjacent positions of the pixels layer.

Labeling Error	NN (%)		MRF (%)	
Component	FP	FN	FP	FN
Iris	1.12 ± 0.29	9.06 ± 1.80	0.17 ± 0.03	2.61 ± 0.51
Sclera	1.61 ± 0.49	5.17 ± 0.83	0.19 ± 0.03	3.60 ± 0.82
Eyebrows	2.20 ± 0.40	6.93 ± 0.95	0.79 ± 0.28	2.25 ± 0.46
Eyelashes	1.47 ± 0.38	5.12 ± 1.13	0.93 ± 0.23	0.62 ± 0.53
Hair	3.16 ± 0.56	6.74 ± 1.27	1.26 ± 0.30	3.09 ± 0.88
Skin	4.10 ± 1.03	4.09 ± 0.69	2.63 ± 0.43	3.86 ± 1.01
Glasses	1.08 ± 0.22	5.03 ± 1.45	0.06 ± 0.01	0.60 ± 0.09

Table 1. Average pixel labelling errors per component, when considering exclusively the $\arg\max_j p\Big(\omega_j|\eta_j\big(\gamma(x,y)\big)\Big)$ value (NN column) and with the proposed MRF model (MRF column).

It should be noted that α_i were found in an empirical and independent way, i.e., no exhaustive evaluation of combined configurations was carried out, nor any parameter optimization algorithm was used, which also points for the robustness of the proposed model against sub-optimal parameterizations. Table 1 gives the error rates per class, when considering exclusively the first phase of our model (maximum of the posterior probabilities, column "NN") and the full processing chain (MRF optimization, column "MRF"). In this table, FP stands for the false positives rate, whereas FN refers to the false negatives rate. In all cases, it is evident that the MRF substantially lowered the labeling error rates, essentially by imposing smoother responses and constraining the range of biologically acceptable solutions.

As the machine learning algorithm described in this paper is *supervised*, it is important to perceive its variations in performance with respect to the amount of learning data used to create the classification models and the prior unary / pairwise potentials. To this end, performance was compared while varying the number of images used in learning, and keeping constant the number of images used in performance evaluation (to assure comparable bias / variance scores). Figure 5 expresses the results: the horizontal axis gives the number of learning images used and the vertical axis is the corresponding pixel classification error, with the corresponding 95% confidence intervals. We observed that when more than 35 images were used in learning, the pixel classification errors tend to converge. This is evident in terms of the absolute error values and of the narrowness of the confidence intervals.

Figure 4. Examples of the segmented periocular regions. "Hair" class is represented by the pink color, "Eyebrows" appear in yellow, "Iris" in green, "Sclera" in blue, "Glasses" in blueberry and "Eyelashes" in gray. Pixels classified as "Skin" are transparent.

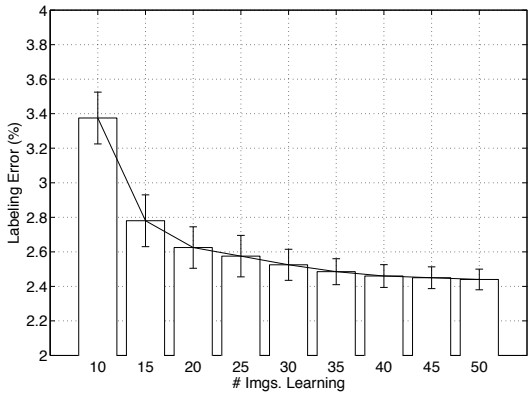


Figure 5. Variations in labelling errors with respect to the number of images used in the learning phase of the algorithm.

4.2. Periocular Biometrics Performance

To exemplify the usefulness of periocular segmentation algorithms, one *all-against-all* matching experiment was designed, using the method of Park *et* al. [13] and two different strategies to define the ROI: as baseline, the iris center was the unique reference for the ROI (upper-left image in Fig. 6). Next, according to the labels provided by the MRF, the center of mass of the cornea was used to define the ROI, which is obviously less sensitive to changes in gaze. Also, regions labelled as hair and glasses were disregarded from the recognition phase, considering that they likely suffer of significant variations among samples of a subject (upper-right image in Fig. 6). The Receiver Operating Characteristic curves for both variants are compared in the bottom plot of Fig. 6 and turn evident the benefits attained due to data segmentation (Equal error rate of 0.128 for the classical ROIs and 0.095 for the improved ROIs configuration). The improvements were substantial in all regions of the performance space, having at some operating points increased the system sensitivity over 10%. It should be stressed that no particular concerns were taken in optimizing the recognition method for the used data set, meaning that the focus was putted much more in the performance gap between both recognition schemes than in the recognition errors in absolute values, which are out of the scope of this paper.

5. Conclusions and Further Work

In this paper we have proposed an algorithm for *one-shot* labelling of all the components in the periocular region: iris, sclera, eyelashes, eyebrows, hair, skin and glasses. Our solution is composed of two major phases: 1) a group of local classification models gives the posterior probabilities for each pixel and class considered; 2) this *appearance*-based information is fused to geometrical constraints and shape priors to feed a two-layered MRF. One layer represents *pixels*, and analyzes the local data appearance while enforcing smoothness of the solutions. The second layer represents *components*, and assures that solutions are biologically plausible. By minimizing the MRF energy, the label of each pixel is found, yielding solutions that are robust against changes in scale, subjects' pose and gaze and dynamic lighting conditions.

As further directions for this work, our efforts are focused in estimate gaze / pose from the labelled data, in order to compensate for deviations before the recognition process.

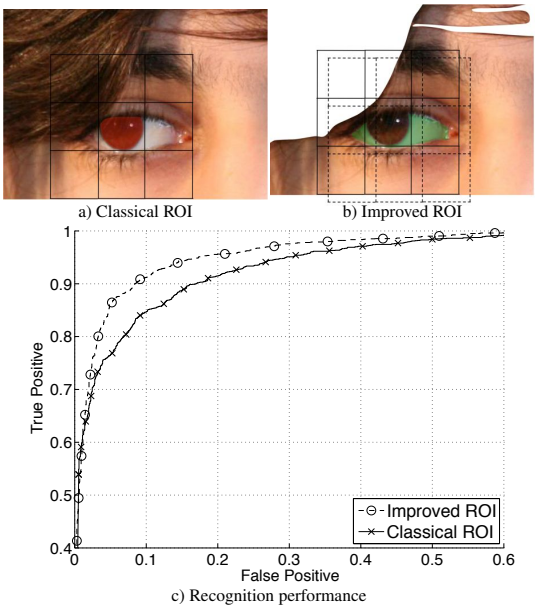


Figure 6. Improvements in periocular recognition performance due to the semantic categorization (labeling) of each pixel in the periocular region.

References

[1] A. Albiol, L. Torres and E. Delp. Optimum color spaces for skin detection. In Proceedings of the *International Conference on Image Processing*, vol. 1, pag. 122–124, 2001.

[2] C. Basca, M. Talos and R. Brad. Randomized Hough Transform for Ellipse Detection with Result Clustering. in Proceedings of the *The International Conference on Computer as a Tool EUROCON*, vol. 2, pag. 1397–1400, 2005.

[3] A. Besbes, N. Komodakis, G. Langs and N. Paragios. Shape Priors and Discrete MRFs for Knowledge-based Segmentation. in Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pag. 1295–1302, 2009.

[4] S. Bharadwaj, H. Bhatt, M. Vatsa and R. Singh. Periocular biometrics: When iris recognition fails. In Proceedings of the 4^{th} *IEEE International Conference on Biometrics: Theory Applications and Systems*, pag. 1–6, 2010.

[5] B. Caputo, S. Bouattour and H. Niemann. Robust appearance-based object recognition using a fully connected Markov random field In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, vol., 3, pag. 565–568, 2002.

[6] S. Crihalmeanu and A. Ross. Multispectral scleral patterns for ocular biometric recognition. *Pattern Recognition Letters*, vol. 33, no. 14, pag. 1860–1869, 2012.

[7] P. Felzenszwalb and D. Huttenlocher. Efficient Belief Propagation for Early Vision. *International Journal of Computer Vision*, vol. 70, no. 1, pag. 41–54, 2006.

[8] B. Glocker, D. Zikic, N. Komodakis, N. Paragios and N. Navab. Linear Image Registration Through MRF Optimization. In Proceedings of the *IEEE International Symposium on Biomedical Imaging*, pag. 422–425, 2009.

[9] K. Oh, B-S. Oh, K-A. Toh, W-Y. Yau and H-L. Eng. Combining sclera and periocular features for multi-modal identity verification. *Neurocomputing*, vol. 128, pag. 185–198, 2014.

[10] Z. Kato and T.C. Pong. A Markov random field image segmentation model for textured images. *Image and Vision Computing*, vol. 24, pag. 1103–1114, 2006.

[11] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pag. 1568–1583, 2006.

[12] T. Ojala, M. PietikŁinen and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Proceedings of the 12^{th} *IAPR International Conference on Pattern Recognition*, pag. 582–585, 1994.

[13] U. Park, R. Jillela, A. Ross and A. Jain. Periocular Biometrics in the Visible Spectrum. *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pag. 96–106, 2011.

[14] U. Park, A. Ross and A. Jain. Periocular biometrics in the visible spectrum: A feasibility study. In Proceedings of the *IEEE* 3^{rd} *International Conference on Biometrics: Theory, Applications, and Systems*, pag. 1–6, 2009.

[15] H. Proença. Iris Recognition: On the Segmentation of Degraded Images Acquired in the Visible Wavelength. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pag. 1502–1516, 2010.

[16] U. Park, A. Ross and A. Jain. Matching highly non-ideal ocular images: an information fusion approach. In Proceedings of the *IEEE* 5^{th} *International Conference on Biometrics*, pag. 446–453, 2012.

[17] 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, pag. 3751–3765, 2013.

[18] T. Tan, X. Zhang, Z. Sun and H. Zhang. Noisy iris image matching by using multiple cues. *Pattern Recognition Letters*, vol. 33, pag. 970–977, 2012.

[19] C. Wang, M. Gorce and N. Paragios. Segmentation, Ordering and Multi-Object Tracking using Graphical Models. In Proceedings of the *in* 12^{th} *International Conference on Computer Vision*, pag. 747–754, 2009.

[20] D. Woodard, S. Pundlik, P. Miller, R. Jillela and A. Ross. On the fusion of periocular and iris biometrics in non-ideal imagery. In Proceedings of the *in* 20^{th} *International Conference on Pattern Recognition*, pag. 201–204, 2010.

[21] D. Woodard, S. Pundlik, P. Miller and J. Lyle. Appearance-based periocular features in the context of face and non-ideal iris recognition. *Signal, Image and Video Processing*, vol. 5, pag. 443–455, 2011.

Joint Head Pose/Soft Label Estimation for Human Recognition *In-The-Wild*

Hugo Proença, *Senior Member, IEEE*, João C. Neves, *Member, IEEE*, Silvio Barra, Tiago Marques, and Juan C. Moreno

Abstract—Soft biometrics have been emerging to complement other traits and are particularly useful for poor quality data. In this paper, we propose an efficient algorithm to estimate human head poses and to infer soft biometric labels based on the 3D morphology of the human head. Starting by considering a set of pose hypotheses, we use a learning set of head shapes synthesized from anthropometric surveys to derive a set of 3D *head centroids* that constitutes a metric space. Next, representing queries by sets of 2D head landmarks, we use projective geometry techniques to rank efficiently the joint 3D head centroids/pose hypotheses according to their likelihood of matching each query. The rationale is that the most likely hypotheses are *sufficiently* close to the query, so a good solution can be found by convex energy minimization techniques. Once a solution has been found, the 3D head centroid and the query are assumed to have similar morphology, yielding the soft label. Our experiments point toward the usefulness of the proposed solution, which can improve the effectiveness of face recognizers and can also be used as a privacy-preserving solution for biometric recognition in public environments.

Index Terms—Soft biometrics, visual surveillance, homeland security, privacy-preserving recognition

1 Introduction

In biometrics research, one of the most challenging goals is the development of recognition systems that work in unconstrained (outdoor) scenarios and do not assume the subjects' willingness to be recognized. In such conditions, the acquired data has poor quality, with faces partially occluded, blurred, or misaligned (Fig. 1).

The idea behind soft biometrics is to obtain "*characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals*" [16]. These characteristics not only complement strong biometric traits, but they also prune the set of identities for a query. Soft biometrics can also be regarded as a response to privacy/ethical issues in using biometrics in public places: it makes it possible to ignore the large majority of the identities in the scene and attempt positive recognition (e.g., with a face recognizer) only for the subjects with soft labels similar to the identities on a watch-list.

This paper describes an algorithm to infer jointly human head poses and soft labels in an efficient way based on poor-quality data. During the learning phase, anthropometric head surveys feed a stochastic process that generates a set of synthetic 3D head meshes representing the major features of a population. Such elements are the input of a self-organizing map that obtains a discretized representation of the feature space, i.e., a matrix of *centroid* heads with a key property; it preserves the topological properties of the input space and enables us to define the closeness of its elements (i.e., the similarity of head shapes). Considering the wildness of the data, we also generate a set of pose hypotheses. Next, all combinations of joint poses/head shape hypotheses are grouped and indexed using as a criterion the proximity of their projected head landmarks.

In classification, having a query represented by a set of head image landmarks (detected as described in [18] or [8]), we rank the set of hypotheses in approximate logarithmic time according to the similarity between the query and the joint pose/head shape 2D projections. The idea is that the most likely hypothesis is *sufficiently* close to the solution so that only slight changes in its parameterization are required to match the query faithfully. This way, local minima are neglected and convex optimization techniques are used to reach acceptable solutions. A convergence test determines whether the process stops or the next hypothesis is considered. The method described here uses some insights from [37] and [30], namely in the generation of the set of hypotheses and in using projective geometry techniques to evaluate them.

The remainder of this paper is organized as follows: Section 2 summarizes the related work. Sections 3 and 4 give a detailed description of the learning and classification phases of the proposed algorithm. Section 5 describes the experiments carried out and discusses the corresponding results. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Soft Biometrics

According to [40], soft biometric traits are classified into three families: 1) global traits, which regard demographic information (e.g., age, gender, and ethnicity); 2) body traits, which are concerned with the subjectõs somatotype, i.e.,

- *H. Proença, J.C. Neves, T. Marques, and J.C. Moreno are with the IT: Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, Covilhã, Portugal. E-mail: {hugomcp, jcneves, tmarques, jcmb}@di.ubi.pt.*
- *S. Barra is with the Università degli Studi di Cagliari, Italy. E-mail: silvio.barra@unica.it.*

Manuscript received 20 Feb. 2015; revised 1 Oct. 2015; accepted 15 Jan. 2016. Date of publication 26 Jan. 2016; date of current version 10 Nov. 2016. Recommended for acceptance by M. Tistarelli. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2016.2522441

Fig. 1. Examples of images acquired by a visual surveillance system, composed by a wide-view camera feeding a pan-tilt-zoom device that collects data from moving and at-a-distance targets (up to 40 meters away).

their overall appearance (height or body volume); and 3) head traits, which analyze the regions that humans instinctively use to identify others (e.g., hair or eye color, nose or neck thickness, and ear shape/size).

Regarding global traits, Heckathorn et al. [11] measured lengths of wrists and forearms. Using the concept of *interchangeability of indicators*, they argued that combining multiple low accuracy measurements yields a highly accurate indicator. Jain and Park [17] used demographic information (gender and ethnicity) and facial marks (scars, moles and freckles) to improve face image matching and retrieval performance. An extended version of this work can be found in [32].

In terms of body traits, Lucas and Henneberg [23] concluded that, upon the availability of accurate anthropometric measurements, the body is actually more distinctive than the face when distinguishing humans. Previously, other works (e.g., Rice et al. [36]) concluded that identification based on body measurements can be as accurate as using the face. Moustakas et al. [29] suggested a framework based on height and stride length information to increase the effectiveness of a gait recognition system, integrating soft labels directly in the estimation of the matching score instead of the traditionally used score-level fusion. Drosou et al. [7] proposed a probabilistic framework for improving the recognition performance via soft labels (global and body-based), modelling the systematic intrinsic error of each measurement (e.g., due to clothing).

Finally, most works in the head traits family analyze the discriminability of hair/facial hair styles and lengths. Dass et al. [6] pre-aligned the images based on the position of the eyes and, using agglomerative clustering techniques, defined five groups of hairstyles according to hair density in image patches. Hewig et al. [13] observed that the typical hair styles are heavily correlated with global traits (gender and age), which might also be useful for identification.

A noteworthy conclusion was drawn by Reid et al. [35]: *comparative* descriptors (relative magnitude between subjects' measurements) have more discriminatory power than the absolute values themselves, and are particularly advantageous in terms of stability. Detailed information about soft biometrics can be found in two comprehensive surveys by Kim et al. [25] and Reid et al. [34].

2.2 Head Pose Estimation

The existing methods for head pose estimation can be divided into two main groups: 1) generative, by fitting parametric models to the query; and 2) discriminative, which are model-free and search for correspondences between image features and known pose configurations.

Generative models consider prior information about human kinematics and anthropometry to reduce the number of plausible configurations for a query. In this family of approaches, appearance template methods (e.g., fed by Gabor descriptors [38]), flexible models based on the elastic graph matching (e.g., [27]) or active appearance models (e.g., [42]) can be highlighted. Model fitting methods, based on generic 3D face [1] and ellipsoidal [41] shapes, are examples of this family of algorithms, which focus on the idea of mapping a set of 3D face models onto the images, based on a group of 2D-3D correspondences. Textured triangular meshes [28] or cubic polynomials [45] can be used in such mapping. In this model-driven family, the work of Krinidis et al. [26] shares some insight with the algorithm proposed in this paper, specifically by inferring the equations that govern the face deformation model, fed by the tracking module.

Discriminative models are usually holistic, and consider the whole image of the head/face for estimation, instead of local landmarks. Li et al. [22] estimated local image gradients, reduced dimensionality by an analysis of principal components and used a support vector regression machine to infer poses. Other similar approaches used manifold embedding algorithms (e.g., [43]) and non-linear regression methods (e.g., based on convolution networks [31]). A representative approach in this family is the work of Huang and Trivedi [14], who used a skin-tone edge-based detector to feed a tracker module based on Kalman filter and a hidden Markov model to infer poses.

Refer to the surveys published by Murphy-Chutorian and Trivedi [5], Ba and Odobez [3] and Zhang and Gao [46] for detailed information about head pose estimation and its taxonomy.

3 Proposed Method: Learning Phase

For comprehensibility, we use the following notation: matrices are represented by capitalized bold fonts and vectors appear in bold. The subscripts denote indexes. All vectors are column-wise. The ring symbol (e.g., $\mathring{\mathbf{x}}$) denotes 2D (image) positions, while 3D positions in the Euclidean space appear in regular font (e.g., $\mathbf{x}$). The hat symbol (e.g., $\hat{\mathbf{x}}$) denotes an estimate and all the hard thresholds are denoted by the κ symbol.

3.1 Generation of Synthetic 3D Head Shape Models

Young [44] reported 22 head dimensions from a random, composite of females and males in an adult population. The author claims these dimensions are able to describe the essential morphological properties of a human head, with 17 of these also being considered in previous surveys (e.g., [12]). Based on data from 195 females and 172 males, this study provides a set of summary statistics (minimum, maximum, mean, standard deviation, coefficient of variation, symmetry and kurtosis) for every type of measurement. In most cases, the landmarks are internal bone features, with paired surface landmarks defining lines in planes from which perpendicular distances are taken. The leftmost part of Fig. 2 illustrates some of the dimensions provided in this survey, while Table 1 lists the types of lengths we consider in this paper (at left) and their levels of linear correlation (rightmost matrix).

We generate the 3D head shape models randomly, starting from a single mesh that is iteratively deformed, according to the target distances between the pairs of vertices. Let $\mathbf{x}_i$ be one 3D vertex and $\mathbf{n_i}$ the normal to the surface

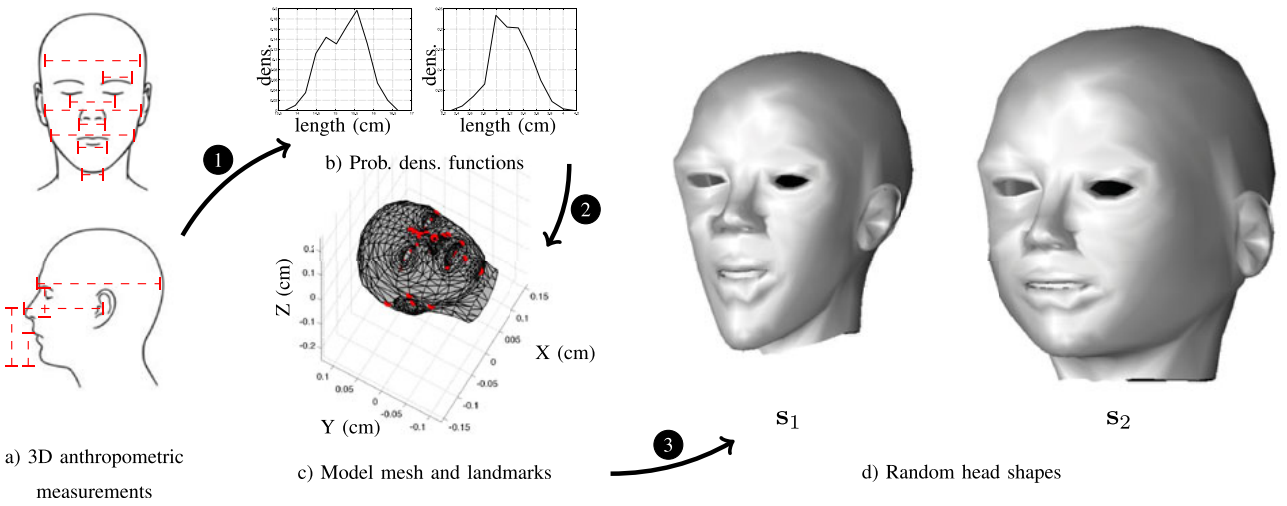


Fig. 2. Overview of the stochastic process that generates an arbitrary number of 3D head shapes (meshes). Based on anthropometric surveys (marker 1), a set of probability density functions for head lengths is defined (marker 2), and used to iteratively deform a *base* mesh, enabling to obtain head shapes of evidently different appearance (marker 3).

at that point. Let $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, $\mathbf{n}_{ij} = \mathbf{n}_i - \mathbf{n}_j$ ($\mathbf{x}_\cdot, \mathbf{n}_\cdot \in \mathbb{R}^3$) and let l_{ij} be the target length (Euclidean distance) between $\mathbf{x}_i$ and $\mathbf{x}_j$. The goal is to find the magnitude of displacement α_{ij} on both vertices with respect to their normal vectors ($\mathbf{x}^{\text{new}} = \mathbf{x}^{\text{old}} + \alpha\mathbf{n}$), such that the resulting distance l_{ij} follows the probability density functions reported in the anthropometric head survey:

$$||\mathbf{n}_{ij}^T\mathbf{x}_{ij}\ \alpha_{ij}^2 + 2\mathbf{x}_{ij}^T\mathbf{n}_{ij}\ \alpha_{ij} + \mathbf{x}_{ij}^T\mathbf{x}_{ij} - l_{ij}||_2 = 0, \tag{1}$$

being $||.||_2$ the $\ell - 2$ norm. Rearranging (1) in matrix form we have:

$$|||[\mathbf{n}_{ij}^T\mathbf{x}_{ij}, 2\mathbf{x}_{ij}^T\mathbf{n}_{ij}, \mathbf{x}_{ij}^T\mathbf{x}_{ij}]\ [\alpha_{ij}^2, \alpha_{ij}, 1]^T - l_{ij}||_2 = 0, \tag{2}$$

which represents one constraint of the head shape model. Let $\mathbf{c}_{ij} = [\mathbf{n}_{ij}^T\mathbf{x}_{ij}, 2\mathbf{x}_{ij}^T\mathbf{n}_{ij}, \mathbf{x}_{ij}^T\mathbf{x}_{ij}]$, $\boldsymbol{\alpha}_{ij} = [\alpha_{ij}, \sqrt{\alpha_{ij}}, 1]^T$. $\mathbf{C}$ is the block diagonal matrix that yields from the concatenation of all $\mathbf{c}$ elements, while $\boldsymbol{\alpha}$ and $\mathbf{l}$ concatenate the remaining terms:

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_{ij} & 0\ 0\ 0 & \dots & 0\ 0\ 0 \\ 0\ 0\ 0 & \mathbf{c}_{i'j'} & \dots & 0\ 0\ 0 \\ \vdots & \vdots & \vdots & 0\ 0\ 0 \\ 0\ 0\ 0 & 0\ 0\ 0 & 0\ 0\ 0 & \mathbf{c}_{i''j''} \end{bmatrix} \Bigg\} n \times 3n$$
$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_{ij} \\ \boldsymbol{\alpha}_{i'j'} \\ \vdots \\ \boldsymbol{\alpha}_{i''j''} \end{bmatrix} \Bigg\} 3n \times 1, \mathbf{l} = \begin{bmatrix} l_{ij} \\ l_{i'j'} \\ \vdots \\ l_{i''j''} \end{bmatrix} \Bigg\} n \times 1, \tag{3}$$

being the unknowns $\boldsymbol{\alpha}$ found by:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} (\mathbf{C}\boldsymbol{\alpha} - \mathbf{l})^T\ (\mathbf{C}\boldsymbol{\alpha} - \mathbf{l}), \quad \text{s.t. } |||[\alpha_{ij}, \dots, \alpha_{i''j''}]|||_\infty \leq \kappa_1, \tag{4}$$

where κ_1 avoids anatomically bizarre solutions and guarantees that the solution closest to the initial configuration is preferred in the quadratic system ($\kappa_1 \approx 0.1$ in our experiments). According to this formulation, (4) is a constrained optimization problem with inequality constraints that can be solved as described in [4]. Once the $\hat{\boldsymbol{\alpha}}$ values are found, the coordinates of the corresponding vertices are updated, with similar distortions (weighted by a Gaussian kernel) applied to neighbouring vertices to enforce smoothness in the resulting mesh. The rightmost images in Fig. 2 are examples of the different meshes that can result from this stochastic process.

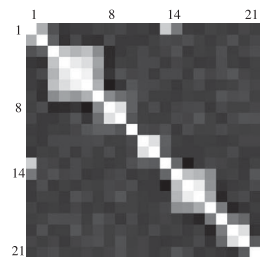
3.2 Head Shape Hypotheses

Let $\mathbf{s} = [\mathbf{x}_1^T, \dots, \mathbf{x}_{t_v}^T]^T$ be a vector representing one head shape, given as a triangulated mesh of 3D vertices. $\mathbf{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_{t_m}\}$ is the set of meshes used for learning purposes, generated as described in Section 3.1. Evidently, there is some correlation between the $\mathbf{x}_i$ elements in each mesh, which can be attenuated by representing meshes in the principal components (PC) space:

$$\mathbf{s}^* = (\mathbf{s} - \mathbf{s}_0)\mathbf{T}_{pc}, \tag{5}$$

being $\mathbf{s}_0$ the $3t_v$-dimensional mean of the elements in $\mathbf{S}$ and $\mathbf{T}_{pc}$ the PC transformation matrix. This way, it is possible to describe each mesh in a feature space of a much lower dimension than the $3t_v$, which is important for the sake of computational effectiveness. In our case, the head models have $t_v = 957$, with 50 PC coefficients being able to represent over 99.9 percent of their variability.

TABLE 1
Types of Anthropometric Measurements Considered in This Paper and Their Levels of Linear Correlation

Measurements	Pearson Correlation
Cranium: {1- Head circumference; 2- Head breadth; 3- Head length; 4- Biorbital breadth; 5- Biectocanthus breadth; 6- Bipupil breadth; 7- Nasal bridge breadth; 8- Bialar breadth; 9- Bicheilion breadth; 10- Bitragion breadth; 11- Bizygomatic breadth; 12- Bigonial breadth}; **Face**: {13- Sellion-menton length; 14- Sellion-supramentale length; 15- Sellion-stomion length; 16- Sellion-subnasion length}; **Nose**: {17- Midnasal bridge height; 18- Pronasale height (Maxilloare plane); 19- Pronasale height (Sellion-promentale plane); 20- Sellion height (medial cants plane); 21- Sellion height (lateral orbital plane)}	

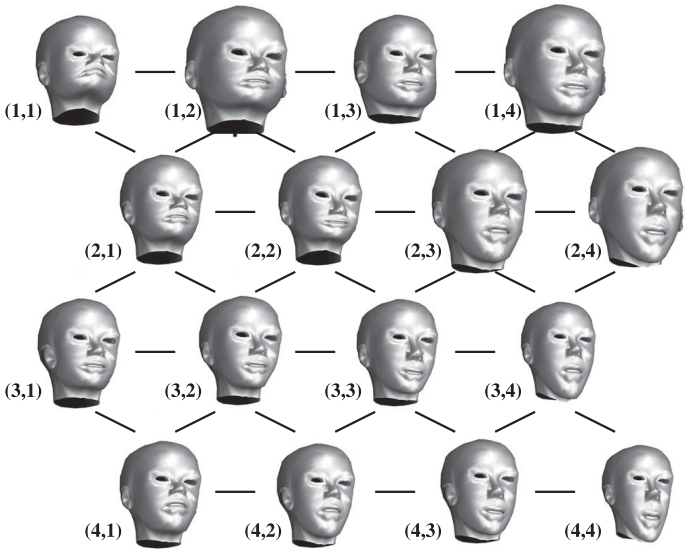


Fig. 3. Representation of the 3D head centroids resulting of a 4×4 SOM. Note the similarity in size/shape between adjacent elements, rooted in the preservation of the topological properties of the input space that this kind of maps offers.

Let $\mathbf{S}^*$ represent the shape hypotheses in the PC space. The next step is the inference of a set of prototypes that intrinsically represent *head shape similarity*, with self-organizing maps [20] (SOMs) of size $t_c \times t_c$ being considered a good choice for the following reasons: 1) SOMs obtain an ordered mapping between the 50D input space and a 2D output space, where each element represents one head prototype; 2) prototypes in the output space are topologically ordered, i.e., neighbor prototypes feature similar head shapes; 3) SOM prototypes reflect the variations in density in the input space, i.e., densely populated regions in the input space (where the most frequent head shapes fall) are represented by the largest number of prototypes; and 4) SOMs are known to be particularly suitable to model non-linear input spaces, such as our input feature space. In practical terms, the SOM output space is a similarity graph, which is important in order to label degraded data: even if a query is not mapped directly to the same cell as the enrolment sample with a corresponding identity, it should be mapped to a neighboring cell. Fig. 3 illustrates the head prototypes (cells) that are used as soft labels.

3.3 3D Head Shape Covariance

Let $\mathbf{s}_{c_i}$ be the head shape centroid corresponding to the i^{th} cell in the SOM, and let $\{\mathbf{s}_{c_{i1}} \ldots, \mathbf{s}_{c_{iw}}\}$ be the shape samples associated with $\mathbf{s}_{c_i}$. For all the elements in $\mathbf{s}_{\cdot}$ that correspond to head landmarks, the displacement between the 3D positions in the samples and in the centroid were measured $(\mathbf{x}_{c_{ij}} - \mathbf{x}_{c_i})$, obtaining a set of 3D vectors from where the mean and covariance matrix were taken. This captures the spread of the 3D data and is used in the algorithm convergence test to discriminate between genuine/spurious query landmarks. To illustrate this point, Fig. 4 plots the 99 percent confidence ellipsoids for the *right ear lobe*, *center of right cornea* and *nose apex* landmarks.

3.4 Pose Hypotheses

Let $\mathbf{p} = \{\mathbf{R}, \mathbf{t}\}$ be a camera pose configuration, with $\mathbf{R}$ being the rotation matrix and $\mathbf{t}$ the translation vector, i.e., $\mathbf{p}$ is a 6D vector accounting for three components of rotation (yaw, pitch and roll) and three of translation (t_x, t_y and t_z). Let $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{t_p}\}$ be a set of pose hypotheses, created randomly using uniformly distributed random numbers for all six degrees of freedom. Given the relatively large number of elements generated ($\approx$ 100,000), a set of pose prototypes is also obtained. In this case, as there are no requirements about the concept of *similar* poses, such prototypes can be found simply by the k-means algorithm, yielding $t_{\tilde{p}}$ pose vectors ($t_{\tilde{p}} \ll t_p$).

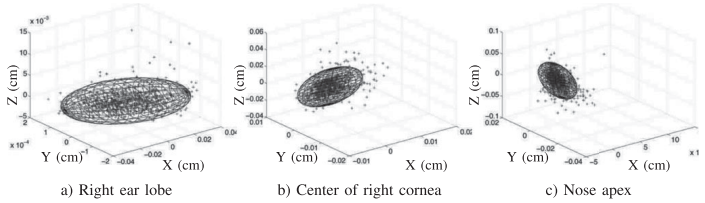


Fig. 4. Examples of the 99 percent confidence ellipsoids that represent the deviations of the positions of landmarks in the head shape samples with respect to their centroid. These values are used in the convergence test of the algorithm to discriminate between genuine/spurious head landmarks.

3.5 Joint Head Shapes/Pose Hypotheses Indexing

Given a set of $t_{\tilde{p}}$ pose and t_c^2 head shape hypotheses, during classification it is required to find the *best* joint pose/shape configuration, which is the most likely match to the query. Theoretically, there are a total of $t_{\tilde{p}} t_c^2$ possibilities, but exploring all by brute-force is prohibitive in terms of time complexity. Moreover, not all the query landmarks will be genuine, and both false negatives and false positives are expected. Given such constraints, a forest of binary trees was created, one per type of landmark, where the hypotheses are grouped (k-means) in leaves according to their neighborhood of one landmark projection, given by the *world-to-image* function:

$$f_{w \rightarrow i}(\mathbf{x}, \mathbf{p}) = \frac{1}{\upsilon} \mathbf{A}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}, \tag{6}$$

being $\mathbf{x}$ the vertices of $\mathbf{s}$, υ the scalar projective parameter, $\mathbf{A}$ the internal camera matrix, and $\mathbf{p}$ = {$\mathbf{R}$ (rotation), $\mathbf{t}$ (translation)} the pose parameters. This way, each tree keeps, within its leaves, the indices of the hypotheses that have similar 2D projections of a landmark. Later, in classification, the position of every query landmark is used in the corresponding tree to obtain the indices of the complying hypotheses. By repeating the process for all landmarks and accumulating the complying indices, the hypotheses are ranked in descending order according to the frequency with which they appear in leaves, so that the most likely (those with the highest number of landmarks close to the query) will be evaluated first.

The retrieval process is illustrated in Fig. 5, and has a time complexity $\mathcal{O}(t_l \log(t_{\tilde{p}} t_c^2))$, t_l being the number of query landmarks. This roughly logarithmic time complexity is important for generating large sets of hypotheses without substantially compromising the time cost of retrieval.

4 Classification Phase

Let $\mathring{\mathbf{q}} = \{\mathring{\mathbf{q}}_1, \ldots, \mathring{\mathbf{q}}_{t_q}\}$ be a set of 2D head landmarks in a query image. We assume that the *type* of each landmark

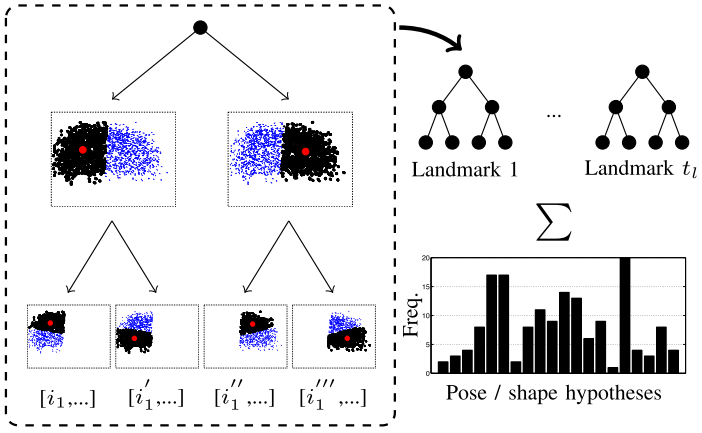


Fig. 5. Data structure that indexes the joint pose/shape hypotheses, grouped according to the similarity of their landmark projections. In retrieval, the indices of the hypotheses complying the query landmarks are accumulated, such that the most voted hypotheses will be evaluated first.

$\tau(\mathring{\mathbf{q}}_i)$ is known, i.e., the anatomic region corresponding to each $\mathring{\mathbf{q}}_i$ is given as input. This is a readily satisfied assumption, using the state-of-the-art techniques for head/face landmark detection (e.g., [18], [8], or [33]).

Using the trees described in Section 3.5, the most likely joint pose/head shape hypothesis for the query is obtained and its pose configuration subsequently optimized. Assuming that the pose hypothesis $\mathbf{p}$ is *relatively* close to the query pose, the idea is to perform only small adjustments in its parameterization to better fit the query:

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}} d\big(f_{w\rightarrow i}(\mathbf{s},\mathbf{p}), \mathring{\mathbf{q}}\big), \tag{7}$$

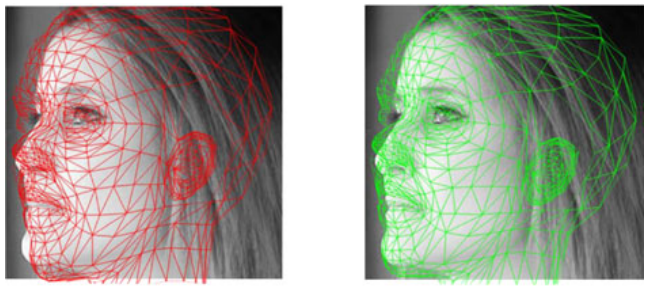
where $f_{w\rightarrow i}(\mathbf{s},\mathbf{p}) = f_{w\rightarrow i}(\mathbf{x},\mathbf{p}), \forall \mathbf{x}\in\mathbf{s} = \mathring{\mathbf{s}}$ and $d(.,.)$ is the function that measures the similarity between two sets of landmarks:

$$d(\mathring{\mathbf{s}},\mathring{\mathbf{q}}) = \frac{1}{\nu(\mathring{\mathbf{q}})}\sum_{i=1}^{\nu(\mathring{\mathbf{q}})}\min_{\mathring{q}_j|\tau(\mathring{q}_j)=\tau(\mathring{\mathbf{x}}_i)} d(\mathring{\mathbf{x}}_i,\mathring{\mathbf{q}}_j), \tag{8}$$

where $d(\mathring{\mathbf{x}},\mathring{\mathbf{q}}) = ||\mathring{\mathbf{x}}-\mathring{\mathbf{q}}||_2$ and $\nu(\mathring{\mathbf{q}})$ is the function that counts the number of distinct types of landmarks in $\mathring{\mathbf{q}}$. Essentially, (8) sums the distances between projections of 3D head vertices and their closest query landmarks of the corresponding type.

The optimization process is regarded as convex and unconstrained, with all the advantages inherent to it in terms of computational cost. We use a derivative-free algorithm proposed by Lagarias et al. [21], due to its proven effectiveness in relatively low dimensionality problems (six in our case). Having an initial pose hypothesis $\mathbf{p}$, the algorithm generates a sample of seven points around $\mathbf{p}$ and iteratively discards the point with the maximum value of the cost function (8), replacing it with a new point generated either by reflection, expansion, contraction or shrinkage of sample points. As Fig. 6 illustrates, this process enables us to better fit the pose hypothesis to the query data by only slightly adjusting the initial configuration.

Having an optimized estimate of pose $\hat{\mathbf{p}}$, the final step is the evaluation of the reasonability of the $\{\hat{\mathbf{p}},\mathbf{s}\}$ solution, either stopping the algorithm or continuing to the next hypothesis. This evaluation is carried out in the 3D space by



a) Initial pose / shape hypothesis $\mathbf{p}$ b) Shape with optimized pose $\hat{\mathbf{p}}$

Fig. 6. Pose refinement, according to a convex optimisation paradigm. Assuming that the initial hypothesis $\mathbf{p}$ is a good approximation of the solution, the probability of falling in local minima is relatively short. $\hat{\mathbf{p}}$ is the optimized configuration.

inferring the most likely 3D positions for the query landmarks. Let $\mathring{\mathbf{q}} = (x, y)$ be one image landmark corresponding to one vertex in $\mathbf{s}$. There is a ray in the Euclidean 3D space from where elements are projected into $\mathring{\mathbf{q}}$, which is given by the *image-to-world* function:

$$f_{i\rightarrow w}(\mathring{\mathbf{q}},\hat{\mathbf{p}}) = \mathbf{R}^T\mathbf{A}^{-1}\upsilon\begin{bmatrix}\mathring{\mathbf{q}}\\1\end{bmatrix} - \mathbf{R}^T\mathbf{t}, \tag{9}$$

with $\mathbf{A}$ being the internal camera parameters, $\mathbf{R}$ and $\mathbf{t}$ its extrinsic parameters (obtained from $\hat{\mathbf{p}}$) and υ being the scalar projective parameter. The shortest distance between the ray and the corresponding vertex in $\mathbf{s}$ is the most *optimistic* location of $\mathring{\mathbf{q}}$ in the 3D space:

$$\hat{\mathbf{q}} = \mathbf{x}_r + \mathbf{v}_r^T \odot \frac{(\mathbf{x}-\mathbf{x}_r)^T\mathbf{v}_r}{||\mathbf{v}_r||^2}, \tag{10}$$

being $\odot$ the point-by-point multiplication operator, $\mathbf{x}_r, \mathbf{v}_r$ the 3D point and vector defining the ray (given by (9)). Fig. 7 illustrates the rationale behind this step, where the 3D positions $\hat{\mathbf{q}}_.$ from where the query landmarks $\mathring{\mathbf{q}}_.$ might have been projected are estimated based on $\{\hat{\mathbf{p}},\mathbf{s}\}$.

According to (10), only the query landmarks $\hat{\mathbf{q}}^*$ that are the most likely to be genuine are selected, providing the minimum $||\hat{\mathbf{q}}_. - \mathbf{x}_.||_2$ values (per type of landmark). Henceforth, all the remaining landmarks are deemed to be spurious and are discarded. Finally, given the set of remaining landmarks and their most plausible 3D positions, ϕ evaluates the reasonability of such positions by checking if

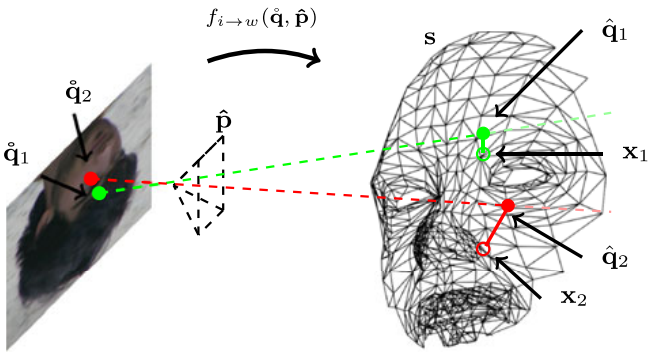


Fig. 7. Finding the 3D positions in the Euclidean space from where the query landmarks might have been projected, according to a pose $\hat{\mathbf{p}}$ and shape $\mathbf{s}$ estimates. The $||\hat{\mathbf{q}}_. - \mathbf{x}_.||_2$ values are used to discriminate between the spurious (in red) and genuine (in green) query landmarks.

Fig. 8. Examples of the data sets used in the empirical validation of the proposed method. The upper row regards the AFLW data set, whereas the bottom rows are from the LFW and SCface sets.

misalignments are inside the prediction interval ellipsoid, obtained as described in Section 3.3:

$$\begin{aligned}&\phi(\hat{\mathbf{q}}_i^*|\mathbf{x}_i,\mathbf{x}_{c_i},\boldsymbol{\Sigma}_i)\\&=(\hat{\mathbf{q}}_i^*-\mathbf{x}_i-\mathbf{x}_{c_i})^T\boldsymbol{\Sigma}_i^{-1}(\hat{\mathbf{q}}_i^*-\mathbf{x}_i-\mathbf{x}_{c_i})-\chi_3^2(0.99),\end{aligned}\tag{11}$$

with $\mathbf{x}_{c_i}$ as the position of the shape centroid, $\boldsymbol{\Sigma}_i$ as the covariance matrix and $\chi_3^2(0.99)$ as the quantile function for probability 99 percent of the chi-squared distribution with three degrees of freedom. In practical terms, this function checks if it is likely to observe a $\hat{\mathbf{q}}_i^*-\mathbf{x}_i$ misalignment between a sample landmark and its centroid, returning a positive value if the misalignment falls inside the covariance error ellipsoid (Fig. 4) and a negative value otherwise. Finally, a solution is *acceptable* if a sufficient number of landmarks is deemed genuine, i.e., $H_{all}() \geq \kappa_2$:

$$H_{all}(\hat{\mathbf{q}}^*|\mathbf{x},\mathbf{x}_{c_i},\boldsymbol{\Sigma})=\frac{1}{\nu(\hat{\mathbf{q}}^*)}\sum_{i=1}^{\nu(\hat{\mathbf{q}}^*)}H\big(\phi(\hat{\mathbf{q}}_i^*|\mathbf{x}_i,\mathbf{x}_{c_i},\boldsymbol{\Sigma}_i)\big),\tag{12}$$

where κ_2 is the convergence threshold, $\nu(\hat{\mathbf{q}}^*)$ is the number of query landmarks, and H is the Heaviside function:

$$H(z)=\begin{cases}1, & \text{if } z\geq 0\\ 0, & \text{if } z<0.\end{cases}\tag{13}$$

5 Results and Discussion

Three well known data sets were selected for our experimental evaluation. The Annotated Facial Landmarks in the Wild [19] (AFLW) set was used to evaluate the results of the pose estimation phase. It has 25,993 color images, each one annotated with a 21-point markup on visibility. In this set, we considered exclusively samples with pose angles in the intervals yaw $\pm\pi/4$, pitch $\pm\pi/2$, and roll $\pm\pi/5$, according to the plausibility of observing such poses in visual surveillance scenarios. The soft biometric labels were evaluated using the Labeled Faces in the Wild [15] (LFW) and in the SCface [10] sets, selected due to the wildness of their data. Out of the 9,164 images in the LFW set, 670 were disregarded due to extremely poor performance of the head landmark detector, resulting in 8,494 samples from 1,574 subjects. For the SCface set, we exclusively used the third sample from cameras 1-5 (650 images from 130 subjects), which have the maximal resolution acquired at visible wavelengths. Fig. 8 shows some images from the data sets considered. In all the experiments below, the thresholds were set to $\kappa_1 = 0.01$ and $\kappa_2 = 0.9$.

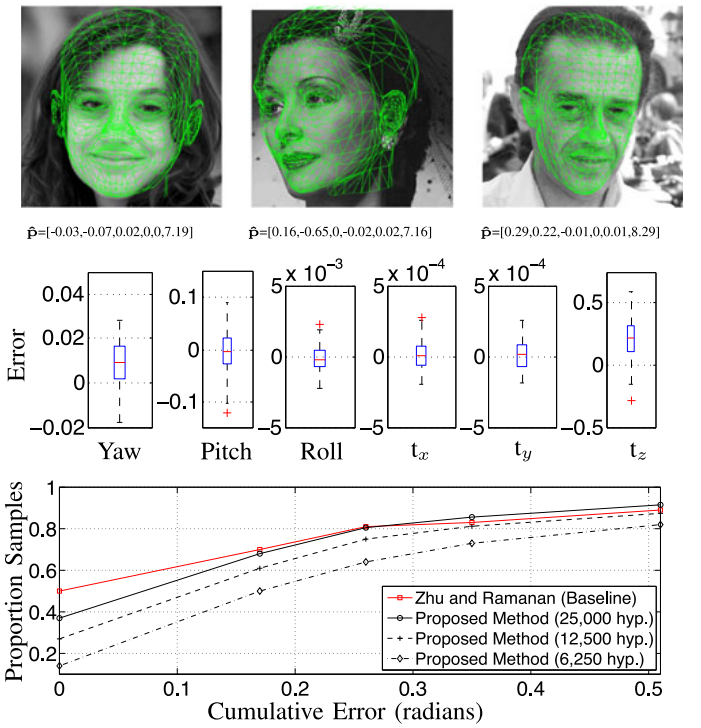


Fig. 9. Upper row: examples of pose estimates in images from the AFLW data set. Second row: boxplot of the pose estimation errors for the six degrees-of-freedom: {yaw, pitch, roll} rotation angles (in radians), plus the $\{t_x, t_y, t_z\}$ translation values. Bottom row: performance comparison with respect to a state-of-the-art pose estimator [48] in a subset of the AFLW set.

5.1 Pose Estimation

Let $\mathbf{p} \in \mathbb{R}^6$ be the ground-truth pose of a sample and $\hat{\mathbf{p}}$ be the pose configuration found by our algorithm. In Fig. 9, we give the box plots of the $\mathbf{p}-\hat{\mathbf{p}}$ values for each of the six pose degrees of freedom, showing the median of the errors (horizontal solid line) and their first and third quartile values (top and bottom of the box marks). The upper and lower whiskers are denoted by the horizontal lines outside each box, and the outliers are denoted by crosses. The upper row exemplifies three queries and the corresponding poses found by the algorithm. In these experiments we used 25,000 joint poses/head shape hypotheses, i.e., $t_{\tilde{p}} = 1,000$, $t_c^2 = 25$, indexed in binary trees of height 10 (≈ 50 hypotheses per leaf).

Overall, upon the availability of a sufficient number of pose hypotheses, the algorithm obtained a visually pleasant approximation of the query poses for the large majority of the cases. Objectively, we compared the performance of our pose estimator, with 6,250, 12,500 and 25,000 joint head shapes/pose hypotheses ($t_{\tilde{p}} = \{250, 500, 1{,}000\}$, $t_c^2 = 25$), to a state-of-the-art method due to Zhu and Ramanan [48], using the data set these authors supply.[1] The cumulative error curves are given in the bottom plot of Fig. 9, with the best configuration in our solution attaining performance close to the state-of-the-art, but using a much lower (and unfiltered) number of facial landmarks than the baseline. Overall, the gap in performance between both methods was the largest for low error values (where a larger number of landmarks would be particularly useful), and the results tended to converge for large cumulative errors which

1. http://www.ics.uci.edu/~xzhu/face/

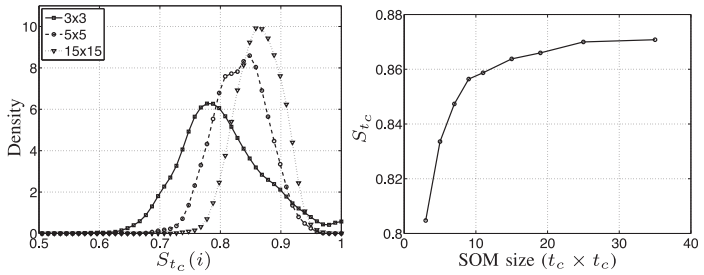


Fig. 10. Left: Probability density functions of the stability of labels per subject ($S_{t_c}(i)$). Right: Variations in the overall stability S_{t_c} with respect to the number of shape centroids considered.

correspond to rough pose estimates. For large cumulative errors - over $\frac{\pi}{4}$ - our method (with $t_{\tilde{p}} = 1{,}000$) attains better pose estimates than the baseline. We note that errors increase substantially when a reduced number of pose hypotheses are generated, particularly for $t_{\tilde{p}}$ values below 250. However, it should also be noted that generating large sets of hypotheses in the learning phase is not a concern, as the indexing strategy used accounts for several thousands of hypotheses without significantly increasing the temporal complexity of retrieval.

5.2 Soft Labels' Stability

The stability of the proposed soft labels varies per subject and depends of the number of SOM centroids. We define the stability of the ith subject as:

$$S_{t_c}(i) = 1 - \frac{1}{\sqrt{2} t_c t_i} \sum_{a=1}^{t_i} ||\mathbf{b}_{ia} - \bar{\mathbf{b}}_i||_2, \tag{14}$$

with $\mathbf{b}_{ia} \in \mathbb{N}^2$ being the ath sample label for the ith subject, $\bar{\mathbf{b}}_i$ being the subject centroid label ($\bar{\mathbf{b}}_i = \frac{1}{t_i} \sum_{a=1}^{t_i} \mathbf{b}_{ia}$), t_i being the number of samples of the subject and t_c denoting the number of columns/rows in the SOM (only square SOMs were considered).

For a set of subjects, a summary of their stability is given by $S_{t_c} = \frac{\sum_{i=1}^{t_s} S_{t_c}(i)\ t_i}{\sum_{i=1}^{t_s} t_i}$, t_s being the number of subjects. Fig. 10 depicts the stability of labels in the LFW data set, with respect to the number of centroids. The left plot gives three probability density functions for the $S_{t_c}(i)$ values using three typical SOM sizes. The right plot gives the group stability S_{t_c}, again as function of the SOM size.

The $S_{t_c}(i)$ values varied from around 0.63 (worst case for small maps) to 1 in the LFW set, with the optimal value observed for subjects with head shapes associated with cells in a SOM corner. Also, by using small SOMs (e.g., 3×3) the probability of obtaining near optimal stability values (all samples of a subject associated to the same cell) is increased, but so is also obtaining many more low stability values. Note that in small maps even small misalignments correspond to large normalized distances.

Overall, the summary stability S_{t_c} varied in direct correspondence with the number of cells in the SOM, converging for values around 0.87 in maps with more than 20×20 cells. Note that (14) provides relative distances with respect to the size of the SOM, i.e., values equal to 1 occur when two labels are separated by $\sqrt{2}\ t_c$ (a SOM diagonal). This explains why

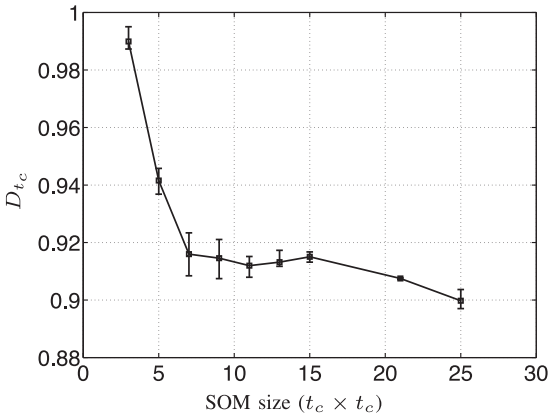


Fig. 11. Relationship between the labels' discriminability and the dimension of the SOMs used.

the stability values increase for larger SOMs, even though small maps should intuitively provide the maximum stability.

5.3 Soft Labels' Discriminability

The discriminability of labels was evaluated based on the flatness of the histogram that counts the number of subject centroids per cell, considering that discriminating labels should spread subjects evenly across the SOM cells. This is measured by an entropy function:

$$D_{t_c} = -\sum_{i=1}^{t_c^2} p_i \log_{t_c^2} p_i, \tag{15}$$

where p_i is the empirical probability that a subject centroid is associated with the ith cell of the SOM. In this case, the subject centroid labels were rounded to their closest cell. Being $D_{t_c} \in [0, 1]$, values close to 1 denote flat histograms, where subjects are spread evenly across the SOM cells. Values close to 0 are the non-interesting case, where most subjects are associated with a reduced number of cells.

Fig. 11 expresses the D_{t_c} values with respect to the SOM dimensions, having attained a maximum for the smallest maps (3×3), with an approximately equal number of subject centroids per cell. As the number of cells increased, some of the cells started to have too few centroids, while others attracted the elements in that region, yielding a more uneven distribution of the number of elements per cell.

Fig. 12 gives examples of the associations between the queries and the 3D head centroids for the LFW data set, using a 10×10 SOM. In each row, the leftmost image is the 3D head shape centroid (label) and the remaining images illustrate samples associated with that cell. Note the evident similarity between the major head features of the subjects and the centroids: at the upper-left extreme in the SOM, the (1,1) cell represents the largest heads with a round shape. At the other extreme, the (10,10) cell represents the most longitudinal heads with salient chins and extent maxillae. In this kind of mapping, cells in the corners provide the most easily distinguishable features (under visual inspection), while central cells are not so obviously distinguishable with respect to neighbors (note the high similarity between elements in cells (3,2) and (3,3)). Moreover, as the central region represents the most densely populated region of the feature space, a larger number of prototypes is used here, which accounts for the higher similarity between neighbouring prototypes.

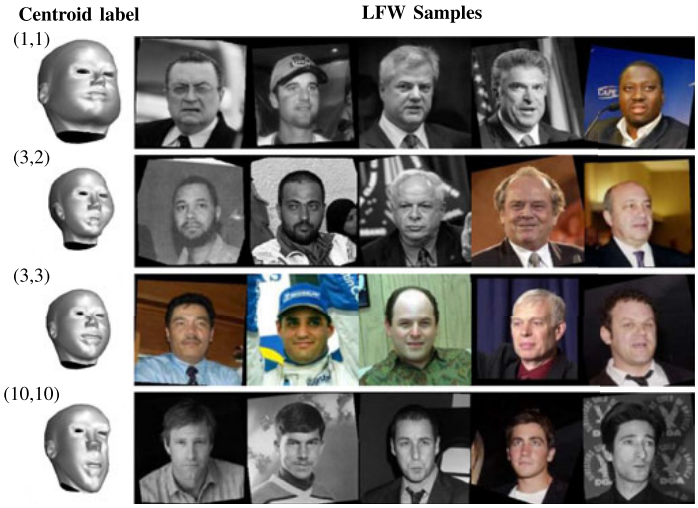


Fig. 12. Examples of associations between samples of the LFW data set and the head shape centroids of a SOM with 10×10 cells.

5.4 Robustness to Clutter

Poor-quality data queries are expected to be cluttered, i.e., with misplaced landmarks not corresponding to the anatomical region they are supposed to represent. This section addresses the effects of such cluttered input in the algorithm performance, which are two-fold: 1) increase the number of head shapes/pose hypotheses explicitly evaluated before convergence; and 2) decrease the convergence rate of the algorithm, which occurs when a solution is not found after evaluating the maximum number of hypotheses (100 in our experiments). Let p_s be the proportion of spurious landmarks with respect to the accurate detections (e.g., $p_s = 0$ represents a non-cluttered input and $p_s = 1$ denotes a balanced number of spurious/genuine landmarks). Using images of the AFLW set (with landmarks confirmed by human observers), cluttered inputs were simulated, by adding landmarks away from their true position (random x, y coordinates uniformly distributed over the entire image space, $\mathcal{U}(0,1)$, with coordinates normalized in the [0,1] interval) or by changing the position of a landmark (again, by generating uniformly distributed displacements over the image space, $\mathcal{U}$(0,1)).

As illustrated in Fig. 13, the algorithm convergence rate decayed with respect to p_s, but only slightly for values below one, which is readily achieved by state-of-the-art head landmark detectors. For larger p_s values, the convergence rate of the algorithm decays evidently and, for $p_s > 5$, the algorithm loses its effectiveness (bottom right plot). In terms of the number of hypotheses explicitly evaluated, an approximately direct linear relationship with respect to p_s was observed (bottom left plot). The top image in Fig. 13 illustrates a query with $p_s = 4$ and the output of the algorithm, where the landmarks deemed genuine (with $\phi() \geq 0$) appear in green and the spurious landmarks are denoted by the color red.

5.5 Soft Labels Standalone Performance: The Watch-List Problem

An important surveillance task is the watch-list problem: authorities have an explicit list of criminals (the *watch-list*) they want to locate or track among a population. Given a query, the goal is to detect occurrences of watch-list

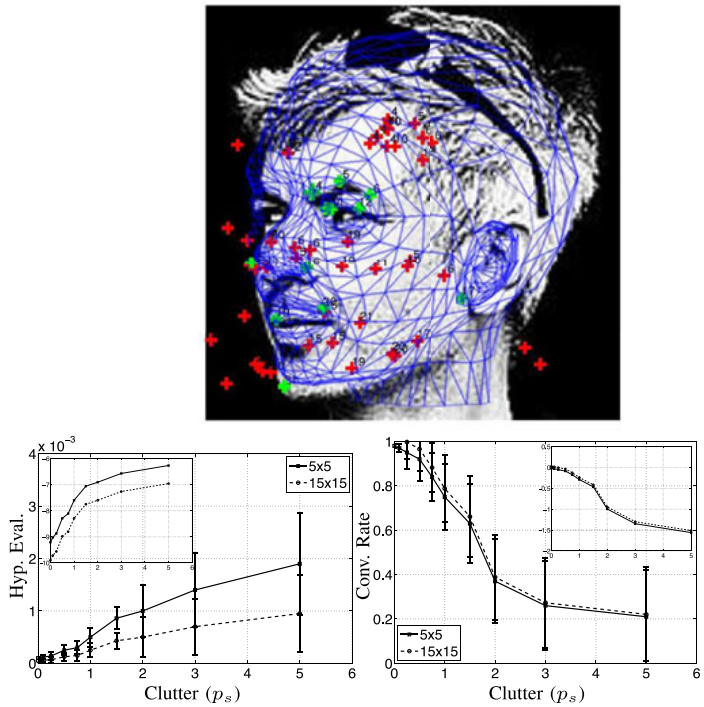


Fig. 13. Top: illustration of a query sample with spurious head landmarks ($p_s = 4$), where the proposed method was still able to correctly estimate the pose and the soft biometric label. Bottom-left plot: effect of the proportion of spurious landmarks in the number of joint pose/head shape hypotheses explicitly evaluated before convergence (given in linear and log scales). Bottom-right plot: decay in the convergence rate with respect to the proportion of spurious correspondences (linear and log scales).

elements without revealing the identities of any other subjects to the central authorities, which is considered a privacy-preserving policy.

The metric space formulation of labels is particularly suitable for handling this type of problem. By assigning a cell to each element in the watch-list, the topological properties of the input space ensure that any query assigned with cells located *sufficiently* far from the watch-list cell does not correspond to the criminal's identity. This is illustrated in the left diagram in Fig. 14. Depending on the radius δ used (which dictates the relationship between the hit/penetration rates), most of the identities in the watch-list can be confidently rejected. The plot given at the right side of Fig. 14 shows the probability density functions of observing distances d between intra-subject samples, which is the key for this watch-list formulation. Values are given for SOMs

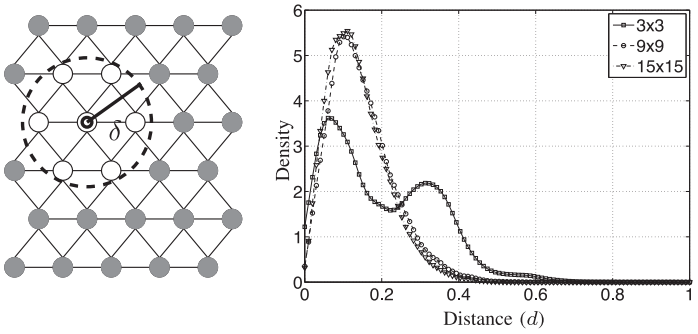


Fig. 14. At left: insight of the *negative identification* concept, used in the watch-list problem. All labels farther than δ of a query correspond to identities that can be rejected. At right: probability density functions of observing distances d between intra-subject labels (values regard the LFW data set).

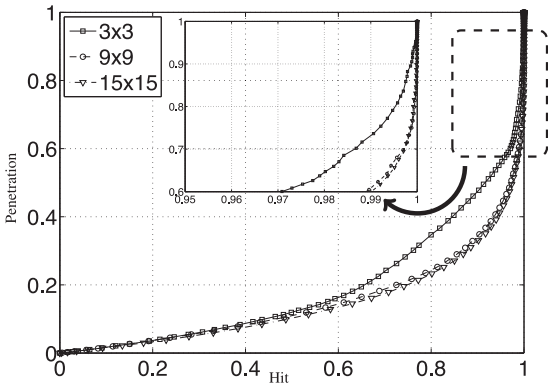


Fig. 15. Hit/penetration plots for the LFW data set, using SOMs of dimensions 3 × 3 (continuous line), 9 × 9 (dashed line) and 15 × 15 (dot-dashed line).

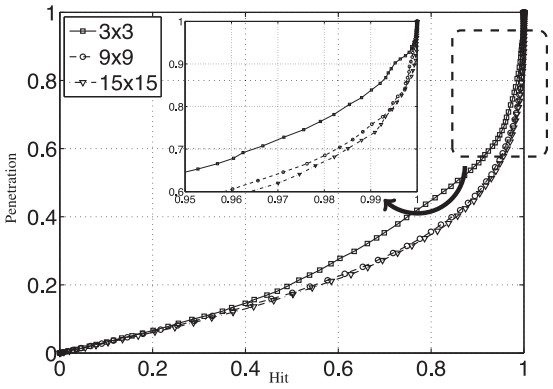


Fig. 16. Hit/penetration plots for the SCface data set, using SOMs of dimensions 3 × 3 (continuous line), 9 × 9 (dashed line) and 15 × 15 (dot-dashed line).

of three different sizes and enable us to conclude that there is a minimal probability of observing large distances (> 0.5) between intra-subject labels.

The suitability of the soft labels for watch-list identification is confirmed in the results given in Fig. 15, which expresses the hit/penetration values for the LFW set, using SOMs of dimensions 3 × 3 (continuous line with square marks), 9 × 9 (dashed line with circular marks), and 15 × 15 (dot-dashed line with triangular marks). The performance lines of the largest SOMs almost overlap and enable to reject over 50 percent of the identities for a query, keeping hit rates close to 99 percent.

Fig. 16 gives the hit/penetration values obtained for the SCface set, which are worse than the LFW values. This was justified by the small image resolution in the set, making the detection of head landmarks an extremely difficult task. Also, poses variations in this set are constrained to pitch angles (yaw and roll angles close to 0, pitch values in $[\pi/40, \pi/10]$), which led us to use only 100 pose prototypes. However, the key factor behind the relatively poor performance was that, in data of such reduced resolution, even small inaccuracies in landmark detection lead to large deviations in the 3D model positions inferred, which considerably reduced the stability of labels.

5.6 Fusion of Soft/Strong Traits: Recognition Performance

This section addresses the effectiveness of the soft labels to provide auxiliary information to a strong biometric expert. As in the previous sections, the LFW was used as main data set, having chosen the evaluation mode (*unsupervised*) that provides the lowest recognition performance among all protocols.[2] As a baseline, we considered the face recognition method due to Arashloo and Kittler [2] based on two reasons: 1) this method is among the best performers in the unsupervised (training free) LFW evaluation mode; and 2) it integrates well known techniques in a typical biometric recognition processing chain that could be easily applied to other traits (i.e., the ocular or the ear regions). It uses a multi-layered graphical model that measures the geometric distortion between image pairs, fed by the Daisy [39] feature descriptor. In classification, multi-resolution LBPs, image registration techniques and the cosine similarity yield the pairwise similarity score.

Note that the purpose of these experiments is not to obtain a system that outperforms the face recognition state-of-the-art, but to show that the proposed type of weak trait can be fused with strong systems and still improve the recognition performance with respect to the baseline. From this perspective, the *relative* performance between the ensemble and the baseline is most important than the absolute effectiveness rates. Also, note that other improvements in performance with respect to the baseline could be obtained by properly using the landmarks information provided by the soft expert inside the face recognition engine. However, that will be an attempt to improve a *specific* face recognizer, which is out of the scope of this paper.

The face and soft biometric experts were fused at the score level, learning a linear discriminant that projects both scores into the subspace that maximizes the Fisher discriminant ratio (found in a disjoint set composed by 10 percent of the available pairwise comparisons). Let ϵ_f be the pairwise similarity score returned by the face recognition expert and ϵ_s be the score returned by the soft expert:

$$\epsilon_s = \frac{1 + \operatorname{erf}\left(\kappa\left(\frac{\|\mathbf{b}_1 - \mathbf{b}_2\|_2}{\sqrt{2}t_c} - 0.5\right)\right)}{2}, \tag{16}$$

where erf() is a transfer function (error function) with sigmoid shape, $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{N}^2$ are the labels (of a $t_c \times t_c$ SOM) associated to the image pairs and κ is the parameter that controls the shape of the transfer function ($\epsilon_s \in [0, 1]$). Results are summarized in the Receiver Operating Characteristic curves of Fig. 17: the black line gives the baseline performance of the face expert, and the colored lines are the results attained by the ensemble, for three different shapes of transfer functions ($\kappa \in \{1, 2, 4\}$), with larger values corresponding to those farther from linear shapes. When compared to the baseline, the improvements in performance were maximized when the transfer function had the most pronounced sigmoid shape ($\kappa = 4$), i.e., when small misalignments between $\mathbf{b}_1$ and $\mathbf{b}_2$ were not excessively penalized. On the other hand, for roughly linear transfer functions ($\kappa \approx 1$), the performance of the ensemble was even slightly worse than the baseline.

Analyzing in detail the $\kappa = 4$ ensemble, we concluded that improvements in performance were due to reducing

2. http://vis-www.cs.umass.edu/lfw/results.html

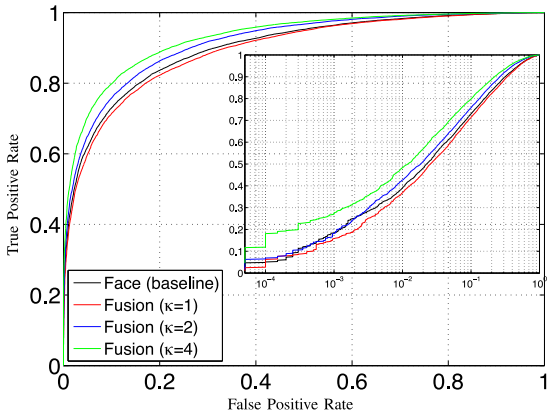


Fig. 17. Comparison between the recognition performance attained by a face recognition system in standalone mode and when using also the soft biometric labels as auxiliary information. Results are given for the LFW data set and regard the *unsupervised* evaluation mode.

the variability of intra-subject scores, typically by improving the pairwise scores when both samples had largely different poses, with the face recognition expert showing a particular sensitivity to such covariates (cases where the graphical model was not able to infer the appropriate deformation parameters). Conversely, we observed that the impostors' score distributions in the baseline and in the ensemble were almost equal.

5.7 Effect of Facial Expressions

Considering that facial expressions may significantly distort the head morphology (Fig. 18), this section addresses the effect of facial expressions on the soft labels from three perspectives. Initially, it reports the deviations in the SOM cells associated with intra-subject samples with neutral/non-neutral expressions. Next, it compares the labels' stability/discriminability in three different scenarios: 1) with 3D head shapes and queries having *neutral* expression; 2) with *neutral* 3D shapes against queries of *unconstrained* (*neutral* and *non-neutral*) expressions; and 3) with *unconstrained* 3D shapes and queries. Finally, it evaluates the suitability of the soft labels recognizing facial expressions.

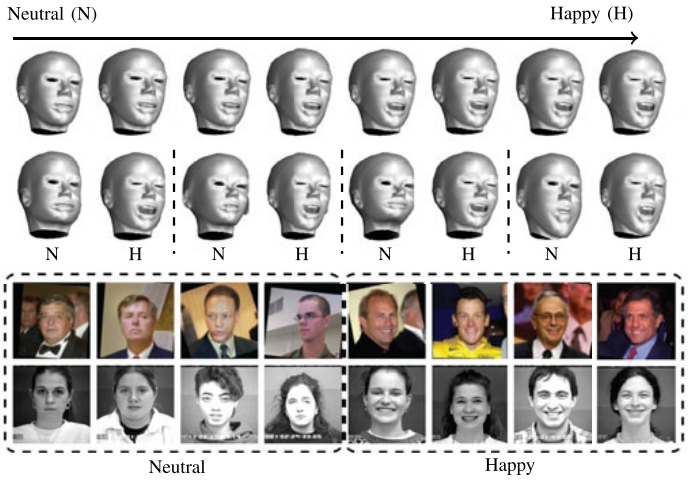


Fig. 18. Upper row: variations in the head shape appearance with respect to the levels of evidence of a facial expression (*happy*). Second row: pairwise samples of 3D head shapes with *neutral*/*happy* facial expression. Bottom row: division of the samples from the LFW and Cohn-Kanade data sets into two disjoint groups, according to their facial expression.

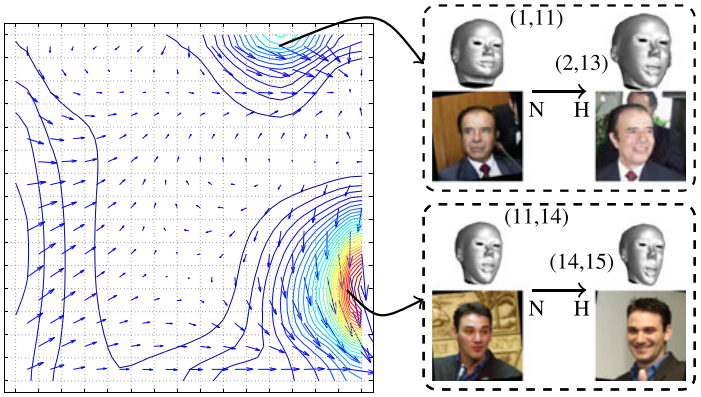


Fig. 19. At left: velocity plot representing the predominant intra-subject displacements in labels with respect to changes in facial expression from *neutral* to *happy* ($\mathbf{b}_i^{(h)} - \mathbf{b}_i^{(n)}$), using 3D head prototypes of exclusively *neutral* expression. At right: samples associated to the SOM regions with the largest movement slopes, i.e., where facial expressions imply the largest misalignments between the positions of soft labels in the SOM.

All the experiments were conducted using the previously mentioned LFW set, with images divided into disjoint groups, according to the facial expressions considered. Additionally, the Extended Kohn-Canade (CK+) [24] set was selected, being one of the most popular sets in this research topic. In terms of the facial expressions considered, we constrained the analysis to the *neutral* and *happy* expressions, due to two reasons: 1) the recognition of the remaining types of facial expressions (e.g., fear, disgust or sadness) implies the detection of action units that depend of an excessively large number of facial landmarks that cannot be detected in poor quality data; and 2) the LFW set has a small number of samples with other facial expressions (apart from *neutral* and *happy*), as they are unlikely in visual surveillance scenarios. In these experiments SOMs had 15×15 cells, maintaining all the κ_i values used previously.

Initially, only 3D head shapes of *neutral* expression were generated, with queries grouped per individual and per facial expression. For each subject, the centroid labels for *neutral* $\mathbf{b}_i^{(n)}$ and *happy* $\mathbf{b}_i^{(h)}$ expressions were found. The left plot in Fig. 19 gives the velocity plot corresponding to the $\mathbf{b}_i^{(h)} - \mathbf{b}_i^{(n)}$ misalignments, showing the average magnitude/direction of vectors representing the typical movements in SOM labels when expressions change from *neutral* to *happy*. It is evident that movements vary across the maps, with central regions being more stable than regions near the corners. Overall, movements converge in the bottom-right corner that represents the most elongated faces (with the largest deformations in the head shape due to the *happy* expression). The rightmost part of Fig. 19 gives two examples of *neutral*/*happy* head shapes falling in the SOM regions where the largest deviations were observed.

In addition, to perceive the decrease in soft labels effectiveness due to facial expressions, Fig. 20 compares the labels' stability/discriminability for three distinct configurations: 1) using 3D head shapes and queries exclusively of *neutral* expression; 2) using *neutral* head shapes and *unconstrained* queries (i.e., samples with *neutral*/*non-neutral* expressions); and 3) using *unconstrained* head shapes and queries. Results are given in terms of the hit/penetration plots and show that facial expressions consistently decrease the effectiveness of soft labels. However, such degradation

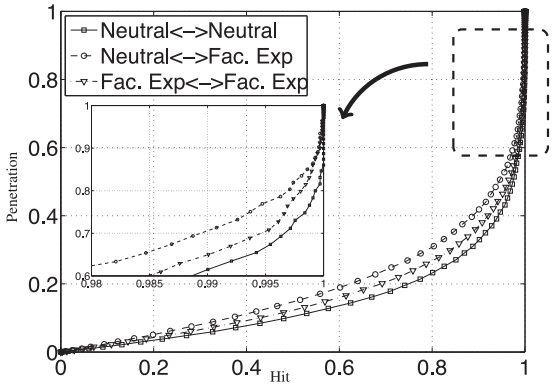


Fig. 20. Decreases in soft labels performance with respect to data of *non-neutral* facial expression. Using queries with *non-neutral* facial expression decreases the effectiveness of the soft labels, which can be counterbalanced if 3D head shapes with *non-neutral* expressions are also used (Fac. Exp. ↔ Fac. Exp. series). Results are given for 15 × 15 SOMs.

is counter-balanced if 3D shape hypotheses with *non-neutral* expressions are also generated, yielding results that are not too far off the baseline *neutral* against *neutral* configuration (at the expense of an increase in the computational burden of the labelling task by doubling the number of head shape hypotheses).

Finally, the suitability of the proposed method recognizing facial expressions in multi-pose data was assessed. We doubled the number of 3D head hypotheses, having generated for each *neutral* head shape a corresponding *happy* expression (second row in Fig. 18). Joint head shapes/pose hypotheses were clustered and indexed in the same way as before. Next, for each SOM cell $\mathbf{s}_{c_i}$, the number of *neutral*/*happy* 3D head shape hypotheses associated with it was assumed to give the class likelihood $p(\mathbf{s}_{c_i}|\theta)$ in that region of the feature space, with $\theta \in \{\text{“Neutral”}, \text{“Happy”}\}$. Then, any query assigned to $\mathbf{s}_{c_i}$ was classified in terms of facial expression according to the Bayesian paradigm, with the posterior probability for a facial expression given by $p(\theta|\mathbf{s}_{c_i}) \propto p(\mathbf{s}_{c_i}|\theta)\ .\ p(\theta)/p(\mathbf{s}_{c_i})$. Under this formulation, and using equal priors per class, queries are considered to have *neutral*/*happy* expressions according to the most frequent expression of the 3D head shape hypotheses associated with that cell.

The left plot in Fig. 21 illustrates the power of cells in a 15 × 15 SOM to discriminate facial expressions, showing the $|\mathbf{s}_{c_i}^{(n)}|/|(\mathbf{s}_{c_i}^{(n)}| + |\mathbf{s}_{c_i}^{(h)}|)$ per cell, $|\mathbf{s}_{c_i}^{(\cdot)}|$ being the number of 3D head shapes of *neutral* (n)/*happy* (h) expression associated with a cell. Values around 0.5 denote the non-interesting cases, i.e., cells with poor discriminating power (the number of *neutral* and *happy* elements is balanced). The right side of this same figure gives the confusion matrices for the LFW and CK+ sets, showing the mean and standard deviation performance values when repeating the recognition tests, using each time 85 percent of the available samples in a bootstrapping-like strategy. The results are below the state-of-the-art [9] method, mostly due to the poor discriminating cells with classification performance only slightly better than random. In our view, results would be improved if facial models with more facial landmarks are used, which in poor quality data would be hard to infer without filtering techniques (e.g., graphical models to obtain the optimum configuration from a set of candidate landmarks). Note that filtering landmarks would violate one constraint in this

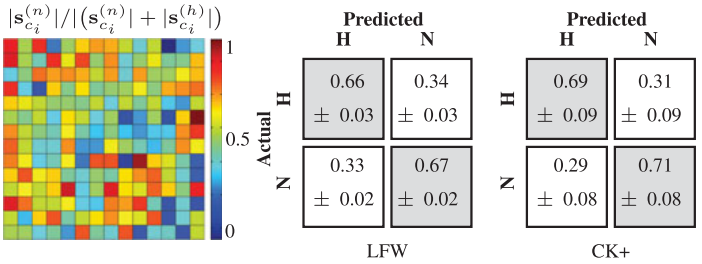


Fig. 21. At left: power of each SOM cell $\mathbf{s}_{c_i}$ to discriminate between *neutral* and *happy* expressions, expressed by the proportion of *neutral* head centroids associated to each cell. Red/orange cells represent predominantly *neutral* regions of the SOM space, whereas blue cells represent predominantly *happy* regions. Green/yellow cells have a balanced number of shapes per expression, making them particularly weak to discriminate between both classes. At right: confusion matrices for discriminating *neutral*/*happy* expressions in the LFW and CK+ sets.

paper: using exclusively *non-filtered* landmarks to enable real-time processing.

6 Conclusion

In this paper, we proposed a method to infer jointly human head poses and soft biometric labels based on the 3D morphology of the human head (the joint lengths between particular positions on the head). Using learning data from anthropometric surveys, a set of typical 3D head shapes (the labels) was inferred. Next, we described an algorithm to associate labels to low quality query samples, where subjects appear partially occluded and in varying poses. Using projective geometry techniques, we efficiently ranked a set of joint poses/head shape hypotheses, and iteratively evaluated the most likely hypothesis. The idea is to explicitly evaluate only a few hypotheses before the algorithm convergence, which is the key for the reduced temporal cost of the whole process.

The experiments were carried out using challenging data sets and support the usefulness of the soft biometric labels in two different ways: 1) coupled with a strong biometric classifier (e.g., a face recognizer), the resulting ensemble offers consistent improvements in performance over the strong expert alone; and, more importantly 2) these labels accord the concept of privacy-preserving recognition. In public environments, there are ethical/privacy issues behind the covert recognition of every subject passing-by. If soft labels are used, the system can confidently ignore the large majority of the identities in a scene and perform positive recognition only for a small subset of the subjects (those with soft labels similar to the watch-list elements).

Acknowledgments

This work was supported by FCT project UID/EEA/50008/2013.

References

[1] K. H. An and M. Chung, “3D head tracking and pose-robust 2D texture map-based face recognition using a simple ellipsoid model,” in *Proc. Int. Conf. Intell. Robots Syst.*, 2008, pp. 307–312.

[2] S. Arashloo and J. Kittler, “Efficient processing of MRFs for unconstrained-pose face recognition,” in *Proc. 2013 IEEE 6th Int. Conf. Biometrics: Theory, Appl. Syst.*, 2013, pp. 1–8.

[3] S. Ba and J.-M. Odobez, “Evaluation of multiple cue head pose estimation algorithms in natural environments,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1330–1333.

[4] R. Byrd, M. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optimization*, vol. 9, no. 4, pp. 877–900, 1999.

[5] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[6] J. Dass, M. Sharma, E. Hassan, and H. Ghosh, "A density based method for automatic hairstyle discovery and recognition," in *Proc. 4th Nat. Conf. Comput. Vis., Pattern Recog., Image Process. Graph.*, 2013, pp. 1–4.

[7] A. Drosou, D. Tzovaras, K. Moustakas, and M. Petrou, "Systematic error analysis for the enhancement of biometric systems using soft biometrics," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 833–836, Dec. 2012.

[8] B. Efraty, C. Huang, S. Shah, and I. Kakadiaris, "Facial landmark detection in uncontrolled conditions," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–8.

[9] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.

[10] M. Grgic, K. Delac, and S. Grgic, "SCface - surveillance cameras face database," *Multimedia Tools Appl. J.*, vol. 51, no. 3, pp. 863–879, 2011.

[11] D. Heckathorn, R. Broadhead, and S. Sergeyev, "Anthropometry of flying personnel-1950," Wright Air Develop. Center, Wright Patterson Air Force Base, OH, US, Tech. Rep. 52-321, 1954.

[12] H. Hertzberg, G. Daniels, and E. Churchill, "A methodology for reducing respondent duplication and impersonation in samples of hidden populations," in *Proc. Annu. Meeting Am. Sociol. Assoc.*, 1997, pp. 543–564.

[13] J. Hewig, R. Trippe, H. Hecht, T. Straube, and W. Miltner, "Gender differences for specific body regions when looking at men and women," *J. Nonverbal Behaviour*, vol. 32, no. 2, pp. 67–78, 2008.

[14] K. Huang and M. Trivedi, "Robust real-time detection, tracking and pose estimation of faces in video streams," in *Proc. Int. Conf. Pattern Recog.*, 2004, pp. 965–968.

[15] G. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[16] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" in *Proc. SPIE*, 2004, vol. 5404, pp. 561–572.

[17] A. K. Jain and U. Park, "Facial marks: Soft biometric for face recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 37–40.

[18] S. Jaiswal, T. Almaev, and M. Valstar, "Guided unsupervised learning of mode specific models for facial point detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 370–377.

[19] M. Koestinger, P. Wohlhart, P. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. 1st IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2144–2151.

[20] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.

[21] J. Lagarias, J. Reeds, M. Wright, and P. Wright, "Convergence properties of the Nelder-Mead simplex method in low dimensions," *SIAM J. Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[22] Y. Li, S. Gong, J. Sherrah, and H. Liddell, "Support vector machine based multi-view face detection and recognition," *Image Vis. Comput.*, vol. 1, no. 5, pp. 413–427, 2004.

[23] T. Lucas and M. Henneberg, "Comparing the face to the body, which is better for identification?" *Int. J. Legal Med.*, 2015, pp. 1–8.

[24] P. Lucey, J. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, 2010, pp. 94–101.

[25] M.-G. Kim, H.-M. Moon, Y. Chung, and S. Pan, "A survey and proposed framework on the soft biometrics technique for human identification in intelligent video surveillance system," *J. Biomedicine Biotechnol.*, article 614146, http://dx.doi.org/10.1155/2012/614146, 2012.

[26] M. Krinidis, N. Nikolaidis, and I. Pitas, "3-D head pose estimation in monocular video sequences using deformable surfaces and radial basis functions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 161–272, Feb. 2009.

[27] N. Kruger, M. Potzsch, and C. von der Malsburg, "Determination of face position and pose with a learned representation based on labeled graphs," *Image Vis. Comput.*, vol. 15, no. 8, pp. 665–673, 1997.

[28] B. Kwolek, "Model based facial pose tracking using a particle filter," in *Proc. Geometric Modell. Imaging - New Trends Conf.*, 2006, pp. 203–208.

[29] K. Moustakas, D. Tzovaras, and G. Stavropoulos, "Gait recognition using geometric features and soft biometrics," *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 367–370, 2010.

[30] F. Moreno-Noguer, V. Lepetit, and P. Fua, "Pose priors for simultaneously solving alignment and correspondence," in *Proc. Eur. Conf. Comput. Vis.*, 2008, part II, pp. 405–418.

[31] R. Osadchy, M. Miller, and Y. LeCun, "Synergistic face detection and pose estimation with energy-based models," *J. Mach. Learn. Res.*, vol. 8, pp. 1197–1215, 2007.

[32] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," *IEEE Trans. Inform. Forensics Security*, vol. 9, no. 3, pp. 406–415, Sep. 2010.

[33] V. Rapp, T. Senechal, K. Bailly, and L. Prevost, "Multiple kernel learning SVM and statistical validation for facial landmark detection," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recog.*, 2011, pp. 265–271.

[34] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross, "Soft biometrics for surveillance: An overview," *Handbook of Statistics*, vol. 31. Boston, MA, US: Newnes, 2013, pp. 327–351.

[35] D. A. Reid, M. S. Nixon, and S. Stevenage, "Soft biometrics; human identification using comparative descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1216–1228, Jun. 2014.

[36] A. Rice, P. J. Phillips, and A. O'Toole, "The role of the face and body in unfamiliar person identification," *Appl. Cognitive Psychol.*, vol. 27, no. 6, pp. 761–768, 2013.

[37] J. Sánchez-Riera, J. Östlund, P. Fua, and F. Moreno-Noguer, "Simultaneous pose, correspondence and non-rigid shape," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1189–1196.

[38] J. Sherrah, S. Gong, and E.-J. Ong, "Face distributions in similarity space under varying head pose," *Image Vis. Comput.*, vol. 19, no. 12, pp. 807–819, 2001.

[39] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[40] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. Nixon, "Soft biometrics and their application in person recognition at a distance," *IEEE Trans. Inform. Forensics and Security*, vol. 9, no. 3, pp. 464–475, Mar. 2014.

[41] J. Tu, T. Huang, and H. Tao, "Accurate head pose tracking in low resolution video," in *Proc. Int. Conf. Automatic Face Gesture Recog.*, 2006, pp. 573–578.

[42] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2D/3D active appearance models," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 2, pp. 535–542.

[43] J. Wu and M. Trivedi, "A two-stage head pose estimation framework and evaluation," *Pattern Recog.*, vol. 41, no. 3, pp. 1138–1158, 2008.

[44] J. W. Young, "Head and face anthropometry of adult U.S. civilians," Office of Aviation Medicine, Federal Aviation Administration, DOT/FAA/AM-93/10, 1993.

[45] C. Zhang and F. Cohen, "3-D face structure extraction and recognition from images using 3-D morphing and distance mapping," *IEEE Trans. Image Process.*, vol. 11, no. 11, pp. 1249–1259, Nov. 2002.

[46] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recog.*, vol. 42, no. 11, pp. 2876–2896, 2009.

[47] Z. Zhu and Q. Ji, "Robust real-time face pose and facial expression recovery," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recog.*, 2006, vol. 1, pp. 681–688.

[48] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Comput. Soc. Comput. Vis. Pattern Recog.*, 2012, pp. 2879–2886.

Hugo Proença (SM'12) received the BSc degree in 2001, the MSc degree in 2004 and the PhD degree in 2007. He is an associate professor at the University of Beira Interior and has been researching mainly about biometrics and visual-surveillance. He is the coordinating editor of the *IEEE Biometrics Council Newsletter* and the area editor (ocular biometrics) of the *IEEE Biometrics Compendium Journal*. He is a member of the Editorial Board of the *International Journal of Biometrics* and served as guest editor of special issues of the *Pattern Recognition Letters*, *Image and Vision Computing and Signal*, *Image and Video Processing* journals. He is a senior member of the IEEE.

João C. Neves (M'15) received the BSc and MSc degrees in computer science from the University of Beira Interior, Portugal, in 2011 and 2013, respectively. He is currently working toward the PhD degree from the same university in the area of biometrics. His research interests include computer vision and pattern recognition, with a particular focus on biometrics and surveillance. He is a member of the IEEE.

Silvio Barra received the BSc degree (cum laude) and the MSc degree (cum laude) in computer science from the University of Salerno, in 2009 and 2012, respectively. Since December 2012, he has been currently working toward the PhD degree at the University of Cagliari. His main research interests include pattern recognition, biometrics and video analysis, and analytics.

Tiago Marques was in management at the Instituto de Economia e Gestão, Lisbon, Portugal, between 2008 and 2011. He is currently working toward the undergraduate degree at the Computer Science Department, Universidade da Beira Interior and a researcher at the SOCIA Lab. He was at the *Induscria*, a Creative Industry's cross-platform association where he coordinated projects between multiple companies, in 2012.

Juan C. Moreno received the BSc and MSc degrees in mathematics from the Central University of Venezuela, in 2004, and the Simón Bolívar University, in 2008, respectively. In 2013, he received the PhD degree in mathematics at the University of Coimbra. He is a research fellow at the Pattern and Image Analysis Group of the University, Beira Interior, Portugal. His interests revolve around mathematical based methods for image processing, computer vision, and pattern recognition problems.

Image and Vision Computing 55 (2016) 22–25

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Visible-wavelength iris/periocular imaging and recognition surveillance environments☆,☆☆

Hugo Proença*, João C. Neves

IT: Instituto de Telecomunicações, University of Beira Interior, Portugal

ARTICLE INFO

Article history:
Received 3 February 2016
Accepted 29 March 2016
Available online 8 April 2016

Keywords:
Visual surveillance
Non-cooperative recognition
Iris/periocular recognition

ABSTRACT

Visual surveillance cameras have been massively deployed in public urban environments over the recent years, as a crime prevention and law enforcement solution. This fact raised the interest in developing automata to infer useful information from such crowded scenes (from abnormal behavior detection to human identification). In order to cover wide outdoor areas, one interesting possibility is to combine wide-angle and pan–tilt–zoom (PTZ) cameras in a master–slave configuration. The use of fish-eye lenses allows the master camera to maximize the coverage area while the PTZ acts as a foveal sensor, providing high-resolution images of the interest regions. This paper addresses the feasibility of using this type of data acquisition paradigm for imaging iris/periocular data with enough discriminating power to be used for biometric recognition purposes.

1. Biometrics in surveillance environments

Recent attacks in crowded urban environments reduced the perception of safety in modern societies, while the citizens' tolerance to reasonable risks has been also decreasing. There are now growing needs of assuring the safety of people, particularly in places/events that concentrate large crowds, which are naturally perceived as those with the highest risk (due to e.g., 2001 New York 9/11, 2004 Madrid train bombing, 2013 Boston marathon and 2015 Paris events). To counterbalance this fear, visual surveillance is now deployed massively worldwide. The amount of surveillance cameras running has grown astonishingly in the recent years, with more than 5.9 million CCTV cameras reported only in the United Kingdom [1]. However, contrary to popular belief, there are still no fully automatic techniques to identify subjects without requiring their participation in data acquisition, and the automated understanding of data is most times reduced to action recognition. For every identification attempt, it is still required some kind of human intervention in the process. Even though national/international authorities have lists of potentially harmful individuals, it is particularly difficult for humans to confirm whether such elements are among a crowd. As an example, the *TIDE: Terrorist Identities Datamart Environment* from the U.S. National Counterterrorism Center has over 745,000 people listed in the database which authorities are willing to arrest, but only a small proportion of these was actually detected in visual surveillance systems.

One interesting possibility is using coupled wide-angle and PTZ devices, which are able to acquire high resolution images on arbitrary scene locations. In this kind of configuration, a master–slave paradigm is usually adopted, i.e., the wide-angle camera covers the whole scene and provides data both for detecting and tracking subjects, also supplying 3D cues about the position to where the PTZ camera should be pointed to. While several advantages of this paradigm can be outlined, inter-camera calibration is the major bottleneck of this configuration, since determining the mapping function from image coordinates to pan–tilt parameters requires depth information. A solution to this problem is described in [2] and illustrated in Fig. 1: by inferring the subjects' height h, the depth ambiguity problem can be avoided and a univocal correspondence between positions in the wide-angle image data (x_s, y_s) and in the 3D physical coordinate system (X_p, Y_p, Z_p) can be obtained, enabling to infer the pan–tilt angle (θ_p, θ_t) values required to center the PTZ device at a particular position in the scene.

Another obvious difference between the operating requirements of systems working in visual surveillance scenarios and the traditional stop-and-stare protocol currently used is that in the former type of environments the number of targets usually exceeds the available active cameras, which demands schedule techniques to not only maximize the number of targets imaged but also the number of

☆ This paper has been recommended for acceptance by Sinisa Todorovic, PhD.
☆☆ This work was supported by FCT: Fundação Ciência e Tecnologia project UID/EEA/50008/2013.
* Corresponding author at: University of Beira Interior, R. Marques D'Avila e Bolama, 6201-001 Covilhã, Portugal.
E-mail addresses: hugomcp@di.ubi.pt (H. Proença), jcneves@di.ubi.pt (J. Neves).

http://dx.doi.org/10.1016/j.imavis.2016.03.015
0262-8856/

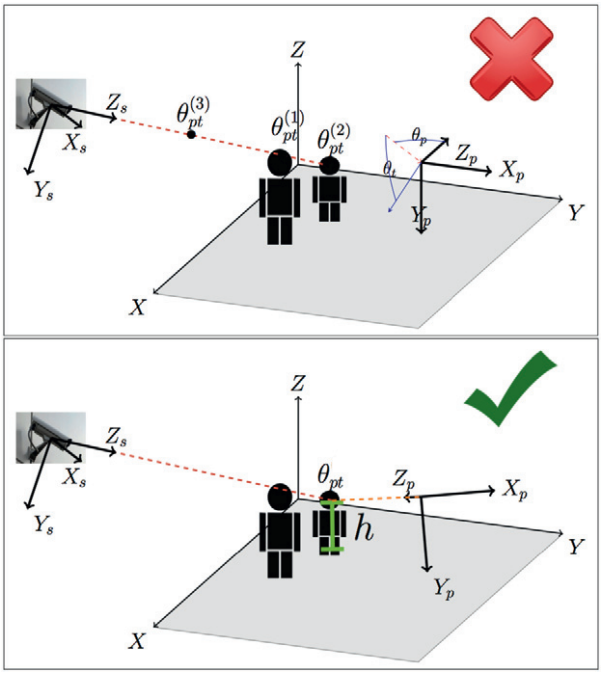


Fig. 1. Top image: illustration of the *depth ambiguity* problem, with several physical positions in the scene correspond to the same pan–tilt angles of the PTZ device ($\theta_{pt}^{(i)}$). Bottom image: by estimating subjects' height (h), it is possible to establish a bijection between physical 3D positions and pan–tilt angles θ_{pt}.

shots taken from each one. This is a variant of the classical *optimal tour* finding problem, which exhaustive solution has time cost $O(n!)$, being n the number of targets in the scene. Although brute-force might be feasible for a reduced number of targets, the real-time nature of the problem prohibits its use for more than six targets. Several works have presented solutions to this problem (e.g., [3] and [4]), where the contextual and dynamic scene information is considered to find the optimal sequence of targets (Fig. 2).

2. Related work

In order to consider an image with *acceptable* quality, the iris recognition standards recommend a resolution of at least 100 pixels across the iris diameter (ISO/IEC 2004) and an in-focus image. Also, sufficient near infrared (NIR) illumination should be ensured (more than 2 mW/cm^2) without harming human health (less than 10 mW/cm^2 according to the international safety standard IEC-60852-1). The space volume in front of the image acquisition system where these constraints met is usually referred as the capture volume of the system. Commercial iris recognition systems achieve extremely low error rates, yet imposing highly restrictive capture volumes that reduce the workability in less constrained scenarios. In recent years, several attempts to relax the constraints of iris recognition systems have been made, exploiting innovative strategies to increase both the capture volume and the stand-off distance, i.e., the distance between the front of the lens and the subjects.

Current strategies to perform the acquisition of iris data in less constrained conditions can be divided into two families, depending of whether they use (or not) magnification devices. In terms of the approaches that make no use of magnification devices, the Iris-on-the-Move [5] system is notable for having significantly decreased the cooperation levels required for image acquisition, allowing subjects continuous movement through a portal equipped with NIR illuminators. Another well known commercial device is the LG IrisAccess4000, where image is acquired at-a-distance, provided that subjects' gaze point at a specific direction.

Magnification devices, such as PTZ cameras, extend the system stand-off distance while providing enough resolution for reliable iris recognition. Wheeler et al. [6] introduced a system to acquire iris data at a resolution of 200 pixels from cooperative subjects at 1.5 m, using a PTZ camera assisted by two wide view cameras. Dong et al. [7] also proposed a PTZ-based system, that images iris data up to distances of 3 m with more than 150 pixels across the iris diameter. Yoon et al. [8,9] relied on a light stripe to determine the 3D position, avoiding the use of an extra wide camera. The *Eagle Eye* system [10] uses one wide view camera and three close view cameras, for capturing simultaneous images of both irises. This system has a stand-off distance of about 5 m with a operational range of 3 m × 2 m × 3 m. This system uses a bi-ocular setup, that enables to recover the 3D world position of the subject by stereo reconstruction. Depth information cues are used both for pan/tilt angles estimation and for getting focused data.

Despite being considered more reliable, the use of two wide-angle cameras significantly increases the system cost and limits its flexibility. To address this problem, various commercial solutions were introduced: Mitsubishi corporation developed a scheme where depth is estimated using the disparity between facial features [11]. Yoo et al. [12] combined the wide-view and narrow-view cameras with a beam splitter to simultaneously acquire facial and iris images. This integrated dual-sensor enables the same ray to be mapped to same position in both cameras sensors, avoiding the need for depth estimation.

3. Challenges

Most of the current iris recognition systems require that the iris is illuminated in the NIR wavelength band. Although this wavelength has the major advantage (with respect to the visible band) of avoiding corneal reflections from the surrounding light, the use of NIR illuminators highly restricts the workability of iris recognition in less constrained scenarios: the irradiance of the illuminators decreases quadratically as the stand-off distance increases, implying the use of extremely powerful illuminators for acquiring the rich details of the iris from large distances. As such, from our viewpoint non-cooperative iris recognition at such large stand-off distances such as in typical surveillance scenarios should be performed in the visible spectrum. Also, we believe that the use of magnification devices (PTZ) cameras is the most efficient solution to acquire iris with sufficient quality for recognition purposes.

We recently described a system—named QUIS-CAMPI—for acquiring high-resolution face imagery at large distances (up to 50 m) [2,13], but here we discuss its usability for unconstrained acquisition of iris/periocular data. This system uses a PTZ camera with full-HD resolution (1920 × 1080) and 30× optical zoom, corresponding to an angle-of-view of 2.1°. We used this framework to acquire close-up shots of the ocular region at standoff distances of 10 and 15 m, with examples for six subjects being illustrated in Fig. 3. Apart from the typical variability factors of unconstrained scenarios (e.g., occlusions due to eyelids and reflections, poorly focused and off-angle data), the resolution across the iris is a key factor for the reasonability of using a system as QUIS-CAMPI for iris recognition in unconstrained scenarios. Using as baseline the standard that recommends 100 pixels across the iris, it can be seen that we are able to get only about 60% and 40% of that resolution respectively at 10 m and 15 m stand-off distances.

In order to determine the maximum stand-off distance that can be afforded without compromising iris quality, we investigate how

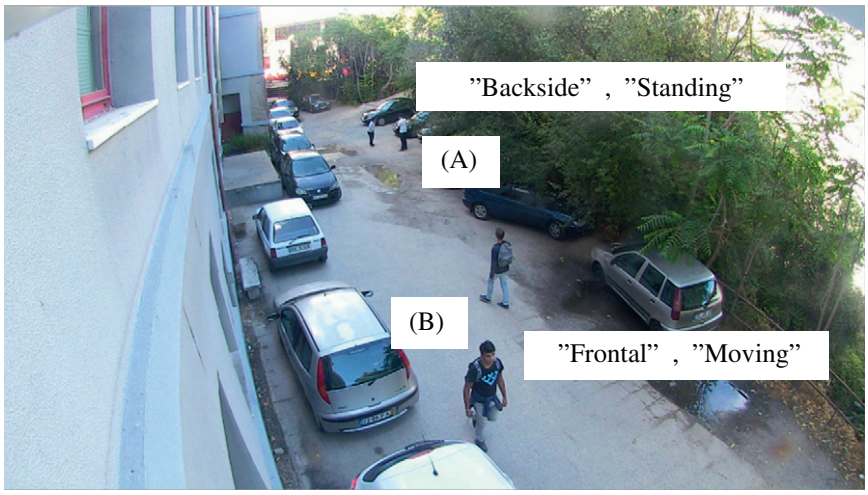


Fig. 2. Target selection: having n subjects in the scene, it is important to consider the scene context and dynamics (e.g., number of samples taken previously from each subject, subjects' position, velocity and perspective) to find the *optimal PTZ tour*, i.e., the data acquisition order. In this example, (B) is evidently a better target than (A).

the number of pixels across the iris ($\psi(d)$) is affected by the stand-off distance (d), angle of view (α) and camera resolution horizontal resolution (ω). Assuming that the human iris has an average diameter of 1.2 cm, the number of pixels across the iris is given by:

$$\psi(d,\alpha,\omega) = \frac{1.2 \times \omega}{2d\tan(\frac{\alpha}{2})}. \quad (1)$$

This relation is depicted in Fig. 4 when using the PTZ camera in QUIS-CAMPI (blue) and when using the next generation of PTZ cameras (4K). For comprehensibility, the chart is divided into three regions according to the quality of data with respect to the resolution factor. It is evident that state-of-the-art PTZ cameras are not sufficient to image iris at large stand-off distances, but it is worth noting that 4 K PTZ cameras should be available soon, and that their maximum optical zoom is also expected to increase. As illustrated in Fig. 4 such type of devices should have an obvious impact in the resulting image resolution, and allow the iris imaging with reasonable quality up to stand-off distances of over 15 m.

Another noteworthy possibility is the use of periocular region as the main biometric trait, which has been advocated as an interesting possibility to increase the robustness of iris recognition in visible-light data. The idea is to compensate for the degradation in the iris data by also considering the discriminating information in the surroundings of the eye (eyelids, eyelashes, eyebrows and skin texture). Even though further empirical validations are required to confirm the reasonability of using the periocular region as biometric trait in this type of data, it is known that periocular recognition is much less demanding in terms of data resolution than iris recognition.

4. Conclusions

Developing automata able to perform biometric recognition in crowded scenes and without explicitly requiring any active human

Fig. 3. Examples of facial data acquired in surveillance environments with the corresponding iris/periocular regions. The upper row contains samples acquired 10m away from the subjects, while the bottom row illustrates images acquired from 15m away.

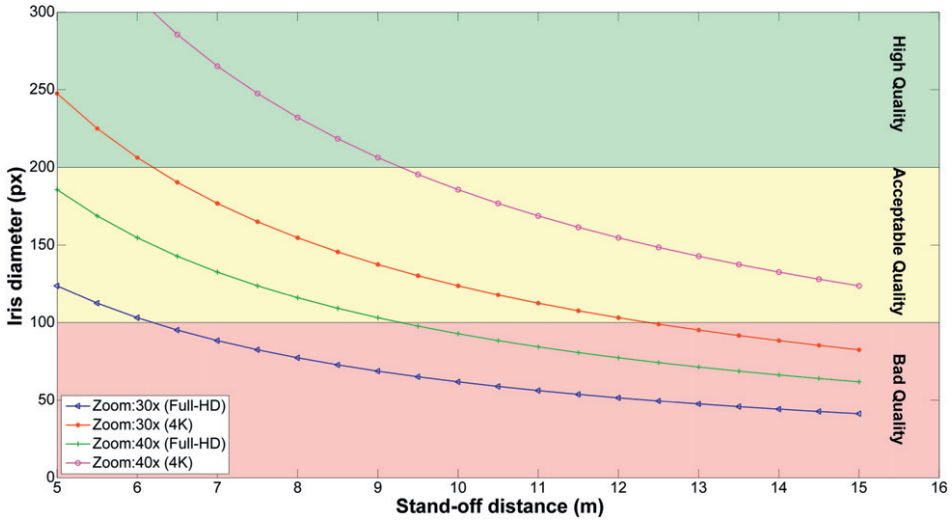


Fig. 4. Relation between the number of pixels across the iris and the stand-off distance when using a different magnification zoom and resolution for the PTZ camera. Considering the current standards as baseline, current PTZ cameras (blue line) do not appropriately image iris data at large stand-off distances. On contrast, 4K resolution cameras (pink line) extend the maximum stand-off distance of the system up to 15 m away from the subjects.

effort in the data acquisition process is an ambition that dates back—at least—to 1949, as a result of the widely famous George Orwell's *Big Brother* character. Even though such type of machine raises evident concerns from the ethical/privacy protection perspectives, it is also obvious that it will constitute a valuable law enforcement/security tool. Among several alternatives, one interesting possibility for such kind of system is to use coupled wide-angle and PTZ devices, that not only cover large outdoor areas, but are also able to acquire high-resolution data from moving subjects and large distances. In this paper we discussed some of the major differences between the processing chains of such type of non-cooperative recognition systems and of the current biometrics operating mode. Also, we illustrated the variations in the resulting facial/iris data with respect to the subjects stand-off distance factor and speculate about the suitability of using the periocular region as main biometric trait in such conditions.

References

[1] D. Barret, One surveillance camera for every 11 people in Britain, says CCTV survey, 2014. the Telegraph (, July 10, 2006. Retrieved June 17, 2014)http://www.telegraph.co.uk/technology/10172298/.
[2] J.C. Neves, J.C. Moreno, S. Barra, H. Proença, Acquiring High-resolution face images in outdoor environments: a master–slave calibration algorithm, IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems (BTAS), 2015. pp. 1–6.
[3] N. Kariotoglou, D. Raimondo, S. Summers, J. Lygeros, A stochastic reachability framework for autonomous surveillance with pan–tilt–zoom cameras, 50th IEEE Conference on Decision and Control and European Control Conference, 2011. pp. 1411–1416.
[4] J.C. Neves, H. Proença, Dynamic camera scheduling for visual surveillance in crowded scenes using Markov random fields, 12th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), 2015. pp. 1–5.
[5] J.R. Matey, O. Naroditsky, K. Hanna, R. Kolczynski, D.J. Lolacono, S. Mangru, M. Tinker, T.M. Zappia, W.Y. Zhao, Iris on the move: acquisition of images for iris recognition in less constrained environments, Proc. IEEE 94 (11) (2006) 1936–1947.
[6] F.W. Wheeler, G. Abramovich, B. Yu, P.H. Tu, et al. Stand-off iris recognition system, IEEE International Conference on Biometrics: Theory, Applications and Systems, (BTAS), 2008. pp. 1–7.
[7] W. Dong, Z. Sun, T. Tan, A design of iris recognition system at a distance, Chinese Conference on Pattern Recognition, (CCPR), 2009. pp. 1–5.
[8] S. Yoon, H.G. Jung, J.K. Suhr, J. Kim, Non-intrusive iris image capturing system using light stripe projection and pan–tilt–zoom camera, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007. pp. 1–7.
[9] S. Yoon, H.G. Jung, K.R. Park, J. Kim, Nonintrusive iris image acquisition system based on a pan–tilt–zoom camera and light stripe projection, Opt. Eng. 48 (3) (2009) 1–15.
[10] F. Bashir, P. Casaverde, D. Usher, M. Friedman, Eagle-eyes: a system for iris recognition at a distance, IEEE Conference on Technologies for Homeland Security, IEEE 2008, pp. 426–431.
[11] G. Guo, M. Jones, P. Beardsley, A system for automatic iris capturing, MERL TR2005-044, 1. 2005.
[12] J-H. Yoo, B. Kang, A simply integrated dual-sensor based non-intrusive iris image acquisition system, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015. pp. 113–117.
[13] J.C. Neves, G. Santos, S. Filipe, E. Grancho, S. Barra, F. Narducci, H. Proença, Quis-Campi: extending in the wild biometric recognition to surveillance environments, ICIAP Workshops, 2015. pp. 59–68.

Journal of Classification 34:85-107 (2017)
DOI: 10.1007/s00357-017-9223-0

Fusing Vantage Point Trees and Linear Discriminants for Fast Feature Classification

Hugo Proença

Instituto de Telecomunicações (IT), University of Beira Interior, Portugal

João C. Neves

University of Beira Interior, Portugal

Abstract: This paper describes a classification strategy that can be regarded as a more general form of nearest-neighbor classification. It fuses the concepts of *nearest neighbor*, *linear discriminant* and *Vantage-Point* trees, yielding an efficient indexing data structure and classification algorithm. In the learning phase, we define a set of disjoint subspaces of reduced complexity that can be separated by linear discriminants, ending up with an ensemble of simple (weak) classifiers that work locally. In classification, the closest centroids to the query determine the set of classifiers considered, which responses are weighted. The algorithm was experimentally validated in datasets widely used in the field, attaining error rates that are favorably comparable to the state-of-the-art classification techniques. Lastly, the proposed solution has a set of interesting properties for a broad range of applications: 1) it is deterministic; 2) it classifies in time approximately logarithmic with respect to the size of the learning set, being far more efficient than nearest neighbor classification in terms of computational cost; and 3) it keeps the generalization ability of simple models.

Keywords: Vantage-point tree; Linear discriminants; Nearest neighbor classification.

Corresponding Author's Address: H. Proença, Instituto de Telecomunicações (IT), University of Beira Interior, Tel.: +351 275 242 081, Fax: +351 275 319 899, email: hugomcp@di.ubi.pt.

Published online: 20 March 2017

1. Introduction

Supervised feature classification (or simply classification) is the assignment of a category (class) to an input value, on the basis of a learning set. For n pairs $(\vec{x}_i, y_i)$, $\vec{x}_i$ is represented in a d-dimensional space Ω and y_i its label (class), the goal is to find a function $f : \Omega^d \rightarrow \mathbb{N}$ that maps feature points into labels. In this context, an *ensemble* $f_e : \Omega^{d \times m} \rightarrow \mathbb{N}$ combines the output of multiple classifiers and seek to improve performance, when compared to individual elements. Ensemble classifiers are widely seen in the literature (e.g., Kuncheva, 2004 and several performance evaluation initiatives were conducted (e.g., Bauer and Kohavi, 1999; Banfield et al., 2007; Dietterich, 2000; Alpaydin, 1999; Demsar, 2006).

Under a computational perspective, the burden of classification is a primary concern for various domains (e.g., computer vision), due to extremely large amounts of data or to very demanding temporal constraints. Here, the turnaround time of classification is usually more concerning than the one of learning, as the latter is carried out off-line and a reduced number of times. As an example of demanding temporal requirements for classification, the defect detection in industrial environments can be referred, where a high number of frames per second must be processed (Kumar, 2008).

This paper proposes an efficient ensemble algorithm based on the concepts of *Vantage-Point* trees (Yianilos, 1993) and linear discriminant analysis, and is from now on designated as *Vantage-Point Classification* (VPC). The idea accords the philosophy of boosting and combines a set of base (weak) classifiers learned from feature subspaces and positioned in leaves of tree. The insight is that, regardless the complexity of the feature space, a linear discriminant is able to separate classes at a sufficiently deep level in the tree. Classification results from weighted voting of the base classifiers, selected according to the distance between the unlabelled sample and the centroid of each subspace. We come out with a solution that has two interesting properties: it attains classification accuracy similar to the state-of-the-art techniques in different problems, and it is efficient in terms of the computational cost of classification.

Hence, the major findings reported in this paper are: 1) for a broad range of the problems considered, VPC obtains better performance than state-of-the-art individual and ensemble models; 2) improvements in classification accuracy were observed along with a decrease in the computational cost of classification, when compared to related models (e.g., neural networks and K-nearest neighbors). Also, it should be stressed that experiments were carried out in datasets widely used in the classification domain (*UCI Machine Learning Repository, Univ. California*), that vary in terms of different criteria: binary/n-ary classification, discrete/continuous features,

balanced/unbalanced prior probabilities and densely/sparsely populated feature spaces.

The remainder of this paper is organized as follows: Section 2 summarizes the most relevant ensemble classification methods. Section 3 provides a description of the VPC algorithm. Section 4 presents and discusses the results. Finally, the conclusions are given in Section 5.

2. Related Work

Sharing several properties with the method proposed in this paper, Ting et al. (2011) proposed the concept of *Feating*, a generic ensemble approach that is claimed to improve the predictive accuracy of both stable and unstable classification models. As in our case, their original concept is that *"a local model formed from instances similar to one we wish to classify will often be more accurate than a global model formed from all instances"* (Frank, Hall, and Pfahringer, 2003). The idea is to divide the feature space into a set of disjoint subspaces, according to user-specified features that control the level of localization. Their trees use, at a given level, the same attribute for feature subspace division; which does not happen in our case, and we claim to be a much more intuitive variant, i.e., the feature that best divides a subspace is not guaranteed to optimally divide another subspace, even if both spaces are represented at the same level of the tree. Another major difference is that in Feating, all models in the ensemble are used for every query (as in Bagging), while in VPC only the models that regard the closets subspaces to the query instances are used, in a weighted way. Comparing the results observed for our strategy and the results reported in Ting et al. (2011), a major advantage is the ability of our ensemble model to get error rates comparable to state-of-the-art classification algorithms, while keeping a relatively short ensemble size, i.e., without building classification trees that are impracticable for most situations.

According to Canuto et al. (2007), there are two major ensemble categories: 1) hybrid ensembles, that combine different types of algorithms; and 2) non-hybrid ensembles, where a single type of algorithm is replicated multiple times. The most popular schemes are non-hybrid and apply a base algorithm to permutated training sets. Among these, *Bagging* and *Boosting* are the most prominent strategies. Originally proposed by Breiman (1996), Bagging (Bootstrap aggregating) builds multiple models for a problem, each one based on a subset of the learning data. Then, voting combines the outputs, improving stability and accuracy when compared to base models. Kuncheva and Rodríguez (2007) replaced each classifier by a *mini-ensemble* of two classifiers and a random linear function (*oracle*). In classification, the oracle decides which classifier to use. The classifiers and oracle are trained

together, so that each classifier is pushed into a different region of the feature space. Also, Hothorn and Lausen (2005) construct a set of classifiers for different bootstrap samples. Then, the predictions of such classifiers in each element of the bootstrap are used for a classification tree. This tree implicitly selects the most effective predictors, which authors consider to *bundle* the predictions for the bootstrap sample. Classification is done by averaging the predictions from a set of trees. In a related topic of tree-based regression models, Ceci, Appice and Malerba (2003) used two basic operations (pruning and grafting) to obtain simpler structures for the regression tree. The core of these simplification operations is to put aside some data for independent pruning, which was observed in practice to improve classification performance.

The idea of Boosting resulted from the stochastic discrimination theory (Kleinberg, 1990), a branch that studies the ways to divide the feature space for class discrimination. This algorithm combines several weak classifiers, each one with high bias and low variance. Experiments point out that is possible to meet high accuracy far before using all the weak classifiers. A relevant boosting method was due to Shapire (1990), but the most well-known is the Adaboost variant (Freund and Schipire, 1995), that increases adaptability by tweaking subsequent classifiers in favor of instances misclassified by previous ones.

Ho (1995) suggested the notion of *Random Forest*, from where a generalization was proposed (*Random Subspace:* Ho, 1998), later known as *Attribute Bagging* (Bryll, Gutierrez-Osuna and Quek, 2003). The idea is to build an ensemble of classifiers, each one using a subset of the available features. In classification, voting produces the final answer. In this context, Zaman and Hirose (2013) enlarged the feature space of the base tree classifiers in a random forest, by adding features extracted from additional predictive models, having empirically concluded that such *hybrid* random forests can be a more efficient tool than the traditional forests for several classification tasks.

Domingos (1996) described an algorithm based on rule induction and instance-based learning, considering individual instances as maximally specific rules, and then devising an algorithm to gradually fuse instances into more general rules. The proposed algorithm was considered an inductive learning approach that produces specialized rules that span the entire feature space, by searching for the best mixture of instances and increasingly augment abstraction of rules, yielding a more general-form of nearest neighbor classification.

Particularly interested in problems with a reduced amount of learning data, Lu and Tan (2011) sought for a subspace that minimizes the within-class to between-class distances ratio. To enlarge the amount of learning

data, they used a linear model to interpolate pairs of prototypes, simulating variants of the available samples. Bock, Coussement and Poel (2010) used generalized additive models as base classifiers for ensemble learning, proposing variants based on bagging and random subspaces. Classification yields from average aggregation.

In a related topic, instance selection aims at obtain a subset of the learning data, so that models on each subset have similar performance to the attained in the complete set. García-Pedrajas (2009) used instance selection in boosting processes, optimizing the training error by the weighted distribution of instances erroneously classified by previous models. Yu et al. (2012) divided the feature space into disjoint subspaces. Then, defined a neighborhood graph in each subspace and trained a linear classifier on this graph, used as base classifier of the ensemble. In classification, the majority-voting rule is used. Starting from models learned by random space and bootstrap data samples, Yan and Tešić (2007) estimated the decision boundaries for each class, concluding that a few shared subspace models are able to support the entire feature space. This scheme is claimed to reduce redundancy while enjoying the advantages of being built from simple base models.

Considering that—typically—the performance of classifier ensembles is maximized in case of Uniform distributions of observations, Jirina and Jirina Jr. (2013) suggested a transformation on the data space that approximates the distribution of observations in the feature space into a uniform distribution, at least in the neighborhood of a query observation. Their transformation is based on a scaling exponent that relates distances between pairs of points in a multivariate space.

As described in the next section, the VPC model shares some of the above referred foundations: similar to the concept behind boosting, we divide the feature space into subspaces, pushing each base classier into disjoint regions of the feature space. Similar to Lu and Tan (2011), the within-to-between class distances proportion is used to determine the number of divisions of the feature space. Then, by preserving neighborhoods between subspaces, for a given query we are able to select a subset of the base-classifiers in a computationally effective way.

3. Vantage Point Classification

Figure 1 illustrates the key idea behind the VPC scheme: the feature space is divided into subspaces Ω_i, according to the distance of elements to pivots p_i. Each subspace (leaf) is simple so that a linear discriminant Φ_i separates classes with a reduced expected error. In classification, for a query $\vec{x}$, only the closest subset of the leaves vote, according to the distances between $\vec{x}$ and p_i. This schema creates a set of spaces Ω_i where classification is done locally.

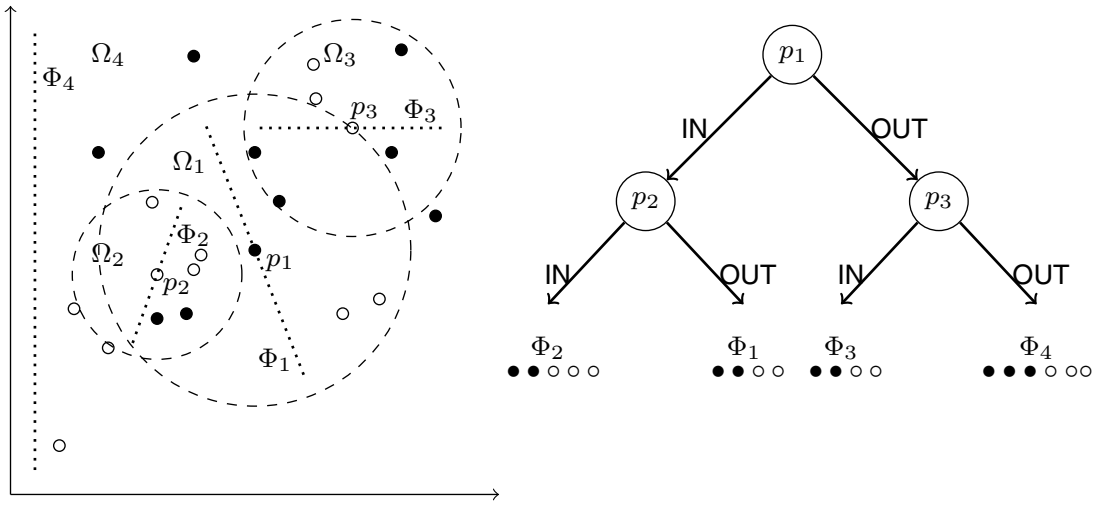


Figure 1. Illustration of the idea behind VPC. At left, the dotted line segments denote the projections Φ_i found for compact spaces Ω_i and the dashed circles denote the median distance between pivots p_i and elements on that space. This value is used to separate elements at each side of the vantage point tree. The right figure gives the VPC data structure corresponding to this feature space.

3.1 Learning

The learning process starts by evaluating if a linear projection Φ separates the feature space with a misclassification rate lower than γ. Let Ω be a d-dimensional feature space containing n instances $\vec{x}_i$ with labels y_i. According to Johnson and Wichern (1988), multiple discriminant analysis is a natural extension of the Fisher linear discriminant, having the within-class matrix given by:

$$\hat{\Sigma_w} = \sum_{c=1}^{k} \sum_{\vec{x}_i | y_c} (\vec{x}_i - \bar{\vec{x}}_{y_c})(\vec{x}_i - \bar{\vec{x}}_{y_c})^T, \tag{1}$$

where k is the number of classes, $\vec{x}_i|y_c$ denotes the elements in the cth class and $\bar{\vec{x}}_{y_c}$ is the centroid of these elements. The scatter matrix is given by:

$$\hat{\Sigma_s} = \sum_{c=1}^{k} n_{y_c} (\bar{\vec{x}} - \bar{\vec{x}}_{y_c})(\bar{\vec{x}} - \bar{\vec{x}}_{y_c})^T, \tag{2}$$

being n_{y_c} the number of training samples in class y_c. $\bar{\vec{x}}$ is the dataset mean vector. A linear transformation, Φ, is obtained by solving the generalized eigenvalue system:

$$\hat{\Sigma_s}\Phi = \lambda \hat{\Sigma_w}\Phi, \tag{3}$$

where λ is a scalar that is usually called the generalized eigenvalue of Σ_s and

Σ_w. Classification is done in the transformed space according to a distance function $\xi(.,.)$. Each query element $\vec{x}$ is classified by:

$$\hat{k_i} = \arg\min_k \xi(\vec{x}\Phi, \bar{x}_{y_k}\Phi). \tag{4}$$

Our stopping criterion for the division of Ω considers the error rate in the learning set, given by:

$$e(\Omega) = \frac{\sum_i \mathbb{I}_{\{k_i \neq y_c\}}}{n}, \tag{5}$$

where $\mathbb{I}$ is an indictor function that evaluates if the predicted and true class values are the same. When $e(\Omega) \leq \gamma$, we consider that Φ appropriately discriminates the feature space and the node is considered a leaf, with support $s(\Omega) = n$. Otherwise, Ω is divided into two halves, according to a pivot.

Note that the term *appropriately*, in terms of discrimination, is used to indicate subspaces where a linear discriminant attain classification error lower than γ. For $\gamma = 0$, the term is equivalent to linearly separable. In practice, for most cases the optimal performance is attained when $\gamma > 0$, i.e., stopping the division of subspaces when the number of elements per class is still much higher than the dimension of the feature space (otherwise, the algorithm would simply return the pseudo-inverse of the Fisher linear Discriminant at a leaf). $\gamma > 0$ values are regarded as a soft margins, i.e., we allow a few mistakes (some points - outliers or noisy examples might be on the wrong side of the linear discriminant), but most times obtain a solution that better separates the bulk of data.

If the ith element in Ω is used as pivot, the remaining elements with indexes $j \in \{1, \ldots, n\}, j \neq i$, span through its left or right branch, depending of the distance $\xi(\vec{x}_i, \vec{x}_j)$. Let Y_{ic}^0 and Y_{ic}^1 be the sets of labels of class c in the left and right branches, when using the ith element as pivot. The *suitability* of that pivot $s(i)$ is equal the support of the corresponding discriminant:

$$s(i) = -\sum_{j=0}^{1}\sum_{c=1}^{k} \frac{|Y_{ic}^{(j)}|}{n-1} \log_2\Big(\frac{|Y_{ic}^{(j)}|}{n-1}\Big), \tag{6}$$

where $|.|$ denotes set cardinality. The best pivot minimizes (6), i.e., is the one that puts all elements of each class in different branches of the tree:

$$\hat{i} = \arg\min_i s(i). \tag{7}$$

Let $d_{\hat{i}j} = \xi(\vec{x}_j, \vec{x}_{\hat{i}}), i = 1, \ldots, n, i \neq \hat{i}$ be the $n-1$ distances between the pivot and the remaining elements and let $d_{\hat{i}}^*$ be the median value of $\{d_{\hat{i}j}\}$. The jth element of the training set is included in the left Ω^L or right Ω^R

subsets, according to:

$$\begin{cases} \{\vec{x}_j, y_j\} \in \Omega^L & \text{if } d_{\hat{i}j} \geq d^*_{\hat{i}} \\ \{\vec{x}_j, y_j\} \in \Omega^R & \text{if } d_{\hat{i}j} \leq d^*_{\hat{i}}. \end{cases} \tag{8}$$

The process is repeated for Ω^L and Ω^R in a way similar to Ω, until the stopping criterion is verified for all the subspaces. In practice, the optimal performance of the VPC model is attained when $\gamma > 0$, i.e., stopping the learning process before having all elements of a single class, which contributes to avoid overfitting.

3.2 Classification

Classification is done by traversing the VPC tree and accumulating the support values of the class predicted at each leaf. Let $\vec{x}$ be an unlabelled element and γ^* the radius of the query. The classification of $\vec{x}$ is given by (9), where $l(\Omega)$ is an indicator function that discriminates between leaves ($l(\Omega) = 1$) and non-leaves ($l(\Omega) = 0$) nodes:

$$c(\vec{x}, \Omega) = \begin{cases} \vec{0} & \text{, if } l(\Omega) = 1 \ \wedge \xi(\vec{x}, \vec{x}_p) > \gamma^* \\ s(\Omega).\vec{v}_i & \text{, if } l(\Omega) = 1 \ \wedge \xi(\vec{x}, \vec{x}_p) \leq \gamma^* \\ \{s(\Omega^L) + s(\Omega^R)\} & \text{, if } l(\Omega) = 0 \ \wedge (d^* - \gamma^*) \leq \xi(\vec{x}, \vec{x}_p) \\ & \quad \leq (d^* + \gamma^*) \\ s(\Omega^L) & \text{, if } l(\Omega) = 0 \ \wedge \xi(\vec{x}, \vec{x}_p) \leq (d^* - \gamma^*) \\ s(\Omega^R) & \text{, if } l(\Omega) = 0 \ \wedge \xi(\vec{x}, \vec{x}_p) \geq (d^* + \gamma^*) \end{cases}, \tag{9}$$

being $\vec{x}_p$ the pivot of a node and $\vec{v_i}$ a unit vector with a single non-zero component at the ith position (corresponding to the ith predicted class). This way, $c(\vec{x}, \Omega)$ returns a vector with k elements, each one containing the accumulated support for the predicted class, i.e., $\vec{s} = \{s_1, \ldots, s_k\}$. The response given by the ensemble corresponds to the position where $c(\vec{x}, \Omega)$ is maximum:

$$\hat{i} = \arg\max_i \{s_i\}. \tag{10}$$

For comprehensibility, Algorithm 1 details the VPC classification scheme in terms of the computational steps. As input, the method receives the Vantage Point tree with a linear discriminant in each leaf. For a query element $\vec{x}$, the binary tree is traversed down to leaves. In each leaf, a linear discriminant predicts the class and the accumulation of the support values gives the final response.

Algorithm 1 VPC Classification Scheme

Require: Vantage Point Tree $\boldsymbol{V}$, unlabelled instance $\vec{x} \in \mathbb{R}^d$, radius query γ^*.

```
Get current node: c ← V.root;
Support values for each class: s⃗ ← [0, ..., 0]
if leaf(c) then
    Accumulate support: s⃗ ← LDA(c, x)
    return s⃗
end if
Get pivot: p ← c.pivot
Get median distance: m ← c.median
if distance(x⃗, p) ≤ m − γ then
    Accumulate support: s⃗ ← s⃗ + VPC(cn.left, x⃗, γ*)
end if
if distance(x⃗, p) ≥ m + γ then
    Accumulate support: s⃗ ← s⃗ + VPC(cn.right, x⃗, γ*)
end if
return s⃗
```

3.3 Usability And Completeness

According to the theory of classification ensembles, it is particularly important that the ensemble is fully *usable* and *complete*. Let Ω be the d-dimensional feature space with Ω_i compact subspaces. The usability of the projection Φ_i was approximated by:

$$P\big(\vec{x} \in \Omega_i\big) \approx \frac{\sum_j \mathbb{I}_{\{(x_j^{(1)},\ldots,x_j^{(d)}) \in \Omega_i\}}}{n}, \tag{11}$$

where $\mathbb{I}_{\{.\}}$ is the indicator function and n is the number of learning instances. Based on the stopping criterion used for learning, $P\big(\vec{x} \in \Omega_i\big) > 0,\ \forall i \in \{1, \ldots, k\}$, guaranteeing that every Φ_i is usable and that the ensemble is fully usable.

As described in Section 3.1, it can be stated that:

$$\Omega^d = \bigcup_i \Omega_i^d, \tag{12}$$

assuring that the ensemble completely covers the feature space (provided that $n \geq d$). Similarly, as:

$$\Omega_j^d \cap \Omega_i^d = \emptyset, \forall i, j \mid i \neq j, \tag{13}$$

the domains of each discriminant Φ_i are disjoint and the full diversity of the ensemble is also assured, i.e., every discriminant Φ_i is obtained from completely disjoint data with respect to the remaining discriminants, which reduces the probability of obtaining correlated models. The fact of using

disjoint data for building each discriminant might augment the probability of overfitting. As a counterbalance, we note that several nodes usually vote for a query, which reduces the probability of overfitting.

3.4 Computational Complexity

Here we analyze the time complexity, which can refer to the *learning* or *classification* phases. But, as the latter phase is done on-line and requires repeated execution, our efforts were concentrated in keeping low the complexity of the classification phase.

The learning phase has two major steps: 1) create the VPC tree; and 2) obtain a linear discriminant for each leaf node. Let n be the number of learning instances. Determining one pivot per node takes $O(n^2)$. The insertion always takes place at the deepest level, with complexity $O(h)$, being h the height of the tree (if the tree is balanced, $h = lg(n)$). Next, learning a linear discriminant for each leaf involves three major steps: 1) singular value decomposition to obtain the within and scatter matrices; 2) compute eigenvectors; and 3) solve the final linear system, which has $O(n\, d \min(n,d) + \min(n,d)^3)$ (Cai, He and Han, 2008) temporal complexity, being d the dimension of the feature space.

In classification, the complexity depends of the radius γ^*, varying between $O(n)$ (when all leaves vote) and $O(lg(n))$ (when a single leaf votes). Our experiments confirm that optimal performance is attained when a reduced number of nodes vote for the ensemble, yielding a temporal complexity around $O(\alpha\ lg(n))$, being α the number of nodes voting ($1 \leq \alpha \leq n$). At each leaf, the temporal complexity of classification is $O(d(k-1)X)$, being k the number of classes. Hence, the time complexity of classification is $O(\alpha d(k-1)) + O(\alpha lg(n))$. Keeping moderate values for k and d, the predominant term is clearly $O(\alpha lg(n))$. Keeping in mind that usually each node of the tree represents more than one training instance (due to the parameter $\gamma > 0$), it follows that $\alpha \ll n$, reducing the complexity to approximately logarithmic.

3.5 Bias-Variance Tradeoff

In VPC, the tradeoff between bias and variance depends of the number of classifiers in the ensemble and of the number of votes per query. The former is determined by γ and the latter by γ^*. Both γ_I and γ_q are in direct proportion to bias, and inversely correlated to variance. As illustrated in Figure 2, high values for γ and γ^* reduce the number of leaves, from where a large proportion is used in classification. At the other extreme, as the values of γ and γ^* decrease, more leaves are created, and smaller proportions of

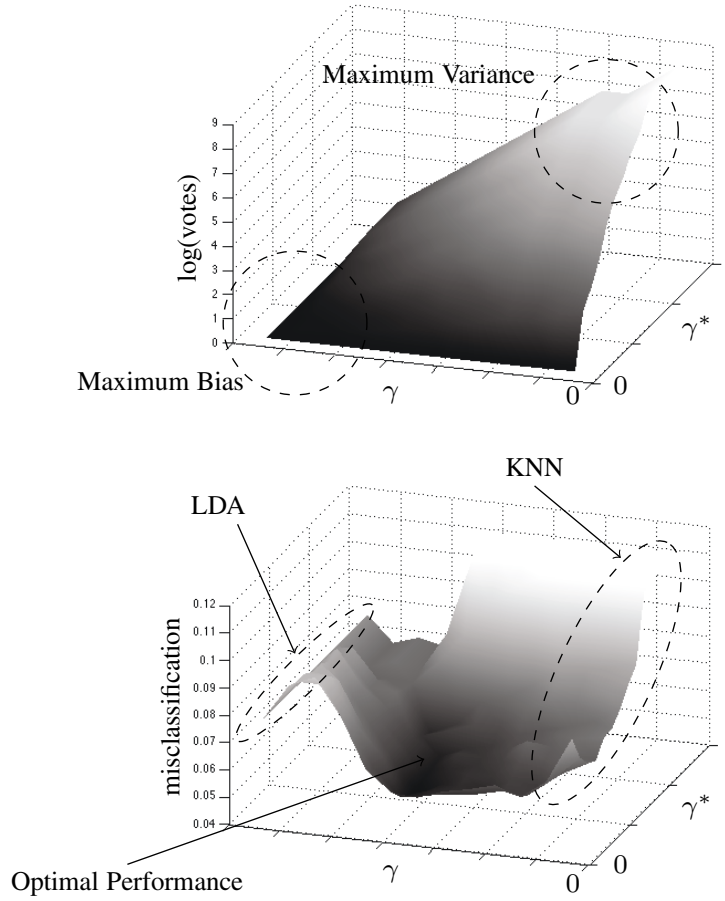


Figure 2. Top plot: effect of the γ and γ^* parameters in the number of classifiers that vote in each query. Bottom plot: corresponding misclassification rates, with the optimal configuration highlighted. The configurations that are equivalent to linear discriminant analysis and k-nearest neighbors appear inside dashed ellipses.

these are used in each query. This corresponds to getting purer—and less biased—estimates. However, a tree with more leaves reduces the sample size per classifier and increases the potential estimation error, increasing the variance of the model.

It is interesting to note that, under specific parameterizations, VPC is equivalent to linear discriminant analysis (LDA) or to k-nearest neighbors (KNN) (Figure 2). It is equivalent to LDA in cases where the tree is composed by a single node (large γ values). Oppositely, for $\gamma = 0$, each leaf of the tree offers perfect separability between classes, which in practice reduces to the nearest neighbor rule. However, as this example illustrates ("Optimal Performance" arrow), one of the key findings reported in this paper is that optimal performance is most times attained for intermediate values of γ and γ^*.

4. Experiments and Discussion

4.1 Comparison Terms

Eight well known classification techniques were used as comparison terms. Four individual models (k-nearest neighbors: Cover and Hart, 1967; linear discriminant analysis: Duda, Hart and Stork, 2000; neural networks: Moller, 1993; and support vector machines: Cortes and Vapnik, 1995) and four ensembles: Bagging with classification trees (CART), quadratic and pseudo-linear weak classifiers, Boosting and Random Spaces with quadratic discriminants and decision tree weak classifiers and the Random Forest (Breiman, 2001) with decision trees as weak classifiers.

The selected algorithms were considered to represent the state-of-the-art in terms of classification. Even though several variants exist, the ones used are the most frequently reported in the literature and in previous performance evaluation initiatives (e.g., Bauer and Kohavi, 1999; Dietterich, 2000; and Demsar, 2006). This way, our idea is that by transitivity, it is possible to compare the performance of VPC against any other classification model that shared any comparison term used in this paper.

Table 1 describes the parameters tested in the optimization of each algorithm. A set of *parameterizations* was tested, by an exhaustive grid combination of parameters in the given range. Additionally, for non-deterministic methods, each parameterization was tested 10 times and the best performance taken. Regarding the VPC method, the Euclidean distance ($\ell 2$-norm) was used in all cases as $\xi(.,.)$ function. All the results correspond to implementations in the MATLAB environment.

4.2 Synthetic Datasets

Performance started to be analyzed on synthetic bi-dimensional datasets (Figure 3). All regard binary classification problems, ranging from linearly separable (Problem A) to complex decision environments: with continuous/discontinuous boundaries (problems B and C) and balanced/unbalanced prior probabilities (problems D and E). VPC denotes the proposed method, LDA the linear discriminant analysis, NN stands for neural networks and KNN for k-nearest neighbor classification. For each problem/algorithm, the decision boundaries appear in black. An immediate conclusion is the suitability of VPC to handle all classes of problems tested, both linearly separable (problem A), with complex decision boundaries (problems C-E) and different levels of prior probabilities per class (problem E). In these experiments, the smoothness of the decision boundaries of VPC was lower than for NN and KNN, which was explained by the parameters used (low values of γ and γ^* were used in this example).

Table 1. Variants of each classification algorithm evaluated and corresponding parameters/intervals evaluated in the optimization process.

Algorithm	**Parameters Optimization**
LDA	-
Neural Networks (NN)	Learning Algorithm: Levenberg-Marquardt backpropagation, scaled-conjugate gradient, gradient descend with adaptive learning rate/momentum; Topology: neurons hidden layer; Learning stopping criteria: validation checks, performance, and epochs.
KNN	Number neighbors: [1, d]
SVM	Kernel type: linear, polynomial and sigmoid; kernel degree: [1, 4]; γ kernel functions: [0.5/d, 2*d]
Bagging (BAG)	Number ensemble classifiers: [2, 2*d], Weak learners: classification trees (CART), quadratic/pseudo-linear
Boosting (BOS)	Number ensemble classifiers: [2, 2*d]; Weak learners: quadratic discriminants and decision trees; Type learning algorithm: multi-class AdaBoost, RUSBoost (Seiffert et al., 2008)
Random Subspaces (RSP)	Number ensemble classifiers: [2, 2*d]; Weak learners: quadratic discriminants and decision trees, Gini's diversity index/maximum deviance split criteria.
Random Forests (RFO)	Number weak classifiers: [2, 2*d]; Number of variables selected per split: [1, $\sqrt{d}$].

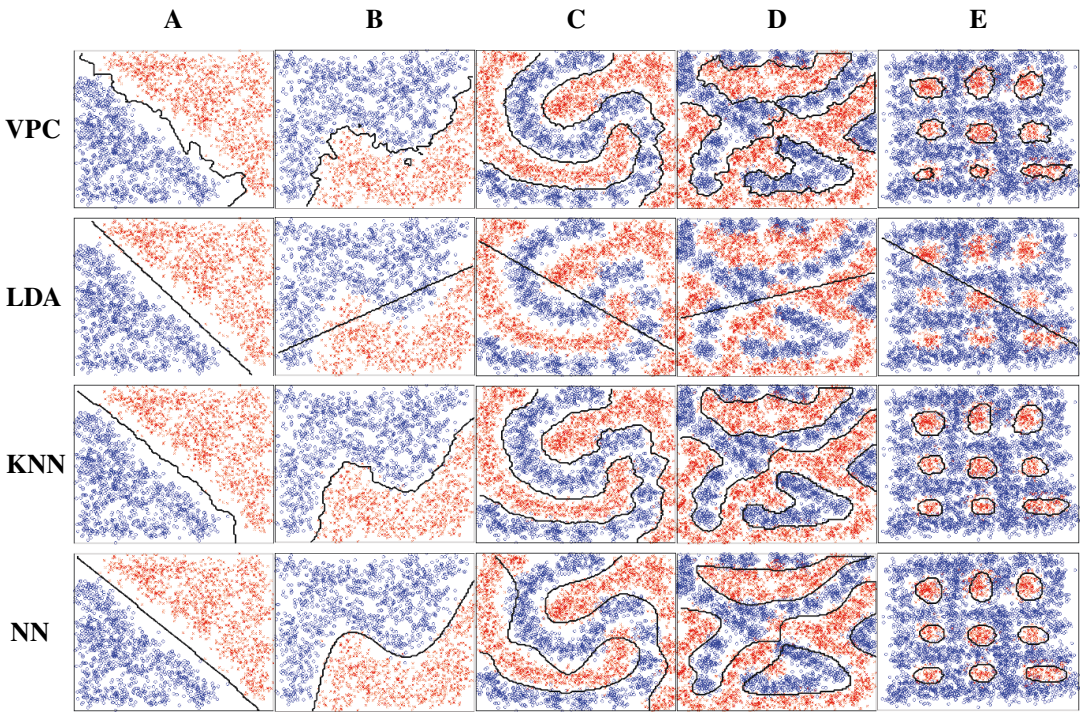


Figure 3. Results attained by the VPC for bi-dimensional synthetic datasets, when compared to linear discriminant analysis (LDA), neural networks (NN) and k-nearest neighbors (KNN) classification methods.

Table 2. Datasets of the *UCI Machine Learning Repository* (Univ. California) used in the performance evaluation of VPC.

ID	Data Set	Instances (Training/Test)	Features	Classes	Prior Probs (%).
BC	**Breast Cancer Wiscosin (Original)**	683 (10 × 615/68)	9	2	65.00, 35.00
HS	**Haberman's Survival**	306 (10 × 276/30)	3	2	74.00, 26.00
IS	**Image Segment**	2310 (10 × 2079/231)	19	7	(balanced)
IR	**Iris**	150 (10 × 135/15)	4	3	(balanced)
IT	**Isolet**	7797 (10 × 7 018/779)	617	26	(balanced)
LR	**Letter Recognition**	20 000 (10 × 18 000/2 000)	16	26	3.94,3.83,3.68,4.03,3.84,3.87,3.86, 3.67,3.77,3.64,3.69,3.81,3.96,3.91, 3.77,4.01,3.91,3.79,3.74,3.98,4.06, 3.82,3.76,3.94,3.93,3.67
MF	**Multifeature Digit**	2 000 (10 × 1 800/200)	649	10	(balanced)
MU	**Musk (Version 2)**	6 598 (10 × 5 939/659)	168	2	84.59,15.41
PB	**Page Blocks**	5 473 (10 × 4 926/547)	10	5	89.77,6.01,0.51,1.61,2.10
SK	**Skin Segmentation**	245 057 (10 × 220 552/24 505)	3	2	20.75,79.25
SP	**Spambase**	4 601 (10 × 4 141/460)	57	2	60.60,39.40
ST	**Statlog (Shuttle)**	58 000 (10 × 52 200/5 800)	9	7	78.60,0.08,0.29,15.35,5.63,0.01, 0.02

4.3 UCI - Machine Learning Repository

Performance was also compared in the *University of California, Irvine: Machine Learning Repository*[1] datasets, which are freely available and widely known in the field of classification. The used sets are summarized in Table 2, and were selected according to four criteria: 1) containing multivariate or univariate features; 2) suitable for classification tasks; 3) exclusively with numeric attributes; and 4) without missing values. We give the number of instances, features, classes and prior probabilities per class. Values inside parenthesis denote the amounts of data selected for learning/testing purposes. For all sets, 10-fold cross validation was adopted and features were rescaled to [0,1] interval according to the *min-max* rule.

1. http://archive.ics.uci.edu/ml/

Table 3. Results obtained for the datasets of the *UCI Machine Learning Repository*. The mean errors are given, together with the standard deviations observed in the 10-fold cross validation procedure. Cells in bold highlight the best algorithm per data set.

Dataset	VPC	LDA	NN	SVM	KNN	BAG	BOS	RSP	RFO
BC	**2.48 ± 1.83**	3.68 ± 2.71	3.38 ± 2.78	2.94 ± 2.19	2.50 ± 1.71	3.68 ± 2.79	2.97 ± 2.06	6.62 ± 2.88	2.91 ± 1.95
HS	20.33 ± 5.54	28.00 ± 6.52	24.00 ± 4.66	25.33 ± 7.57	23.33 ± 4.16	23.67 ± 4.57	**16.70 ± 4.98**	25.67 ± 6.49	23.06 ± 6.06
IS	3.58 ± 1.07	11.26 ± 1.56	5.37 ± 5.29	7.01 ± 1.73	**3.12 ± 1.29**	8.18 ± 1.63	5.17 ± 2.03	21.17 ± 2.13	3.58 ± 1.52
IR	2.00 ± 1.22	4.00 ± 4.66	4.67 ± 6.32	4.00 ± 5.62	3.33 ± 4.71	**1.33 ± 2.81**	2.67 ± 3.44	4.00 ± 4.66	3.69 ± 1.02
IT	5.73 ± 0.72	5.73 ± 0.72	6.78 ± 6.89	**3.07 ± 0.84**	8.97 ± 1.27	5.35 ± 0.83	12.91 ± 2.28	30.90 ± 1.79	11.03 ± 2.02
LR	9.01 ± 0.63	37.49 ± 1.09	26.20 ± 5.17	17.65 ± 0.66	**3.89 ± 0.29**	11.21 ± 0.64	11.30 ± 0.69	49.03 ± 0.72	12.47 ± 1.28
MF	1.75 ± 0.67	1.75 ± 0.68	2.35 ± 2.96	1.55 ± 0.83	1.65 ± 0.58	**1.05 ± 0.69**	7.91 ± 0.97	9.25 ± 2.12	2.28 ± 0.80
MU	2.40 ± 0.38	6.43 ± 0.97	**0.93 ± 0.45**	5.13 ± 0.76	3.11 ± 0.37	3.25 ± 0.68	3.35 ± 0.60	11.90 ± 1.55	2.57 ± 0.94
PB	**3.32 ± 0.70**	7.35 ± 1.48	3.33 ± 0.77	7.11 ± 1.20	4.04 ± 0.97	5.60 ± 1.10	4.52 ± 1.45	7.77 ± 0.73	3.48 ± 0.53
SK	0.06 ± 0.01	6.60 ± 0.14	0.31 ± 0.53	1.02 ± 0.06	**0.04 ± 0.01**	1.64 ± 0.08	1.64 ± 0.08	17.45 ± 0.27	1.70 ± 0.22
SP	6.34 ± 1.31	9.48 ± 2.02	6.17 ± 1.74	9.57 ± 1.82	8.85 ± 1.03	10.80 ± 1.91	6.13 ± 1.02	34.28 ± 3.71	**4.36 ± 0.80**
ST	0.05 ± 0.02	17.48 ± 0.82	0.42 ± 0.02	3.08 ± 0.22	0.06 ± 0.03	5.52 ± 0.29	1.59 ± 0.32	14.66 ± 1.21	**0.03 ± 0.00**

Note that the used sets are heterogenous from different perspectives, ranging from *easy* problems (such as SK and ST), to extremely *hard* (such as the HS), due to low feature-to-instance ratio and classes overlapping. Also, for some problems large amounts of data are available (e.g., SK), while others have a reduced number of instances available (e.g., IR).

Results are given in Table 3 and a first evidence is that VPC only got the *best* performance among all algorithms in two different problems (BC and PB). However, the most important observation is that, for all the remaining cases, VPC was among the best half of the algorithms. Also—as expected—KNN got the best results in problems with low feature-to-instance ratio, that correspond to densely populated feature spaces (e.g., SK).

Not only VPC, but also NN, SVM, KNN, Bagging and Random Forest got the 1^{st} rank in some problem. Among the ensemble algorithms, Bagging, Boosting and Random Forest outperformed all the remaining algorithm in some problem. In opposition, random subspaces got particularly hazardous results in low dimensionality datasets, where the projection into feature subspaces does not keep enough discriminating information.

To perceive the classes of problems where each algorithm got the best results, their relative effectiveness was tested. Differences in performance were validated in terms of statistical significance using Student t-tests (at the 95% level), assuming that errors are normally distributed. Having performance scores of two algorithms (vectors $\vec{v}_1$ and $\vec{v}_2$, length 10), a t-test t_e was carried out, stating as null hypothesis H_0 that "$\vec{v}_1$ *and* $\vec{v}_2$ *are independent random samples from normal distributions with equal means and unknown variances*". The alternative hypothesis was that means are different:

$$t_e = \frac{\text{abs}(\mu_{\vec{v}_1} - \mu_{\vec{v}_2})}{\sqrt{\frac{\sigma^2_{\vec{v}_1} + \sigma^2_{\vec{v}_2}}{10}}}, \tag{14}$$

Table 4. Summary of the *algorithm-to-algorithm* relative performance, for datasets of the *UCI Machine Learning Repository* (represented in each cell in the same order as in Table 2). The symbol "○" denotes that the algorithm in the column is *better* than the algorithm in the row, with statistical significance at 95% level. "●" denotes *worse* performance and "·" corresponds to differences in results without statistical significance. Each cell in the bottom row summarizes the total of "○", "·" and "●" cases for an algorithm.

Alg.	VPC	LDA	NN	SVM	KNN	BAG	BOS	RSP	RFO
VPC	-	··●···●···●●●·●●●	·····●···○···●	··●···○●···●●●·●●	····●○···●···●·	··●···●···●●●●●	····●●●●···●···●	···●···●●●●●●●●	···●●●····●○○
LDA	-		·····○·○○○·○	··○···○○···○·○	···○···●○·○○○·○	·····○···○···○	·○○···●○●○·○○○	···●···●●●●···●●○	·· ○·○○○·○○○○○
NN	·····○···●····○	·····●···●●●···●●	-	·····○···●●···●	·····○···●····○	·····○···●●●●●	·○●●···●●···●···●	···●···●●●●●●●●	····●○···●···●○
SVM	··○···●○···○○○·○	···●···●●····●···●	·····●···○○···○	-	···○···●○···○○○·○	····●○···○···●·	····●○●○○···●··	···●···●●●●···●●●	··○···●○·○○○○○
KNN	·····○●···○····	···●···○●···●●●·●●	·····●···○···●	···●···○●···●●●·●●	-	···●···○●····●···●	····○○●····●···●	···●···●●●●●●●●	·····●····●○○
BAG	··○···○·○○○○○	·····●···●···●·	·····●···○○○○○	····○●···●···○·○	···○···●○····○·○	-	····●···●····○○	···●···●●●●●●●●	···○···●····○·○○
BOO	·····○○○○·○·○	···○···○●○●···●·○	···○·○●○○·○○○	···○···○●○●···○·○	···○·○○○···○·○	···○···○···○····	-	·····○●○○●···●●○	···●·····○····○
RSP	··○···○○○○○○○○	···○···○○○○···○○●	···○·○○○○○○○○	···○···○○○○···○○○	···○···○○○○○○○○	···○···○○○○○○○○	···○···○·○○○○●	-	○·○···○○○○○○○○
RFO	···○○○○···○●●	··●···●●···●●●●●	····○●···○·○○●	···●···○●···●●●●●	·····○····○●●	···●···○····●·●●	·○·····●····●	●···●···●●●●●●●●	-
Total	43, 48, 5	12, 46, 38	35, 52, 9	25, 45, 26	36, 51, 9	22, 51, 23	23, 48, 25	4, 27, 65	37, 46, 13

where abs(.) denotes the absolute value, $\mu_{\vec{v}_1}$ and $\mu_{\vec{v}_2}$ are the means and $\sigma_{\vec{v}_1}$ and $\sigma_{\vec{v}_2}$ the standard deviations. Every time $t_e > 2.10$, H_0 was rejected and assumed that both algorithms actually have different performance.

Table 4 summarizes the *algorithm-to-algorithm* comparison in the datasets evaluated (in the same order in each cell as in the rows of Table 2). Symbol "○" denotes cases where the algorithm in the column got better results (with statistical significance) than the one in the row. Oppositely, symbol "●" denotes worse performance and "·" denotes results without statistical significance. As main conclusion, the best performance of VPC among all is evident: only for three problems VPC got worse results than any other algorithm: *Musk* (worse than neural networks), *Isolet* (worse than support vector machines) and *Letter Recognition* (worse than k-nearest neighbors). The bottom row of the table gives the summary statistics, showing the number of "○", "·" and "●" cases. VPC, NN, SVM, KNN and Random Forests got overall positive balance, meaning that they were *better* more times than *worse*. VPC attained maximal balance ("○" - "●") value (38), followed by KNN (27) and NN (26) algorithms. Interestingly, the performance attained by two of the ensemble strategies (Boosting and Random Subspaces) was poorer than the observed for individual algorithms. When compared to LDA, in no case did VPC get worse performance, and in the IT dataset results were exactly equal, corresponding to a case where the VPC learning process stopped at the tree root (yielding the LDA).

Regarding the ensemble algorithms, Random Forests got the best results, followed by Bagging, in accordance to the results given by Banfield et al. (2007). Boosting got smaller errors than Bagging in half of the problems. The Random Forest algorithm outperformed all the remaining in two problems (*Spambase* and *Statlog*) that share the property of class imbalance. Random Subspaces got especially bad results in low dimensionality problems (e.g., Skin) and Boosting was among the best algorithms for problems with a reduced number of classes (e.g., *Musk*).

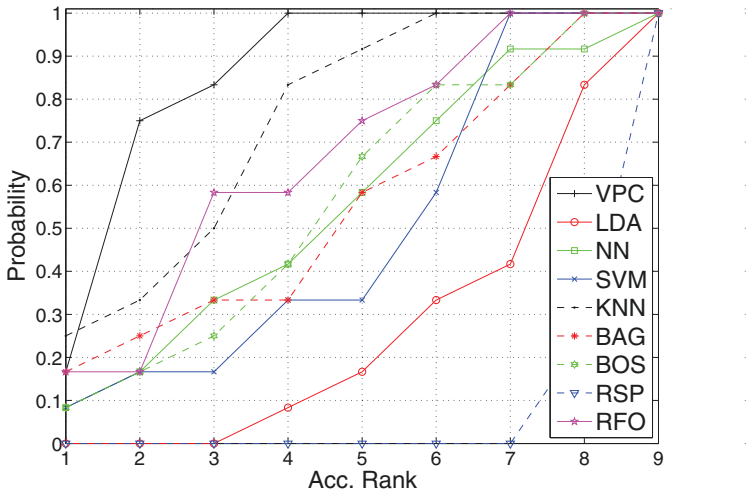


Figure 4. Accumulated probabilities of performance ranks observed for the eight algorithms in the datasets of the *UCI-Machine Learning Repository*

Demsar (2006) suggested that the fairest way to compare algorithms is to use their average ranking on multiple datasets and cross-validation accuracy. Hence, the order-rank of the algorithms in each problem was compared and the results illustrated in Figure 4, showing the accumulated probabilities in terms of ranks, i.e., the probabilities that an algorithm is among the top-k rank (*Acc. Rank* axis). Here, the *best* algorithm appears most close to the upper-left corner, which enables to intuitively visualise the relative effectiveness.

It is evident that VPC was the algorithm with ranks closest to the upper-left corner, followed by KNN and Random Forest. Next, a group of four algorithms (NN, SVM, Bagging and Boosting) got similar results. LDA appears next and the Random Subspaces algorithm got the worst results. Another particularly interesting property of VPC is that—for all problems—got performance among the top-half algorithms. This had only happened in about 85% for KNN, and around 60% of the problems for the Random Forest, which we consider a substantial difference. All the remaining algorithms were among the top-half in less than 50% of the problems.

Further attention should be given to the levels of correlation between the responses given by the best algorithms, in order to anticipate the advantages of using *meta-ensembles*, i.e., the improvements in performance that might result of fusing at the score level VPC, KNN, NN, SVM and Bagging classifiers.

4.4 Major Performance Covariates

According to the results given above, it is particularly important to perceive the factors that most evidently affect the performance of VPC with respect to the competitors. For such, we decided to compare the results of VPC to KNN and Random Forest classification strategies, with respect to

Figure 5. Examples of images used in the GTSR set released by the *Institute für Neuroinformatik*, in the scope of a competition held at the 2011 International Joint Conference on Neural Networks.

two different factors: 1) balance of classes prior probabilities; and 2) feature spaces dimension. The choice of the comparison terms used was motivated by the overall ranking of algorithms in the experiments above, summarized in Figure 4.

All results reported in this section were based on data from the German traffic sign recognition benchmark (Stallkamp et al., 2012) (Figure 5) was used. This data set is a multi-class image classification challenge held at the International Joint Conference on Neural Networks (IJCNN) 2011. There are 43 classes in the data set and more than 50 000 images [2], divided into disjoint training and test sets. Using the available meta-data that defines a region-of-interest for each sample, a set of 1,300 features per sample was extracted, scanning all 3×3 patches of the scale-normalized images (40×40 pixels) with the highly popular Local Binary Patterns descriptor (Ojala, Pietikainen, and Harwood, 1966). According to a bootstrapping-like strategy, random samples with 80% of the available learning and test data were drew, and the effectiveness of the measured, yielding the results given in the sub-sections below, where we plot the average performance plus the first and third quartile of the results, as a confidence interval.

4.4.1 Balance of Classes Prior Probabilities

To perceive the effect that the balance of classes priors has in VPC performance, we selected the most frequent classes in the GTSR data set (classes 1, 2, 12, 13 and 38). Next, we drew multiple samples of the learning and test sets, each one comprising two classes and varying the proportion of elements per class. Results are given in the left plot of Figure 6, for VPC (solid line), KNN (dashed line) and Random Forests (dashed-dotted

2. http://benchmark.ini.rub.de/

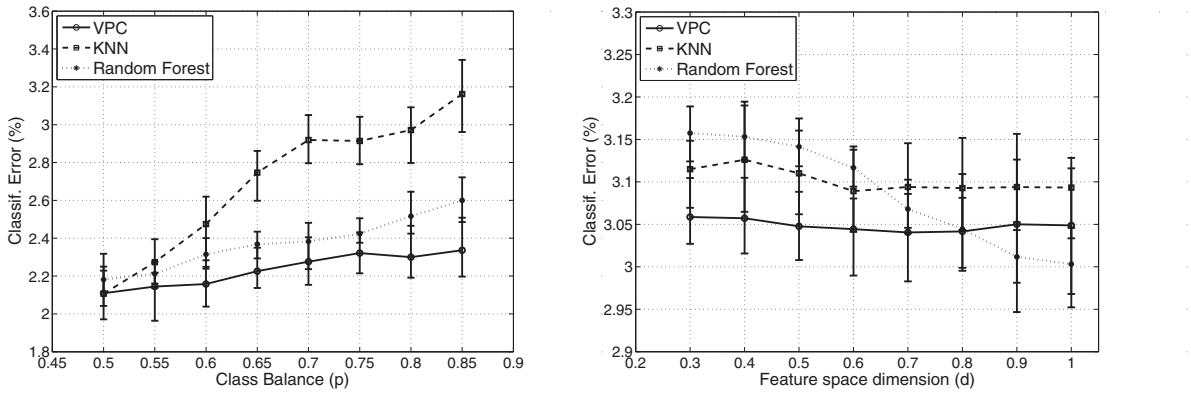


Figure 6. At left: Variations in performance with respect to the balance of classes prior probabilities. At right: Variations in performance with respect to the dimensionality of the feature spaces.

lines) algorithms, with the corresponding first and third quartile performance values observed (p is the proportion of elements the less frequent class) It is obvious that the gap between VPC and the competitors tends to increase directly in proportion to the levels of class unbalance, which accords our previous observations. Without surprise, KNN showed a larger deterioration in performance than the remaining algorithms with respect to this factor, whereas Random Forest showed a relatively small decrease in classification effectiveness.

4.5 Feature Spaces Dimension

Further, we compared the classification effectiveness of VPC, KNN and Random Forests with respect to the dimension of the feature space, which also correlates to the density of the learning feature space (as the number of used instances was kept constant). In this experiment, all classes of the GTSR set were considered, using feature subsets composed by 10 to 100% of all the available features, chosen randomly. The results are given in the right plot of Figure 6 (d represents the proportion of features considered), and appear to confirm that the performance of VPC with respect to competitors is maximized for problems of reduced and moderate dimensionality, i.e., corresponding to more densely populated feature spaces. For large dimensionality problems, the pivots chosen at each node of the classification tree were observed to decrease the representativity of the corresponding elements on that node. Noting that the confidence intervals associated with each algorithm largely overlap, it is still possible to perceive an inverse tendency between the performance of VPC/KNN and the Random Forest algorithm, that is the unique where the rule *"the more features the better"* appears to apply. Even due to different reasons, this does not holds for VPC and KNN,

which is known to suffer from the *irrelevant features* issue that happens often in high dimensionality problems.

5. Conclusions

This paper proposes a classification strategy that accords the idea of Boosting and is based on a *Vantage-Point* tree that recursively divides the feature space into compact subspaces (leaves) that are separated by weak classifiers (linear discriminants). By preserving the neighborhood of subspaces, the binary data structure is traversed in a computationally efficient way and only a reduced number of leaves vote for the response of the ensemble, which yields the low computational cost of classification.

The resulting ensemble classifies in temporal cost of approximately $O(lg(n))$. Also, in terms of accuracy, it attains results similar to the state-of-the-art in most of the problems tested. The computational cost/accuracy balance is regarded in a particularly positive way due to the broad range of problems considered (binary/n-ary classification, discrete/continuous features, balanced /unbalanced priori probabilities, with densely/sparsely populated datasets).

In terms of the results observed, we highlight the following conclusions:

- Even though the proposed method (VPC) outperformed all the other algorithms in a relatively short proportion of the problems (2/12), the interesting property is that, for all problems VPC was among the best algorithms, which clearly did not happened for any of the remaining comparison terms.
- With respect to its competitors, the best results of VPC were observed for problems with unbalanced priori probabilities per class. This was explained by the fact that VPC classifies (locally) in subspaces, so that the prior probabilities in the complete feature space do not bias each local classifier.
- Also in terms of relative effectiveness, VPC is particularly suitable for problems of moderate dimensionality, where the vantage-point retrieval scheme works better. In very large dimensionally problems, as the $\ell 2$-norm was used to obtain the distance between feature points, pivots decrease the representativity of all the elements in the corresponding node.
- When compared to KNN, the major advantage of VPC is its smaller sensitivity to irrelevant features, which can be naturally disregarded by the linear discriminants at the tree leaves.
- It should be noted that in all the problems considered, the number of training instances n was always much higher than the dimensionally

d of the feature space. Hence, as further work, we plan to analyze the effectiveness of VPC in feature spaces with such high dimensionality and relatively reduced amount of learning data ($n \approx d$, or even $n < d$).

References

ALPAYDIN, E. (1999), "Combined 5 × 2 cv F-Test for Comparing Supervised Classification Learning Algorithms", *Neural Computation, 11(8)*, 1885–1892.

BANFIELD, R., HALL, L., BOWYER, K., and KEGELMEYER, W. (2007), "A Comparison of Decision Tree Ensemble Creation Techniques", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(1)*, 173–180.

BAUER, E., and KOHAVI, R. (1999), "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants", *Machine Learning, 36(1-2)*, 105–139.

BOCK, K., COUSSEMENT, K., and POEL, D. (2010), "Ensemble Classification Based on Generalized Additive Models", *Computational Statistics and Data Analysis, 54*, 1535–1546.

BREIMAN, L. (1996), "Bagging Predictors", *Machine Learning, 24(2)*, 123–140.

BREIMAN, L. (2001), "Random Forests", *Machine Learning, 45(1)*, 5–32.

BRYLL, R., GUTIERREZ-OSUNA, R., and QUEK, F. (2003), "Attribute Bagging: Improving Accuracy of Classifier Ensembles by Using Random Feature Subsets", *Pattern Recognition, 36(6)*, 291–1302,

CAI, D., HE, X., and HAN, J. (2008), "Training Linear Discriminant Analysis in Linear Time", in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pp. 209–217.

CANUTO, A., ABREU, M., OLIVEIRA, L., XAVIER JR., J., and SANTOS, A. (2007), "Investigating the Influence of the Choice of the Ensemble Members in Accuracy and Diversity of Selection-Based and Fusion-Based Methods for Ensembles", *Pattern Recognition Letters, 28(4)*, 472–486.

CECI, M., APPICE, A., and MALERBA, D. (2003), "Comparing Simplification Methods for Model Trees with Regression and Splitting Nodes", in *Proceedings of the Fourteenth International Symposium on Methodologies for Intelligent Systems,* Lecture Notes in Artificial Intelligence Vol. 2871, pp. 49–56.

CORTES, C., and VAPNIK, V. (1995), "Support Vector Networks", *Machine Learning, 20*, 1–25.

COVER, T., and HART, P. (1967), "Nearest Neighbor Pattern Classification", *IEEE Transactions on Information Theory, 13(1)*, 21–27.

DEMSAR, J. (2006), "Statistical Comparisons of Classifiers over Multiple Data Sets", *Journal of Machine Learning Research, 7*, 1–30.

DIETTERICH, T. (2000), "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization", *Machine Learning, 49(2)*, 139–157.

DOMINGOS, P. (1996), "Unifying Instance-Based and Rule-Based Induction", *Machine Learning, 24*, 141–168.

DUDA, R., HART, P., and STORK, D. (2000), *Pattern Classification* (2nd ed.), Wiley Interscience, ISBN 0-471-05669-3.

FRANK, E., HALL, M., and PFAHRINGER, B. (2003), "Locally Weighted Naive Bayes", in *Proceedings of the 19th conference on Uncertainty in Artificial Intelligence*, San Mateo, pp. 249–256.

FREUND, Y., and SCHAPIRE, R. (1995), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", in *Proceedings of the 2nd European Conference on Computational Learning Theory*, pp. 23–37.

HO, T.K. (1995), "Random Decision Forests", in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282.

HO, T.K. (1998), "The Random Subspace Method for Constructing Decision Forests" *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8)*, 832–844.

HOTHORN, T., and LAUSEN, B. (2005), "Building Classifiers by Bagging Trees", *Computational Statistics and Data Analysis, 49*, 1068–1078.

JIRINA, M., and JIRINA JR., M. (2013), "Utilization of Singularity Exponent in Nearest Neighbor Based Classifier", *Journal of Classification, 30(1)*, 3–29.

JOHNSON, R., and WICHERN, D. (1988), *Applied Multivariate Statistic Analysis* (2nd. ed.), Englewood Cliffs NJ: Prentice Hall Inc.

KLEINBERG, E.M. (1990), "Stochastic Discrimination", *Annals of Mathematics and Artificial Intelligence, 1*, 207–239.

KUMAR, A. (2008), "Combining Pattern Classifiers: Methods and Algorithms", *IEEE Transactions on Industrial Electronics, 55(1)*, 348–363.

KUNCHEVA, L. (2004), *Combining Pattern Classifiers: Methods and Algorithms*, Hoboken NJ: John Wiley & Sons.

KUNCHEVA, L., and RODRÍGUEZ, J. (2007), "Classifier Ensembles with a Random Linear Oracle", *IEEE Transactions on Knowledge and Data Engineering, 19(4)*, 500–508.

LU, J., and TAN, Y-P. (2011), Nearest Feature Space Analysis for Classification", *IEEE Signal Processing Letters, 18(1)*, 55–58.

MOLLER, M. (1993), "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning", *Neural Networks, 6*, 525–533.

OJALA, T., PIETIKAINEN, M., and HARWOOD, D. (1996), "A Comparative Study of Texture Measures with Classification Based on Feature Distributions", *Pattern Recognition, 29*, 51–59.

GARCÍA-PEDRAJAS, N. (2009), "Constructing Ensembles of Classifiers by Means of Weighted Instance Selection", *IEEE Transactions on Neural Networks, 20(2)*, 258–277.

SCHAPIRE, R. (1990), "The Strength of Weak Learnability", *Machine Learning, 5(2)*, 197–227.

SEIFFERT, C., KHOSHGOFTAAR, T., HULSE, J., and NAPOLITANO, A. (2008), "RUSBoost: Improving Classification Performance When Training Data Is Skewed", in *Proceedings of the 19th International Conference on Pattern Recognition*, pp. 1-4.

STALLKAMP, J., SCHLIPSING, M., SALMEN, J., and IGEL, C. (2012), "Man vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition", *Neural Networks, 32*, 323–332.

TING, K.M., WELLS, J.R., TAN, S.C., TENG, S.W., and WEBB, G.I. (2011), "Feature-Subspace Aggregating: Ensembles for Stable and Unstable Learners", *Machine Learning, 82*, 375–397.

VIOLA, P.A., and JONES, M.J. (2004), "Robust Real-Time Face Detection", *International Journal of Computer Vision, 57(2)*, 137–154.

YAN, R., and TEŠIĆ, J. (2007), "Model-Shared Subspace Boosting for Multi-Label Classification", in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 834–843.

YIANILOS, P. (1993), "Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces", in *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics,* pp. 311–321.

YU, G., ZHANG, G., YU, Z., DOMENICONI, C., YOUC, J., and HANA, G. (2012), "Semi-Supervised Ensemble Classification in Subspaces", *Applied Soft Computing, 12*, 1511–1522.

ZAMAN, M., and HIROSE, H. (2013), "DF-SVM: A Decision Forest Constructed on Artificially Enlarged Feature Space by Support Vector Machine", *Artificial Intelligence Review, 40*, 467–494.

Soft Biometrics: Globally Coherent Solutions for Hair Segmentation and Style Recognition Based on Hierarchical MRFs

Hugo Proença, *Senior Member, IEEE*, and João C. Neves, *Member, IEEE*

***Abstract*— Markov Random Fields (MRFs) are a popular tool in many computer vision problems and faithfully model a broad range of local dependencies. However, rooted in the Hammersley–Clifford theorem, they face serious difficulties in enforcing the global coherence of the solutions without using too high order cliques that reduce the computational effectiveness of the inference phase. Having this problem in mind, we describe a multi-layered (hierarchical) architecture for MRFs that is based exclusively in pairwise connections and typically produces globally coherent solutions, with 1) one layer working at the local (*pixel*) level, modeling the interactions between adjacent image patches; and 2) a complementary layer working at the *object* (hypothesis) level pushing toward globally consistent solutions. During optimization, both layers interact into an equilibrium state that not only segments the data, but also classifies it. The proposed MRF architecture is particularly suitable for problems that deal with biological data (e.g., biometrics), where the reasonability of the solutions can be objectively measured. As test case, we considered the problem of hair / facial hair segmentation and labeling, which are soft biometric labels useful for human recognition *in-the-wild*. We observed performance levels close to the state-of-the-art at a much lower computational cost, both in the segmentation and classification (labeling) tasks.**

***Index Terms*— Soft biometrics, visual surveillance, homeland security.**

I. Introduction

IN VISUAL surveillance / biometrics research, the development of systems to work in unconstrained data acquisition protocols and uncontrolled lighting environments is a major ambition. The images resulting of such conditions are degraded in multiple ways, such as blurred, shadowed, of poor resolution, with subjects off-angle and partially occluded (Fig. 1). In these cases, soft biometrics can be seen as an identity retrieval tool that attenuates the decrease in performance of the classical biometric traits (e.g., the face or the iris).

The descriptions of the facial hair and hair styles are among the most effective soft biometric traits reported in the literature [24]. In this scope, the pioneer analysis methods were designed to work exclusively in good quality images of frontal subjects. Regardless recents attempts to increase the robustness (e.g., [29]), the ambition of working effectively in images acquired in typical visual surveillance conditions remains to be achieved.

Fig. 1. Examples of images captured by an outdoor visual surveillance system, with *unconstrained* acquisition conditions and protocols. Images have typically poor resolution and are often blurred, with subjects partially occluded and under varying poses.

Markov Random Fields (MRFs) are a classical tool for many computer vision problems, from image segmentation [13], image registration [8] to object recognition [4]. Among other strengths, they provide non-causal models with isotropic behaviour and faithfully model a broad range of local dependencies. On the other way, they hardly guarantee globally coherent solutions without using too high order cliques that compromise the computational effectiveness of the inference phase. Having this problem in mind, in this paper we propose a multi-layered (hierarchical) MRF that does not use high order cliques but still typically reaches globally coherent solutions. As test case, we consider the hair / facial hair style analysis, and describe an inference process composed of two phases:

1) three supervised non-linear classifiers run at the pixel level and provide the posterior probabilities for each image position and class of interest: *hair*, *skin* and *background*. Each classifier detects one component based on texture and shape image statistics;
2) the posteriors based on data *appearance* are combined with geometric constraints and a set of model hypotheses to feed the MRF, composed of a *segmentation* and a *classification* layer. One layer discriminates locally the classes of interest, while the other infers the soft biometric labels that describe the query's facial hair and hair styles.

The key idea is to combine the strengths of MRFs with groups of synthetic hypotheses that are projected onto the input plane and guarantee the global consistency (biological coherence) of the solution. The proposed model

Manuscript received April 21, 2016; revised October 31, 2016; accepted February 27, 2017. Date of publication March 9, 2017; date of current version April 13, 2017. This work was supported by the FCT Project under Grant UID/EEA/50008/2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Matti Pietikainen.

The authors are with the Department of Computer Science, IT: Instituto de Telecomunicações, University of Beira Interior, 6201-001 Covilhã, Portugal (e-mail: hugomcp@di.ubi.pt; jcneves@di.ubi.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2017.2680246

TABLE I

SUMMARY OF THE MOST RELEVANT METHODS TO PERFORM AUTOMATED DETECTION, SEGMENTATION AND CLASSIFICATION OF FACIAL HAIR AND HAIR STYLES. Y, P, R , A AND C STAND FOR THE DATA VARIATION FACTORS EACH METHOD CLAIMS TO HANDLE: DEVIATIONS IN YAW, PITCH AND ROLL ANGLES, UNALIGNED DATA AND NON-EXISTENCE OF HAIR COLOR CONSTRAINTS

Method	Year	Type	Working Mode		Data Variability					Color Sp.	Summary
			Segm.	Class.	Y	P	R	A	C		
Yacoob and Davis [30]	2006	H	✓	✓	✗	✗	✗	✗	✓	RGB	Gabor kernels (hair texture), dominant color, anthropometric statistics
Lee *et* al. [17]	2008	H	✓	✗	✗	✗	✗	✗	✓	RGB, LAB	Graphical model
Lipowezky *et* al. [18]	2008	H	✓	✗	✗	✗	✗	✗	✓	LAB, HSV	Seed detection: EDISON algorithm, edge analysis, Haar filtering, region growing: k-means clustering
Rouset and Coulon [22]	2008	H	✓	✗	✗	✗	✗	✗	✓	HSV, YCbCr	Frequency analysis (Gaussian filtering), color analysis (local deviations)
Zhang *et* al. [33]	2008	H	✓	✗	✓	✓	✓	✗	✓	HSV, XYZ	Gaussian mixture model-based density estimation
Zhang *et* al. [34]	2009	H	✓	✗	✓	✓	✓	✓	✓	HSV, XYZ	Gaussian mixture density estimation, analysis of skin, hair and head spatial constraints
Julian *et* al. [11]	2010	H	✓	✗	✗	✗	✗	✗	✓	HSV, YCbCr	Histogram analysis, active shape models
Wang and Ai [27]	2013	H	✓	✓	✗	✗	✓	✗	✓	LUV	Informative patches (RankBoost, SVM), graphical model, clustering
Ugurlu [25]	2012	H, FH	✓	✗	✓	✗	✗	✗	✓	HSV	Non-linear supervised local classification
Dass *et* al. [5]	2013	H	✓	✓	✓	✗	✗	✓	✓	HSV, grayscale	Eyes detection (Adaboost), alignment (similarity transform), Otsu thresholding, clustering
Kae *et* al. [12]	2013	H, FH	✓	✓	✓	✓	✓	✗	✓	RGB	Conditional random fields (local consistency), restricted Boltzmann machines (global consistency)
Wang *et* al. [28]	2013	H	✓	✗	✓	✓	✓	✓	✓	RGB	Coarse local likelihood, manifold inference, refined segmentation
Krupka *et* al. [16]	2014	H	✓	✗	✗	✗	✗	✗	✓	RGB	Background modelling (Gaussian mixture), convex hull analysis
Shen and Ai [23]	2014	H	✓	✓	✗	✗	✗	✗	✓	YCbCr	Landmarks detection (ASM), graph-cuts, histogram analysis
Wang *et* al. [29]	2014	H	✓	✓	✓	✓	✓	✓	✗	Grayscale	Thresholding, histogram analysis, line intersection)
Proposed Method	2016	H, FH	✓	✓	✓	✓	✓	✓	✓	RGB, HSV, YCbCr	Non-linear pixel classification, 3D Model projection, hierarchical graphical model, manifold learning

inherits some insights from previous works that used shape priors to constraint the models (e.g., [2]) and multiple layered MRFs (e.g., [26], [20]).

The remainder of this paper is organized as follows: Section II analyzes the related work. Section III details the learning and inference phases of the proposed method. Section V describes our experiments and the conclusions are given in Section VI.

II. RELATED WORK

Table I overviews the literature for facial hair / hair style analysis. Algorithms are classified according to their scope (Hair (H) / Facial Hair (FH), Segmentation / Classification), along with a description of the techniques / color spaces used. The data variability factors considered are enumerated, with Y, P and R denoting deviations in *yaw*, *pitch* and *roll* angles, and A and C referring the abilities to work with unaligned data and unconstrained hair colors. Below, methods are grouped into three families: predominantly generative, discriminative and hybrid.

Lee *et al.* [17] propose a generative model that infers a set of hypotheses for the face, hair, and background regions. In classification, the most reliable pixels are the information source for mixture models that parameterise each component and define the MRF unary costs. Still in the generative family, Shen *et al.* [23] propose a face detector to define the ROI and consider color information (YCbCr space) to feed a MRF used for segmentation. As post-processing, nearest neighbour analysis enforces the homogeneity between adjacent regions. Wang *et al.* [28] formulate the segmentation problem as finding pairs of isomorphic manifolds, using a set of learning images with the corresponding ground-truth, designated as *optimal maps*. Here, queries are represented as combinations of optimal maps. Zhang *et al.* [33] and [34] infer a set of probability density functions of four typical hair colors (XYZ and HSV spaces), learned by the expectation-maximization algorithm. Assuming the statistical independence between color channels, they obtain the likelihood in each color space and use a Bayesian framework to segment hair. Finally, a simple approach is due to Dass *et al.* [5], that segment the hair regions by thresholding and use agglomerative clustering to parameterise five groups of hairstyles, based on the proportion of hair pixels in image patches.

Regarding methods that are predominantly discriminative, Kae *et al.* [12] detect the most homogenous image patches (*super-pixels*), which provide the appearance information to a CRF. To guarantee the global coherence of the hypotheses, a restricted Boltzmann machine encodes the global shape priors and enforces shape constraints. Wang and Ai [27] learn a discriminator between the hair / non-hair regions. In classification, seven hairstyles are considered, with the RankBoost algorithm selecting the most informative patches and defining hairstyle similarity directly on the hair shapes. Under the same paradigm, Rousset and Coulon [22] fuse color (YCbCr space) to frequency information, in order to locally discriminate between hair / non-hair pixels.

Hybrid approaches are typically based in template matching, with the pioneer method due to Yacoob and Davis [30]. These authors use face and eye detectors to define the ROIs. Based on spatial and color information, a set of seeds is inferred and region growing is used based on local homogeneity. Finally, morphologic operators enforce connected components. Julian *et al.* [11] learn a set of shape templates of the upper part of the head, based on the boundary control points. Using principal components analysis, they propose the concept of *eigen shape*, keeping the top variability vectors that represent the 3D head orientation and the face morphology. Hair regions are classified at the pixel level according to a texture-analysis strategy, generating seeds for subsequent finer parameterisations (active contours). Ugurlu [25] use a head pose detector based both in shape and texture, being the latter described statistically in the HSV color space. Wang *et al.* [29] use a head detector that defines a ROI, based in histogram analysis and nearest neighbour rules. The hair length is inferred by line scanning on the segmented hair region. A relevant gap of this work is the fact of being only suitable for handling dark hair subjects. Lipowezky *et al.* [18] start by detecting head landmarks (eyes and mouth) to find the most homogenous image patches. Color information (LAB and YCbCr spaces) is fused to the Canny magnitude and to four texture descriptors (wavelets-based), feeding a region-growing algorithm. Similarly, Krupka *et al.* [16] use a head detector that defines the ROI where the skin is detected and segmented. The differences between the head foreground and the skin pixels provide the estimate of the hair positions. Skin seeds are detected by thresholding, further expanded upon homogeneity.

III. Proposed Method

For comprehensibility, the following notation is adopted: matrices are represented by capitalized bold font and vectors appear in bold. The subscripts denote indexes. All vectors are column-wise. The ring symbol (e.g., $\mathring{\mathbf{x}}$) denotes image positions, while 3D positions appear in regular font (e.g., $\mathbf{x}$).

A. Synthesis of 3D Models

We consider three types of 3D models: 1) head; 2) hair; and 3) facial hair. The head models are generated as described in [19]. Using the Young's [31] head anthropometric survey to obtain a group of probability density functions of human head lengths and a basis 3D mesh, we deform the mesh according to randomly drew target distances between pairs of vertices (l_{ij}). Let $\mathbf{x}_i$ be one 3D vertex and $\mathbf{n_i}$ the normal to the surface at that point. Let $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, $\mathbf{n}_{ij} = \mathbf{n}_i - \mathbf{n}_j$ ($\mathbf{x}_\cdot, \mathbf{n}_\cdot \in \mathbb{R}^3$) and let l_{ij} be the target length (Euclidean distance) between $\mathbf{x}_i$ and $\mathbf{x}_j$. This yields a system of linear equations with inequality constraints, enabling to find (using [3]) the magnitude of the displacement α_{ij} on both vertices with respect to their normals ($\mathbf{x}^{\text{new}} = \mathbf{x}^{\text{old}} + \alpha\mathbf{n}$), such that their distance is l_{ij} and $||\boldsymbol{\alpha}||_\infty \leq \kappa_0$ (to avoid anatomically bizarre solutions). The top row of Fig. 2 illustrates our population of head shapes.

Let $\mathbf{s}_s = [\mathbf{x}_1^T, \ldots, \mathbf{x}_{t_v}^T]^T$ be a vector representing one head shape, given as a triangulated mesh of t_v vertices. Considering a set of head shapes $\mathbf{S}_s = \{\mathbf{s}_{s,1}, \ldots, \mathbf{s}_{s,t_m}\}$, there is evidently strong correlation between the $\mathbf{x}_i$ elements in those meshes, which is attenuated if they are represented in the principal components (PC) space:

$$\mathbf{s}_s^* = (\mathbf{s}_s - \bar{\mathbf{s}_s})\mathbf{T}_{pc}, \tag{1}$$

where $\bar{\mathbf{s}_s}$ is the 3 t_v-dimensional mean of the elements in $\mathbf{S}_s$ and $\mathbf{T}_{pc}$ is the PC transformation matrix. This way, each mesh is represented in a feature space of a much lower dimension than the 3 t_v, which accounts for the computational effectiveness of the whole method. In our case, the head models have $t_v = 957$, but 50 PC coefficients represent over 99.9% of the variability.

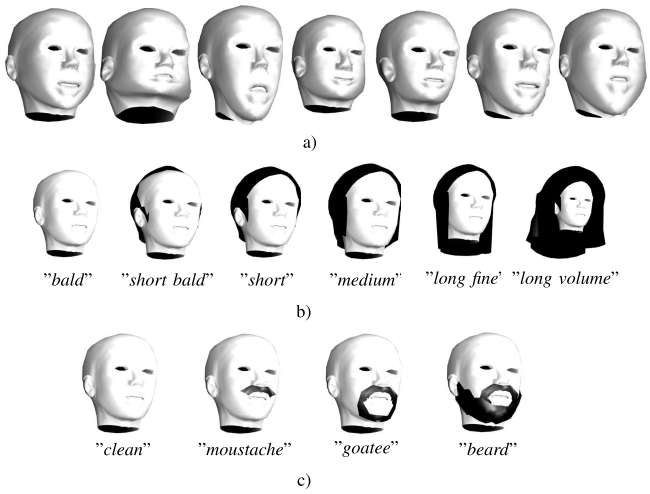


Fig. 2. Illustration of the 3D head shapes, hair and facial hair models that are used as the hypotheses considered in this paper. (a) Head models ($\mathbf{S}_s$). (b) Hair models ($\mathbf{S}_h$). (c) Facial hair models ($\mathbf{S}_f$).

Regarding the hair / facial hair models, we use the concept of *hair mesh* from Yuksel *et al.* [32] and consider hair / facial hair classes as particular cases of polygonal mesh modelling. For simplicity, we keep a short number of hypotheses for each class: $\mathbf{S}_h$={"*bald*", "*short bald*", "*short*", "*medium*", "*long fine*", "*long volume*"} for the hair and $\mathbf{S}_f$={"*clean*", "*moustache*", "*goatee*", "*beard*"} for the facial hair. As previously, all models are generated by deforming iteratively a basis 3D mesh (examples are shown at the bottom rows of Fig. 2).

B. Pose Hypotheses

We also consider a set of pose hypotheses. Let $\mathbf{p} = \{\mathbf{R}, \mathbf{t}\}$ be a camera pose configuration, with $\mathbf{R}$ being the rotation matrix and $\mathbf{t}$ the translation vector, i.e., $\mathbf{p}$ is a 6D vector accounting for three components of rotation (yaw, pitch and roll) and three of translation along the orthogonal axes t_x, t_y and t_z. $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_{t_p}\}$ is the set of t_p pose hypotheses uniformly distributed over all the six degrees of freedom.

C. Joint Head Shape / Pose Hypotheses Indexing

Given a set of t_p pose and t_s head shape hypotheses, it is required to find the best joint pose / head shape configuration, which will most likely match the query. To avoid exploring by brute-force all $t_p t_s$ possibilities, a forest of

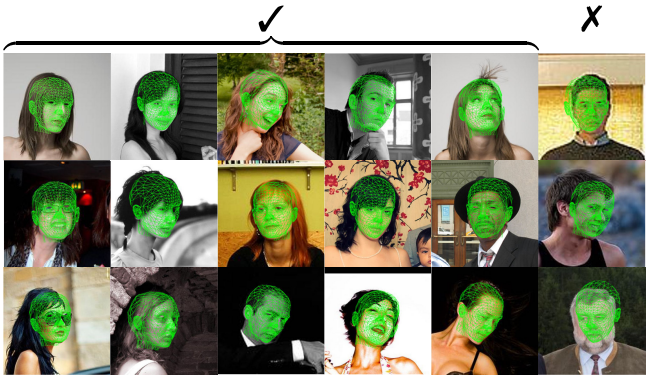


Fig. 3. Successful / failure (rightmost column) estimates of the head shape / pose. Most failed cases are due to ambiguities in various head shape / pose configurations that provide too many overlapped landmark projections.

binary trees is created at learning time, one tree per type of landmark. In these indexing structures, hypotheses are grouped (k-means) in branches according to the neighbourhood of the projected landmark. The *world-to-image* function projects the $\mathbf{x}$ vertices of a head shape hypothesis $\mathbf{s_s}$ according to a pose configuration $\mathbf{p}$:

$$f_{w\to i}(\mathbf{x},\mathbf{p}) = \mathring{\mathbf{x}} = \frac{1}{v}\mathbf{A}[\mathbf{R}|\mathbf{t}]\begin{bmatrix}\mathbf{x}\\1\end{bmatrix}, \tag{2}$$

where v is the scalar projective parameter, $\mathbf{A}$ is the internal camera matrix, and $\mathbf{R}$ and $\mathbf{t}$ are the pose parameters. The retrieval time of the forest is approximately logarithmic with respect to the number of hypotheses, which enables to generate a large set of hypotheses without compromising the time cost of retrieval. Additional details about this data structure are given in [19].

Let $\mathring{\mathbf{q}} = \{\mathring{\mathbf{q}}_1, \ldots, \mathring{\mathbf{q}}_{t_q}\}$ be a set of 2D head landmarks in a query image. We assume that the *type* of each landmark $\tau(\mathring{\mathbf{q}}_i)$ is known, i.e., the anatomic part corresponding to each $\mathring{\mathbf{q}}_i$ is given as input. This is a readily satisfied assumption, using the state-of-the-art techniques for head / face landmark detection (e.g.,[6], [10], or [21]). The position of every query landmark enters in the corresponding binary tree to retrieve the indices of the complying hypotheses. By accumulating the complying indices over all trees, the hypotheses are ranked in descending order according to the likeliness they match the query. Refer to [19] for full details about the way the most likely head shape $\hat{\mathbf{s}}$ and pose $\hat{\mathbf{p}}$ hypotheses are inferred. Fig. 3 gives examples of the head shape / pose estimation inference, using images of the AFLW [15] set. The five leftmost columns contain successful cases, whereas the rightmost column illustrates failure cases, mostly due to ambiguities in various head shape / pose configurations that provide too many overlapped landmark projections.

IV. Soft Labels Inference

After inferring the query head shape $\hat{\mathbf{s}}$ and pose $\hat{\mathbf{p}}$ hypotheses, all hair / facial hair hypotheses are projected according to $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$, to perceive how much they agree with the data appearance terms, which implicitly constitute part of the MRF costs. Fig. 4 gives a cohesive perspective of the two-layered MRF we propose. One layer works at the pixel level

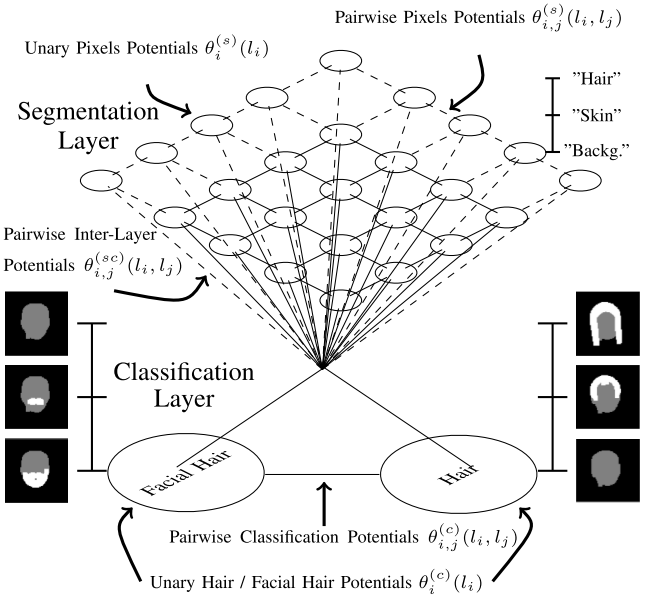


Fig. 4. Structure of the MRF that fuses the data appearance information (upper layer) to global constraints (bottom layer). During optimization, the the network should converge into a balance point where the predominant labels at the segmentation level are biologically plausible and accord globally coherent facial hair / hair hypotheses (at the classification level).

(segmentation layer), with a bijection between image pixels and nodes, each one with three potential labels: *hair*, *skin* and *background*. The other layer (classification) has two nodes that represent the facial hair / hair hypotheses. During model optimization, the interaction between both layers privilege pixel labels that accord a parameterization of the classification nodes and *vice-versa*, forcing the network to converge into an equilibrium state where the configurations at one layer implicitly segment data and the parameterizations in the other layer enforce biologic coherent solutions and describe the facial hair / hair styles.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph representing a MRF of t_v vertices $\mathcal{V}$, linked by t_e edges $\mathcal{E}$. Let t_s be the number of vertices in the *segmentation* layer and t_c the number of vertices in the *classification* layer, such that $t_v = t_s + t_c$. The MRF is a representation of a discrete latent random variable $\boldsymbol{L} = \{L_i\}, \forall i \in \mathcal{V}$, where each element L_i takes one value l_i from a set of labels. Let $\boldsymbol{l} = \{l_1, \ldots, l_{t_s}, l_{t_s+1}, \ldots, l_{t_s+t_c}\}$ represent one configuration of the MRF. In our model, the classification nodes are connected to each other and to all pixel nodes, while the pixel nodes are connected to their horizontal / vertical neighbours. Note that the proposed model does not use high-order cliques. Even though there is a point in Fig. 4 that joins multiple edges, it actually represents overlapped pairwise connections between one classification and one segmentation node.

The energy of a configuration $\boldsymbol{l}$ of the MRF is the sum of the unary $\theta_i(l_i)$ and pairwise $\theta_{i,j}(l_i, l_j)$ potentials:

$$E(\boldsymbol{l}) = \sum_{i\in\mathcal{V}} \theta_i(l_i) + \sum_{(i,j)\in\mathcal{E}} \theta_{i,j}(l_i, l_j). \tag{3}$$

According to this formulation, segmenting / classifying an image is done by inferring the random variables that minimize

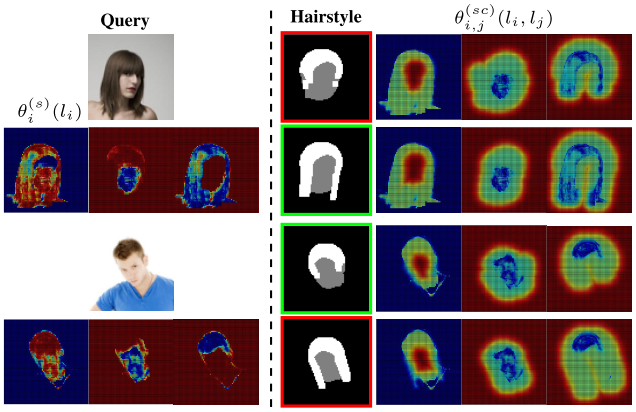


Fig. 5. Rationale for the inter-layer pairwise costs $\theta_{i,j}^{(sc)}(l_i, l_j)$. For two queries, the unary costs (segmentation layer) are shown at the left side, At the right side, having one plausible (green frame) and one non-acceptable (red frame) hair model, the inter-layer pairwise costs encode the reasonability of fitting the data appearance term to the corresponding models (warm colors denote high costs). During inference, the MRF converges into an equilibrium between $\theta_{i,j}^{(sc)}(l_i, l_j)$ and $\theta_i^{(s)}(l_i)$.

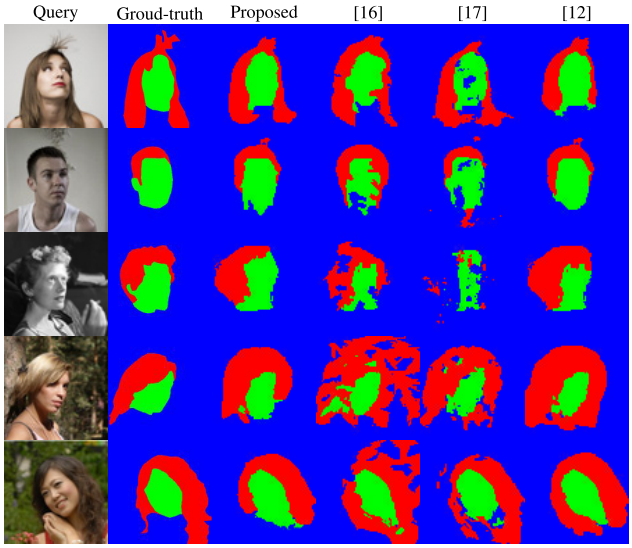


Fig. 6. Typical hair segmentation results obtained by our model (second column), when compared to the methods due to Krupka *et al.* [16] (third column), Lee *et al.* [17] (forth column) and Kae *et al.* [12] (fifth column).

its energy:

$$\hat{\boldsymbol{l}} = \arg\min_{\boldsymbol{l}} E(\boldsymbol{l}), \tag{4}$$

where $\{\hat{l}_1, \ldots, \hat{l}_{t_s}\}$ are the labels of the pixels and $\{\hat{l}_{t_s+1}, \ldots, \hat{l}_{t_s+t_c}\}$ specify the parameterizations in the classification nodes.

A. Feature Extraction

The data appearance is analyzed at the pixel level, to distinguish between three components in the image: hair, skin and background (any remaining information). As the red / blue chroma values provide good separability between skin and non-skin pixels [1] and the hair is frequently discriminated by analysing the HSV / RGB triplets (Table I), we extract, for each image pixel, a feature set composed of 81 elements: {red, green and blue channels (RGB); hue, saturation and value channels (HSV); red and blue chroma (yCbCr); LBP from the value channel}, considering the average, standard deviation and range statistics in square patches of side {5, 9, 15} around the central element.

B. Learning

1) Unary Potentials: Let $\gamma : \mathbb{N}^2 \rightarrow \mathbb{R}^{81}$ be the feature extraction function that produces a vector $\gamma(x, y) \in \mathbb{R}^{81}$ for each pixel at position (x, y). Let $\Gamma = [\gamma(x_1, y_1), \ldots, \gamma(x_n, y_n)]^T$ be a $n \times 81$ matrix in a learning set used to create three non-linear binary classification models, one for each component $\omega_i \in$ {"Hair", "Skin", "Background"}. Let $\eta_i : \mathbb{R}^{81} \rightarrow [0, 1]$ be the response of the i^{th} model, regarded as an estimate of the class likelihood $P\Big(\eta_i\big(\gamma(x, y)\big)|\omega_i\Big)$. According to the Bayes rule, and assuming equal priors, the posterior probabilities are given by:

$$P\Big(\omega_i|\eta_i\big(\gamma(x, y)\big)\Big) = \frac{P\Big(\eta_i\big(\gamma(x, y)\big)|\omega_i\Big)}{\sum_{j=1}^{3} P\Big(\eta_j\big(\gamma(x, y)\big)|\omega_j\Big)}. \tag{5}$$

In our model, the unary potentials of the vertices in the segmentation layer are defined as $\theta_i^{(s)}(l_i) = 1 - P\Big(\omega_i|\eta_i\big(\gamma(x, y)\big)\Big)$. The unary potentials in the classification layer correspond to the agreement (exclusive-or) between the index of the maximum posterior probability at each point $I_m(x, y) = \arg\max_j p\Big(\omega_j|\eta_j\big(\gamma(x, y)\big)\Big)$ and the 3D model projections $I_p(x, y)$ obtained by the *world-to-image* function (2):

$$\theta_i^{(c)}(l_i) = \frac{1}{h\ w} \sum_{y=1}^{h} \sum_{x=1}^{w} \big(1 - \delta(I_m(x, y), I_p(x, y))\big), \tag{6}$$

with $\delta(.,.)$ being the Kronecker delta function, h and w the query height and width. The rationale here is to privilege the hair and facial hair models that provide the maximum overlap between the responses of the non-linear models and the projections of the corresponding 3D meshes.

2) Pairwise Potentials: There are three types of pairwise potentials in our model: 1) between segmentation nodes; 2) between classification nodes; and 3) between inter-layer nodes. The pairwise potentials between segmentation nodes $\theta_{i,j}^{(s)}(l_i, l_j)$ correspond to the prior probability of observing labels l_i, l_j in adjacent positions of a learning set, to privilege smooth solutions:

$$\theta_{i,j}^{(s)}(l_i, l_j) = \frac{1}{\kappa_1 + P(\mathfrak{C}(x', y') = \omega_i, \mathfrak{C}(x, y) = \omega_j)}, \tag{7}$$

where $P(.,.)$ is the joint probability, (x', y') and (x, y) are 4-adjacent positions, $\mathfrak{C}(.,.)$ denotes the component label {hair, skin, background} at one position and $\kappa_1 \in \mathbb{R}^+$ avoids infinite costs.

The pairwise potentials between classification nodes $\theta_{i,j}^{(c)}(l_i, l_j)$ consider the prior probabilities of observing two facial hair and hair hypotheses in the learning set (e.g., beards

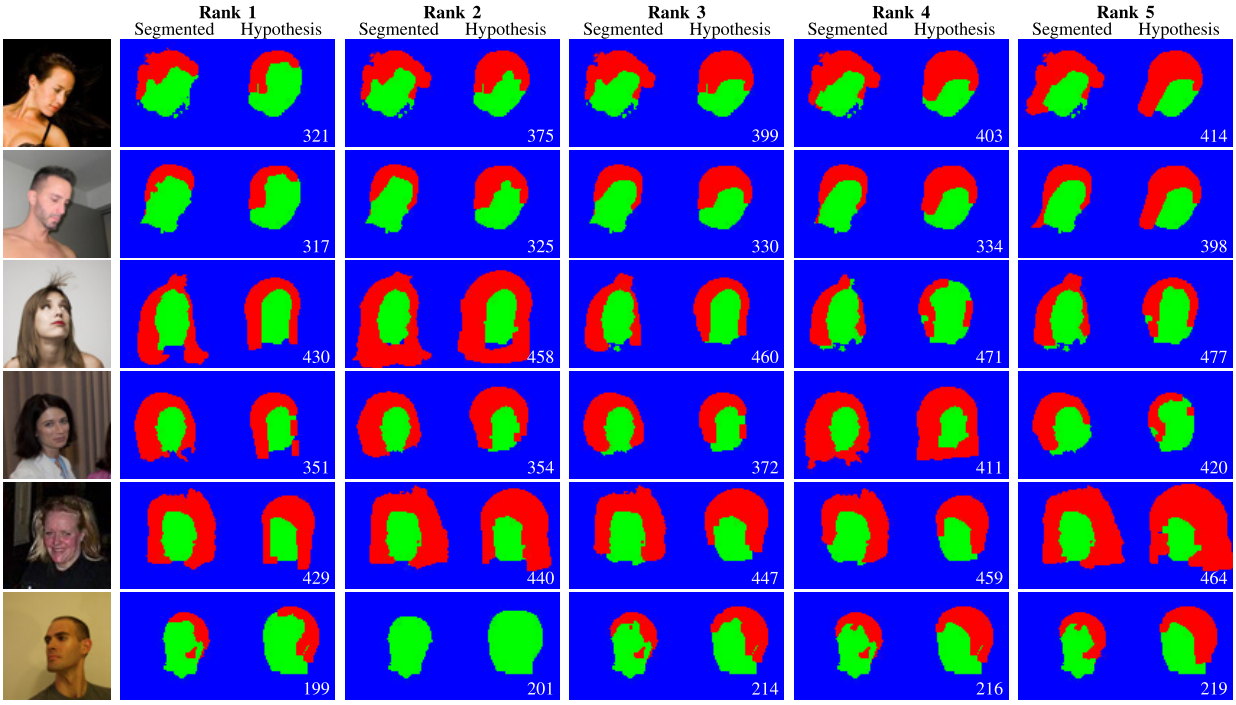


Fig. 7. Hair segmentation / hairstyle inference results obtained by the proposed method. The queries are shown at the leftmost column. For each query, the most likely segmentation and hairstyle models are given in descending likelihood order (from left to right), showing also the cost of the optimal MRF state.

are more probable in bald and short hair than in long / long volume subjects).

The pairwise potentials between inter-layer nodes $\theta_{i,j}^{(sc)}(l_i, l_j)$ enforce the biological plausibility of the solution (8) and privilege the consistency between the configurations in both layers. This is done by penalising parameterisations of pixel nodes that are outside of the polygons defined by the boundaries of the projections of the head, hair and facial hair models (e.g., it is too costly to observe a *hair* pixel and a *bald* hypothesis).

$$\theta_{i,j}^{(sc)}(l_i, l_j) = \begin{cases} 0, & \text{if } \delta\Big(\delta(l_i, \mathfrak{C}'(i, l_j)), \\ & \quad \delta(\psi_t(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j), 0\Big) = 0 \\ \text{erf}\Big(\kappa_2 \ \psi_d(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j)\Big), & \text{otherwise,} \end{cases}$$

where $\psi_t(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j) : \mathbb{N}^2 \times \mathbb{N}^n \rightarrow \{0, 1\}$ is an indicator function that assumes a unit value when the point (x_i, y_i) is inside the polygon defined by vertices $\{\mathbf{x}_j, \mathbf{y}_j\} = \{(x_{j,k}, y_{j,k})\}$. $\psi_d(x_i, y_i, \mathbf{x}_j, \mathbf{y}_j) : \mathbb{N}^2 \times \mathbb{N}^n \rightarrow \mathbb{R}^+$ is the point-to-polygon distance divided by the image diagonal length. erf(.) : $\mathbb{R}^+ \rightarrow [0, 1]$ is a transfer function (error function) with sigmoid shape with κ_2 controlling its shape (larger values lead to farther from linear shapes). Here, $\mathfrak{C}'(i, l_j)$ denotes the component label (hair, skin or background) at the i^{th} image position under the j^{th} joint facial hair / hair hypothesis. Fig. 5 illustrates the rationale of this kind of costs: for two queries, the responses given by the three non-linear classifiers are shown at the left side. The right side shows one plausible (green square) and one unlikely (red square) hair hypothesis, with the corresponding pairwise costs.

V. Results and Discussion

A. Datasets

The LFW [9] was the main dataset used in the empirical validation of our model, due to two reasons: 1) it contains heterogenous images acquired indoor / outdoor, with the degradation factors that are likely in visual surveillance environments; and 2) it has a subset of manually segmented images (the *funnelled* version) into hair, skin and background. Additionally, the AFLW [15] set was considered for evaluating the variations in segmentation performance with respect to errors in the head landmarks detection phase.

B. Model Inference

All our models were optimized using the Loopy Belief Propagation [7] algorithm. Even though it is not guaranteed that it converges to global minimums on non sub-modular graphs (such our models), it provides visually pleasant solutions most of the times. As future work, we plan to evaluate the effectiveness of our method according to more sophisticated energy minimization algorithms (e.g., sequential tree-reweighed message passing [14]).

C. Segmentation

We compared the segmentation accuracy of our method to three baseline methods: 1) a computationally inexpensive method due to Wang *et al.* [29], based on a set of seeds from where the adjacent regions are thresholded; 2) a single layered MRF due to Lee *et al.* [17], which is a particular case of our model, with constant costs in the *objects* layer

TABLE II

COMPARISON BETWEEN THE PIXEL SEGMENTATION PERFORMANCE (AFLW SET). THE SUPERSCRIPTS GIVE THE 95% CONFIDENCE INTERVALS

Method	Overall	Hair	Skin	Background
Proposed Method	$2.87^{\pm 0.03}$	$1.99^{\pm 0.04}$	$3.02^{\pm 0.03}$	$2.95^{\pm 0.03}$
Krupka *et* al. [16]	$4.21^{\pm 0.05}$	$4.05^{\pm 0.06}$	$4.70^{\pm 0.06}$	$4.16^{\pm 0.04}$
Lee *et* al. [17]	$4.23^{\pm 0.05}$	$4.19^{\pm 0.06}$	$5.26^{\pm 0.07}$	$4.09^{\pm 0.04}$
Kae *et* al. [12]	$2.85^{\pm 0.03}$	$2.02^{\pm 0.04}$	$2.98^{\pm 0.04}$	$2.93^{\pm 0.03}$

and in the inter-layer edges; and 3) the method due to Kae *et al.* [12], which we consider the state-of-the-art and has a rationale much similar to our solution: it uses a random field to model the transitions at the pixel level and a restricted Boltzmann machine to enforce globally coherent hypotheses. Fig. 6 illustrates the typical outputs provided by the methods compared: whereas ours and Kae *et al.* methods typically produce similar results, Krupka *et al.* and Lee *et al.* methods are frequently trapped in local minima of their cost functions, due to not enforcing the biological coherence of the solutions. Particularly, Krupka *et al.* produce poor results when the seeds do not faithfully represent the distributions of the components (due to textured data). Finally, by regarding exclusively image appearance, Lee et al.'s method often produces biological unlikely solutions, with discontiguous skin / hair regions with boundaries having too many number of degrees-of-freedom.

More objectively, Table II quantifies the average segmentation performance for the methods evaluated. We got slightly better results than Kae *et al.* for the hair component and worse results for the skin and background, in all cases with differences not being statistically significant (inside the 95% confidence intervals). The method due to Krupka *et al.* ranked third, yet it was the one that most frequently produced biologically inconsistent solutions. Also, this method performed particularly poor in highly textured background images, where the seeds hardly represent the high entropy in the background regions.

Another interesting feature of our method is its ability to rank the plausibility of the hypotheses with respect to the queries. This can be done by optimizing the model iteratively and, at each step, remove the hypothesis considered optimal in the previous iteration. Results of this ordering are shown in Fig. 7, with the top-5 most similar hair hypotheses with respect to queries, along with the segmentation masks for each hypothesis. At the bottom-right corner, the cost of the solution is given, i.e., the cost of fitting the segmentation mask in the corresponding template.

Note that our results depend of the head landmarks to infer the head shape and pose hypotheses (Sec. III-C). Failures at this point introduce a bias in the way hypotheses are projected and in the MRF unary / pairwise costs. Hence, we used a set of ground-truth head landmarks (AFLW set) to perceive the sensitivity to this factor, introducing inaccuracies in landmarks detection by adding random offsets to the accurate landmarks. Results are given in Fig. 8 (the overall accuracy is shown), with respect to the proportion of landmarks inaccurately detected (horizontal axis) and the relative magnitude of the offset

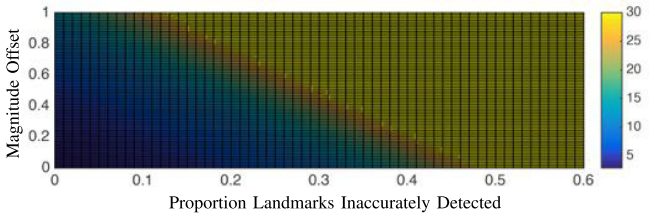


Fig. 8. Variations in segmentation performance of the proposed method with respect to the proportion of head landmarks inaccurately detected (horizontal axis) and the magnitude of these inaccuracies (vertical axis). The overall accuracy is shown.

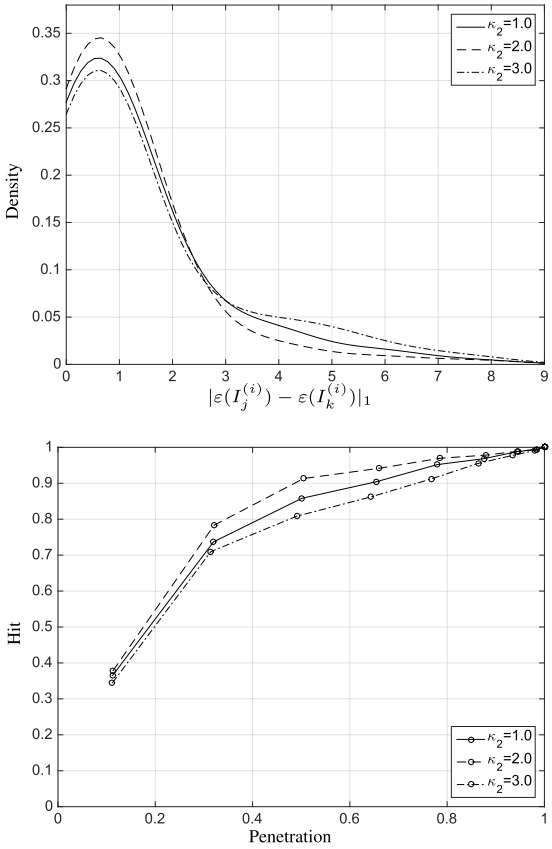


Fig. 9. Top plot: probability density functions for the distance between intra-subject labels $|\varepsilon(I_j^{(i)}) - \varepsilon(I_k^{(i)})|_1$. Bottom plot: hit / penetration plots for the LFW data set.

(i.e., the Euclidean distance between the original and the biased landmark positions, weighted by the image diagonal length, vertical axis). The segmentation performance remained approximately invariant when less than 20% of the head landmarks were inaccurate. Also, the magnitude of the detections offset was observed to play a relatively minor role in segmentation accuracy, but, in practice, the algorithm looses its effectiveness when more than 35% of the landmarks are inaccurate.

D. Identity Retrieval

This section reports the identity retrieval results in the LFW set. A one-dimensional manifold $\mathbf{M}$ for the hair models was inferred using a self-organized map fed by a feature set composed of the concatenation of the mode label in local

3D volumes regularly sampled in the 3D hair models $\mathbf{S}_h$: $\mathbf{M}$:={0: bald, 1: short bald, 2: short, 3: medium, 4: long fine, 5: long volume}. Let $\mathbf{I}_j^{(i)}$ be the j^{th} image from the i^{th} subject ($j = 1, \ldots t_i$), with t_i representing the number of images for that subject. Let $\varepsilon(I_{\cdot}^{(\cdot)})$: $\mathbb{N}^2 \rightarrow \mathbf{M}$ be the inference function (MRF) that associates one query to one hair style in $\mathbf{M}$. $|\varepsilon(I_j^{(i)}) - \varepsilon(I_k^{(i)})|_1$ captures the spread of the intra-subject labels distribution, with the probability density function for this value shown in the upper part of Fig. 9. Results are given with respect to the κ_2 parameter that controls the shape of the transfer function (Sec. IV-B.2). In all cases, it is obvious that large deviation values (> 3, for $\kappa_2 = 2.0$) in intra-subject labels rarely occur, which is the insight for using these labels in identity retrieval. The bottom plot gives the corresponding hit / penetration plots, once again with respect to the κ_2 parameter.

VI. Conclusions

Being a classical tool in computer vision, MRFs traditionally have difficulties in assuring globally coherent solutions without using too-high order cliques that compromise the computational effectiveness of the inference process. In this paper we described a hierarchical architecture for MRFs free of high-order cliques that still enforces globally coherent models. The idea is to have the bottom layer working at the local (pixel) level, while the upper layers work at the hypotheses level, providing possible solutions for the problem. During optimization, all layers interact and converge into an equilibrium state, where the configuration in the bottom layer implicitly segments the data, and the configuration in the other layers correspond to the most likely models. As test case, we considered the segmentation and labelling of hair / facial hair styles in degraded data, which are important soft biometric labels for human recognition *in-the-wild*. Our experiments were carried out in the challenging LFW data set, and we observed performance similar to the state-of-the-art methods, both in the hair segmentation and hairstyle labelling tasks, and at a much lower computational cost. Further, the proposed MRF architecture can be applied with minimal adaptations to other segmentation / classification computer vision problems, particularly in cases where the biological (global) coherence of the solutions can be objectively measured.

References

[1] A. Albiol, L. Torres, and E. J. Delp, "Optimum color spaces for skin detection," in *Proc. Int. Conf. Image Process.*, Oct. 2001, pp. 122–124.

[2] A. Besbes, N. Komodakis, G. Langs, and N. Paragios, "Shape priors and discrete MRFs for knowledge-based segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1295–1302.

[3] R. Byrd, M. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 877–900, 1999.

[4] B. Caputo, S. Bouattour, and H. A. Niemann, "Robust appearance-based object recognition using a fully connected Markov random field," in *Proc. 16th Int. Conf. Pattern Recognit.*, Aug. 2002, pp. 565–568.

[5] J. Dass, M. Sharma, E. Hassan, and H. Ghosh, "A density based method for automatic hairstyle discovery and recognition," in *Proc. 4th Nat. Conf. Comput. Vis. Pattern Recognit. Image Process. Graph. (NCVPRIPG)*, Dec. 2013 , pp. 1–4.

[6] B. Efraty, C. Huang, S. K. Shah, and I. A. Kakadiaris, "Facial landmark detection in uncontrolled conditions," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–8.

[7] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, Oct. 2006.

[8] B. Glocker, D. Zikic, N. Komodakis, N. Paragios, and N. Navab, "Linear image registration through MRF optimization," in *Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro*, Jul. 2009, pp. 422–425.

[9] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[10] S. Jaiswal, T. Almaev, and M. Valstar, "Guided unsupervised learning of mode specific models for facial point detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Jun. 2013, pp. 370–377.

[11] P. Julian, C. Dehais, F. Lauze, V. Charvillat, A. Bartoli, and A. Choukroun, "Automatic hair detection in the wild," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 4617–4620.

[12] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting CRFs with boltzmann machine shape priors for image labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2019–2026.

[13] Z. Kato and T. C. Pong, "A Markov random field image segmentation model for color textured images," *Image Vis. Comput.*, vol. 24, no. 10, pp. 1103–1114, Oct. 2006.

[14] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, Oct. 2006.

[15] M. Köstinger, P. Wohlhart, P. Roth, and H. A. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Nov. 2011, pp. 2144–2151.

[16] A. Krupka, J. Prinosil, K. Riha, and J. Minar, "Hair segmentation for color estimation in surveillance systems," in *Proc. 6th Int. Conf. Adv. Multimedia*, Jan. 2014, pp. 102–107.

[17] K. Lee, D. Anguelov, B. Sumengen, and S. Gokturk, "Markov random field models for hair and face segmentation," in *Proc. 8th IEEE Int. Conf. Automat. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[18] U. Lipowezky, O. Mamo, and A. Cohen, "Using integrated color and texture features for automatic hair detection," in *Proc. IEEE 25th Conv. Elect. Electron. Eng. Israel*, Dec. 2008, pp. 51–55.

[19] H. Proença, J. C. Neves, S. Barra, T. Marques, and J. C. Moreno, "Joint head pose/soft label estimation for human recognition in-the-wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 38, no. 12, pp. 2444–2456, Dec. 2016, doi: 10.1109/TPAMI.2016.2522441.

[20] H. Proença, J. C. Neves, and G. Santos, "Segmenting the periocular region using a hierarchical graphical model fed by texture / shape information and geometrical constraints," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2014, pp. 1–7.

[21] V. Rapp, T. Senechal, K. Bailly, and L. Prevost, "Multiple kernel learning SVM and statistical validation for facial landmark detection," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 265–271.

[22] C. Rousset and P. Y. Coulon, "Frequential and color analysis for hair mask segmentation," in *Proc. 15th Int. Conf. Image Process.*, Oct. 2008, pp. 2276–2279.

[23] Y. Shen, Z. Peng, and Y. Zhang, "Image based hair segmentation algorithm for the application of automatic facial caricature synthesis," *Sci. World J.*, vol. 2014, 2014, Art. no. 748634, doi: 10.1155/2014/748634.

[24] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon, "Soft biometrics and their application in person recognition at a distance," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 3, pp. 464–475, Mar. 2014.

[25] Y. Ugurlu, "Head posture detection using skin and hair information," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 1–4.

[26] C. Wang, M. Gorce, and N. Paragios, "Segmentation ordering and multi-object tracking using graphical models," in *Proc. 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 747–754.

[27] N. Wang and H. Ai, "Hair style retrieval by semantic mapping on informative patches," in *Proc. 1st Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2011, pp. 110–114.

[28] D. Wang, S. Shan, H. Zhang, W. Zeng, and X. Chen, "Isomorphic manifold inference for hair segmentation," in *Proc. 10th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.(FG)*, Apr. 2013, pp. 1–6.

[29] Y. Wang, Z. Zhou, E. Teoh, and B. Su, "Human hair segmentation and length detection for human appearance model," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 450–454.

[30] Y. Yacoob and L. S. Davis, "Detection and analysis of hair," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1164–1169, Jul. 2006.

[31] J. Young, "Head and face anthropometry of adult U.S. civilians. Office of aviation medicine," Federal Aviation Admin., Washington, DC, USA, Tech. Rep. DOT/FAA/AM-93/10, 1993.

[32] C. Yuksel, S. Schaefer, and J. Keyser, "Hair meshes," *ACM Trans. Graph.*, vol. 28, no. 5, p. 166, Dec. 2009.
[33] Z. Zhang, H. Gunes, and M. Piccardi, "An accurate algorithm for head detection based on XYZ and HSV hair and skin color models," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1644–1647.
[34] Z. Zhang, H. Gunes, and M. Piccardi, "Head detection for video surveillance based on categorical hair and skin Colour models," in *Proc. 16th IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 1137–1140.

Hugo Proença (SM'12) received the B.Sc., M.Sc., and Ph.D. degrees from the University of Beira Interior, in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the University of Beira Interior and has been researching mainly about biometrics and visual-surveillance. He is the Coordinating Editor of the IEEE BIOMETRICS COUNCIL NEWSLETTER and the Area Editor (ocular biometrics) of the IEEE BIOMETRICS COMPENDIUM JOURNAL. He is a member of the Editorial Boards of the *Image and Vision Computing* and the *International Journal of Biometrics* journals and served as a Guest Editor of special issues of the *Pattern Recognition Letters*, *Image and Vision Computing*, and *Signal, Image and Video Processing* journals.

João C. Neves (M'15) received the B.Sc. and M.Sc. degrees in computer science from the University of Beira Interior, Portugal, in 2011 and 2013, respectively, where he is currently pursuing the Ph.D. degree in biometrics. His research interests include computer vision and pattern recognition, with a particular focus on biometrics and surveillance.

IRINA: Iris Recognition (even) in Inaccurately Segmented Data

Hugo Proença and João C. Neves
IT - Instituto de Telecomunicações
University of Beira Interior, Portugal
{hugomcp,jcneves}@di.ubi.pt

Abstract

*The effectiveness of current iris recognition systems depends on the accurate segmentation and parameterisation of the iris boundaries, as failures at this point misalign the coefficients of the biometric signatures. This paper describes **IRINA**, an algorithm for **I**ris **R**ecognition that is robust against **INA**ccurately segmented samples, which makes it a good candidate to work in poor-quality data. The process is based in the concept of "*corresponding*" patch between pairs of images, that is used to estimate the posterior probabilities that patches regard the same biological region, even in case of segmentation errors and non-linear texture deformations. Such information enables to infer a free-form deformation field (2D registration vectors) between images, whose first and second-order statistics provide effective biometric discriminating power. Extensive experiments were carried out in four datasets (CASIA-IrisV3-Lamp, CASIA-IrisV4-Lamp, CASIA-IrisV4-Thousand and WVU) and show that IRINA not only achieves state-of-the-art performance in good quality data, but also handles effectively severe segmentation errors and large differences in pupillary dilation / constriction.*

1. Introduction

Iris recognition is a mature technology, with systems successfully deployed in domains such as border controls, computers login and national ID cards. Since the pioneer algorithm [5] proposed in 1993, a long road has been travelled in iris biometrics research [2], with two major weaknesses subsisting:

- accurate segmentation and parameterization of the iris boundaries is required to image normalisation. As most of the iris encoding / matching strategies are phase-based, failures in segmentation lead to bit shifting in the biometric signatures, with a corresponding increase of false rejections;
- false rejections also increase in case of severely dilated / constricted pupils, which cause non-linear deformations in the iris texture that are only partially compensated by the normalisation phase. Pupil movements laterally pressure the iris, with some of the fibers folding underneath others and changing texture appearance.

Note that 1) varying lighting conditions change the levels of pupillary dilation; and 2) less constrained acquisition protocols reduce data quality and make hard to accurately parameterise the iris boundaries. Hence, the *robustness* of recognition can be seen as the major concern behind the method proposed in this paper (IRINA), keeping as main goal to achieve state-of-the-art performance in good-quality data while also handling segmentation inaccuracies and non-linear texture deformations.

A cohesive perspective of IRINA is given in Fig. 1, with a processing chain divided into three phases:

1. we estimate the posterior probabilities that patches from two iris samples *correspond*, even in case of non-linear texture deformations. Starting from a learning set of manually annotated point correspondences that define convex polygons, we densely sample these regions and obtain a large number of patches considered to regard the same biological region. This information feeds a convolution neural network (CNN), that: a) explicitly discriminates between the *corresponding* and *non-corresponding* patches; and b) implicitly learns the typical iris texture deformations;

2. we infer a free-form deformation field (set of 2D vectors) that registrates pairs of samples represented in normalised coordinates. This step is formulated using a discrete Markov random field (MRF), with unary costs provided by the responses of the CNN, and pairwise costs imposing smooth solutions that penalize local gradients of the deformation field. The loopy belief propagation (LBP) algorithm [8] is used to solve the image registration problem;

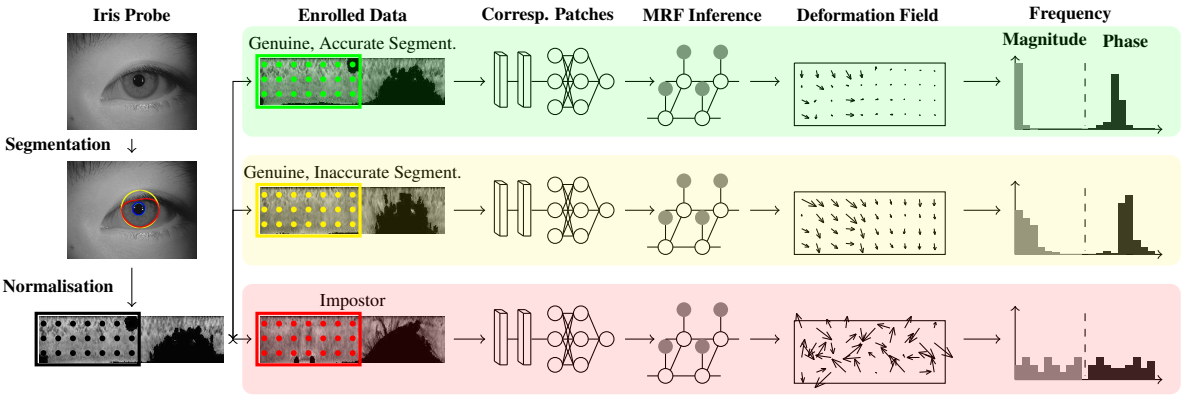


Figure 1. Overview of IRINA's processing chain. The first and second order statistics of the free-form deformation vector fields are the basis for biometric recognition. Deformation fields are exaggerated for illustration purposes.

3. for biometric recognition, the key observation is that *genuine* deformation fields (between samples of the same subject) are composed of 2D vectors with phase and magnitude gradients substantially smaller than those of *impostors*. First and second-order statistics of these vectors provide the discriminating information for biometric recognition.

Belonging to the discriminative family of pattern recognition methods, IRINA's rationale is evidently original with respect to the state-of-the-art, in which the generative paradigm rules: apart from assuming that data is accurately segmented, previous methods consider that no parts of the iris texture appear / vanish due to pupillary dilation. As an example, Thornton *et al*. [28] assume that the iris regions unaffected by pupil dilation still provide enough information for matching (providing the insight for subsequent works [14] and [24]) , while other authors provided (inevitably rough) parameterizations of iris deformations (e.g., [33], [4] and [29]). In a discriminative approach, Ross *et al*. [21] propose an information fusion framework where three distinct feature extraction and matching schemes are fused to handle the significant variability in the input ocular images. Finally, note that our idea of *corresponding* patch is different from the used in keypoint-matching iris recognition algorithms, which analyzed the geometric distribution of perfectly matching pairs of keypoints between two images (e.g., using SIFT descriptors [1]), but fail in case of varying levels of focus, lighting or non-linear iris deformations.

1.1. Iris Recognition

Given the maturity of iris biometrics technology, strides have been concentrated in improving particular features of the recognition process: i) extending the data acquisition volume; ii) improving performance in *less constrained* conditions; iii) augment the human interpretability of results; iv) develop cancellable signatures; and v) provide inter-sensor operability.

In terms of the data acquisition volume, a good example is the *iris-on-the-move* system [17], that acquires data from subjects walking through a portal. For similar purposes, Hsieh *et al*. [13] used wavefront coding and super-resolution techniques. In terms of the recognition robustness, Dong et al. [7] proposed an adaptive personalized matching scheme that highlights the most discriminating features. Pillai *et al*. [19] used the sparse representation for classification algorithm in randomly projected iris patches, claiming to increase the robustness against acquisition artefacts. Yang *et al*. [32] relied in high-order information to perform iris matching, while Alonzo-Fernandez *et al*. [9] focused in the image enhancement phase, proposing a super-resolution method based on PCA and eigen-transformations of local iris patches. Bit consistency is also a concern, with several approaches selecting only parts of the biometric signatures for matching (e.g. [12], [27] and [16]).

Under complementary perspectives, the lack of interpretability hinders the use of iris recognition in forensics [3]. Also, inter-sensor recognition provided the motivation for Pillai et al. [20], which learned transformations between data acquired by different sensors. Cancellable biometrics is a privacy-preserving solution that requires to find hardly invertible transfer functions of the biometric data into different domains: Zhao *et al*. [34] proposed the concept of negative recognition, using only complementary information (p-hidden algorithm) of the biometric data for matching. Finally, according to the growing popularity of CNNs, various approaches based on this paradigm appeared recently in the literature, either for specific phases of the recognition chain (e.g., segmentation [15] or spoofing de-

tection [18]) or for the whole process [10].

1.2. Image Registration

Image registration involves three components: i) a transformation model that maps regions of one image into regions of another; ii) a similarity criterion, that quantifies the nearness between image patches; and iii) an optimization strategy, that finds a global mapping between both images.

Transformation models can be global / local. The first family includes linear transformations such as rotation, scaling, translation and affine. Local transformations allow to warp regions of one image into another, using radial basis functions, physical continuum and large deformation models. The similarity criterion quantifies how much one image patch resembles another one in the reference data, using cross-correlation, mutual information or other distance functions. Similarity can be intensity or feature-based, with the latter family matching the most complex structures as lines and curves, based in spatial and frequency information. Finally, during optimization the set of parameters that optimally match both images are found. Exhaustive search techniques were firstly used here, but later abandoned due to their reduced computational feasibility. Modern approaches use optimization algorithms and gradient-free / gradient based techniques to derive reasonable solutions, which might be sub-optimal in case of non-convex cost functions. For additional information, Sotiras *et al*. [25] provide an overview of the state-of-the-art in image registration.

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the proposed method. In Section 3 we discuss the obtained results and the conclusions are given in Section 4.

2. Proposed Method

2.1. Corresponding Iris Patches

The concept of *corresponding* patches between pairs of iris images is the key to learn the typical non-linear deformations in normalized representations of the iris due to pupillary dilation / constriction and segmentation errors.

Iris recognition systems comprise a normalisation phase [6] that compensates for differences in scale, perspective and pupillary dilation, assuming that iris deformations are linear and limited to the radial direction. This does not compensate for the actual deformations, which are non-linear, radial and angular, with fibers vanishing / appearing for different levels of pupillary dilation [30]. Several authors proposed non-linear iris normalization schemes to attenuate the problem: Wyatt [31] developed a mathematical model to explain how the collagen fibers in the iris deform and Yuan and Shi [33] proposed a scheme based on that model. Also, Clark *et al*. [4] described a theoretical model for the iris dynamics, used subsequently by Tomeo-Reyes *et al*. [29].

To infer the *corresponding* patches, we use pairs of normalized samples from the same subject and manually annotate sets of control points that (by visual inspection) seem to regard the same biological region. These control points define two convex polygons $\mathbf{\Gamma}$ and $\mathbf{\Gamma}'$ and are represented by the coloured dots ($\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$) in the upper part of Fig. 2. Let $\boldsymbol{x}_i = (x_i, y_i)$ and $\boldsymbol{x}'_i = (x'_i, y'_i)$, $i = \{1, \dots n\}$ be the locations of pointwise correspondences. We learn two functions f_r, f_c that establish a dense set of correspondences between positions (rows, columns) in $\mathbf{\Gamma}$ and $\mathbf{\Gamma}'$, $f_r, f_c : \mathcal{N}^2 \rightarrow \mathcal{N}$, such that $\forall\ \boldsymbol{x}'_i \in \mathbf{\Gamma}'$, $\boldsymbol{x}'_i = \big(f_c(\boldsymbol{x}), f_r(\boldsymbol{x})\big)$:

$$f_c(\boldsymbol{x}) = \boldsymbol{\lambda}_c^T[\boldsymbol{\phi}, p(\boldsymbol{x})], \tag{1}$$

$$f_r(\boldsymbol{x}) = \boldsymbol{\lambda}_r^T[\boldsymbol{\phi}, p(\boldsymbol{x})], \tag{2}$$

with $\boldsymbol{\phi} = \big[\phi\big(|\boldsymbol{x} - \boldsymbol{x}_1|_2\big), \dots, \phi\big(|\boldsymbol{x} - \boldsymbol{x}_n|_2\big)\big]$, $|.|_2$ representing the ℓ_2 norm, $\phi(r) = e^{(-r/\kappa)^2}$ being a radial basis function and $p(\boldsymbol{x}) = [1, x, y]$ being a polynomial basis of first degree for a 2-dimensional vector space ($\kappa = 0.1$ was used in all our experiments).

In order to obtain the $\boldsymbol{\lambda}$ coefficients, we define a $n \times n$ matrix $\boldsymbol{A}$, $A_{i,j} = \phi\big(|\boldsymbol{x}_i - \boldsymbol{x}_j|_2\big)$ and $\boldsymbol{P}$ as the $n \times 3$ polynomial basis matrix, such that $\boldsymbol{P} = [p(\boldsymbol{x}_1); \dots; p(\boldsymbol{x}_n)]$. Then, $\boldsymbol{\lambda}_c$ and $\boldsymbol{\lambda}_r$ are given by:

$$\boldsymbol{\lambda}_c = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{0} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}'_c \\ \boldsymbol{0} \end{bmatrix} \tag{3}$$

$$\boldsymbol{\lambda}_r = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{P} \\ \boldsymbol{P}^T & \boldsymbol{0} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}'_r \\ \boldsymbol{0} \end{bmatrix} \tag{4}$$

with $\boldsymbol{x}'_c$=$[x'_1, \dots, x'_n]^T$ and $\boldsymbol{x}'_r$=$[y'_1, \dots, y'_n]^T$ concatenating the horizontal (column) and vertical (row) positions of the control points in $\mathbf{\Gamma}'$.

According to this procedure, we deem that positions $\boldsymbol{x} \in \mathbf{\Gamma}$ correspond biologically to $\boldsymbol{x}' = \big(f_c(\boldsymbol{x}), f_r(\boldsymbol{x})\big) \in \mathbf{\Gamma}'$. As $\mathbf{\Gamma}$ and $\mathbf{\Gamma}'$ have different size and shape, this set of correspondences implicitly encodes the non-linear deformations that affect the iris texture. Finally, we consider patches $\boldsymbol{P}$ (from $\mathbf{\Gamma}$) and $\boldsymbol{P}'$ (from $\mathbf{\Gamma}'$) of 21×21 pixels, cropped from the learning data and centered at each point correspondence.

Using 320 images (from 75 subjects) of the CASIA-IrisV3-Lamp set, 510,000 corresponding $C_{i,j}$ patches ($C_{i,j} = [\boldsymbol{P}_i, \boldsymbol{P}'_j; \boldsymbol{P}'_j, \boldsymbol{P}_i]$) were cropped. Also, using image pairs from different subjects, another 510,000 non-corresponding $\bar{C}_{i,j}$ patches were created. Both were used to train a CNN that extracts high-level texture information and distinguishes between the corresponding / non-corresponding patches. Note that the iris boundaries in this

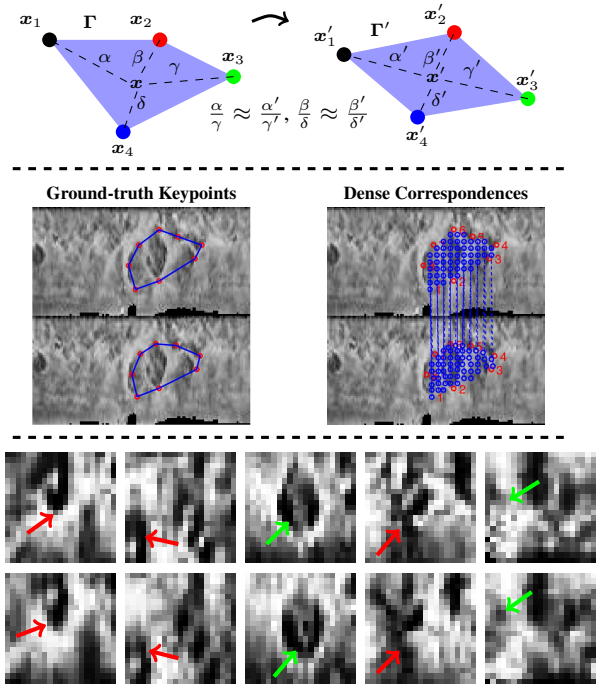


Figure 2. Concept of *corresponding* iris patches. The top part of the figure gives a schema of the way correspondences are found: based on a set of manually marked corresponding control points between two iris samples ($\boldsymbol{x}_.$ and $\boldsymbol{x}'_.$), two polygons ($\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}'$) are defined. Next, for every point inside $\boldsymbol{\Gamma}$, the corresponding position in $\boldsymbol{\Gamma}'$ is found (middle row). The bottom part of the figure shows five pairs of corresponding iris patches, where non-linear deformations (red arrows), and vanishing / emerging regions (green arrows) inside each patch are evident.

learning set (obtained as described in Sec. 3.1) were not manually confirmed, i.e., there are accurately and inaccurately samples in this set, which is important to infer the deformations in the iris texture yielding from segmentation failures.

The CNN input is 42×42 image patches and its architecture (Fig. 3) is composed of six layers (three convolutional plus three fully connected layers): the first convolutional layer uses 32 kernels (3×3), and the next ones are composed of 64 kernels of size $3 \times 3 \times 32$. The responses from these layers feed max-pooling layers (stride equals to 1, given the relatively small size of the input data). Next, there are two fully connected layers, each one with 256 cells. The output is a *soft max* loss corresponding to the probability of two iris patches to *correspond*. Learning was done according to the stochastic gradient descend algorithm, with an initial learning rate of $1e^{-2}$, momentum set to 0.9 and weight decay equals to $1e^{-3}$.

The responses of the CNN enable to obtain the posterior probabilities that two iris patches regard the same biological region. Such information enters a Markov random field (MRF), which energy minimization provides the solution to the image registration problem, used as information source for biometric recognition.

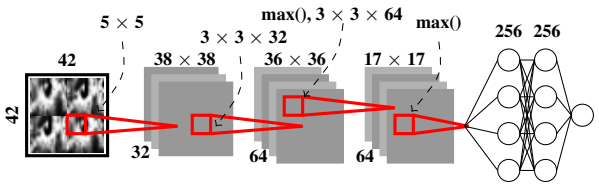


Figure 3. Structure of the convolutional neural network (CNN) used to discriminate between the *corresponding* and *non-corresponding* iris pairwise patches.

2.2. Deformation Field Inference

We consider a free-form transformation model [22] to represent a deformation field, expressed as a set of 2D vectors $\boldsymbol{d} \in \mathcal{Z}^2$ at control points $\mathring{\boldsymbol{x}} \in \mathcal{N}^2$. We superimpose a r $\times$ c regular grid at positions $\boldsymbol{G} = \{\mathring{\boldsymbol{x}}_1, \ldots, \mathring{\boldsymbol{x}}_{|G|}\}$, $|\boldsymbol{G}|$=r.c, over the left half of the normalized images representation (corresponding to the lower part of the iris that is less prune to occlusions and shadows). Also, we assume that deformations at any position $\boldsymbol{d}(\boldsymbol{x})$ can be obtained by interpolating the closest control points deformations [11]:

$$\boldsymbol{d}(\boldsymbol{x}) = \sum_{i=1}^{|G|} \nu(\boldsymbol{x})\, \boldsymbol{d}(\mathring{\boldsymbol{x}}_i), \tag{5}$$

with $\boldsymbol{d}(\mathring{\boldsymbol{x}}_i)$ representing the deformation at the ith control point and $\nu()$ being the interpolation function.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph representing a MRF, composed of a set of t_v vertices $\mathcal{V}$, linked by t_e edges $\mathcal{E}$. In our model, every control point of $\boldsymbol{G}$ is a vertex of $\mathcal{G}$, i.e., $t_v = |\boldsymbol{G}|$ and $t_e = 2.r.c - r - c$, using a typical grid configuration (4-neighborhood). The MRF is a representation of a discrete latent random variable $\boldsymbol{L} = \{L_i\}, \forall i \in \mathcal{V}$, where each element L_i takes one value l_i from a set of labels (each corresponding to a deformation vector $\boldsymbol{d}$). In practical terms, having the ith image patch centered at position $\boldsymbol{x}$, we find its corresponding patch in the second sample at positions $\boldsymbol{x} + \boldsymbol{m}$, $\boldsymbol{m} = (m_1, m_2)$, $m_i \in \{-m_{max}, \ldots m_{max}\}$. We use $m_{max} = 7$ in our experiments (Fig. 4).

Let $\boldsymbol{l} = \{l_1, \ldots, l_{t_v}\}$ be one configuration of the MRF. The energy of $\boldsymbol{l}$ is the sum of the unary $\upsilon_i(l_i)$ and pairwise $\upsilon(l_i, l_j)$ potentials:

$$E(\boldsymbol{l}) = \sum_{i \in \mathcal{V}} \upsilon_i(l_i) + \sum_{(i,j) \in \mathcal{E}} \upsilon(l_i, l_j). \tag{6}$$

According to this formulation, obtaining the deformation model between a pair of images is equivalent to infer the random variables in the MRF that minimize its energy:

$$\hat{\boldsymbol{l}} = \arg\min_{\boldsymbol{l}} E(\boldsymbol{l}), \tag{7}$$

where $\hat{\boldsymbol{l}} = \{\hat{l}_1, \ldots, \hat{l}_{t_v}\}$ ($\hat{l}_i \equiv \boldsymbol{d}_i$) are the labels inferred. In all cases, MRFs were optimized according to the Loopy Belief Propagation [8] algorithm. Even though it is not guaranteed to converge to global minimums on loopy non-sub modular graphs (such as ours), we concluded that the algorithm provides acceptable solutions most of the times.

2.2.1 Unary Costs

Let $\eta(i,j) : \mathbb{N}^2 \rightarrow [0,1]$ be the CNN response for one pair of patches, expressing the likelihood $p\big(\eta(i,j) \mid C_{i,j}\big)$ that the ith patch of one sample corresponds to the jth patch of its counterpart. According to the Bayes rule, and assuming equal priors, the posterior probability functions are given by:

$$p\big(C_{i,j} \mid \eta(i,j)\big) = \frac{p\big(\eta(i,j) \mid C_{i,j}\big)}{\sum_{k=1}^{|M|} p\big(\eta(i,k) \mid C_{i,k}\big)}, \tag{8}$$

with $|M|$ expressing the number of positions in the second image where we search for the position corresponding to the ith patch. This way, the unary costs of the labels in each vertex are defined as:

$$\upsilon_i(l_i) = \alpha \left(1 - p\big(C_{i,j} \mid \eta(i,j)\big)\right), \tag{9}$$

with $\alpha \in [0,1]$ determining the trade-off between the strength of the unary to the pairwise costs in MRF optimization.

2.2.2 Pairwise Costs

In our model, the pairwise costs serve to control the derivatives in the deformation field, i.e., penalise adjacent positions with dramatically different deformation vectors that are not biologically plausible.

Let l represent a deformation vector $\boldsymbol{d} \in \mathcal{Z}^2$ for one control point. For computational purposes, it is important to discretise the solution space, not only limiting the maximum displacement m_{max} allowed for $\boldsymbol{d}$, but also defining an appropriate sampling strategy (dense sampling produces $(2.m_{max}+1)^2$ labels). Based in [11], we use a circular sparse grid with $\frac{\sqrt{2}}{4}.\pi.\ r$ nodes, $r = \{1, \ldots, m_{max}\}$ at positions $x = i_r.\cos(\theta)$, $y = i_r.\sin(\theta)$, $\theta \in [0, 2\pi]$, being i the sampling rate at the r-radius circumference. This sparse sampling strategy reduces over 50% the number of labels without significant decreases in the method performance (leftmost part of Fig. 4).

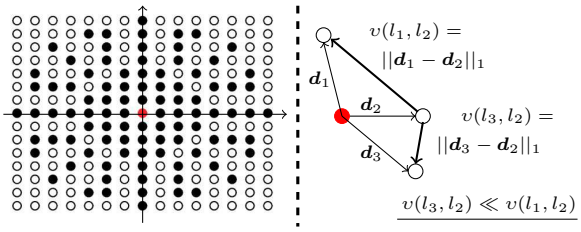


Figure 4. At left: comparison between the number of labels (maximum displacement $m_{max} = \pm 7$) when using a sparse sampling strategy, with respect to the dense sampling variant (solid black points denote the displacements $\boldsymbol{d}_i$ using the sparse sampling strategy, while the white points would have been also considered by the dense sampling strategy). At right: schema of the pairwise cost $\upsilon(l_i, l_j)$ for observing two displacement vectors $(\boldsymbol{d}_1, \boldsymbol{d}_2)$ and $(\boldsymbol{d}_3, \boldsymbol{d}_2)$ in adjacent positions of the deformation field: $\boldsymbol{d}_1$ being farther than $\boldsymbol{d}_3$ from $\boldsymbol{d}_2$ implies that $\upsilon(l_3, l_2) \ll \upsilon(l_1, l_2)$.

Finally, the pairwise cost for labelling two adjacent nodes is defined by:

$$\upsilon(l_i, l_j) = (1 - \alpha)\ |\boldsymbol{d}_i - \boldsymbol{d}_j|_1, \tag{10}$$

being $|.|_1$ the ℓ_1 norm.

2.3. Classification

The biometric recognition task is regarded as a binary classification problem. We use a machine-learned classifier to discriminate between the set of features extracted from positive (*genuine*) and negative (*impostor*) pairwise deformation fields. Let $\hat{\boldsymbol{l}} = \{\hat{l}_1, \ldots, \hat{l}_{t_v}\}$ represent the set of labels returned by the MRF. Each label l_i corresponds bijectively to a free form deformation vector $\boldsymbol{d}_i \in \mathcal{Z}^2$ at a position $\boldsymbol{x}$ of the normalised coordinates space. We extract the histogram of magnitudes and phase angles of $\boldsymbol{d}_i$ and their second-order statistics (local energy and homogeneity) from the magnitude and phase maps (with 6×12 vectors, taken in 3×3 and 5×5 regions, using stride 3 and 5), yielding 34 features that feed the binary discriminant (SVM in our case). A disjoint set from the CASIA-IrisV3-Lamp set (with 3,000 genuine / 3,000 impostor pairwise comparisons) was used as learning data at this point.

3. Results and Discussion

3.1. Datasets and Experimental Setting

IRINA was empirically validated in four iris datasets: CASIA-IrisV3-Lamp, CASIA-IrisV4-Lamp, CASIA-

IrisV4-Thousand [1] and WVU. [2]. Examples are given in Fig. 5, showing the degradation factors of each set: off-angle and occluded irises, glasses, dilated / constricted pupils (all sets) and shadows (WVU). 500 classes (eyes) per data set were used: for all the CASIA-Iris sets, 10 images per class were considered, while for the WVU the number of images per class varied between 2 and 10. All images were successfully segmented according to a coarse-to-fine strategy [23], composed by a form fitting step and a geodesic active contours algorithm. This way, we accurately parameterize the iris boundaries, having the pupillary contour described by shapes of 20 degrees-of-freedom (dof) and the scleric boundary described by 3 dof. At this point, images were normalised into the pseudo polar domain [6] and their right halves were discarded (corresponding to the upper half of the irises in the original representation).

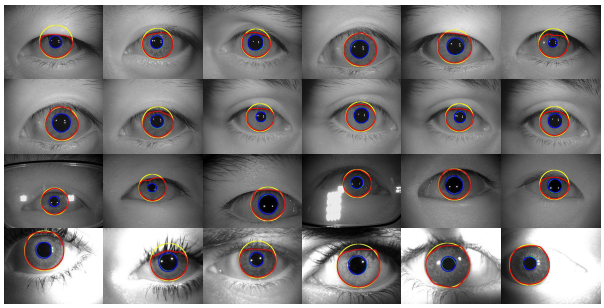


Figure 5. Datasets used in IRINA's empirical validation. From top to bottom rows, images of the CASIA-IrisV3-Lamp, CASIA-IrisV4-Lamp, CASIA-IrisV4-Thousand and WVU sets are shown.

As baselines, the methods due to Yang *et al.* [32] (using the O^{2}PT *iris-only* variant, with block size $w = 2, h = 14$, translation vector $[6, 3]^T$ and neighbourhood 8×8) and Sun and Tan [26] (with dilobe and trilobe filters, Gaussians 5×5, $\sigma = 1.7$, inter-lobe distances $\{5,9\}$ and sequential feature selection) were firstly considered, as both concern about the robustness of recognition to pupillary dilation and to non-linear iris deformations. Also, the method due to Belcher and Du [1] (with 64 bins = 4 (horizontal) $\times$ 4 (vertical) $\times$ 4 (orientation), SIFT descriptors extracted using VLFeat package [3]) was chosen due to the fact of being keypoints-based, even though its results cannot be considered state-of-the-art anymore. Three performance measures are reported: the decidability index (d'), the area under curve (AUC) and the receiver operating characteristic curve (ROC). In all experiments, the pairwise comparisons per dataset were divided into random samples (drew with repetition), each one with 90% of the available pairs. Then, independent performance tests were conducted in each subset, with the obtained results approximating the confidence intervals at each point, according to a bootstrapping-like strategy.

[1] CASIA iris image database, http://biometrics.idealtest.org

[2] West Virginia University iris dataset, http://www.clarkson.edu/citer/research/collections/

[3] http://www.vlfeat.org/

3.2. Learning and Parameter Tuning

It is important to note that the learning data used in the CNN was exclusively composed of CASIA-IrisV3-Lamp images. Using randomly sampled learning / validation and test sets (with 60% / 20% / 20% of the available pairwise comparisons), performance was tuned and all parameters strictly kept for the remaining datasets, meaning that the CASIA-Iris-V4-Lamp, CASIA-IrisV4-Thousand and WVU were used exclusively as test sets. The left plot in Fig. 6 shows the decision environment resulting from the responses $\eta(i,j)$ of the CNN, to distinguish between the *corresponding* $C_{i,j}$ and *non-corresponding* $\bar{C}_{i,j}$ iris patches. The likelihood functions $p\big(\eta(i,j) \mid C_{i,j}\big)$ and $p\big(\eta(i,j) \mid \bar{C}_{i,j}\big)$ in the CASIA-IrisV3-Lamp test set are shown.

In terms of IRINA's parameterisation, the value set to α (9) is the most sensitive, as it expresses the relative weight in the MRF between the unary and the pairwise costs. Here, $\alpha = 1$ corresponds to deformation vectors that are independent of their neighbours (no MRF would be required). In opposition, small α values reduce the local variations in the deformation field, with values below 0.9 imposing constant deformation fields with poor biometric discriminability. The AUC values obtained with respect to the value of α are shown in the right plot of Fig. 6. Note the significantly best performance (and smallest variance) for the CASIA-IrisV3-Lamp among all sets, due to the learning data that fed the CNN (same set, yet with disjoint instances). Based on these results, $\alpha = 96.35$ was used in all our subsequent experiments.

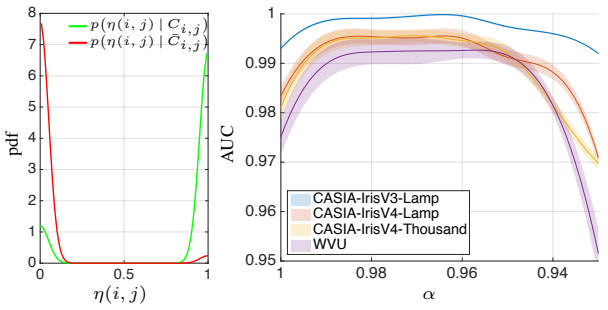


Figure 6. Left plot: decision environment of the responses given by the CNN to distinguish between *corresponding* $C_{i,j}$ and *non-corresponding* $\bar{C}_{i,j}$ iris patches (CASIA-IrisV3-Lamp set). Right plot: variations in recognition performance with respect to the α parameter.

3.3. Accurately Segmented Data

Performance was evaluated in two different settings: at first we used the accurate parameterisations of the iris boundaries, to perceive IRINA's performance in relatively good quality data. Results are given in Fig. 7, comparing the ROC curves (in linear and log scales) for the four methods and four data sets considered. It can be seen that IRINA outperformed its competitors in all cases and regions of the performance space, with exception to a narrow band around FAR $\approx 10^{-3}$ in the CASIA-V4-Thousand. In the remaining cases, IRINA was considerably better than the other methods, at some operating points with reductions in FAR levels over 40% with respect to the second best approach (usually Yang *et al.*). At the other extreme, the method due to Belcher and Du got consistently the worst results in our experiments, due to the difficulties in finding exact keypoint correspondences between images with different levels of focus or pupillary dilation. Overall, IRINA's best performance among all methods is particularly evident in the CASIA-IrisV3-Lamp, where the decreases in the error rates (over the second-best strategy) almost reached one order of magnitude.

The most relevant performance indicators are summarised in Table 1. It should be noted that results reported here should not be directly compared to the last generation of iris recognition evaluation initiatives (International Biometric Group evaluation FRR 2-5%@ FAR $\approx$ 1e^{-6} and Iris Challenge Evaluation FRR 1-3% @ FAR $\approx$ 1e^{-3}), as the average quality of data here is substantially lower than in those contests. Even though, in order to provide easy baselines, IRINA obtained FRR levels at FAR $\approx [1e^{-1}, 1e^{-2}, 1e^{-3}, 1e^{-4}]$ of [0.001, 0.006, 0.021, 0.121] (CASIA-IrisV3-Lamp), [0.012, 0.039, 0.076, 0.084] (CASIA-IrisV4-Lamp), [0.011, 0.054, 0.140, 0.156] (CASIA-Iris-V4-Thousand) and [0.023, 0.080, 0.116, 0.121] (WVU).

Method	AUC	d'	EER
CASIA-IrisV3-Lamp			
IRINA	0.999 $\pm 1e^{-4}$	12.623 $\pm$ 0.716	0.006 $\pm$ 0.001
Yang *et al.*	0.995 $\pm 4e^{-4}$	4.085 $\pm$ 0.590	0.021 $\pm$ 0.004
Sun and Tan	0.989 $\pm 5e^{-4}$	3.239 $\pm$ 0.501	0.044 $\pm$ 0.004
Belcher and Du	0.930 $\pm$ 0.005	2.701 $\pm$ 0.799	0.083 $\pm$ 0.009
CASIA-IrisV4-Lamp			
IRINA	0.995 $\pm$ 0.002	6.623 $\pm$ 0.454	0.026 $\pm$ 0.005
Yang *et al.*	0.993 $\pm 5e^{-4}$	3.629 $\pm$ 0.385	0.028 $\pm$ 0.004
Sun and Tan	0.992 $\pm 4e^{-4}$	3.448 $\pm$ 0.404	0.029 $\pm$ 0.005
Belcher and Du	0.948 $\pm$ 0.007	2.933 $\pm$ 0.696	0.077 $\pm$ 0.011
CASIA-IrisV4-Thousand			
IRINA	0.996 $\pm$ 0.001	6.179 $\pm$ 0.380	0.030 $\pm$ 0.005
Yang *et al.*	0.988 $\pm 6e^{-4}$	2.995 $\pm$ 0.366	0.045 $\pm$ 0.004
Sun and Tan	0.984 $\pm 6e^{-4}$	3.097 $\pm$ 0.583	0.052 $\pm$ 0.006
Belcher and Du	0.901 $\pm$ 0.009	2.104 $\pm$ 0.597	0.097 $\pm$ 0.012
WVU			
IRINA	0.991 $\pm$ 0.002	5.179 $\pm$ 0.361	0.042 $\pm$ 0.008
Yang *et al.*	0.980 $\pm$ 0.001	2.552 $\pm$ 0.185	0.065 $\pm$ 0.008
Sun and Tan	0.967 $\pm$ 0.001	2.210 $\pm$ 0.193	0.098 $\pm$ 0.007
Belcher and Du	0.882 $\pm$ 0.011	2.008 $\pm$ 0.780	0.116 $\pm$ 0.015

Table 1. Comparison between the performance obtained by IRINA with respect to three other strategies.

3.4. Inaccurately Segmented Data

At a second stage, we added two type of errors (translation and scale) to the iris boundaries parameterisations, to perceive the decreases in performance when the iris is inaccurately segmented. Segmentation errors of magnitude up to 21% were randomly generated, with "*magnitude*" expressing the difference between the maximum Euclidean distance between boundary points in the original and in the inaccurate segmentation parameterisation (e.g., for a circular boundary with diameter of 100 pixels, a scale error of magnitude 10% will either change the diameter to 90 or 110 pixels, whereas a translation error will move the boundary 10 pixels in a random direction).

According to our observations, the inaccurate segmentation setting is exactly when the advantages of IRINA with respect to the state-of-the-art are the most evident. The key insight IRINA' s robustness to segmentation failures is illustrated in Fig. 8, showing the deformation fields for genuine image pairwise comparisons, with accurate (green boundaries) and inaccurate (red boundaries) segmentations in the a)-c) rows, and one impostor comparison (bottom row) from the CASIA-IrisV4-Thousand set. Note that the impostor deformation field is almost chaotic, with much larger local derivatives than any genuine deformation field, where local correlation is evident.

The average decreases in performance with respect to segmentation inaccuracies up to 21% are given in Fig. 9 (mean AUC values, with 95% confidence intervals). It can be seen that IRINA almost kept its performance up to segmentation inaccuracies of 12%, and then slightly decreased its results, which could even be attenuated if larger magnitudes in the deformation field m_{max} were tolerated (yet, this would have increased the number of labels in the MRF and the computational cost). In opposition, both Yang *et al.* and Sun and Tan showed substantial decreases in performance even for relatively small segmentation errors, and almost loose any efficiency for errors larger than 15%. Finally, as it is not phase-based, the method due to Belcher and Du proved to be relatively robust against segmentation inaccuracies, but at much lower performance levels than IRINA.

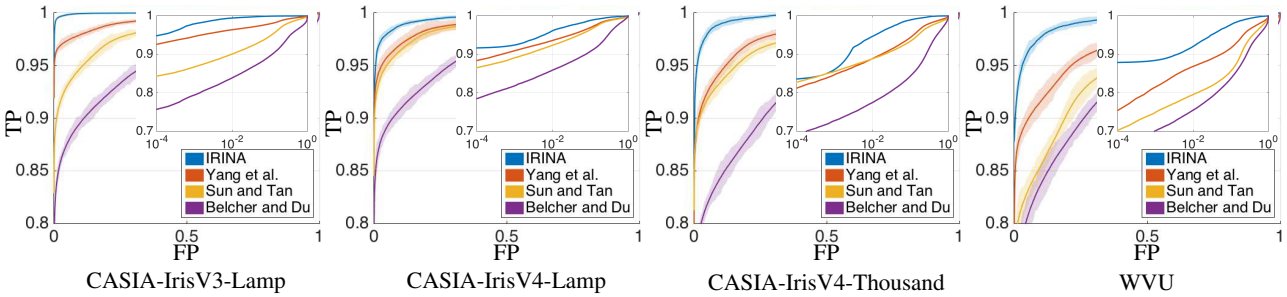


Figure 7. Comparison between the ROC curves obtained for the three methods and four datasets considered. At each operating point, the confidence interval is denoted by the shade region.

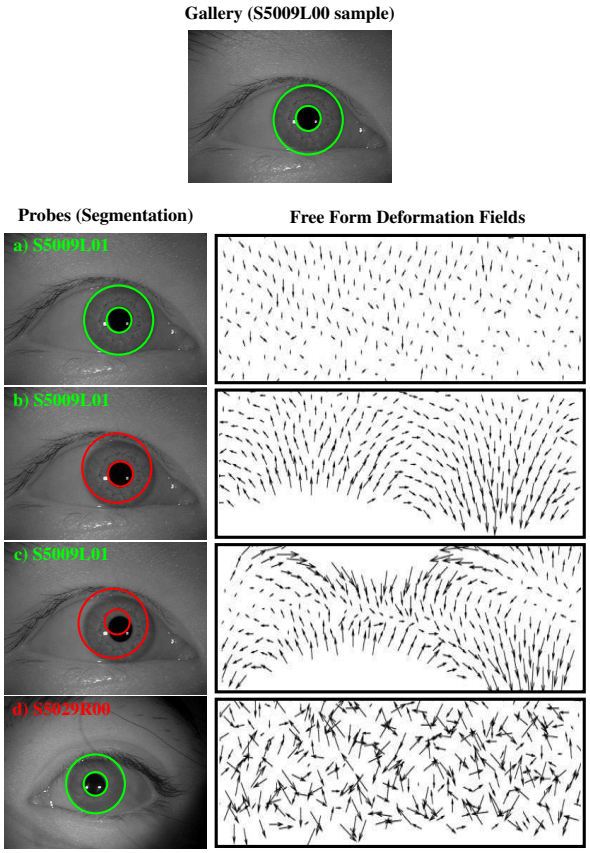


Figure 8. Examples of deformation fields with respect to failures in the segmentation of the iris. a) genuine comparison using an accurately segmented image; b) and c) genuine pairwise comparisons in inaccurately segmented data; d) impostor pairwise comparison. For illustration purposes, circular boundaries (3 dof) are used, as they provide the most evident patterns in the deformation fields.

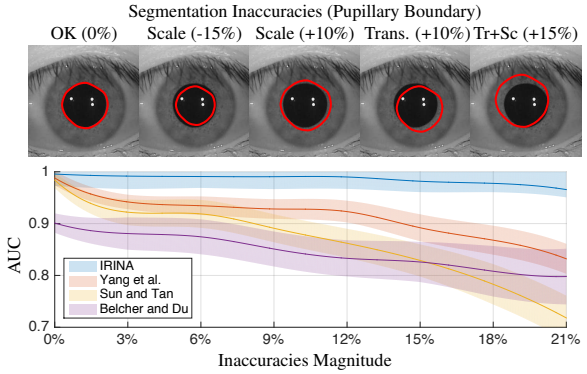


Figure 9. Top row: examples of segmentation inaccuracies and their corresponding magnitudes. Bottom row: AUC values with respect to the magnitude of segmentation inaccuracies (CASIA-IrisV4-Thousand set).

4. Conclusions and Further Work

Iris recognition has limited robustness against failures in segmentation and dilated / constricted pupils. In this paper we proposed an algorithm (IRINA) to cope with such covariates. The idea is to learn the "*corresponding*" patches between pairs of iris samples non-linearly deformed due to segmentation failures and to pupillary dilation. A convolutional neural network is used for such purposes, which provides the information for an image registration step that matches patches of the query iris sample into the enrolled data. A Markov random field infers a free form deformation field (set of 2D vectors), which first and second order statistics provide the discriminating information for biometric recognition. Our experiments show that IRINA not only achieves state-of-the-art performance in good quality data, but also effectively handles severe segmentation errors and large differences in pupillary dilation / constriction.

As current work, we are concentrated in finding alternate strategies to obtain the 2D deformation fields and reduce the computational cost of matching.

Acknowledgements

This work was supported by UID/EEA/50008/2013 research program.

References

[1] C. Belcher and Y. Du. Region-based SIFT approach to iris recognition. *Opt Lasers Engeneering*, vol. 47, no. 1, pag. 139–147, 2009. 2, 6

[2] K. Bowyer, K. Hollingsworth and P. Flynn. A Survey of Iris Biometrics Research: 2008-2010. In M. J. Burge and K. W. Bowyer (eds), *Handbook of Iris Recognition, Advances in Computer Vision and Pattern Recognition*, pag. 15–54, Springer, 2013. 1

[3] J. Chen, F. Shen, D. Chen and P. Flynn. Iris Recognition Based on Human-Interpretable Features. *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 7, pag. 1476–1485, 2016. 2

[4] A. Clark, S. Kulp, I. Herron and A. Ross. A theoretical model for describing iris dynamics. In M. J. Burge and K. W. Bowyer (eds), *Handbook of Iris Recognition, Advances in Computer Vision and Pattern Recognition*, pag. 129–150. Springer, 2013. 2, 3

[5] J. 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, pag. 1148–1161, 1993. 1

[6] J. Daugman. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pag. 21–30, 2004. 3, 6

[7] W. Dong, Z. Sun and T. Tan. Iris matching based on personalized weight map *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pag. 1744–1757, 2011. 2

[8] P. Felzenszwalb and D. Huttenlocher. Efficient Belief Propagation for Early Vision. *International Journal of Computer Vision*, vol. 70, no. 1, pag. 41–54, 2006. 1, 5

[9] F. Alonso-Fernandez, R. Farrugia and J. Bigun. Eigen-Patch Iris Super-Resolution for Iris Recognition Improvement. In proceedings of the *23^{rd} European Signal Processing Conference*, doi: 10.1109/EUSIPCO.2015.7362348, 2015. 2

[10] A. Gangwar and A. Joshi. DeepIrisNet: Deep Iris Representation With Applications in Iris Recognition and Cross-Sensor Iris Recognition. In proceedings of the *International Conference on Image Processing*, pag. 2301–2015, 2016. 3

[11] B. Glocker, N. Komodakis, N. Navab, G. Tziritas and N. Paragios. Dense Registration with Deformation Priors. *Information Processing in Medical Imaging*, vol. 21, pag. 540–551, 2009. 4, 5

[12] Y. Hu, K. Sirlantzis and G. Howells. Exploiting stable and discriminative iris weight map for iris recognition under less constrained environment. In proceedings of the *IEEE International Conference on Biometrics Theory, Applications and Systems*, pag. 1–8, 2015. 2

[13] S-H. Hsieh, Y-H. Li, C-H. Tien and C-C. Chang. Extending the Capture Volume of an Iris Recognition System Using Wavefront Coding and Super-Resolution. *IEEE Transactions on Cybernetics*, doi: 10.1109/TCYB.2015.2504388, 2016. 2

[14] B. Vijaya Kumar, J. Thornton, M. Savvides, V. Boddeti and J. Smereka. Application of Correlation Filters for Iris Recognition. *Handbook of Iris Recognition* , pag. 337–354, 2013. 2

[15] N. Liu, H. Li, M. Zhang, J. Liu, Z. Sun and T. Tan. Accurate Iris Segmentation in Non-cooperative Environments Using Fully Convolutional Networks. In proceedings of the *International Conference on Biometrics*, pag. 1–8, 2016. 2

[16] N. Mahadeo, A. Paplinski and S. Ray. Optimization of Iris Codes for Improved Recognition. In proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pag. 48–55, 2014. 2

[17] J. Matey, O. Naroditsky, K. Hanna, R. Kolczynski, D. Lolanoco, S. Mangru, M. Tinker, T. Zappia and W. Zhao. Iris on the move: Acquisition of images for iris recognition in less constrained environments. *Proceedings of the IEEE*, vol. 94, no. 11, pag. 1936–1947, 2006. 2

[18] D. Menotti, G. Chiachia, W. Schwartz, H. Pedrini, A. Falcão and A. Rocha. Deep Representations for Iris, Face, and Fingerprint Spoofing Detection. *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pag. 864–879, 2015. 3

[19] J. Pillai, V. Patel, R. Chellappa and N. Ratha. Secure and robust iris recognition using random projections and sparse representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pag. 1877–1893, 2011. 2

[20] 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, pag. 73–85, 2014. 2

[21] A. Ross, R. Jillela, J. Smereka, V. N. Boddeti, B. V. K. VijayaKumar, R. Barnard, X. Hu, P. Pauca and R. Plemmons. Matching Highly Non-ideal Ocular Images: An Information Fusion Approach. In proceedings of the *5th IAPR International Conference on Biometrics*, doi: 10.1109/ICB.2012.6199791, 2012. 2

[22] T. Sederberg and S. Parry. A Free-form deformation of solid geometric models. In proceedings of the *13^{th} annual conference on Computer graphics and iterative techniques*, pag. 151–160, 1986. 4

[23] S. Shah and A. Ross. Iris Segmentation Using Geodesic Active Contours. *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pag. 824–836, 2009. 6

[24] J. Smereka, V. Boddeti and B. V. K. Vijaya Kumar: Probabilistic Deformation Models for Challenging Periocular Image Verification. *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 9, pag. 1875–1890, 2015. 2

[25] A. Sotiras, C. Davatzikos and N. Paragios. Deformable Medical Image Registration: A Survey. *IEEE Transactions on Medical Imaging*, vol. 32, issue 7, pag. 1153–1190, 2013. 3

[26] Z. Sun and T. Tan. Ordinal Measures for Iris Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pag. 221–2226, 2009. 6

[27] C-W. Tan and A. Kumar. Accurate Iris Recognition at a Distance Using Stabilized Iris Encoding and Zernike Moments Phase Features. *IEEE Transactions on Image Processing*, vol. 23, no. 9, pag. 3962–3974, 2014. 2

[28] J. Thornton, M. Savvides and B. Vijaya Kumar. A Bayesian approach to deformed pattern matching of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pag. 596–606, 2007. 2

[29] I. Tomeo-Reyes, A. Ross, D. Antwan and C. Vinod. A biomechanical approach to iris normalization. In proceedings of the *International Conference on Biometrics*, pag. 1–9, 2015. 2, 3

[30] Z. Wei, T. Tan and Z. Sun. Nonlinear iris deformation correction based on Gaussian model. In proceedings of the *Advances in Biometric Person Authentication, Lecture Notes in Computer Science*, vol. 4642, pag. 780–789, 2007. 3

[31] H. Wyatt. A minimum wear-and-tear meshwork for the iris. *Vision Research*, vol. 40, pag. 2167–2176, 2000. 3

[32] G. Yang, H. Zeng, P. Li and L. Zhang. High-Order Information for Robust Iris Recognition Under Less Controlled Conditions. In proceedings of the *International Conference on Image Processing*, pag. 4535–4539, 2015. 2, 6

[33] X. Yuan and P. Shi. A non-linear normalization model for iris recognition. In *Advances in Biometric Person Authentication, Lecture Notes in Computer Science*, vol. 3781, pag. 135–141, 2005. 2, 3

[34] D. Zhao, W. Luo, R. Liu and L. Yue. Negative Iris Recognition. *IEEE Transactions on Dependable and Secure Computing*, doi: 10.1109/TDSC.2015.2507133, 2016. 2

Bibliography

[1] M. McCahill and C. Norris, "Cctv in britain," *Center for Criminology and Criminal Justice-University of Hull-United Kingdom*, pp. 1–70, 2002.

[2] J. Matey, O. Naroditsky, K. Hanna, R. Kolczynski, D. LoIacono, S. Mangru, M. Tinker, T. Zappia, and W.-Y. Zhao, "Iris on the move: Acquisition of images for iris recognition in less constrained environments," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1936–1947, 2006.

[3] K. Hanna, P. Burt, S. Peleg, D. Dixon, D. Mishra, L. Wixson, R. Mandlebaum, P. Coyle, and J. Herman, "Fully automated iris recognition system utilizing wide and narrow fields of view," 2004, uS Patent 6,714,665.

[4] NeuroTechnology, "Verilook surveillance," http://www.neurotechnology.com/verilook-surveillance.html, accessed: 2015-10-03.

[5] A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross, "Biometrics: A grand challenge," in *Proceedings of the International Conference on Pattern Recognition*, 2004, pp. 935–942.

[6] J. C. Neves, J. C. Moreno, S. Barra, and H. Proença, "A calibration algorithm for multi-camera visual surveillance systems based on single-view metrology," in *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*, 2015, pp. 552–559.

[7] ——, "Acquiring high-resolution face images in outdoor environments: A master-slave calibration algorithm," in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2015, pp. 1–8.

[8] J. C. Neves, J. Moreno, and H. Proença, "A master-slave calibration algorithm with fish-eye correction," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[9] J. C. Neves, K. Wysoczanska, and H. Proença, "Evaluation of background subtraction algorithms for human visual surveillance," in *Proceedings of the International Conference on Signal and Image Processing Applications*, 2015.

[10] J. C. Neves and H. Proença, "Dynamic camera scheduling for visual surveillance in crowded scenes using markov random fields," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2015, pp. 1–8.

[11] J. C. Neves and H. Proenca, "ICB-RW 2016: International challenge on biometric recognition in the wild," in *Proceedings of the International Conference on Biometrics*, 2016, pp. 1–6.

[12] J. C. Neves, J. C. Moreno, and H. Proença, "QUIS-CAMPI: An annotated multi-biometrics data feed from surveillance scenarios," 2016, submitted to IET Biometrics.

[13] J. C. Neves and H. Proença, "'A leopard cannot change its spots': Improving face recognition using 3D-based caricatures," 2017, submitted to IEEE Transactions on Information Forensics and Security.

[14] ——, "Exploiting data redundancy for error detection in degraded biometric signatures resulting from in the wild environments," in *Proceedings of the International Workshop*

on Biometrics in the Wild, IEEE Conference on Automatic Face and Gesture Recognition, 2017, pp. 981–986.

[15] J. C. Neves, F. Narducci, S. Barra, and H. Proença, "Biometric recognition in surveillance scenarios: A survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 1–27, 2016.

[16] J. C. Neves, J. C. Moreno, S. Barra, F. Narducci, and H. Proença, "Unconstrained data acquisition frameworks and protocols," in *Human Recognition in Unconstrained Environments*, M. D. Marsico, M. Nappi, and H. Proença, Eds. Academic Press, 2017, pp. 1 – 30.

[17] J. C. Neves, G. Santos, S. Filipe, E. Grancho, S. Barra, F. Narducci, and H. Proença, "Quiscampi: Extending in the wild biometric recognition to surveillance environments," in *Proceedings of the International Conference on Image Analysis and Processing Workshops*, 2015, pp. 59–68.

[18] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1198–1211, 2008.

[19] P. KaewTrakulPong and R. Bowden, "A real time adaptive visual surveillance system for tracking low-resolution colour targets in dynamically changing scenes," *Image and Vision Computing*, vol. 21, no. 10, pp. 913 – 929, 2003.

[20] R. Jain and H.-H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 206–214, 1979.

[21] N. McFarlane and C. Schofield, "Segmentation and tracking of piglets in images," *Machine Vision and Applications*, vol. 8, no. 3, pp. 187–193, 1995.

[22] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, 2003.

[23] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[24] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 246–252.

[25] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the International Conference on Pattern Recognition*, 2004, pp. 28–31.

[26] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172 – 185, 2005.

[27] D. Butler, S. Sridharan, and V. Bove, "Real-time adaptive background segmentation," in *Proceedings of the International Conference on Multimedia and Expo*, 2003, pp. 341–344.

[28] J. Wu, J. Xia, J. Chen, and Z. Cui, "Adaptive detection of moving vehicle based on on-line clustering," *Journal of Computers*, vol. 6, no. 10, 2011.

[29] D. E. Butler, V. M. Bove, and S. Sridharan, "Real-time adaptive foreground/background segmentation," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 14, pp. 841–926, 2005.

[30] L. Maddalena and A. Petrosino, "The 3dSOBS+ algorithm for moving object detection," *Computer Vision and Image Understanding*, vol. 122, pp. 65 – 73, 2014.

[31] R. Luque, E. Domínguez, E. Palomo, and J. Muñoz, "A neural network approach for video object segmentation in traffic surveillance," in *Proceedings of the International Conference on Image Analysis and Recognition*, 2008, pp. 151–158.

[32] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 3, pp. 185 – 203, 1981.

[33] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[34] A. Talukder and L. Matthies, "Real-time detection of moving objects from moving vehicles using dense stereo and optical flow," in *Proceedings of the International Conference on Intelligent Robots and Systems*, 2004, pp. 3718–3725.

[35] J. Yao and J.-M. Odobez, "Fast human detection from joint appearance and foreground feature subset covariances," *Computer Vision and Image Understanding*, vol. 115, no. 10, pp. 1414 – 1426, 2011.

[36] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[37] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proceedings of the International Conference on Computer Vision*, 2003, pp. 734–741.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[39] D. Moctezuma, C. Conde, I. de Diego, and E. Cabello, "Person detection in surveillance environment with hogg: Gabor filters and histogram of oriented gradient," in *Proceedings of the International Conference on Computer Vision Workshops*, 2011, pp. 1793–1800.

[40] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis," in *Proceedings of the International Conference on Computer Vision*, 2009, pp. 24–31.

[41] 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, 1996.

[42] L. Zhang, S. Li, X. Yuan, and S. Xiang, "Real-time object classification in video surveillance based on appearance learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[43] X. Wang, T. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of the International Conference on Computer Vision*, 2009, pp. 32–39.

[44] Y. Gurwicz, R. Yehezkel, and B. Lachover, "Multiclass object classification for real-time video surveillance systems," *Pattern Recognition Letters*, vol. 32, no. 6, pp. 805 – 815, 2011.

[45] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proceedings of the European Conference on Computer Vision*, 2004, pp. 69–82.

[46] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 31, no. 6, pp. 645–654, 2001.

[47] V. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *Proceedings of the International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 470–475.

[48] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1208–1221, 2004.

[49] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[50] ——, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *International Journal of Computer Vision*, vol. 82, no. 2, pp. 185–204, 2009.

[51] N. Sprague and J. Luo, "Clothed people detection in still images," in *Proceedings of the International Conference on Pattern Recognition*, 2002, pp. 585–589.

[52] S. Harasse, L. Bonnaud, and M. Desvignes, "Human model for people detection in dynamic scenes," in *Proceedings of the International Conference on Pattern Recognition*, 2006, pp. 335–354.

[53] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.

[54] Q. Zhou and J. Aggarwal, "Object tracking in an outdoor environment using fusion of features and cameras," *Image and Vision Computing*, vol. 24, no. 11, pp. 1244 – 1255, 2006.

[55] Y. Wu and T. Huang, "A co-inference approach to robust visual tracking," in *Proceedings of the International Conference on Computer Vision*, 2001, pp. 26–33.

[56] S. Zhou, V. Krueger, and R. Chellappa, “Probabilistic recognition of human faces from video,” *Computer Vision and Image Understanding*, vol. 91, no. 1–2, pp. 214 – 245, 2003.

[57] J. Shi and C. Tomasi, “Good features to track,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[58] B. Babenko, M.-H. Yang, and S. Belongie, “Robust object tracking with online multiple instance learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[59] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof, “Prost: Parallel robust online simple tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 723–730.

[60] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, “Tracking the invisible: Learning where the object might be,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1285–1292.

[61] D. Comaniciu, V. Ramesh, and P. Meer, “Kernel-based object tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[62] E. Maggio and A. Cavallaro, “Multi-part target representation for color tracking,” in *IEEE International Conference on Image Processing*, 2005, pp. 729–732.

[63] H. Stern and B. Efros, “Adaptive color space switching for tracking under varying illumination,” *Image and Vision Computing*, vol. 23, no. 3, pp. 353 – 364, 2005.

[64] S. J. McKenna, Y. Raja, and S. Gong, “Tracking colour objects using adaptive mixture models,” *Image and Vision Computing*, vol. 17, no. 3–4, pp. 225 – 231, 1999.

[65] S. Hare, A. Saffari, and P. H. S. Torr, “Struck: Structured output tracking with kernels,” in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 263–270.

[66] Z. Kalal, K. Mikolajczyk, and J. Matas, “Tracking-learning-detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[67] T. B. Dinh, N. Vo, and G. Medioni, “Context tracker: Exploring supporters and distracters in unconstrained environments,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1177–1184.

[68] J. Supancic and D. Ramanan, “Self-paced learning for long-term tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2379–2386.

[69] X. Zhang, W. Hu, H. Bao, and S. Maybank, “Robust head tracking based on multiple cues fusion in the kernel-bayesian framework,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1197–1208, 2013.

[70] K. Okuma, A. Taleghani, N. Freitas, J. J. Little, and D. G. Lowe, “A boosted particle filter: Multitarget detection and tracking,” in *Proceedings of the European Conference on Computer Vision*, 2004, pp. 28–39.

[71] W. Zhong, H. Lu, and M.-H. Yang, “Robust object tracking via sparsity-based collaborative model,” in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1838–1845.

[72] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2042–2049.

[73] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2259–2272, 2011.

[74] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1822–1829.

[75] D. Huttenlocher, J. Noh, and W. Rucklidge, "Tracking non-rigid objects in complex scenes," in *Proceedings of the International Conference on Computer Vision*, 1993, pp. 93–101.

[76] Y. Wu and T. Yu, "A field model for human detection and tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 753–765, 2006.

[77] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[78] Z. Liu, H. Shen, G. Feng, and D. Hu, "Tracking objects using shape context matching," *Neurocomputing*, vol. 83, pp. 47 – 55, 2012.

[79] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME – Journal of Basic Engineering*, no. 82, pp. 35–45, 1960.

[80] M. W. Szeto and D. C. Gazis, "Application of kalman filtering to the surveillance and control of traffic systems." *Transportation Science*, vol. 6, no. 4, p. 419, 1972.

[81] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[82] A. Mittal and L. S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *International Journal of Computer Vision*, vol. 51, no. 3, pp. 189–203, 2003.

[83] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[84] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 3, pp. 125–141, 2008.

[85] Y. Wu, H. Ling, J. Yu, F. Li, X. Mei, and E. Cheng, "Blurred target tracking by blur-driven tracker," in *Proceedings of the IEEE Conference on Computer Vision*, 2011, pp. 1100–1107.

[86] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 864–877.

[87] J. Kwon and K.-M. Lee, "Visual tracking decomposition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1269–1276.

[88] J. Xiao, R. Stolkin, and A. Leonardis, “An enhanced adaptive coupled-layer lgtracker++,” in *Proceedings of the International Conference on Computer Vision Workshops*, 2013, pp. 137–144.

[89] W. Hu, X. Zhou, M. Hu, and S. Maybank, “Occlusion reasoning for tracking multiple people,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 114–121, 2009.

[90] Y. Cheng, “Mean shift, mode seeking, and clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[91] J. Kang, I. Cohen, and G. Medioni, “Continuous tracking within and across camera streams,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 267–272.

[92] Q. Zhao and H. Tao, “A motion observable representation using color correlogram and its applications to tracking,” *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 273 – 290, 2009.

[93] J. Huang, S. Kumar, M. Mitra, and W.-J. Zhu, “Spatial color indexing and applications,” in *International Conference on Computer Vision*, 1998, pp. 602–607.

[94] L. Sevilla-Lara, “Distribution fields for tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1910–1917.

[95] M. Felsberg, “Enhanced distribution field tracking using channel representations,” in *Proceedings of the International Conference on Computer Vision Workshops*, 2013, pp. 121–128.

[96] C.-T. Chu, J.-N. Hwang, H.-I. Pai, and K.-M. Lan, “Tracking human under occlusion based on adaptive multiple kernels with projected gradients,” *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1602–1615, 2013.

[97] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, “Parametric correspondence and chamfer matching: Two new techniques for image matching,” in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1977, pp. 659–663.

[98] C. Liu, C. Hu, and J. Aggarwal, “Eigenshape kernel based mean shift for human tracking,” in *Proceedings of the International Conference on Computer Vision Workshops*, 2011, pp. 1809–1816.

[99] D. Huttenlocher, G. Klanderman, and W. Rucklidge, “Comparing images using the hausdorff distance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[100] D. Gavrila and V. Philomin, “Real-time object detection for ldquo;smart rdquo; vehicles,” in *Proceedings of the International Conference on Computer Vision*, 1999, pp. 87–93.

[101] D. Gavrila, “Multi-feature hierarchical template matching using distance transforms,” in *Proceedings of the International Conference on Pattern Recognition*, 1998, pp. 439–444.

[102] ——, “A bayesian, exemplar-based approach to hierarchical shape matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1408–1421, 2007.

[103] S. Munder, C. Schnorr, and D. Gavrila, "Pedestrian detection and tracking using a mixture of view-based shape-texture models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 2, pp. 333–343, 2008.

[104] E. Saber, Y. Xu, and A. M. Tekalp, "Partial shape recognition by sub-matrix matching for partial matching guided image labeling," *Pattern Recognition*, vol. 38, no. 10, pp. 1560 – 1573, 2005.

[105] M. Husain, E. Saber, V. Misic, and S. Joralemon, "Dynamic object tracking by partial shape matching for video surveillance applications," in *Proceedings of the IEEE International Conference on Image Processing*, 2006, pp. 2405–2408.

[106] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proceedings of the British Machine Vision Conference*, 2006, pp. 1–10.

[107] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proceedings of the European Conference on Computer Vision*, 2008, pp. 234–247.

[108] P. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Proceedings of the Advances in Neural Information*, 2005, pp. 1417–1426.

[109] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[110] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE Journal of Oceanic Engineering*, vol. 8, no. 3, pp. 173–184, 1983.

[111] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.

[112] Y. Cai, N. de Freitas, and J. J. Little, "Robust visual tracking for multiple targets," in *Proceedings of the European Conference on Computer Vision*, 2006, pp. 107–118.

[113] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Proceedings of the International Conference on Computer Vision*, 2009, pp. 1515–1522.

[114] B. Leibe, K. Schindler, and L. Van Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *Proceedings of the International Conference on Computer Vision*, 2007, pp. 1–8.

[115] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[116] A. Andriyenko and K. Schindler, "Globally optimal multi-target tracking on a hexagonal lattice," in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 466–479.

[117] H. Jiang, S. Fels, and J. Little, "A linear programming approach for multiple object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[118] J. Berclaz, F. Fleuret, and P. Fua, "Multiple object tracking using flow linear programming," in *Proceedings of the International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009, pp. 1–8.

[119] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.

[120] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1265–1272.

[121] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

[122] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1926–1933.

[123] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3457–3464.

[124] P. Kumar, A. Dick, and T. S. Sheng, "Real time target tracking with pan tilt zoom camera," in *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*, 2009, pp. 492–497.

[125] P. Varcheie and G.-A. Bilodeau, "Active people tracking by a ptz camera in ip surveillance system," in *Proceedings of the IEEE International Workshop on Robotic and Sensors Environments*, 2009, pp. 93–108.

[126] Y. Yao, B. Abidi, and M. Abidi, "3D target scale estimation and target feature separation for size preserving tracking in PTZ video," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 244–263, 2009.

[127] P. D. Z. Varcheie and G.-A. Bilodeau, "Adaptive fuzzy particle filter tracker for a PTZ camera in an IP surveillance system," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 2, pp. 1952–1961, 2011.

[128] B. Tordoff and D. Murray, "Reactive control of zoom while fixating using perspective and affine cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 98–112, 2004.

[129] J. Zhou, D. Wan, and Y. Wu, "The chameleon-like vision system," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 91–101, 2010.

[130] H.-C. Liao and W.-Y. Chen, "Eagle-eye: A dual-PTZ-camera system for target tracking in a large open area," *Information technology and control*, vol. 39, no. 3, 2015.

[131] R. Bodor, R. Morlok, and N. Papanikolopoulos, "Dual-camera system for multi-level activity recognition," in *Proceedings of the International Conference on Intelligent Robots and Systems*, 2004, pp. 643–648.

[132] A. D. Bimbo, F. Dini., G. Lisanti, and F. Pernici, "Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 611–623, 2010.

[133] I. Everts, N. Sebe, and G. A. Jones, "Cooperative object tracking with multiple PTZ cameras," in *Proceedings of the International Conference on Image Analysis and Processing*, 2007, pp. 323–330.

[134] C.-H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi, "Heterogeneous fusion of omnidirectional and ptz cameras for multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1052–1063, 2008.

[135] M. Tarhan and E. Atlug, "A catadioptric and pan-tilt-zoom camera pair object tracking system for UAVs," *Journal of Intelligent & Robotic Systems*, vol. 61, no. 1, pp. 119–134, 2011.

[136] Y. Xu and D. Song, "Systems and algorithms for autonomous and scalable crowd surveillance using robotic ptz cameras assisted by a wide-angle camera," *Autonomus Robots*, vol. 29, no. 1, pp. 53–66, 2010.

[137] Y. Lu and S. Payandeh, "Cooperative hybrid multi-camera tracking for people surveillance," *Canadian Journal of Electrical and Computer Engineering*, vol. 33, no. 3/4, pp. 145–152, 2008.

[138] G. Scotti, L. Marcenaro, C. Coelho, F. Selvaggi, and C. Regazzoni, "Dual camera intelligent sensor for high definition 360 degrees surveillance," *IEE Proceedings of Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 250–257, 2005.

[139] N. Krahnstoever, T. Yu, S.-N. Lim, K. Patwardhan, and P. Tu, "Collaborative Real-Time Control of Active Cameras in Large Scale Surveillance Systems," in *Proceedings of the Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[140] C.-S. Yang, R.-H. Chen, C.-Y. Lee, and S.-J. Lin, "PTZ camera based position tracking in IP-surveillance system," in *Proceedings of the International Conference on Sensing Technology*, 2008, pp. 142–146.

[141] J. A. Fayman, O. Sudarsky, E. Rivlin, and M. Rudzsky, "Zoom tracking and its applications," *Machine Vision and Applications*, vol. 13, no. 1, pp. 25–37, 2001.

[142] Y. O. Kim, J. Paik, J. Heo, A. Koschan, B. Abidi, and M. Abidi, "Automatic face region tracking for highly accurate face recognition in unconstrained environments," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2003, pp. 29–36.

[143] H. Shah and D. Morrell, "An adaptive zoom algorithm for tracking targets using pan-tilt-zoom cameras," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 721–724.

[144] F. Wheeler, R. Weiss, and P. Tu, "Face recognition at a distance system for surveillance applications," in *Proceedings of the IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–8.

[145] X. Zhou, R. Collins, T. Kanade, and P. Metes, "A master-slave system to acquire biometric imagery of human at distance," in *Proceedings of the ACM International Workshop on Video Surveillance*, 2003, pp. 113–120.

[146] H. C. Liao and Y. C. Cho, "A new calibration method and its application for the cooperation of wide-angle and pan-tilt-zoom cameras," *Information Technology Journal*, vol. 7, no. 8, pp. 1096–1105, 2008.

[147] Y. Liu, S. Lai, C. Zuo, H. Shi, and M. Zhang, "A master-slave surveillance system to acquire panoramic and multiscale videos," *The Scientific World Journal*, 2014.

[148] L. You, S. Li, and W. Jia, "Automatic weak calibration of master-slave surveillance system based on mosaic images," in *Proceedings of the International Conference on Pattern Recognition*, 2010, pp. 1824–1827.

[149] U. Park, H.-C. Choi, A. Jain, and S.-W. Lee, "Face tracking and recognition at a distance: A coaxial and concentric PTZ camera system," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1665–1677, 2013.

[150] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 161–195, 1998.

[151] F. Wheeler, A. Perera, G. Abramovich, B. Yu, and P. Tu, "Stand-off iris recognition system," in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2008, pp. 1–7.

[152] W. Dong, Z. Sun, and T. Tan, "A design of iris recognition system at a distance," in *Proceedings of the Chinese Conference on Pattern Recognition*, 2009, pp. 1–5.

[153] S. Yoon, H. G. Jung, J. K. Suhr, and J. Kim, "Non-intrusive iris image capturing system using light stripe projection and pan-tilt-zoom camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–7.

[154] F. Bashir, P. Casaverde, D. Usher, and M. Friedman, "Eagle-eyes: a system for iris recognition at a distance," in *Proceedings of the IEEE Conference on Technologies for Homeland Security*, 2008, pp. 426–431.

[155] G. Guo, M. Jones, and P. Beardsley, "A system for automatic iris capturing," *MERL TR2005-044*, vol. 1, 2005.

[156] J.-H. Yoo and B. J. Kang, "A simply integrated dual-sensor based non-intrusive iris image acquisition system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 113–117.

[157] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle, "Face cataloger: multi-scale imaging for relating identity to location," in *Proceedings of the IEEE conference on Advance Video and Signal Based Surveillance*, 2003, pp. 13–20.

[158] S. Stillman, R. Tanawongsuwan, and I. Essa, "Tracking multiple people with multiple cameras," in *Proceedings of the International Conference on Audio and Video-based Biometric Person Authentication*, 1999.

[159] L. Marchesotti, S. Piva, A. Turolla, D. Minetti, and C. S. Regazzoni, "Cooperative multisensor system for real-time face detection and tracking in uncontrolled conditions," in *Proceedings SPIE, Image and Video Communications and Processing*, 2005.

[160] H. Choi, U. Park, and A. Jain, "Ptz camera assisted face acquisition, tracking & recognition," in *Proceedings of the IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–6.

[161] P. Amnuaykanjanasin, S. Aramvith, and T. H. Chalidabhongse, "Face tracking using two cooporative static and moving cameras," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2005, pp. 1158–1161.

[162] K. Bernardin, F. v. d. Camp, and R. Stiefelhagen, "Automatic person detection and tracking using fuzzy controlled active cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[163] A. Mian, "Realtime face detection and tracking using a single pan, tilt, zoom camera," in *Proceedings of the International Conference in Image and Vision Computing New Zealand*, 2008, pp. 1–6.

[164] S. Venugopalan and M. Savvides, "Unconstrained iris acquisition and recognition using COTS PTZ camera," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 38, pp. 1–20, 2010.

[165] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 1998, pp. 214–219.

[166] H. Faulds, "On the skin-furrows of the hand," *Nature*, vol. 22, no. 574, p. 605, 1880.

[167] W. W. Bledsoe, "The model method in facial recognition," Panoramic Research, Inc., Palo Alto, California, Palo Alto, California, Tech. Rep. PRI 15, 1964.

[168] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1991, pp. 586–591.

[169] P. Belhumeur, J. 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.

[170] 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.

[171] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[172] S. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 627–639, 2007.

[173] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[174] C. H. Chan, M. Tahir, J. Kittler, and M. Pietikäinen, "Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1164–1177, 2013.

[175] J. Heikkila, E. Rahtu, and V. Ojansivu, "Local phase quantization for blur insensitive texture description," in *Local Binary Patterns: New Variants and Applications*, 2014, pp. 49–84.

[176] V. Ojansivu, E. Rahtu, and J. Heikkila, "Rotation invariant local phase quantization for blur insensitive texture analysis," in *Proceedings of the International Conference on Pattern Recognition*, 2008, pp. 1–4.

[177] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[178] 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.

[179] B. He, D. Xu, R. Nian, M. van Heeswijk, Q. Yu, Y. Miche, and A. Lendasse, "Fast face recognition via sparse coding and extreme learning machine," *Cognitive Computation*, vol. 6, no. 2, pp. 264–277, 2014.

[180] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.

[181] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3499–3506.

[182] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[183] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild." in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 787–796.

[184] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification." in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[185] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition.” in *Proceedings of the British Machine Vision Conference*, 2015.

[186] I. Masi, A. Tran, T. Hassner, J. Leksut, and G. Medioni, “Do we really need to collect millions of faces for effective face recognition?” in *Proceedings of the European Conference on Computer Vision*, 2017, pp. 579–596.

[187] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[188] Y. Zhong, J. Chen, and B. Huang, “Toward end-to-end face recognition through alignment learning,” *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1213–1217, 2017.

[189] W. Gong, M. Sapienza, and F. Cuzzolin, “Fisher tensor decomposition for unconstrained gait recognition,” in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2013.

[190] A. Dantcheva, P. Elia, and A. Ross, “What else does your biometric data reveal? A survey on soft biometrics,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2015.

[191] D. Matovski, M. Nixon, S. Mahmoodi, and J. Carter, “The effect of time on gait recognition performance,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 543–552, 2012.

[192] Y. Ran, Q. Zhen, R. Chellapa, and T. M. Strat, “Applications of a simple characterization of human gait in surveillance,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 4, pp. 1009–1020, 2010.

[193] I. Venkat and P. D. Wilde, “Robust gait recognition by learning and exploiting sub-gait charcterisitics,” *International Journal of Computer Vision*, vol. 91, no. 1, pp. 7–23, 2010.

[194] K. Moustakas, D. Tzovaras, and G. Stavropoulos, “Gait recognition using geometric features and soft biometrics,” *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 367–370, 2010.

[195] S.-U. Jung and M. Nixon, “On using gait to enhance frontal face extraction,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1802–1811, 2012.

[196] S. D. Choudhury and T. Tjhajadi, “Gait recognition based on shape and motion analysis of silhouette contours,” *Computer Vision and Image Understanding*, vol. 117, no. 12, pp. 1770–1785, 2013.

[197] ——, “Robust view-invariant multiscale gait recognition,” *Pattern Recognition*, vol. 48, no. 3, pp. 798–811, 2015.

[198] W. Kusakunniran, “Recognize gaits on spatio-temporal feature domain,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 9, pp. 1416–1423, 2014.

[199] L. Maddalena and A. Petrosino, “A self-organizing approach to background subtraction for visual surveillance applications,” *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1168–1177, 2008.

[200] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[201] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1937–1944.

[202] "Vot2015 challenge," http://www.votchallenge.net/vot2015/, accessed: 2015-12-21.

[203] F. Lv, T. Zhao, and R. Nevatia, "Self-calibration of a camera from video of a walking human," in *Proceedings of the International Conference on Pattern Recognition*, 2002, pp. 562–567.

[204] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

[205] A. W. Senior, A. Hampapur, and M. Lu, "Acquiring multiscale images by pan-tilt-zoom control and automatic multicamera calibration," in *Proceedings of the IEEE Workshop on Application of Computer Vision*, 2005, pp. 433–438.

[206] Q. Li, Z. Sun, S. Chen, and Y. Liu, "A method of camera selection based on partially observable markov decision process model in camera networks," in *Proceedings of the American Control Conference*, 2013, pp. 3833–3839.

[207] N. Kariotoglou, D. Raimondo, S. Summers, and J. Lygeros, "A stochastic reachability framework for autonomous surveillance with pan-tilt-zoom cameras," in *Proceedings of the IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 1411–1416.

[208] S.-N. Lim, L. Davis, and A. Elgammal, "Scalable image-based multi-camera visual surveillance system," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2003, pp. 205–212.

[209] C. J. Costello, C. P. Diehl, A. Banerjee, and H. Fisher, "Scheduling an active camera to observe people," in *Proceedings of the ACM International Workshop on Video Surveillance and Sensor Networks*, 2004, pp. 39–45.

[210] A. D. Bimbo and F. Pernici, "Towards on-line saccade planning for high-resolution image sensing," *Pattern Recognition Letters*, vol. 27, pp. 1826–1834, 2006.

[211] F. Qureshi and D. Terzopoulos, "Planning ahead for ptz camera assignment and handoff," in *Proceedings of the International Conference on Distributed Smart Cameras*, 2009, pp. 1–8.

[212] F.-Z. Qureshi and D. Terzopoulos, "Surveillance camera scheduling: a virtual vision approach," *Multimedia Systems*, vol. 12, no. 3, pp. 269–283, 2006.

[213] S.-N. Lim, L. Davis, and A. Mittal, "Task scheduling in large camera networks," in *Proceedings of the Asian Conference on Computer Vision*, 2007, pp. 397–407.

[214] A. Ilie and G. Welch, "Online control of active camera networks for computer vision tasks," *ACM Transactions Sensor Networks*, vol. 10, no. 2, pp. 1–40, 2014.

[215] P. F. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 41–54, 2006.

[216] G. Antonini, M. Bierlaire, and M. Weber, "Discrete choice models of pedestrian walking behavior," *Transportation Research Part B: Methodological*, vol. 40, no. 8, pp. 667–687, 2006.

[217] T. Robin, G. Antonini, M. Bierlaire, and J. Cruz, "Specification, estimation and validation of a pedestrian walking behavior model," *Transportation Research Part B: Methodological*, vol. 43, no. 1, pp. 36–56, 2009.

[218] Y. Cai, G. Medioni, and T. B. Dinh, "Towards a practical ptz face detection and tracking system," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 2013, pp. 31–38.

[219] B. Klare, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a." in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1931–1939.

[220] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abeljaoued, and E. Mayoraz, "Comparison of face verification results on the xm2vtfs database," in *Proceedings of the International Conference on Pattern Recognition*, 2000, pp. 858–863.

[221] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostin, F. Cardinaux, S. Marcel, S. Bengio, C. Sanderson, N. Poh, Y. Rodriguez, J. Czyz, L. Vandendorpe, C. McCool, S. Lowther, S. Sridharan, V. Chandran, R. P. Palacios, E. Vidal, L. Bai, L. Shen, Y. Wang, C. Yueh-Hsuan, L. Hsien-Chang, H. Yi-Ping, A. Heinrichs, M. Muller, A. Tewes, C. von der Malsburg, R. Wurtz, Z. Wang, F. Xue, Y. Ma, Q. Yang, C. Fang, X. Ding, S. Lucey, R. Goss, and H. Schneiderman, "Face authentication test on the banca database," in *Proceedings of the International Conference on Pattern Recognition*, 2004, pp. 523–532.

[222] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 947–954.

[223] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 831–846, 2010.

[224] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "The good, the bad, and the ugly face challenge problem," *Image and Vision Computing*, vol. 30, no. 3, pp. 177–185, 2012.

[225] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan, and S. Weimer, *Overview of the Multiple Biometrics Grand Challenge*, 2009, pp. 705–714.

[226] M. Grgic, K. Delac, and S. Grgic, "SCface-surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 863–879, 2011.

[227] H. Maeng, S. Liao, D. Kang, S.-W. Lee, and A. K. Jain, "Nighttime face recognition at long distance: Cross-distance and cross-spectral matching," in *Proceedings of the Asian Conference on Computer Vision*, 2012, pp. 708–721.

[228] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 365–372.

[229] H. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets." in *Proceedings of the IEEE International Conference on Image Processing*, 2014, pp. 343–347.

[230] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 529–534.

[231] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops*, 2011, pp. 81–88.

[232] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng, "The challenge of face recognition from digital point-and-shoot cameras," in *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[233] F. S. Samaria, "Face recognition using hidden markov models," Ph.D. dissertation, University of Cambridge, 1994.

[234] A. Martinez and R. Benavente, "The ar face database," Tech. Rep. 24, 1998.

[235] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[236] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) database," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 46–51.

[237] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition*, 2008, pp. 1–8.

[238] H. Proença, S. Filipe, R. Santos, J. Oliveira, and L. Alexandre, "The UBIRIS.v2: A database of visible wavelength 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.

[239] C. Padole and H. Proença, "Compensating for pose and illumination in unconstrained periocular biometrics," *International Journal of Biometrics*, vol. 5, no. 3/4, pp. 336–359, 2013.

[240] F. Juefei-Xu and M. Savvides, "Unconstrained periocular biometric acquisition and recognition using COTS PTZ camera for uncooperative and non-cooperative subjects," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 2012, pp. 201–208.

[241] G. B. H. E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2014-003, 2014.

[242] T. Y. Wang and A. Kumar, "Recognizing human faces under disguise and makeup," in *Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis*, 2016, pp. 1–7.

[243] L. Li, T. Nawaz, and J. Ferryman, "PETS 2015: Datasets and challenge," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2015, pp. 1–6.

[244] T. Wang, S. Gong, X. Zhu, and S. Wang, "Proceedings of the person re-identification by video ranking," in *European Conference Computer Vision*, 2014, pp. 688–703.

[245] R. Fisher, "Caviar dataset," http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/, 2005.

[246] R. Vezzani and R. Cucchiara, "Video surveillance online repository (ViSOR): An integrated framework," *Multimedia Tools Applications*, vol. 50, no. 2, pp. 359–380, 2010.

[247] M. Hofmann, S. M. Schmidt, A. Rajagopalan, and G. Rigoll, "Combined face and gait recognition using alpha matte preprocessing," in *Proceedings of the International Conference on Biometrics*, 2012, pp. 1–8.

[248] D. Muramatsu, H. Iwama, Y. Makihara, and Y. Yagi, "Multi-view multi-modal person authentication from a single walking image sequence," in *Proceedings of the International Conference on Biometrics*, 2013, pp. 1–8.

[249] S. Sarkar, J. Phillips, Z. Liu, I. Vega, P. Grother, and K. Bowyer, "The HumanID gait challenge problem: Data sets, performance, and analysis," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[250] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1511–1521, 2012.

[251] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, 2014.

[252] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. Nixon, "Soft biometrics and their application in person recognition at a distance," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 464–475, 2014.

[253] C. Wu, S. Agarwal, B. Curless, and S. Seitz, "Multicore bundle adjustment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3057–3064.

[254] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild." in *Proceedings of the IEEE International Conference on Computer Vision*, 2012, pp. 2879–2886.

[255] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the International Conference on Machine Learning*, 2007, pp. 209–216.

[256] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 498–505.

[257] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[258] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher Vector Faces in the Wild," in *Proceedings of the British Machine Vision Conference*, 2013.

[259] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[260] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy – automatic naming of characters in TV video," in *Proceedings of the British Machine Vision Conference*, 2006.

[261] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligence Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[262] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[263] D. Cox and N. Pinto, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops*, 2011, pp. 8–15.

[264] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2707–2714.

[265] H. Proença and J. C. Neves, "Creating synthetic iriscodes to feed biometrics experiments," in *Proceedings of the IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*, 2013, pp. 8–12.

[266] H. S. G. Pussewalage, J. Hu, and J. Pieprzyk, "A survey: Error control methods used in bio-cryptography," in *Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery*, 2014, pp. 956–962.

[267] F. Hao, R. Anderson, and J. Daugman, "Combining crypto with biometrics effectively," *IEEE Transactions on Computers*, vol. 55, no. 9, pp. 1081–1088, 2006.

[268] S. Noto, P. L. Correia, and L. D. Soares, "Analysis of error correcting codes for the secure storage of biometric templates," in *Proceedings of the International Conference on Computer as a Tool*, 2011.

[269] F. J. MacWilliams and N. J. A. Sloane, *The theory of error correcting codes*. Elsevier, 1977.

[270] J. Bringer, H. Chabanne, G. Cohen, B. Kindarji, and G. Zemor, "Optimal iris fuzzy sketches," in *Proceedings of the International Conference on Biometrics: Theory, Applications, and Systems*, 2007, pp. 1–6.

[271] S. Kanade, D. Camara, E. Krichen, D. Petrovska-Delacretaz, and B. Dorizzi, “Three factor scheme for biometric-based cryptographic key regeneration using iris,” in *Proceedings of the Biometrics Symposium*, 2008, pp. 59–64.

[272] S. Kanade, D. Petrovska-Delacretaz, and B. Dorizzi, “Cancelable iris biometrics and using error correcting codes to reduce variability in biometric data,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 120–127.

[273] E. Maiorana, D. Blasi, and P. Campisi, “Biometric template protection using turbo codes and modulation constellations,” in *Proceedings of the International Workshop on Information Forensics and Security*, 2012, pp. 25–30.

[274] S. H. Moi, P. Saad, N. A. Rahim, and S. Ibrahim, “Error correction on iris biometric template using reed solomon codes,” in *Proceedings of the Asian International Conference on Mathematical/Analytical Modelling and Computer Simulation*, 2010, pp. 209–214.

[275] T. G. Dietterich and G. Bakiri, “Solving multiclass learning problems via error-correcting output codes,” *Journal of Artificial Intelligence Research*, vol. 2, no. 1, pp. 263–286, 1995.

[276] T. Windeatt and G. Ardeshir, “Boosted ecoc ensembles for face recognition,” in *Proceedings of the International Conference on Visual Information Engineering*, 2003, pp. 165–168.

[277] J. Kittler, R. Ghaderi, T. Windeatt, and J. Matas, “Face verification via error correcting output codes,” *Image and Vision Computing*, vol. 21, no. 13–14, pp. 1163 – 1169, 2003.

[278] L. Masek, “Recognition of human iris patterns for biometric identification,” Master’s thesis, The University of Western Australia, 2003.

[279] “Casia iris image database,” http://biometrics.idealtest.org/.

[280] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[281] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.

[282] A. Burton, S. Wilson, M. Cowan, and V. Bruce, “Face recognition in poor-quality video: Evidence from security surveillance.” *Psychological Science*, vol. 10, no. 3, pp. 243–248, 1999.

[283] J. Kaufmann and S. Schweinberger, “Distortions in the brain erp effects of caricaturing familiar and unfamiliar faces.” *Brain Research*, vol. 1228, pp. 177–188, 2008.

[284] M. Itz, S. Schweinberger, and J. Kaufmann, “Effects of caricaturing in shape or color on familiarity decisions for familiar and unfamiliar faces.” *PLOS ONE*, vol. 11, no. 2, pp. 1–19, 2016.

[285] G. Rhodes, S. Brennan, and S. Carey, “Identification and ratings of caricatures implications for mental representations of faces.” *Cognitive Psychology*, vol. 19, no. 4, pp. 473–497, 1987.

[286] K. Lee and D. Perrett, "Presentation-time measures of the effects of manipulations in colour space on discrimination of famous faces." *Perception*, vol. 26, no. 6, pp. 733–752, 1997.

[287] ——, "Manipulation of colour and shape information and its consequence upon recognition and best-likeness judgments." *Perception*, vol. 29, no. 11, pp. 1291–312, 2000.

[288] K. Lee, G. Byatt, and G. Rhodes, "Caricature effects, distinctiveness, and identification: Testing the face-space framework." *Psychological Science*, vol. 11, no. 5, pp. 379–385, 2000.

[289] G. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images." in *Proceedings of the International Conference on Computer Vision*, 2007, pp. 1–8.

[290] I. Shlizerman, S. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[291] V. Bruce, M. Burton, and N. Dench, "What's distinctive about a distinctive face?" *The Quarterly Journal of Experimental Psychology Section A*, vol. 47, no. 1, pp. 119–141, 1994.

[292] R. Mauro and M. Kubovy, "Caricature and face recognition." *Memory and Cognition*, vol. 20, no. 4, pp. 433–440, 1992.

[293] P. Benson and D. Perrett, "Visual processing of facial distinctiveness." *Perception*, vol. 23, pp. 75–93, 1994.

[294] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about." *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.

[295] S. Brennan, "The dynamic exaggeration of faces by computer," *Leonardo*, vol. 18, no. 3, pp. 170–178, 1985.

[296] P. Liao and T. Li, "Automatic caricature generation by analyzing facial features." in *Proceedings of the Asian Conference on Computer Vision*, 2004.

[297] Z. Mo, J. Lewis, and U. Neumann, "Improved automatic caricature by feature normalization and exaggeration." in *ACM SIGGRAPH Sketches*, 2004, p. 57.

[298] C. Tseng and J. Lien, "Synthesis of exaggerative caricature with inter and intra correlations." in *Proceedings of the Asian Conference on Computer Vision*, 2007, pp. 314–323.

[299] T. Lewiner, T. Vieira, D. Martinez, A. Peixoto, V. Mello, and L. Velho, "Interactive 3d caricature from harmonic exaggeration." *Computers and Graphics*, vol. 35, no. 3, pp. 586–595, 2011.

[300] L. Clarke, M. Chen, and B. Mora, "Automatic generation of 3d caricatures based on artistic deformation styles." *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 6, pp. 808–821, 2011.

[301] M. Sela, Y. Aflalo, and R. Kimmel, "Computational caricaturization of surfaces." *Computer Vision and Image Understanding*, vol. 141, pp. 1–17, 2015.

[302] P. Li, Y. Chen, J. Liu, and G. Fu, "3d caricature generation by manifold learning." in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2008, pp. 941–944.

[303] J. Liu, Y. Chen, C. Miao, J. Xie, C. Ling, X. Gao, and W. Gao, "Semi-supervised learning in reconstructed manifold space for 3d caricature generation." in *Computer Graphics Forum*, 2009, pp. 2104–2116.

[304] J. Xie, Y. Chen, J. Liu, C. Miao, and X. Gao, "Interactive 3d caricature generation based on double sampling." in *Proceedings of the ACM International Conference on Multimedia*, 2009, pp. 745–748.

[305] X. Han, C. Gao, and Y. Yu, "Deepsketch2face: A deep learning based sketching system for 3d face and caricature modeling." *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–12, 2017.

[306] S. B. Sadimon, M. S. Sunar, D. Mohamad, and H. Haron, "Computer generated caricature: A survey," in *Proceedings of the International Conference on Cyberworlds*, 2010, pp. 383–390.

[307] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space." in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 113–120.

[308] B. F. Klare, S. S. Bucak, A. K. Jain, and T. Akgul, "Towards automated caricature recognition." in *Proceedings of the International Conference on Biometrics*, 2012, pp. 139–146.

[309] B. Abaci and T. Akgul, "Matching caricatures to photographs." *Signal, Image and Video Processing*, vol. 9, no. 1, pp. 295–303, 2015.

[310] S. Ouyang, T. Hospedales, Y. Song, and X. Li, "Cross-modal face matching: beyond viewed sketches." in *Proceedings of the Asian Conference on Computer Vision*, 2014, pp. 210–225.

[311] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees." in *Proceedings of the IEEE International Conference on Computer Vision*, 2014, pp. 1867–1874.

[312] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks." *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[313] Z. Zhang, P. Luo, C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 918–930, 2016.

[314] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces." in *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, 1999, pp. 187–194.

[315] S. Romdhani and T. Vetter, "Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior." in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 986–993.

[316] G. Hu, P. Mortazavian, J. Kittler, and W. Christmas, "A facial symmetry prior for improved illumination fitting of 3d morphable model." in *Proceedings of the International Conference on Biometrics*, 2013, pp. 1–6.

[317] G. Hu, C.-H. Chan, J. Kittler, and B. Christmas, “Resolution-aware 3d morphable model.” in *Proceedings of the British Machine Vision Conference*, 2012, pp. 1–10.

[318] P. Mortazavian, J. Kittler, and W. Christmas, “3d morphable model fitting for low-resolution facial images.” in *Proceedings of the International Conference on Biometrics*, 2012, pp. 132–138.

[319] P. Huber, G. Hu, R. Tena, P. Mortazavian, W. P. Koppen, W. Christmas, M. Ratsch, and J. Kittler, “A multiresolution 3d morphable face model and fitting framework.” in *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[320] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, “A 3d face model for pose and illumination invariant face recognition,” in *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance*, 2009, pp. 1–8.

[321] P. Rautek, I. Viola, and M. E. Groller, “Caricaturistic visualization.” *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1085–1092, 2006.

[322] M. Roberts, “A unified account of the effects of caricaturing faces.” *Visual Cognition*, vol. 6, pp. 1–42, 1999.

[323] Y. Lipman, O. Sorkine, D. Cohen-Or, D. Levin, C. Rossl, and H. P. Seidel, “Differential coordinates for interactive mesh editing.” in *Proceedings of the International Conference on Shape Modeling and Applications*, 2004, pp. 181–190.

[324] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rossl, and H. P. Seidel, “Laplacian surface editing.” in *Proceedings of the Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, 2004, pp. 175–184.

[325] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, “Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization.” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2144–2151.

[326] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge.” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[327] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, “Localizing parts of faces using a consensus of exemplars.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[328] A. Tran, T. Hassner, I. Masi, and G. Medioni, “Regressing robust and discriminative 3d morphable models with a very deep neural network.” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[329] J. Klontz, B. Klare, S. Klum, E. Taborsky, M. Burge, and A. K. Jain, “Open source biometric recognition.” in *Proceedings of the IEEE Conference on Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[330] D. Wang, C. Otto, and A. K. Jain, “Face search at scale,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1122–1136, 2017.

[331] J. Chen, V. Patel, and R. Chellappa, “Unconstrained face verification using deep cnn features.” in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2016, pp. 1–8.