Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Doctor of Philosophy

TITLE Pose-Tolerant Face Recognition

PRESENTED BY Ramzi Abi Antoun

ACCEPTED BY THE DEPARTMENT OF

Electrical and Computer Engineering

ADVISOR, MAJOR PROFESSOR	DATE
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

DEAN	DATE

Pose-Tolerant Face Recognition

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in the
Department of Electrical and Computer Engineering

Ramzi Abi Antoun

B.S., Electrical & Computer Engineering, Carnegie Mellon University
M.S., Electrical & Computer Engineering, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

May 2013

THESIS COMMITTEE

Professor Marios Savvides, Advisor

Department of Electrical and Computer Engineering

Carnegie Mellon University

Professor Vijayakumar Bhagavatula

Department of Electrical and Computer Engineering

Carnegie Mellon University

Professor Takeo Kanade

Robotics Institute

Carnegie Mellon University

Professor Arun Ross

Computer Science & Engineering

Michigan State University

To Violette and Elia Abi Antoun

ABSTRACT

Automatic face recognition performance has been steadily improving over years of active research, however it remains significantly affected by a number of external factors such as illumination, pose, expression, occlusion and resolution that can severely alter the appearance of a face and negatively impact recognition scores. The focus of this thesis is the pose problem which remains largely overlooked in most real-world applications. Specifically, we focus on *one-to-one* matching scenarios where a query face image of a random pose is matched against a set of "mugshot-style" near-frontal gallery images. We argue that in this scenario, a 3D face-modeling geometric approach is essential in tackling the pose problem. For this purpose, we utilize a recent technique for efficient synthesis of 3D face models called 3D General Elastic Model (3DGEM). It solved the pose synthesis problem from a single frontal image, but could not solve the pose correction problem because of missing face data due to self-occlusion. In this thesis, we extend the formulation of 3DGEM and cast this task as an occlusion-removal problem. We propose a sparse feature extraction approach using subspace-modeling and ℓ_1-minimization to find a representation of the geometrically 3D-corrected faces that we show is stable under varying pose and resolution. We then show how pose-tolerance can be achieved either in the feature space or in the reconstructed image space. We present two different algorithms that capitalize on the robustness of the sparse feature extracted from the pose-corrected faces to achieve high matching rates that are minimally impacted by the variation in pose. We also demonstrate high verification rates upon matching non-frontal to non-frontal faces. Furthermore, we show that our pose-correction framework lends itself very conveniently to the task of super-resolution. By building a multiresolution subspace, we apply the same sparse feature extraction technique to achieve single-image superresolution with high magnification rates. We discuss how our layered framework can potentially solve both pose and resolution problems in a unified and systematic approach. The modularity of our framework also keeps it flexible, upgradable and expandable to handle other external factors such as illumination or expressions. We run extensive tests on the MPIE dataset to validate our findings.

ACKNOWLEDGEMENTS

I would like to thank my academic advisor Prof. Marios Savvides for his support, guidance, patience and help. I have been associated with Prof. Savvides one way or another for a decade. In the fall of 2003, I was a junior undergraduate student when he was the teaching assistant for 18-396 Signals and Systems taught by his advisor Prof. Bhagavatula. A couple of years later he was directing my master's research project to build a robot that navigates a closed space, track and follow faces (more on the robot story later). I was then one of the first three students to enroll at the newly formed biometrics lab, along with Ramu Bhagavatula and Sung Won Park. This thesis is the culmination of all those years of unwavering support and patience.

I would also like to thank my committee members, Prof. Vijayakumar Bhagavatula, Prof. Takeo Kanade and Prof. Arun Ross for serving on my committee and for their comments, guidance and suggestions. It is not any Ph.D. candidate who gets to have such a distinguished thesis panel and I am very grateful for it.

I have been at Carnegie Mellon since the very beginning. As Khalid Harun once observed, I was there when they built the Fence[1]. For over seven years, I have had the pleasure and honor to interact with great students who have to come to do research in the Biometrics Lab. Not only were all of them great minds to work and interact with, but overall great people as well. I would like to thank all of them, in particular Utsav Prabhu, Jingu Heo, Keshav Seshadri, for contributing ideas, theory and code that have made my research possible. Additional thanks to Keshav, Utsav and Shreyas for proof-reading this thesis. Should you find typographical errors please direct your anger at them. This is the rest of the list, starting chronologically with the Ph.Ds. Sung Won Park, Yung-Hui Li, Jingu Heo, Ramu Bhagavatula, Dhruv Batra, Keshav Seshadri, Shreyas Venugopalan, Utsav Prabhu, Juefei (Felix) Xu, Sekhar Bhagavatula, Khoa Luu, Thi Hoang Ngan (Nancy) Le. For those who haven't graduated yet, I hope you find the patience and strength to see that day, and include me

[1] http://en.wikipedia.org/wiki/Carnegie_Mellon_University_traditions

in your acknowledgement pages. I will miss working with you until the wee hours of the morning, and hope the friendships last beyond the confines of university halls.

I have also had the great pleasure to work with an almost uncountable number of master's and undergraduate students in the lab, chronologically Hung-Chi Lai, Sheethal Bhat, Kavya Patil, Ted Trebaol, Mukta Gore, Taihei Munemoto, Khalid Harun, Neha Agrawal, Kyle Neblett, Madhu Bhagavatula, Joe Heyman, Nick Vandal, Brendan Jou, Jameson Merkow, Arti Chhajta, Siddhesh Mehra, Unni Prasad, Sasikanth Bendapudi, Siddhanth Deshpande, Abhinandan Krishnan, Shreyas Bethur (aka Shreyas v2.0), Shyama Asokan, Yiting Xie, Yogesh Nagaraj, Amit Krishnan, Aaron Jaech, Siying (Diana) Hu, Miriam Cha, Martin Jaszewski, Murium Iqbal, James Marsanico, Greg Lew, Nolan Hergert, Sandeep Chakravarthula, Divya Hariharan. Apologies to whoever I forgot to list, it is definitely not intentional. All of whom I have enrolled in the Iris Access System and created accounts for on the servers. Remember that if it were not for me you would all be standing outside knocking on the door waiting for someone to let you in. (By the way, I have not deleted you from the Iris system so you can come back in anytime). I hope you're all rich and famous by now.

A shout-out to other non-biometric grad students, especially Divya Sharma and Raja Sambasivan, who were as confused as I was about the purpose of grad school and life in general. I know that Raja is graduating with me, so that will leave Divya to explore all the answers by herself.

The alert reader will have noticed the name Bhagavatula mentioned several times in these acknowledgement pages. This is because I have had the distinction and pleasure to work at some point during my tenure at the Biometrics Lab with all of my committee member Prof. Bhagavatula's sons, Ramu, Sekhar and Madhu. To the best of my knowledge, I am so far the only person in academia who has worked under a Professor and then with *all of his children* in the *same* lab.

I would also like to thank our colleagues, neighbors and CyLab room-mates at the Parallel Data Laboratory (PDL), Greg Economou, Michael Stroucken, Mitch Franzos, Zisimos Economou, who over the years have come to our rescue when our servers randomly went into deep slumbers never

to wake up again. It it weren't for them, this thesis would be missing 80% of its results and graphs. Your mind-bending linux wizardry have kept our machines running. Hats off to you.

Throughout my long years at the Biometrics Lab, we all benefited from *outstanding* support and help courtesy of the staff of Carnegie Mellon CyLab. Thank you Cathy Schaefer, Gene Hambrick, Linda Whipkey, Tina Yankovich, Michael Balderson, Helen Conti, Samantha Stevick, Rachael Swetnam, Ivan Liang, Kelley Conley, and Megan Kearns. Special thanks as well to Karen Lindenfelser. CyLab is a much friendlier place because of her. Her snacks, funny rants/emails and puzzles kept a smile on everyone's face.

I would also like to thank the staff of the ECE undergraduate and graduate office for their help and support. Thank you Reenie Kirby, Janet Peters, Elaine Lawrence and Samantha Goldstein for always being there for us. I hope to remain your favourite trouble maker!

Thank you as well for the staff of the CMU Office of International Education, especially Neslihan Ozdoganlar, Jennifer McNabb and Linda Gentile, for churning no fewer than *nine* I-20s during my extended stay at CMU, and for their expert advice when dealing with the less-than-friendly US immigration and embassies.

Thank you Chriss Swaney at the CMU Media Relations for getting us media exposure, making us almost famous and for bringing celebrities into our lab.

The following is a list of people from outside of school who liked me (I think) and supported me without ever understanding what it is I do, or why it is important, and why it is worth all those years. Some even bet money that I would never leave school before I reach retirement age. Chronologically, Diab Haddad (along with the rest of the Haddad family, Lara, Wissam, Elie and Jamilé), Abir Tebbo, Patricia Abou Chahine, Marita Chakhtoura, Nadine Jarjur, Ziad Rohayem. Thank you for all your blind support and patience. Your friendship and indifference provided a support system that is vital to surviving grad school. You were great friends and I love you all. Special thanks well for the rest of the Pittsburgh gang, starting chronologically with Cheuk Lai (Charlie) Choi, Alex Shkolnik, Michael Abdel Malek, Loubna El Abbadi, Carolyn Eissa, Deemah

Altahawi. I owe you my sanity.

I would like to apologize to my advisor for not finishing the robot I started building seven years ago. You can find the bits and piece lying in random drawers of the lab. You can probably recoup some of the money by auctioning the parts on eBay.

In time, I will look back at 2013 as the year I finally graduated, and also the year Keshav finally managed to pass his driving test. He had been diligently working on it since 2011 and I am glad to have witnessed this historical feat.

No thesis acknowledgement is complete without thanking people who do not expect to be acknowledged. So here are the random people I would like to thank. Back in the summer of 2006 AD, I had to travel back home to Beirut to obtain a new student visa to start my Ph.D. program. A few hours after my interview at the US embassy, war breaks out, the airport is bombed, and the US embassy closes up shop and hastily evacuates (*with* my passport) in a chaotic helicopter evacuation reminiscent of Operation Frequent Winds at the fall of Saigon in 1975. A few weeks later, a Syrian war-profiteering gentleman cabdriver by the unassuming name of Mwaffaq (which in Arabic means lucky) offers to drive me across the border for a nominal sum of 550$ (the normal going-rate at the time being around 50$). If it wasn't for this 6 hours trip I endured in a yellow Renault 12 Diesel listening to an exposé on the military might of the Syrian regime, I would not have reported to school on time. So thank you lucky Mwaffaq.

For the duration of grad school, I rented a small apartment in a Victorian house built around 200 BC at the intersection of Wightman and Forbes. The prehistoric wooden floor squeaked and creaked loud enough to make this house worthy of the next Addams Family movie sequel. For most of my studies, I rarely went home before 2 A.M causing an exodus of renters from the apartment below, and a 20% overall decrease in rent for that apartment to offset the inconvenience of my sleeping habits. Thank you all my neighbors who endured this for the last seven years.

Last but never least, and on a more serious note, none of this would have been possible had my parents not believed in me and thought I was school-worthy and decided to send me overseas to

an expensive school thousands of miles away from home at the tender age of 18. Your love and prayers are the only things that prevented me from giving up and quitting before I was 19. I just wish you realize I am over 30 now and stop calling me every day. Here's to you.

And now, in the words of Nick Vandal,

Mahalo, !@?#%¿&

This work has been partly funded by Carnegie Mellon CyLab.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Despite years of active research, automatic face recognition in images remains an unsolved problem for a number of reasons. Firstly, faces are non-rigid objects that deform non-linearly with the slightest change in expression. Secondly, they are captured at a distance which makes them prone to illumination changes and cast shadows. Moreover, even the slightest movement of the head often results in a noticeable perspective change. All of these factors significantly alter the appearance of the face, which can severely impact face recognition performance. Thirdly, being a non-obtrusive biometric that requires little cooperation from the subject, face recognition also needs to handle other external factors such as occlusion due to articles of clothing, hair obstructing parts of the face, eyewear, facial hair, etc. Face recognition becomes even more challenging in uncooperative scenarios such as surveillance footage or faces in a crowd, where all the above factors are magnified and present simultaneously.

While all of these factors have been studied, the pose problem is the only one that we believe is still largely unsolved, particularly in trying to solve face verification tasks. Most commercially deployed face recognition systems do not explicitly correct for the pose the same way that they preprocess the image to normalize illumination, and merely assume that the test faces are frontal or near-frontal. They sometimes employ simple 2D affine or projective transforms to align a small

set of fiducial points, which does not take into consideration the 3D structure of the face and in the process add other detrimental artifacts to the resulting 2D pose-corrected face. In this thesis, we propose to analyze and understand the pose problem thoroughly, and propose a novel solution.

Pose-tolerant feature extraction becomes even more critical for face recognition systems that rely on a single view of a subject for matching or enrollment, which is the main focus of this thesis. Such *one-to-one* systems take a single query facial image of an arbitrary pose and try to match it to a single or set of presumed frontal "mugshot" gallery images. The widespread availability of affordable imaging devices such as security cameras and mobile phone cameras is generating an ever-increasing number of faces that might need to be automatically processed, and in these uncontrolled environments the query faces are usually non-frontal. In the case of surveillance footage, the acquisition devices are usually placed higher up on a wall or ceiling to offer them a vantage viewpoint, and as a result the observed faces will exhibit high degrees of pitch and yaw, where traditional and naive 2D-based automatic pose-correcting measures (such as affine and projective transforms) quickly breakdown. As a matter of fact, perfectly frontal face images are very hard to come by, even in controlled environments where the subject is cooperating with the system. It is commonly observed that most people adopt a slight head tilt when their picture is taken. In fact, the posture of the head is part of the human body language, which along with the posture of the rest of the body, gestures and eye movements enables humans to subconsciously send and receive signals for what is known as non-verbal communication. It is believed that non-verbal communication constitutes a very large part of human communication [2]. Even though body language is not an exact science, it has been thoroughly studied and analyzed over the years, and in the psychology community the posture of the head is believed to suggest certain meanings [3, 4] as follows: head tilted to one side is generally accepted as a non-threatening gesture, submissiveness or thoughtfulness. It signals possible interest or vulnerability. A head tilted downwards is generally a signal of criticism, disapproval and admonishment if it comes from a position of authority. A chin pointed up suggests pride, defiance or confidence. It usually exposes the neck, which is a signal

of strength and resilience. The point that comes across is that even in fully-cooperative scenarios perfectly frontal faces are the exception rather than the norm, and face recognition engines need to systematically handle pose variations.

In practice, the pose problem is often compounded with another equally disruptive problem, that of low-resolution in the observed image. This problem plagues most surveillance footage, where the quality of the imaging sensors is usually poor, and ultra-wide angle lenses are attached to the sensor to increase the field of view but in return reduces the number of pixels apportioned for every observed face. Therefore, it is important for an automatic face recognition engine to be able to handle low resolution faces in conjuncture with the variations in pose. This thesis aims to provide answers for both problems.

1.1 The Big Picture

The architecture we are proposing in this thesis to handle pose and resolution variations remains a general, modular and extendable framework. The gist of our approach to handling pose variations is a 3D-face modeling step to geometrically correct the viewpoint of the face and a sparse feature-extraction step to map the pose-corrected faces to a domain that is stable under pose variations. Most pose-tolerant approaches in the literature eschew 3D modeling for its complicated implementation and high-computational cost and instead attempt to solve the surrogate problem of learning the relationship between different viewpoints using standard statistical methods. In this thesis, we go back to the original problem of matching off-angle faces which we believe at heart is a 3D problem, and we show how we can efficiently accomplish the 3D modeling and pose-correction using simple and easily reproducible techniques. The sparsity component of our approach is crucial to handle the underdetermined systems which arise as a by-product of pose-correction, as the amount of missing information representing self-occluded pixels increases. We borrow on ideas from the field of Compressive Sensing to find a stable solution in highly under-

determined systems, and specifically rely on ℓ_1-minimization for its sparsity-inducing properties. Finally, we show that this framework effortlessly lends itself to handling low-resolution faces as well. We show how we can modify our framework to achieve *single-image* super-resolution to enhance the visual cues of a low-resolution face and show how we could potentially achieve both pose-tolerance *and* resolution-tolerance by developing a unified theoretical framework to handle both challenges simultaneously.

When designing a face recognition system, it is important to build a modular and flexible system of interconnected modules with minimal dependence between them. Our framework keeps this in mind and relies on a sequence of independent modules to perform various tasks, such as face detection, face pose estimation, facial landmarking, 3D modeling, etc. One could potentially use a different implementation for a given module, or fuse modules together, or insert a new module to handle a different external face factor such as illumination.

1.2 Outline of this Thesis

For the remainder of this document, we refer to the techniques that take a non-frontal test face and generate an equivalent frontal-looking face as *pose correction*. On the other hand, we refer to the techniques that take a near-frontal face and generate an equivalent face of an arbitrary non-frontal viewpoint as *pose synthesis*. This dissertation is structured as follows. Chapter 2 will give a brief overview of some recent efforts to tackle pose-tolerance in face recognition. We argue that we need an essential 3D modeling step and for that we review a fast and efficient 3D face-modeling technique in Section 2.2.2 called 3DGEM. Chapter 3 will extend the formulation of 3DGEM to handle non-frontal images when building a 3D model. The core of our pose-correction approach is fitting a 3D model to the non-frontal test image, rendering the 3D model at a frontal viewpoint and extracting features that are stable under pose variations. We develop different shape-free representations where we use ℓ_1-minimization to obtain a sparse feature vector. Even though the algorithms

we develop later only rely on this feature extraction step, we also show how we could synthesize an equivalent frontal-looking face that we could feed into traditional face recognition algorithms. Chapter 4 will apply the pose-correction technique of Chapter 3 to improve verification rates when matching a non-frontal test image to a near-frontal gallery image. We benchmark all the different image representations in two separate but related domains. We also show that we can readily exploit the qualities of the sparse feature extraction in the coefficient domain to significantly improve the verification rates. We present two different algorithms that can capitalize on the discriminative power available in the sparse pose-tolerant feature vector. We also investigate the use of shape information in a pose-correction framework, and compare the results of pose-correction to pose-synthesis. Chapter 5 will apply some of the techniques of Chapter 3 towards the problem of super-resolution. We rely on sparse feature extraction in a shape-free representation to enhance the spatial resolution of low resolution faces, with magnification rates as high as 16x, and starting from faces with as few as 8 pixels between the eyes. We also analyze the impact of low resolution non-frontal faces on the pose-correction technique presented in this dissertation, and offer future ideas on how to solve the pose and resolution problem simultaneously. Chapter 6 will conclude this work by summarizing the results and findings and discussing future research ideas.

Appendix A presents a quick primer on Compressed Sensing and ℓ_1-minimization as it is relevant to this thesis. Appendix B derives the SimBoost algorithm used in Chapter 4. Appendix C presents a brief overview of advanced correlation filters and the Class-dependent Feature Analysis (CFA) algorithm that we used in Chapter 4. Finally Appendix D includes more experimental results and tables and graphs as well as results on a different dataset of face images to show that our methods generalizes well on different images.

Chapter 2

Background

Pose-tolerance in face recognition has been researched for a number of years. The adverse effects of non-frontal poses in face recognition were quantitatively assessed over a decade ago [5]. A recent comprehensive review by Zhang and Gao [6] categorizes and compiles all the different approaches in the literature. In this comprehensive study, they conclude that all of the surveyed methods come with limitations and fall short of solving the pose problem, and that "continuing efforts are still necessary towards ultimately reaching the goal of pose-invariant face recognition and achieving the full advantage of being passive for face recognition". We next include a brief literature survey of most well-known techniques to tackle face recognition in the presence of non-frontal poses.

2.1 Brief Literature Review

The earliest efforts included recording the test subjects at different possible angles and used a statistical model for each angular bin [7, 8]. TensorFaces [9] were used to interpolate an unseen view, however TensorFaces are very computationally and memory expensive. To overcome this limitation, simple 2D warping methods were applied on frontal faces to generate equivalent non-

frontal images such as in [10]. Beyond simple 2D warping techniques, most efforts fall under two main families of algorithms. One family of algorithms makes use of an explicit 3D geometric transformation, while the other family avoids the 3D modeling step, which we call the statistical modeling approach. This thesis follows the 3D geometric approach to solve the pose problem, but we briefly review both approaches in the next section.

2.1.1 Statistical Modeling Approaches

A common family of pose-tolerant approaches in face recognition can be categorized under *Statistical Modeling Approaches*. In these methods, the relationship between frontal and non-frontal image are treated as a statistical machine learning model problem, without including an explicit 3D mapping step. This relationship could be either learned in the image space or some other feature space mapped from the image representation. In [11], Gabor jets were extracted from several (hand-selected) locations on the face and their responses were transformed (by multiplying with a transformation matrix) when the viewpoint changed. The authors showed how the rank-1 identification rates between different viewpoints favourably improved when they applied this kind of transformation to the Gabor jets. Gabor jets play a central role in the Elastic Bunch Graph Matching (EBGM) algorithm [12], and it has since been extended to handle non-frontal viewpoints by modifying the Gabor jets extraction step [13]. In [14], the relationship between training pairs of frontal and non-frontal images are learned using Neural Nets (NN). The authors combine this prior knowledge with Hidden Markov Models (HMM) to synthesize profile faces from frontal mugshot images. Recent statistical approaches attempt to map input features to a pose-invariant feature space. In [15], the authors developed a Bayesian classifier based on Gaussian Mixtures Models (GMM) to extend a frontal face model with artificially synthesized models corresponding to non-frontal views. The synthesis techniques, based on Maximum Likelihood Linear Regression, learn how the frontal face models are related to the non-frontal models. In [16], inspired from a view-based Linear Discriminant Analysis (LDA), several locally linear transformations are simul-

taneously extracted to maximize inter-class separation while minimizing intra-class separation in a transformation domain that corresponds to the soft-clustering of the data. The number of clusters is dictated by the number of training viewpoints.

In the most prominent global statistical approach, Eigen Light-Fields, Gross *et al.* [17] unified all possible appearances of faces in different poses within a framework of light field. The authors make use of the redundancy of the light-field to represent face images in different poses using a single set of eigenvectors and eigenvalues to capture identity-change variations. Pose-invariance is treated as a missing data problem: the test image and gallery image are assumed to belong to a larger data vector containing all possible poses. The missing information can be estimated based on a prior knowledge of the joint multivariate Gaussian probability distribution of the complete dataset.

All of the previous statistical approaches are mainly global approaches as they model the full face. Local statistical approaches have been suggested to improve the overall performance. In [18] and [19] for instance, a probabilistic formulation is adopted to measure similarities between face patches belonging to different viewpoints.

One of the most prominent local statistical approaches is Tied Factor Analysis (TFA) [20]. Prince *et al.* assume that all face images of a single person lie on a manifold in the "observation space", and they can be generated from the same vector in a high-dimensional "identity space". Their approach is a generative method where they learn the mapping from the "identity space" to the "observation space" rather than the other way around. They learn the parameters of the generative model that maximizes the joint likelihood of the observed data and the corresponding data in identity space. However they cannot observe the identity vector directly, and can only infer a posterior distribution over them for some fixed parameter vector. To solve this chicken-and-egg problem, they resort to an iterative approach using Expectation-Maximization (EM) algorithm on a set of training images of known poses. Having learned the parameters that describe the mapping from the identity space to the observed space, they can now synthesize what a face will look like

at a different pose. However when matching a probe image of an arbitrary pose, they follow a Bayesian rule that assigns the test image to the class that maximizes the likelihood of it belonging to that class, which takes into account the evidence with respect to all the other classes. Despite this last classification rule which clearly corresponds to score normalization, their rank-1 identification rates remain modest. To further improve the performance, they resort to a local-patch based approach. Given 14 non-occluded points control points on the face, they extract Gabor-like information from a 5×5 grid centered around the control points. Every element of that grid contains a square patch of the image, and they extract information that represents gradient, scales and mean intensities that they combine into one feature vector. In training, they learn 14 separate tied factor analyzers, one for each grid. Each feature gets its own likelihood estimation for each possible subject. When matching a testing face, they treat each of these likelihoods independently and take the product to calculate the final classification Bayes rule. This local approach significantly increases the complexity of the algorithm, but enables them to increase their rank-1 identification rates. Their results include test angles up to $\pm 90^{\circ}$, and compare favourably against results from [17] and [21]. Note that when handling half-profile and full-profile faces, some of their control points are placed on the ears and the hairline of the subjects, which typically are not included in 3D face modeling.

A more recent effort ([22]) aimed to solve the pose and resolution problem simultaneously. Tensor analysis is used to estimate the approximate pose and the rough locations of the facial landmarks. SIFT features are then extracted from the fiducial points. The authors use Multidimensional scaling (MDS) to map the SIFT descriptors extracted from both high and low resolution images to a common space where the distances from each other approximate the distances had all descriptors been extracted from high-resolution images. This mapping is learned on training data that provided the SIFT feature vector at different resolutions and viewpoints. The transformation matrix responsible for the mapping is obtained by minimizing a cost function that also invokes the iterative majorization algorithm. The dimensionality of the combined SIFT features extracted around fiducial points is reduced with PCA. The authors train on 100 randomly sampled subjects from

MPIE and include all the different illuminations available in their training. They report improved results on 3x magnification rates and $\pm 30^{\circ}$ yaw variation.

2.1.2 Geometric Approach with 3D Modeling

3D face modeling was suggested as a way to handle pose. Some of the earliest efforts [23, 24] required multiple viewpoints to create a basic 3D model. In [24], photometric stereo techniques are employed to synthesize novel viewpoints and illumination patterns. Their approach sequentially estimates lighting conditions and surface gradients using singular value decomposition (SVD). Other methods made use of multiple images without an explicit 3D step [25]. In [26], a geometrical mapping is adopted to map a face image after estimating its pose onto the surface of a 3D ellipsoid and the recognition is performed on the surface of the ellipsoid. The multiple image requirement requirement made the 3D geometric approach unattractive. A breakthrough was achieved when Blanz and Vetter [27] relaxed this requirement and generated a 3D model from a single 2D image with a method they called 3D Morphable Models (3DMM). In [21], the authors proposed the use of 3DMM (reviewed below) to correct for the pose problem in a related approach to what we are proposing in this thesis. Part of the results was included in [5] and showed that explicit 3D correction helped the verification results (measured on 2 test yaw angles, and 2 test pitch angles) favourably. In [28], a variant of the 3DMM formulation is presented to estimate the texture information and illumination coefficients using spherical harmonics models. The spherical harmonics representation offers a low-dimensional linear subspace that can accurately approximate a convex Lambertian object under a wide variety of lighting conditions.

2.2 3D Face Modeling Primer

The ill-posed problem of reconstructing 3D face models from 2D images has been well studied over the last few years [29]. "Shape-from-X" techniques rely on either multiple images separated

temporally or by pose. For the requirement of the *one-to-one* face matching application of this paper, we require a technique which can generate a 3D model from a single input image. We briefly review the two most common techniques to achieve 3D face models from a single 2D input image.

2.2.1 3D Morphable Models

Blanz and Vetter proposed a 3D Morphable Model (3DMM) [27] based on image-based reconstruction and prior knowledge of face shapes and textures. During training, an offline 3D face model is learned for a number of 3D scans. Using a variant of Optical Flow, they find the full correspondence between vertices. The structure of a face is then captured in a shape vector $\mathbf{s} \in \Re^{3N}$ (containing x, y and z coordinates of N vertices of a face) and a texture vector $\mathbf{t}$ (containing the R,G,B color values of the mean-warped face image). 3DMM builds two separate models: a shape model such that any shape vector $\mathbf{s}$ can be represented by $\mathbf{s} = \bar{\mathbf{s}} + \sum_i \alpha_i \mathbf{s}_i$. The second model is a texture model, such that a given texture vector can be expressed as $\mathbf{t} = \bar{\mathbf{t}} + \sum_i \beta_i \mathbf{t}_i$. The purpose of the 3D model is to parametrize the face representation, such that we can synthesize a new face image by changing the parameters of both models simultaneously. During the testing phase, 3DMM presented an algorithm to fit the Morphable Model to an input 2D face. The first step is to initialize the Morphable Model (which may require human input), and then iteratively estimate the model parameters of the testing image by minimizing the mean squared error between the input pixel intensities and the reconstructed intensities using a stochastic gradient descent. The gradient descent converges to a final estimate of all of the model parameters, such as shape and texture coefficients, 3D orientations, ambient light parameters, color offsets and contrasts, etc. Altering the pair of the shape and texture parameter vectors in a constrained fashion enables 3DMM to modify other facial attributes such as expressions or the perceived weight and gender of the test subject. Altering the ambient lighting and color parameters enables 3DMM to modify other shading attributes such as illumination.

The 3DMM formulation is accurate, however it is computationally expensive and requires a complicated iterative optimization of the model parameters. For our purpose of pose-correction, we require a rapid 3D prototyping technique and for that we rely on another state-of-the-art yet simpler and less computationally intensive 3D modeling technique we review in the next section.

2.2.2 Classical 3D Generic Elastic Models for Pose Synthesis

3D Generic Elastic Models (3DGEM) were first introduced by Heo and Savvides [30] as an efficient and computationally inexpensive real-time method for generating near-accurate 3D models from a single *frontal* face image. It was also shown that near-accurate 3D models are sufficient for robust rank-1 identification accuracy at off-pose angles [31], by rotating the frontal target image to match the probe viewpoint, making the 3DGEM technique invaluable at pose-invariant face identification.

Let $\mathbf{p}' = [x'_p, y'_p]^{\mathrm{T}}$ represent the location of any single observed landmark point on an image, corresponding to the "actual" 3D point represented as $\mathbf{p} = [x_p, y_p, z_p]^{\mathrm{T}}$. The aim of 3D structure reconstruction can be formulated as a very ill-posed problem of recovering $\mathbf{p}$ from a single observation $\mathbf{p}'$. These points are related by the well-known camera projection equation:

$$\mathbf{p}' = \mathbf{K} \cdot [\mathbf{R}|\mathbf{t}] \cdot \mathbf{p} \tag{2.1}$$

where $\mathbf{p}'$ and $\mathbf{p}$ represent the 2D and 3D points in homogeneous coordinates respectively, $\mathbf{K}$ represents the 3×4 intrinsic camera matrix, $\mathbf{R}$ and $\mathbf{t}$ represent the 3D rotation and translation components respectively. Since any translation component can be compensated for by merely re-centering the frame of reference for the face, we can represent the problem as a simpler projection equation:

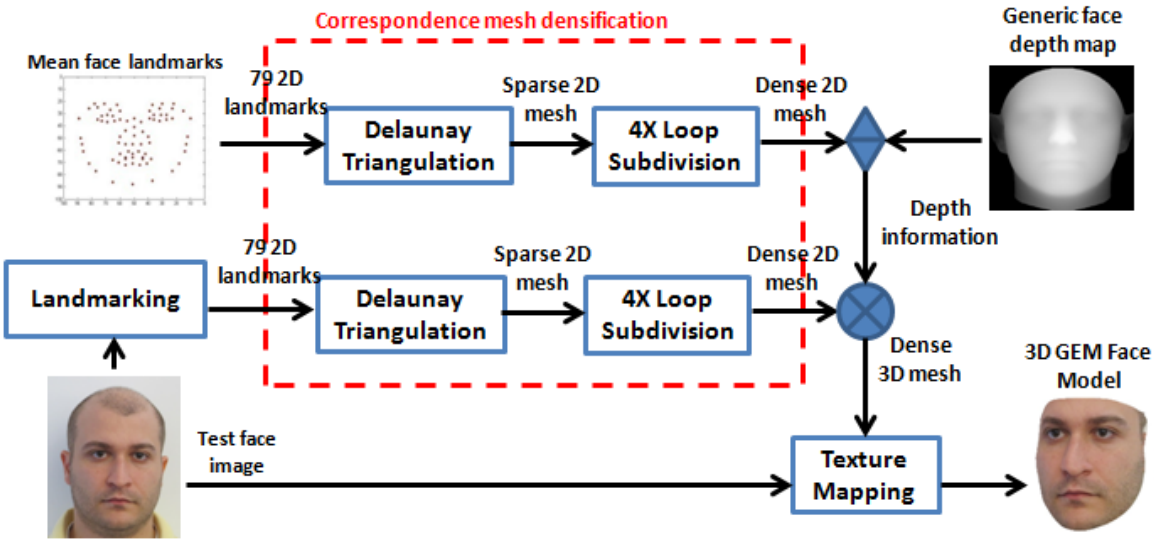
$$\mathbf{p}' = \mathbf{K} \cdot \mathbf{R} \cdot \mathbf{p} \tag{2.2}$$

without using homogeneous coordinates, i.e. $\mathbf{K}$ and $\mathbf{R}$ are the 2×3 projection matrix and 3×3 rotation matrix respectively. In the case of frontal 3D GEM, the rotation matrix $\mathbf{R}$ is clearly an identity matrix, and the internal camera matrix is assumed to be an orthographic projection (i.e. it simply drops the z component without altering the (x, y) components), $\mathbf{K} = [I_{2\times2}|0]$.

The underlying assumption in the 3DGEM technique is that structural information of depth z of a human face is not significantly discriminative across different subjects, provided the faces are perfectly aligned at all 2D fiducial landmarks (x, y). Given a perfect pixel correspondence at every facial feature on the face, Heo and Savvides showed that the depth information for these aligned correspondences does not vary significantly among a particular ethnic and gender group. Consequently, if this dense 2D (x, y) spatial configuration of an input face image can be reliably localized, then a mean generic depth map can be elastically deformed based on the 2D fiducial facial features and the mean depth values $\bar{z}$ can be assigned to corresponding points to generate a complete 3D structure.

We obtain this dense 2D spatial structure of the face image by first localizing certain predefined fiducial points on the face either manually or by utilizing an automatic technique such as the Modified Active Shape Model (MASM) [32]. A triangular mesh is constructed from these points, which is then subdivided multiple times by means of the loop subdivision algorithm [33] to obtain a dense 2D point set representing the deformations in the face. An elastic deformation function ϕ can be constructed from these points which can deform a mean 2D structure $(\bar{x}, \bar{y})$ to the identified shape, i.e. $(x, y) = \phi(\bar{x}, \bar{y})$.

A mean generic depth map represented as $\hat{z} = \bar{z}|\phi(\bar{x}, \bar{y})$ is constructed from a set of aligned 3D face models obtained from the USF Human ID database [27]. This depth map serves as an index for depth association, corresponding depth values from this map are assigned to the dense 2D point set obtained from the test image using the learned elastic deformation function ϕ to obtain a dense 3D structure $\mathbf{p} = [x', y', \hat{z}]^{\mathrm{T}}$ for the face. Heo and Savvides [34] have shown that the mean depth map captured most of the variations and there was little gain by using more basis vectors to



(a) 3DGEM flowchart

(b) 3DGEM output

Figure 2.1: 3DGEM pose synthesis example. The different images were obtained by rendering the 3D model obtained in the Figure 2.1a at different yaw angles.

represent the variations.

After a texture mapping step, a dense 3D model of the input face is obtained, which can be rendered at any pose. The flowchart of the complete algorithm outlined above to construct these 3D models from frontal 2D input images is depicted in Figure 2.1a. Examples of synthetic viewpoints generated by rendering the 3D model at different angles are depicted in Figure 2.1b.

Chapter 3

Pose Correction

In section 2.2.2 we showed how could take a near-frontal image and generate an arbitrary viewpoint. However the practical problem that needs attention is the inverse problem, that of taking an arbitrary non-frontal query image and match it to a database of images. Most face gallery databases (such as passport photos and DMV photos databases) consist of frontal or near-frontal faces, and it would currently be too expensive to rotate all faces in the database to match the viewpoint of a query image of an arbitrary angle. Instead, we argue that it is more efficient to leave the gallery unperturbed and process the query image to correct its pose back to frontal, where more face features are visible. Figure 3.1 shows a flowchart of our proposed pose correction approach. However, our proposed approach implies building a 3D model from a non-frontal image. Section 3.1 offers an extension of 3DGEM to handle non-frontal input images. It is important to note that correcting the pose of non-frontal images will introduce pixels that are "self-occluded" for being on the nonvisible farside of the face. Section 3.2 will introduce a novel approach to extract a sparse feature vector from the image which we will show is stable under changing pose and with an arbitrary number of missing pixel dimensions. We will use this feature vector extensively in Chapter 4 to boost verification rates. Finally in section 3.4, we will show that given this sparse feature vector, we can synthesise a pose-corrected face with no missing pixel dimensions from which we can generate a

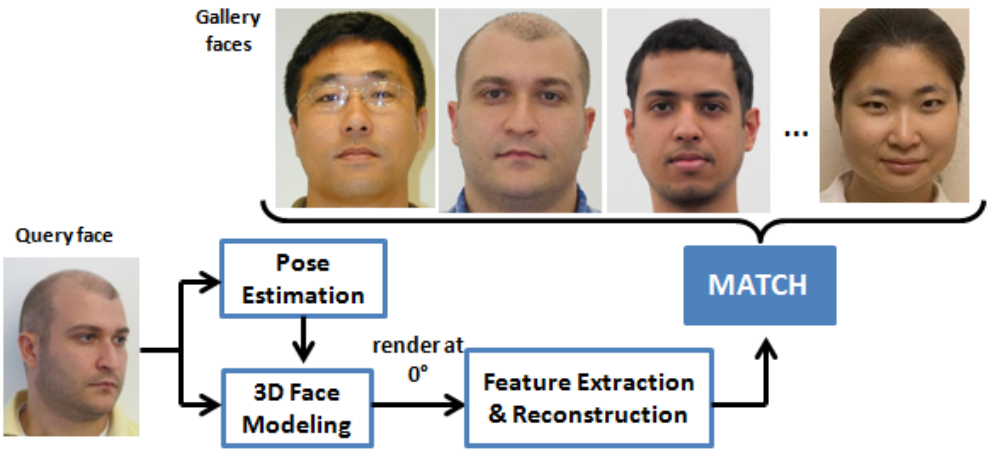


Figure 3.1: To match a non-frontal query image to a set of nearly frontal gallery images, we propose a methodology to correct the pose of the query image. Firstly, 3D modeling is adopted to render an equivalent frontal-looking face. Secondly, a sparse feature representation is extracted from the 3D-corrected face so that we can match it to the frontal gallery images in a low-dimensional feature space that is pose-tolerant.

traditional crop to feed into current commercial face matchers.

3.1 Construction of 3DGEM Structure from Non-Frontal Images

While the 3DGEM technique outlined in the previous paragraph has been proved to be efficient and reasonably accurate, its formulation only allows the generation of 3D models from a single *frontal* 2D face image. We now extend the functionality of 3DGEM to construct near-accurate 3D models from a single *non-frontal* image as well.

In the case of non-frontal images, the primary impediment stems from the fact that the observed 2D landmark points in the image do not directly represent the (x, y) variations of the face in the 3D space, as the rotation matrix $\mathbf{R}$ is no longer an identity matrix. In this case, we can rewrite

Equation 2.2 as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3.1}$$

where the $\mathbf{K}$ and $\mathbf{R}$ matrices have been written in terms of their individual components. Under the GEM assumption, the value of z can be assigned from a generic depth map as $\hat{z} = \bar{z}|\phi(\bar{x}, \bar{y})$. Incorporating this value and simplifying the above equation, we obtain the values of x and y as:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}^{-1} \begin{bmatrix} x' - r_{13}\hat{z} \\ y' - r_{23}\hat{z} \end{bmatrix} \tag{3.2}$$

It is important to note that the 3D structure of the face can be recovered only if the rotation matrix $\mathbf{R}$ is known. We compute this rotation matrix by using a commercial pose estimation engine which gives us the pitch, yaw and roll angles of the face in the image.

There are two caveats to 3D reconstruction in this manner: firstly, the technique relies on accurate localization of a sparse set of landmarks on the input face image, some of which may be occluded due to the pose. Secondly, since the texture details of some regions of the face are distorted due to the pose angle, it may be difficult to accurately obtain the texture of the 3D face structure.

We address the first issue by developing a new version of MASM which is able to localize the 79 points on the face across a wide range of input face poses (from $-40°$ to $+40°$). In this range of poses, we find that all 79 points are reliably visible or can be approximated to a reasonable accuracy.

Once the sparse set of 2D landmarks are localized in the input image, we densify this point set by means of triangulation and multiple levels of loop subdivision. The corresponding 3D structure

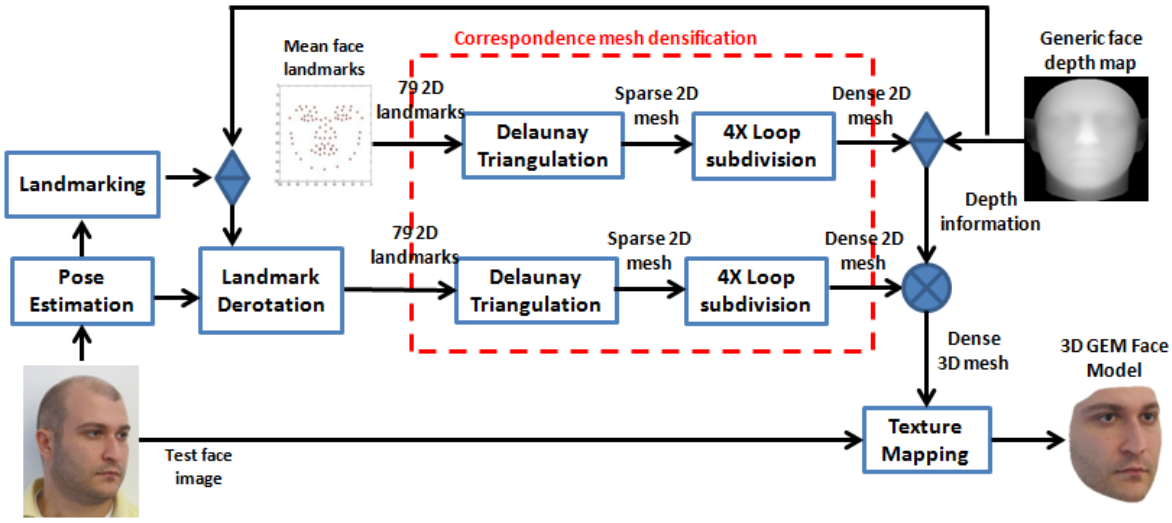


Figure 3.2: Non-frontal 3DGEM rendering flowchart.

is then generated and texture-mapped by sampling and interpolating the image at the corresponding points. This 3D model can then be rendered at a frontal viewpoint to obtain a "pose-corrected" image of the face. The flowchart of the modified algorithm to construct these 3D models from 2D input images is depicted in Figure 3.2.

3.2 Subspace Modeling to Extract Features Stable Under Pose Variations

In the previous section we showed how we can use 3D modeling to change the viewpoint of a face by taking into account the 3D structure of the human face. However this 3D correction step, which is an approximation of the true 3D face model, will inevitably introduce errors, especially after we render the model at novel and arbitrary angles. There are two main sources of 3DGEM errors. The first is at the 3D model formation stage, which is inherent to 3DGEM (such as the generic depth assumption, or the accuracy of landmarks, etc.). The second source of error is induced at the rotation/de-rotation stage, where we might be amplifying the errors of the first stage by arbitrarily varying the posture of the 3D model (such as the problem of dealing with missing/occluded pixel

information on the far side of the face). Hence, relying on 3D modeling alone is not sufficient to achieve reliable pose-tolerance. Instead, we aim to extract a feature vector from pose-corrected faces that will minimally change when the viewpoint changes. In order to achieve this, we rely on subspace modeling after the 3D step to extract features. Specifically, we can build a subspace from a very large number of representative training face images which offer variations in gender, expressions, illuminations, external features (such as eyewear, facial hair, scars, etc.) which will let us represent any unseen test image. The following subsection will detail how we use this subspace to extract a pose-tolerant feature vector.

3.2.1 Shape-Free Representation for Accurate Texture Analysis

Since we are interested in obtaining a feature vector from mostly texture information, it is beneficial to isolate the texture from shape when building the subspace. We define "shape" by the x and y coordinates of a specific set of predefined landmarks on a face. The shape information can be discounted by having all faces adopt the same shape before we model the subspace. In other words, for a true shape-free representation, all facial features (eyebrows, eyes, nose, mouth, etc.) should have the same dimensions and locations for all faces. Once this is accomplished, the only discriminative factor left between the shape-normalized faces is the texture. For this purpose, a global transformation, such as an affine transformation of the entire face, is not enough, since it multiplies every pixel of the input image by the same transformation matrix. Instead, local transformations, where every triangle bounded by any three control points can be transformed independently, are needed.

One approach is to warp all faces to a particular mean shape using a piecewise-linear warping scheme as in [35]. This method however will introduce discontinuities and other artifacts which negatively impact the subspace modeling of the texture information. An alternative approach is to make use of 3DGEM, and build a 3D generic structure and render all face textures with this common structure to normalize the shape of faces as depicted in Figure 3.3. This will essentially

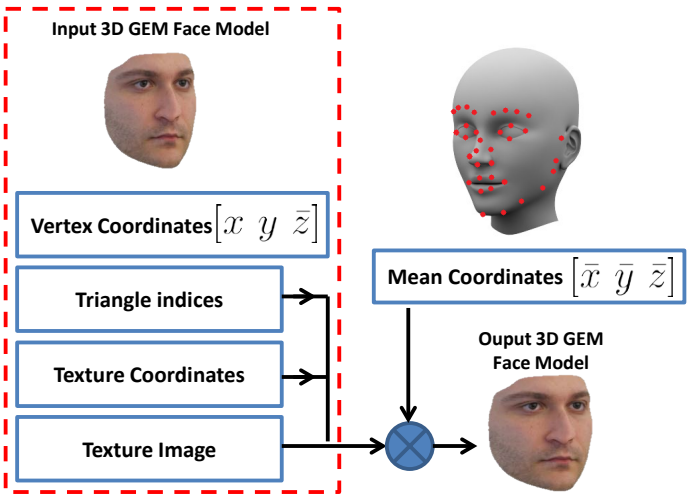


Figure 3.3: Shape-free generation module.

render all the faces given a common shape denominator just like the previous method, but 3DGEM will ensure a much smoother rendering free of high frequency discontinuities and artifacts. Figure 3.4 depicts three frontal faces from the MPIE database [36] with a traditional crop using eye coordinates, and their equivalent shape-free representation.

With this shape-free representation, we can set out to build a subspace that accurately models all the texture variation. Since the purpose is not to learn how to discriminate between the training faces but to represent the texture feature-space accurately, simple representations such as PCA [37] or NMF [38] built from several thousands of face images transformed in the shape-free domain is sufficient to capture the principal texture variation. We will employ PCA for its energy compaction properties which will become evident once we discuss the sparse feature extraction step.

3.2.2 3D Occlusion Detection for Pose-Corrected Faces

3DGEM relies on texture mapping to render the model at given angle. When the model is built on a frontal face, the observed texture contains enough information for texture mapping to generate a smooth and artifact-free image. However, when the model is built on a non-frontal face, the pixel information on the far-side of the face is mostly unobserved due to self-occlusion, and as a

(a) Frontal images of the first 3 subjects in the MPIE database.

(b) The corresponding traditional 128×128 pixels crops of the images in (a).

(c) The corresponding shape-free representations of the images in (a).

Figure 3.4: Comparison between traditional face crop and shape-free representation. The latter is essential for feature extraction and reconstruction in the presence of missing texture information.

result, texture mapping will invariably introduce stretching artifacts when the model is rendered at a frontal or opposite angle. The advantage of using 3D modeling is that we can relatively easily detect the pixels which are likely to exhibit stretching artifacts after pose-correction. This can be inferred in 3D, where we represent the face surface by a triangular mesh. Every triangle is defined by three control points obtained from the initial 79 landmark points and their subsequent subdivisions. Since we have the coordinates of the vertices of all triangles in the mesh, we can

(a) Original face image with varying yaw.

(b) Pose-corrected images in shape-free representation with occlusion detection.

Figure 3.5: An MPIE subject with (a) the original image from different angles (b) the corresponding pose-corrected shape-free images depicting the pixels beyond a certain angle from the viewing direction, and which will exhibit high stretching.

obtain the direction of the surface normal, and the angle between that normal and the viewing direction. If that angle exceeds a given threshold (60° in this case), then this pixel is likely to belong to a region that will exhibit stretching artifacts (due to self-occlusion), and hence not be representative of the true texture value at that location. Figure 3.5 illustrates the different levels of stretching-prone pixels for pose-corrected face images taken from different angles.

Note that in Figure 3.5 this method results in finding pixels prone to exhibit stretching even in the frontal faces. This is because even in a frontal face, the normal to the side of the nose surface deviates over 60° from the viewing direction. This will not be an issue, as we will be reconstructing these missing pixels via subspace modeling anyway.

3.2.3 The Duality of Pose Correction and Occlusion

By carefully observing Figure 3.5, we can clearly see that pose correction (after the 3D de-rotation step) translates into an occlusion removal problem: the more severe the out-of-plane rotation in Figure 3.5a, the wider the vertical occlusion "band" that appears in the shape-free representation in Figure 3.5b and the greater the number of missing pixels. Therefore our 3D-modeling and de-rotation step has reformulated the original pose-correction problem into an occlusion challenge.

Our feature-extraction task now becomes a problem of coping with a varying number of missing pixels (or dimensions). While making use of symmetry is an obvious option to overcome this problem and estimate the occluded part of the face, it is not the most accurate way. Faces are understood to be asymmetrical to the point where facial asymmetry has been used as a biometric [39]. Moreover, faces become even more asymmetrical when expressions are involved, as neuropsychology research has shown that humans display more expressions on left side of the face [40, 41]. Studies have tried to link the uneven display of expressions on the face to hemispherical specialization of the brain and handedness, as right-handed subjects are believed to express emotions differently than left-handed subjects [42, 43].

We now explain one way we can theoretically fit a feature vector on data with missing dimensions (corresponding to missing pixels on the face) using a subspace of trained faces. This approach is a much better way to tackle the pose-correction modeling problem than assuming facial symmetry. To simulate a finer varying degree of occlusion (finer than what is provided by MPIE's non-frontal angles), we switch to using traditional frontal faces and artificially remove a vertical band of increasing width to approximate the occlusion pattern. For the remainder of this subsection, we assume we have a built a large subspace of training faces that are representative of the occluded face that needs to be reconstructed and is represented by a PCA subspace [37] using a basis of N eigenfaces. It is important to note that our method can be applied to other subspaces, and we only show PCA as one simple example to demonstrate the pose-tolerance feature extraction process.

Let the test image be vectorized into a vector $\mathbf{x}$ of dimensionality d and represented by $\mathbf{x} = \mathbf{V}\mathbf{c} + \mathbf{m}$, where $\mathbf{V}$ is the matrix of vectorized eigenfaces, $\mathbf{c}$ is the vector of coefficients, and $\mathbf{m}$ is the vectorized mean face. The problem is to now estimate the values of the missing pixels in a partially occluded face by utilizing the subspace. This is a "missing data" problem, and the following derivation details how we can overcome it.

3.2.4 ℓ_2-norm Feature Extraction in Subspace with Missing Dimensions

For notation simplicty, let $\mathbf{x}'$ be the vector of active pixels of $\mathbf{x}$. $\mathbf{x}'$ is of size d' and $d' < d$ pixels. Similarly, let $\mathbf{m}'$ be the mean of active mean pixels. For notational simplicity, we can also introduce $\mathbf{x}'_\mathbf{c} = \mathbf{x}' - \mathbf{m}'$, which represents the centered version of $\mathbf{x}'$ of size d'. $\mathbf{V}'$ is the matrix of active rows that are in $\mathbf{V}$. We need to solve for the coefficient vector $\mathbf{c}$, so in the case of an overdetermined system ($d' > N$), we can minimize the following cost function $J(\mathbf{c})$:

$$J(\mathbf{c}) = \|\mathbf{x_c}' - \mathbf{V}'\mathbf{c}\|_2^2 \tag{3.3}$$

$$= (\mathbf{x}'_\mathbf{c} - \mathbf{V}'\mathbf{c})^{\mathrm{T}} (\mathbf{x}'_\mathbf{c} - \mathbf{V}'\mathbf{c}) \tag{3.4}$$

$$= \mathbf{x}'^{\mathrm{T}}_\mathbf{c}\mathbf{x}'_\mathbf{c} - \mathbf{x}'^{\mathrm{T}}_\mathbf{c}\mathbf{V}'\mathbf{c} - \mathbf{c}^{\mathrm{T}}\mathbf{V}'^{\mathrm{T}}\mathbf{x}'_\mathbf{c} + \mathbf{c}^{\mathrm{T}}\mathbf{V}'^{\mathrm{T}}\mathbf{V}'\mathbf{c} \tag{3.5}$$

Setting the gradient $\nabla\mathbf{J}(\mathbf{c})$ with respect to $\mathbf{c}$ to zero, we obtain:

$$2\mathbf{V}'^{\mathrm{T}}\mathbf{V}'\hat{\mathbf{c}}_{\text{me}} = 2\mathbf{V}'^{\mathrm{T}}\mathbf{x_c}' \tag{3.6}$$

$$\hat{\mathbf{c}}_{\text{me}} = (\mathbf{V}'^{\mathrm{T}}\mathbf{V}')^{-1}\mathbf{V}'^{\mathrm{T}}\mathbf{x_c}' \tag{3.7}$$

$$\hat{\mathbf{c}}_{\text{me}} = (\mathbf{V}'^{\mathrm{T}}\mathbf{V}')^{-1}\mathbf{V}'^{\mathrm{T}}(\mathbf{x}' - \mathbf{m}') \tag{3.8}$$

The vector $\hat{\mathbf{c}}_{\text{me}}$ represents the *minimum-error* PCA projection coefficients of the image $\mathbf{x}'$ which includes occluded pixels. Observe that the dimensionality of $\hat{\mathbf{c}}_{\text{me}}$ is the same as the number of eigenfaces. Therefore, we can reconstruct a new image $\mathbf{x}$ of size d from $\mathbf{c}$ using the following:

$$\hat{\mathbf{x}}_{me} = \mathbf{V}\hat{\mathbf{c}}_{me} + \mathbf{m} \tag{3.9}$$

What we have achieved with Equations 3.8 and 3.9 is the obtaining of projection coefficients from an input image with missing dimensions, and the reconstructing of a full-dimension image

from those coefficients with no missing pixels. When there is no occlusion ($d' = d$), Equation 3.8 degenerates into a standard a PCA formulation and $\hat{\mathbf{x}}_{\text{lse}} = \mathbf{x}$. This simple reconstruction step enables us to visually assess the quality of the fit. If the coefficient vector we solve for is a "bad" solution, then the reconstruction would not produce a natural-looking smooth face with no artifacts. Conversely, if we can get to the "true" coefficient vector that explains all the observed faces across poses, then the reconstructed faces would look identical no matter what the input pose is.

Even though Equation 3.5 yields a reasonable reconstruction as shown in Figures 3.6a and 3.6b, it is obvious that the method described above suffers from a major problem: it is struggling to reconstruct the same face as the occlusion increases. Figures 3.6a and 3.6b show that as the occlusion grows to occupy half of the face, the reconstruction contains artifacts. This is because the matrix to invert in Equation 3.8 is of size $N \times N$ but of rank d'. As the occlusion grows bigger, d' grows smaller and $\mathbf{V'}^{\mathrm{T}}\mathbf{V'}$ becomes rank deficient when N exceeds d'. So accordingly we have to limit N to be smaller than d'.

Alternatively, instead of minimizing the least squares of the error vector, as in Equation 3.5, we can opt for an underdetermined ($d' < N$) to minimize the following cost function:

$$\min J(\mathbf{c}) = \|\mathbf{c}\|_2 \text{ subject to } \mathbf{V'c} = \mathbf{x'_c} \tag{3.10}$$

The Euclidean norm allows us to elegantly solve the equation using Lagrange multipliers. We define the Lagrangian:

$$\mathcal{L}(\mathbf{c}, \lambda) = \|\mathbf{c}\|_2 + \lambda^{\mathrm{T}}(\mathbf{V'c} - \mathbf{x'_c}) \tag{3.11}$$

The optimality conditions occur when the following system of equations holds. First set the gradient of $\mathcal{L}(\mathbf{c}, \lambda)$ with respect to $\mathbf{c}$ to 0:

$$\frac{\partial \mathcal{L}(\mathbf{c}, \lambda)}{\partial \mathbf{c}} = 2\mathbf{c} + \mathbf{V'}^{\mathrm{T}}\lambda = 0 \tag{3.12}$$

and the gradient with respect to λ to 0:

$$\frac{\partial \mathcal{L}(\mathbf{c}, \lambda)}{\partial \lambda} = \mathbf{V}'\mathbf{c} - \mathbf{x}'_{\mathbf{c}} = 0 \tag{3.13}$$

Equation 3.12 yields the following:

$$\hat{\mathbf{c}}_{mn} = -\frac{1}{2}\mathbf{V}'^{\mathrm{T}}\lambda \tag{3.14}$$

Plugging this solution into $\mathbf{V}'\mathbf{c} = \mathbf{x}'_{\mathbf{c}}$ from Equation 3.13 leads to:

$$\mathbf{V}'\hat{\mathbf{c}}_{mn} = -\frac{1}{2}\mathbf{V}'\mathbf{V}'^{\mathrm{T}}\lambda \Rightarrow \lambda = -2(\mathbf{V}'\mathbf{V}'^{\mathrm{T}})^{-1}\mathbf{x}'_{\mathbf{c}} \tag{3.15}$$

Inserting this solution to Equation 3.14 yields the following answer:

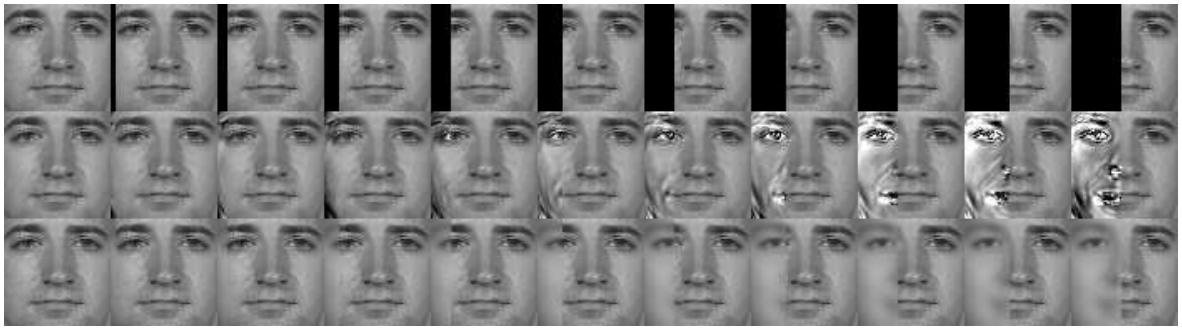
$$\hat{\mathbf{c}}_{mn} = -\frac{1}{2}\mathbf{V}'^{\mathrm{T}}\lambda = \mathbf{V}'^{\mathrm{T}}(\mathbf{V}'\mathbf{V}'^{\mathrm{T}})^{-1}(\mathbf{x}' - \mathbf{m}') \tag{3.16}$$

The vector $\hat{\mathbf{c}}_{\mathrm{mn}}$ represents the *minimum-norm* in ℓ_2 that can be used analogously to Equation 3.5 to reconstruct a new image $\hat{\mathbf{x}}_{\mathrm{mn}} = \mathbf{V}\hat{\mathbf{c}}_{\mathrm{mn}} + \mathbf{m}$.

To compare the different reconstructions, we set up a toy example where we introduce an artificial occlusion on a test face, as mentioned earlier, and try to reconstruct the occlusion using the methods described above. The test face belongs to a subject not contained in the training set. We consider different image resolutions to cover the overdetermined case ($d' > N$) and underdetermined case ($d' < N$). We compare the resulting minimum-norm and least-square error reconstructions for increasing levels of occlusion for the underdetermined case ($d' < N$). Figures 3.6a and 3.6b depict the occluded face, the least-square error reconstructions (after reducing the number of eigenfaces to keep Equation 3.8 feasible) and the minimum ℓ_2-norm reconstruction, for images of size 100×100 pixels and 64×64 pixels respectively.

As depicted in Figures 3.6a and 3.6b, both methods struggle to reconstruct the full face when

(a) 100×100 pixels



(b) 64×64 pixels

Figure 3.6: Reconstruction of occluded faces of different sizes. For each size, the top row depicts the occlusion level. The middle row depicts the ℓ_2 least-square-error solution while the bottom row depicts the ℓ_2 minimum norm solution.

the level of occlusion increases. The least-square error reconstruction, which we will refer to as "L2-PCA" for the remainder of this document, suffers from numerical instability, while the minimum-norm reconstruction degenerates towards the mean face in the dimensions that are occluded.

3.2.5 ℓ_1-norm Sparse Feature Extraction in Subspace with Missing Dimensions

To address the challenges mentioned at the end of subsection 3.2.4, we need to find a solution that doesn't suffer from the invertibility problem and can reconstruct the faces when the size of the subspace increases. One possible way, is to find a sparser solution for the coefficient vector c using ℓ_1-minimization. A sparse solution will allow us to represent individual faces in clusters of eigenvector bases since the face data is more likely to be multimodal, while PCA assumes unimodality. Moreover, in ℓ_1-minimization literature, underdetermined problems are usually the norm rather than the exception, and N can largely exceed d'.

ℓ_1-minimization has been studied since the 1970s, and has been known to return a sparse solution. Initially used in reflection seismology in geophysics, it has experienced a very high-profile resurgence with the emergence of Compressed Sensing (CS) theory [44] where, under some assumptions on the sensing matrix, it has been shown to be an efficient equivalent to the so-called "ℓ_0-norm"approach to recover the sparsest solutions to certain undetermined system of linear equations [45]. *Compressed Sensing* corresponds to the idea of encoding a large sparse signal using a relatively small number of measurements, and minimizing the ℓ_1-norm (or one of the many variants) to decode the original signal. It is attractive for numerous applications because it reduces the number of measurements required to reconstruct a given amount of information. The trade-off is the need to develop a non-trivial decoding process. See Appendix A for a brief overview of CS and ℓ_1-minimization solvers.

The standard $\ell 1$-minimization problem solves the following convex program:

$$\min J(\mathbf{c}) = \|\mathbf{c}\|_1 \text{ subject to } \mathbf{V}'\mathbf{c} = \mathbf{x}'_{\mathbf{c}} \tag{3.17}$$

Equation 3.17, known as *Basis Pursuit* (BP) [46], finds the vector with smallest ℓ_1 norm of vector $\mathbf{c}$ defined as $\|\mathbf{c}\|_1 = \sum_{i=1}^{d} |c_i|$.

As the results in [45] show, if a sufficiently sparse $\mathbf{c_0}$ exists such that $\mathbf{V}'\mathbf{c_0} = \mathbf{x}'_{\mathbf{c}}$, then Equation 3.17 will find it. In the presence of noise, Equation 3.17 becomes:

$$\min \|\mathbf{c}\|_1 \text{ subject to } \|\mathbf{V}'\mathbf{c} - \mathbf{x}'_{\mathbf{c}}\|_2 \leq \epsilon \tag{3.18}$$

for a given ϵ and is known as *Basis Pursuit Denoising* (BPDN) [47] in the signal processing community and *Lasso* [48] in the statistical community. Moreover, there are numerous variants of Equations 3.17 and 3.18 that are all sparsity-inducing techniques.

CS theory shows that a sparse signal can be recovered provided the measurement matrix possesses certain desirable properties. To date, random matrices and matrices whose rows are taken from orthonormal matrices have been proven to be desirable. These matrices are invariably dense, which contradicts the usual assumption made by conventional optimization solvers. Dedicated algorithms for such signal reconstruction has been developed since the early days of CS. There are several classes of methods that seek to solve the previous equation, namely *Gradient Projection* (GP) [49], *Homotopy* [50], *Iterative Shrinkage-Thresholding* (IST) [51, 52], *Proximal Gradient* (PG) [49, 53], and *Augmented Lagrange Multiplier* (ALM) [54].

The intrinsic details of specific solvers is beyond the scope of this analysis, however so far a solver based on (ALM) [55] has consistently outperformed other solvers in the case of our feature-extraction and face reconstruction application, and will be used for the remainder of this study (see Appendix A for a brief overview of the ALM solver). We will refer to this method as "L1-PCA". Figures 3.7a and 3.7d depict the reconstructions obtained via L1-PCA and L2-PCA. Figures 3.7

show that the sparse reconstruction is smoother and more visually pleasing at all resolutions and avoids introducing artifacts.

3.3 Pose Tolerance Induced by Sparse Representation

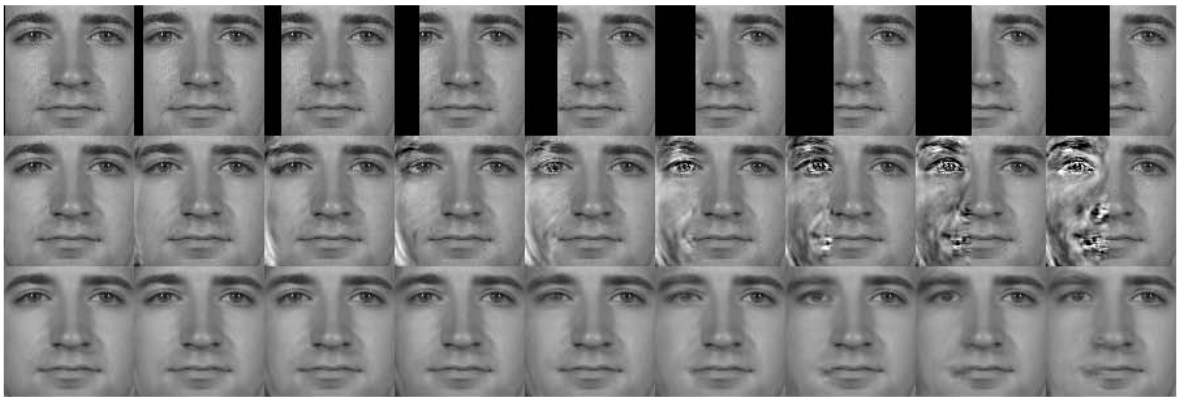
The reconstructions achieved by L1-PCA (and depicted in Figure 3.7) are nothing short of remarkable compared to L2-PCA and given the challenge that nearly half of the input pixels are missing. This highlights the capabilities of the sparse feature extraction induced by ℓ_1-minimization. In the next few subsections we empirically show the advantages of the sparse feature extraction by measuring its resilience to occlusion, which as a reminder, in our case is the dual problem of pose-tolerance.

3.3.1 Sparsity Analysis

In this section, we measure and compare the sparsity or the number of nonzero coefficients of the different feature extraction schemes on the synthetic occlusion toy dataset presented earlier. Figure 3.8a depicts the histogram of the values of the L2-PCA coefficient vector of the synthetic toy example at the 2% occlusion level (which corresponds to the first column of images in figure 3.7). The height of the bin at 0 represents the number of null coefficients. Figure 3.8b depicts the histogram of the values of the L1-PCA coefficient vector. It is very obvious that at this very low occlusion level (2%) and for a dimensionality of 6000 (which corresponds to the number of eigenfaces in our PCA basis), ℓ_1-minimization forced over half of the values in the coefficient vector to be zero, while the L2-PCA coefficient vector is not sparse at all (a mere 28 coefficients ouf of 6000 were set to zero). Figure 3.8d depicts the standard deviations of the distributions in Figures 3.8a and 3.8b as the occlusion level increases to 50%. The coefficient histograms of L2-PCA get wider and wider as the level of occlusion increases, which means that L2-PCA is struggling to fit a coefficient vector that best explains the observed data and is assigning arbitrary weights to

(a) 200×200 pixels

(b) 128×128 pixels



(c) 100×100 pixels

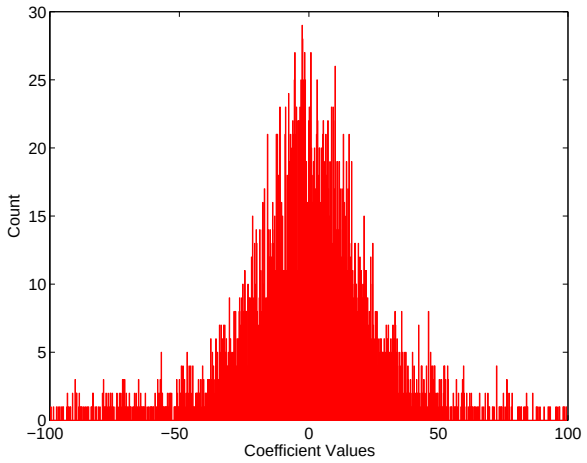
(d) 64×64 pixels

Figure 3.7: Reconstructing occluded faces of different sizes. For each size, the top row shows the occluded input images. The middle row shows the L2-PCA reconstructions. The bottom row shows the L1-PCA reconstructions.

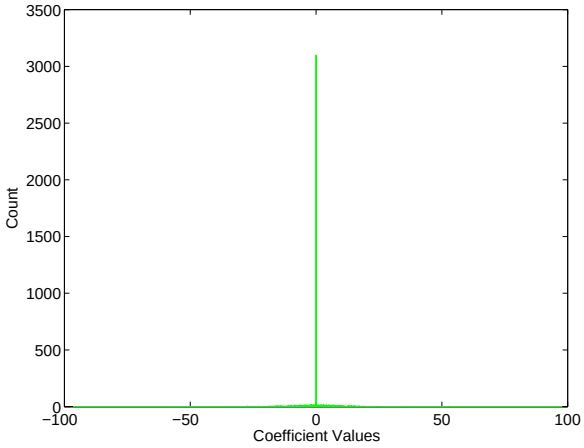
principal components across the entire basis. Figure 3.8d plots the count of null coefficients for the two methods as the level of occlusion increases. With L1-PCA, as the occlusion level increases to cover half of the face, the sparsity level reaches over 60%. This suggests that L1-PCA abstains from assigning weights arbitrarily when the fitting becomes challenging as a result of missing dimensions. This in returns enables it to maintain a natural looking reconstruction, because higher order principal components that represent mostly noise are not assigned a heavier weight as the number of available dimensions drops and the system becomes increasingly underdetermined.

3.3.2 Robustness Analysis

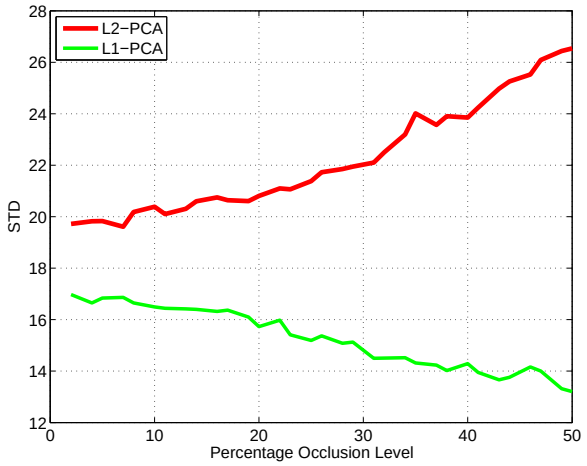
Another advantage of the sparse feature extraction method that we observed is its ability to keep finding a good solution, regardless of the size of the training data or the increasing availability of misleading training faces that might deviate the solution. To illustrate this concept, we consider the half-occluded face from the synthetic dataset, which is the most challenging problem. We then solve for the coefficients using L1-PCA and L2-PCA with a subspace consisting of increasing numbers of training subjects (all different from the test subject). Figure 3.9 compares the MSE between the original and reconstructed images as the number of training subjects in the subspace increase. What we observe is that the MSE initially drops as a result of having a more representative subspace, but then the added degrees of freedom adversely affects L2-PCA, while L1-PCA achieves and maintains a good reconstruction regardless of the number of training subjects in the subspace. This trend justifies our hypothesis that the training data becomes increasingly multimodal and that the unimodal assumptions of PCA hinders the L2-PCA solution. Essentially, it becomes harder to find a "good" and stable solution when the subspace contains a large number of individuals. We observe the same pattern regardless of the size of the image/subspace.



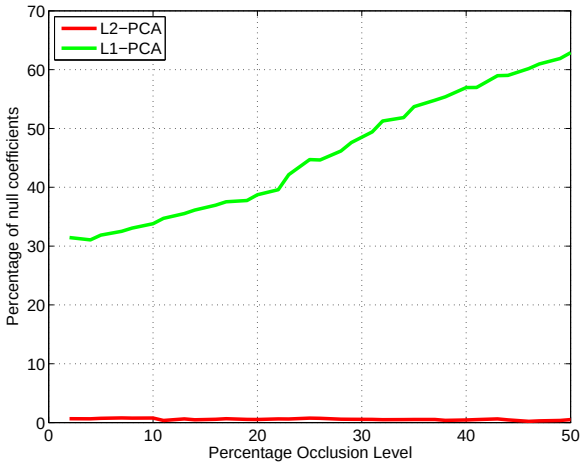
(a) Sparsity histogram of the L2-PCA coefficient at a 2% occlusion level.



(b) Sparsity histogram of the L1-PCA coefficient at a 2% occlusion level.

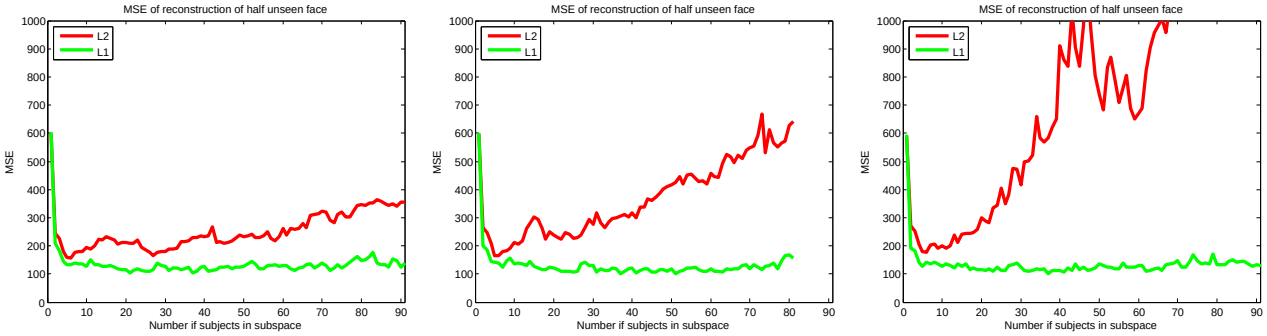


(c) Standard deviation of the distributions of (a) and (b) as a function of the percentage of occlusion level.



(d) Percentage of null coefficients as a function of the percentage of occlusion level.

Figure 3.8: Sparsity Analysis. Figures (a) and (b) plot the histogram of the coefficient values for L2-PCA and L1-PCA respectively at a 2% occlusion level. The bin at 0 represents the number of zero coefficients. Figure (c) plots the standard deviation of the histograms in (a) and (b) as a function of the increasing occlusion level. Figure (d) plots the sparsity or the number of null coefficients as a function of the occlusion level.



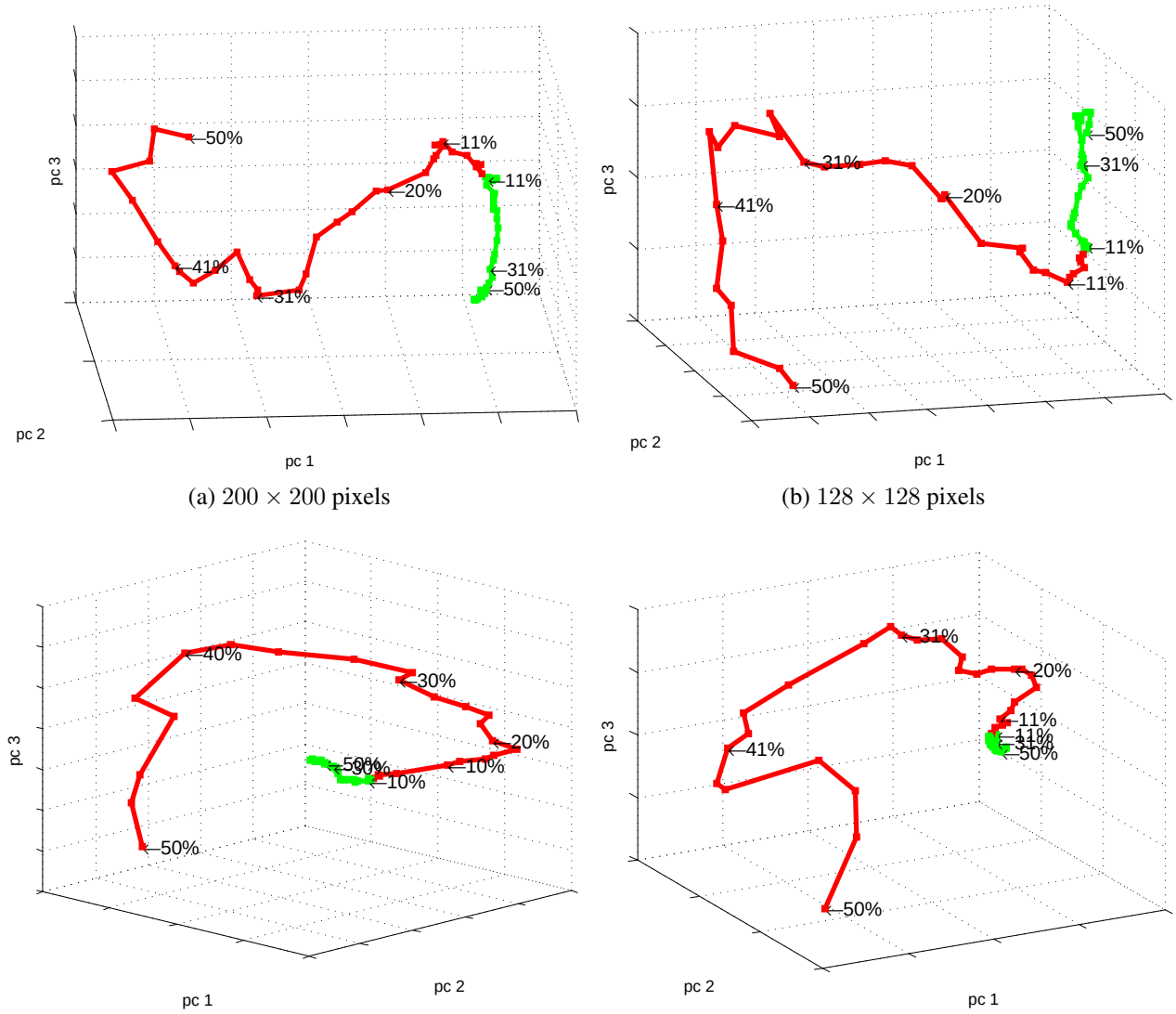
(a) 200 × 200 pixels (b) 128 × 128 pixels (c) 100 × 100 pixels

Figure 3.9: Mean-Square Error between the reconstructed half-face belonging to an unseen subject and the original full face produced by L2-PCA and L1-PCA using an increasing number of subjects in the training subspace.(a), (b) and (c) show the MSE plots at different image resolutions.

3.3.3 Stability Analysis

Achieving pose tolerance in this dual occlusion-space means that the coefficient that we extract from the observed data should minimally change as the number of observed dimensions drops. We empirically observed that the L1-PCA coefficient vector is stable for varying number of missing dimensions (pixels). To confirm this observation, we need to monitor the "drift" of fit as the occlusion increases. We could simply plot the first three coefficients of each method as function of occlusion, but this will not give a holistic view of the change that could be happening further down the higher order dimensions. Instead, to check the overall relative stability of the L1-PCA solution compared to the L2-PCA solution, we build a joint subspace of L2-PCA and L1-PCA coefficients, and project these coefficients onto this subspace. Figure 3.10 depicts the 3D plots of the first three coefficients for both reconstructions as the occlusion increases for varying resolutions.

The clustering of the L1-PCA coefficients, shown in Figure 3.10, in the joint coefficient space implies that the sparse approach, when applied to pose-corrected images, is critical to producing consistent reconstructions (as the yaw angle of the input image varies), thus resulting in a pose-tolerant feature representation. In Chapter 4 we will show that the stability of this sparse feature extraction step is crucial for face recognition. The minimal drift of the feature as the occlusion



(a) 200 × 200 pixels

(b) 128 × 128 pixels

(c) 100 × 100 pixels

(d) 64 × 64 pixels

Figure 3.10: 3D plots of the first three principal components of the L2-PCA (red) and L1-PCA (green) reconstruction coefficients in the joint coefficient space for varying occlusion levels. These plots enable us to compare the evolution of the reconstruction coefficients as the occlusion increases. The L1-PCA coefficients (in green) seem to be better clustered (yielding a more invariant set of features with less variation) than their L2-PCA counterparts (in red) implying a higher level of pose-tolerance. The different subplots (a),(b),(c) and (d) correspond to different image resolutions. The occlusion levels are indicated along the 3D curve.

increases (which represents increasingly non-frontal viewpoints) means that it can effectively represent the identity of the subject. The clustering of the feature vector in coefficient space thus translates into a clustering of the feature vector in identity space, as coefficients extracted from different subjects (and different viewpoints) will be maximally separated in the feature space. Section 4.3.2 will investigate this concept further.

3.3.4 A Compressed Sensing Perspective

ℓ_1-minimization is the workhorse of Compressed Sensing (briefly reviewed in Appendix A). A great deal of theoretical work has been undertaken to prove when and why ℓ_1-minimization is able to recover a signal. From a compressed sensing perspective, it is assumed that an image $\mathbf{x}$ of size d needs to be recovered from d' measurements (such that $d' < d$). It is also assumed $\mathbf{x}$ has a sparse linear expansion in a basis $\mathbf{\Psi}$ such that $\mathbf{x} = \mathbf{\Psi c}$ and that $\mathbf{c}$ is S-sparse (i.e contains S non-zero coefficients). Assume only d' total measurements of $\mathbf{x}$ are observed, where every measurement i is obtained by the inner product of the i^{th} row $\boldsymbol{\phi}_i$ of a measurement matrix $\mathbf{\Phi}$ with the image, given by $\langle \boldsymbol{\phi}_i, \mathbf{x} \rangle$. The coherence of the two basis, given by $\mu(\mathbf{\Psi}, \mathbf{\Phi}) = d. \max_{i,k} |\langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_k \rangle|$, where $\boldsymbol{\psi}_k$ is the k^{th} column of $\mathbf{\Psi}$, was introduced in [56] to represent how "distant" the two basis are from each other. It was shown in [56] that if d' is greater than $\mu\left(\mathbf{\Psi}, \mathbf{\Phi}\right) .S. \log(d)$, then ℓ_1-minimization will recover the image $\mathbf{x}$ with very high probability.

From this incoherence view of compressed sensing, we can justify why our L1-PCA derivation can recover the image. In our pose-correction formulation, the sparse basis $\mathbf{\Psi}$ is the matrix of eigenfaces $\mathbf{V}$. The image $\mathbf{x}$ has a fundamentally sparse basis expansion in $\mathbf{V}$ as most of the energy is concentrated in the first few eigenfaces that carry most of the variation. The measurements are simply the d' un-occluded pixels and the measurement basis $\mathbf{\Phi}$ is therefore the standard canonical basis (it will essentially be an identity matrix missing some of its rows that correspond to the self-occluded pixels). The coherence between these two bases is minimal, and CS theory indicates that despite a significant proportion of missing pixels, we can still reconstruct the original image

$\mathbf{x}$ using ℓ_1-minimization. For that reason, we do not need complex dictionary learning methods to be able to reconstruct the image, as the simple PCA subspace offers us a sparse representation. One could still learn a purpose-specific dictionary that might generate a better reconstruction, but we leave this step as future work, and focus on the combination of PCA representation and ℓ_1-minimization to recover the image with missing dimensions.

3.4 Subspace Modeling to Synthesize Pose-Corrected Faces

The primary purpose of this chapter is to extract a pose-tolerant feature vector from non-frontal images. The 3D-modeling de-rotation step transformed the pose-correction problem to an occlusion problem. In the previous section, we used ℓ_1-minimization to extract a sparse representation from a subspace with missing dimensions and showed that this sparse representation is stable under varying degrees of occlusion. We validated our results by reconstructing the synthetic occlusion toy example and checking the quality of the reconstruction. We can do the same on the pose-corrected face images. As previously mentioned, we solve for the vector $\mathbf{V}\hat{\mathbf{c}}_{\text{L1-PCA}}$ from missing data, but we can then reconstruct a full dimension image:

$$\hat{\mathbf{x}}_{\text{L1-PCA}} = \mathbf{V}\hat{\mathbf{c}}_{\text{L1-PCA}} + \mathbf{m} \tag{3.19}$$

A slightly more accurate reconstruction approach is to avoid reconstructing the observed pose-corrected pixels, and to restrict the reconstruction to the pixels that are self-occluded. Let OCCLUSION denote the set of occluded pixels detected with the technique described in section 3.2.2 and depicted in Figure 3.5b. Let $\hat{\mathbf{x}}_{\text{STRETCH}}$ denote the pose-corrected image rendered by 3DGEM at a frontal viewpoint and with no occlusion detection. Then for every pixel i, we can define:

$$\hat{\mathbf{x}}^{(i)}_{\text{HYBRID}} = \begin{cases} \hat{\mathbf{x}}^{(i)}_{\text{L1-PCA}} & \text{if} \quad i \in \text{OCCLUSION} \\ \hat{\mathbf{x}}^{(i)}_{\text{STRETCH}} & \text{else} \end{cases} \tag{3.20}$$

$\hat{\mathbf{x}}_{\text{HYBRID}}$ represents the hybrid shape-free image that only reconstructs the occluded parts of the pose-corrected face. Figures 3.12, 3.13, and 3.14 compare the hybrid synthesis to the L1-PCA synthesis. In general, the differences are very subtle and only become prominent in the presence of high-frequency information such as facial hairs, moles, eye-glasses, etc. This is due to the fact that L1-PCA's sparse representation tends to suppress the higher-order principal components that are usually responsible for high-frequency components. To make this document more readable, we now introduce labels for the different image representations that we have presented so far and that we will keep encountering in this thesis:

- HOLES: denotes the shape-free representation that contains blank pixels which represent the pixels likely to exhibit rendering artifacts, as detected by the 3D occlusion detection module presented in section 3.2.2.
- STRETCH: denotes the shape-free representation where we disable the occlusion detection module, and let the texture mapping component of 3DGEM estimate all pixel values of the pose-corrected face. As explained earlier, this will inevitably introduce stretching artifacts, because of the scarcity of available pixel information for texture mapping to operate on.
- L1-PCA: denotes the shape-free representation generated by Equation 3.19 which reconstructs the entire face using the sparse feature vector fit on the non-occluded pixels observed in HOLES.
- L2-PCA: denotes the shape-free representation generated by Equation 3.9 which reconstructs the entire face using the minimum-squared error feature vector fit on the non-occluded pixels observed in HOLES.
- HYBRID: denotes the shape-free representation generated by Equation 3.20, which mixes the non-occluded pixels from STRETCH and the occluded pixels from L1-PCA.

The next chapter will empirically show the discriminative properties of the feature vector $\hat{\mathbf{c}}_{\text{L1-PCA}}$ and how we can use it to improve verification rates, but in this section we showcase the

qualitative properties of the L1-PCA and HYBRID reconstructions. The shape-free representation of section 3.2.3 is ideally-suited for synthesis as well as analysis. So the following reconstructions are shown using the shape-free representation.

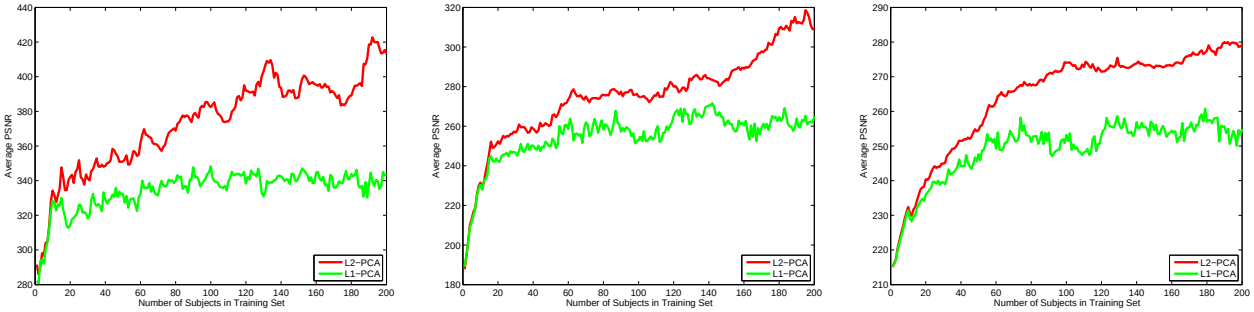
3.4.1 Correcting Yaw

To illustrate our pose-correction reconstruction results, we employ the CMU Multi-PIE (MPIE) database [36]. It offers multiple viewpoints of a total of 337 subjects split into four sessions and spread across a wide of yaw angles. Currently, for yaw-correction, we limit our method to handle the $[-45^\circ \ +45^\circ]$ range. This range guarantees that we still observe both eyes which is what our automatic facial landmarking scheme and 3DGEM currently require. Note that this is not a limitation of our framework, rather one imposed by the landmarking and 3D modeling steps, and in future work, we plan to eliminate this limitation. For given input images of varying yaw, we build a non-frontal 3DGEM model, de-rotate the model and render it at 0°, and detect the pixels likely to be occluded. We then apply the different reconstruction methods mentioned in the previous section. In our comparison we also include the case where we disable the occlusion detection module resulting in a representation we denoted previously as "STRETCH". Figures 3.12c indicates the locations of the pixels likely to be corrupted by 3D rendering artifacts and whose representation we called "HOLES". Figures 3.12d and 3.12e depict the L2-PCA and L1-PCA reconstruction methods respectively. Figure 3.12f depicts the "HYBRID" representation. Figures 3.13 and 3.14 offer the same comparisons for different subjects. Appendix D contains more sample reconstructions for different subject in MPIE and for subjects from the FERET [57] pose database to show this technique is not database-dependent. It is obvious from Figures 3.12, 3.13 and 3.14 that the sparse reconstruction is more visually pleasing, and more natural looking. Beyond 15°, the stretching artifacts induced by standard 3DGEM (as illustrated in Figure 3.12b) become too apparent, and the inconsistency of the L2-PCA reconstruction (Figure 3.12d) introduces unnatural artifacts that are reminiscent of numerical instability.

To confirm the stability of the sparse feature extraction method (described in section 3.3.2), we run a slightly different experiment. While the experiment in section 3.3.2 was performed in the dual occlusion-space, and increased the size of the training set by adding more unseen subjects, the following experiment takes place in the primal pose-correction space. The training subspace always starts by featuring the original **frontal** face of the testing subject, but then we increase the training set by adding more *unseen* subjects. We then check the reconstruction of the **pose-corrected** image of the test subject for the different methods using this increasing training set. We measure the MSE between the pose-corrected image and the original frontal image of the test subject. We repeat this procedure for ten different subjects and average the MSE results. Figure 3.11 plots the MSE between reconstructed pose-corrected images and the original frontal images as we increase the number of training "texture" subjects. Note that both methods start with a very low MSE, because at the first iteration the subspace is built on frontal faces of the test subject only. But at the next iteration, the subspace consists of more unseen frontal faces, and as a result the reconstruction error will increase. However, observe that regardless of the input yaw angle, the L1-PCA average MSE stabilizes after a certain number of training subjects (which include the test subject), while the L2-PCA average MSE keeps climbing with the number of training subjects. This confirms our previous finding (of section 3.3.2) that L1-PCA is less prone to "lose" the correct solution and always seems to converge to the true solution regardless of the number of available misleading exemplars in the training set.

3.4.2 Correcting Pitch

So far, our analysis and experimental results have focused on correcting yaw. However, the same technique can be used to correct the pitch problem. A nonzero pitch angle is very likely to be observed in challenging faces to identify. In a surveillance scenario, the cameras will most probably be placed high up on a wall or a ceiling to provide the camera with a vantage viewpoint. This in return will make subjects appear with a high degree of pitch. Assuming we can accurately estimate



(a) $-45°$ Pose Corrected (b) $-30°$ Pose Corrected (c) $-15°$ Pose Corrected

Figure 3.11: Average PSNR for 10 different faces computed between pose-corrected and frontal shape-free images, for a varying number of subjects in the training set. The training set always starts by having the test subjects, then adds more subjects. (a), (b), (c) show results at different yaw angles.

this pitch angle, the pose-correction technique we described above can be applied with minimal change to generate a frontal-looking corresponding image. However, estimating pitch is slightly more challenging than estimating yaw. This is because the face goes through less deformation when pitch changes compared to yaw, and a $10°$ change in pitch is barely discernible while it is very obvious in the case of yaw change. The MPIE database does not offer as much variation in pitch than it does in yaw, so for the following experiment, we captured our own images to prove the concept. We set up the camera at a known distance from the subject, and increase its height in such a way that the apparent pitch angle of the face is a known value. Figure 3.15a depicts the resulting face image for a pitch range of 0 to 40 with increments of $5°$. The following columns of the same figure depict the shape-free representation with occlusion detection, the resulting L1-PCA and HYBRID reconstructions respectively.

3.4.3 Generating a Standard Facial Crop

After we synthesise a shape-free pose-corrected image, we can go one extra step in adding the shape back in the reconstructed image. This is essentially reversing the process described in section 3.2.3 and depicted in Figure 3.3. We can build a 3DGEM frontal modal from the reconstructed

face image (whose x and y landmark coordinates are known and are common for all reconstructed faces) and then replace the mean coordinates with the actual pose-corrected coordinates of the original input face. This will void the shape-free representation and give the reconstructed faces their individual shape information back[1]. We can then take the reconstructed face with original shape information and crop it to a specific size (for example 128×128 pixels) and using given control points (for example crops aligned using the two eyes). Figure 3.16 compares our pose-corrected traditional crop, to another traditional approach which consists of warping the non-frontal images to a frontal mean shape and then cropping by aligning the face using both eyes. We benchmark against a piecewise-linear local warping scheme (see [35]) which is typically used in Active Appearance Models (AAM) [58] and seems to be one of the very few local warping schemes that can handle the severe distortion due to varying yaw. As we can observe in Figure 3.16, our L1-PCA pose-corrected traditional crop (with shape information) renders a smoother and more realistic face image while local 2D warping introduces severe distortions that essentially renders the image useless for face recognition purposes.

3.5 Summary of Findings and Results

In this chapter we presented a technique to extract a sparse feature vector from non-frontal images and showed that this feature vector is stable under varying poses. We will use this sparse feature vector to perform one-to-one face verification in the next chapter. We also used this sparse feature vector to synthesize a pose-corrected natural-looking image that corresponds to the non-frontal face image. Thus, these are the main contributions of this chapter:

- Extended the 3DGEM formulation to handle non-frontal images and produce a 3D model from a non-frontal input face.
- Introduced an artifact-free shape-free representation based 3DGEM which we showed is

[1]The next chapter will discuss the shape information in detail and its impact on face recognition and its use in this pose-correction framework

ideal for texture analysis and synthesis.

- Introduced a 3D occlusion detection technique that enables us to identify which pixels are likely to be occluded and hence be badly rendered when de-rotating the 3D model to a frontal viewpoint.
- Formulated the pose problem as an occlusion problem. After the 3D de-rotation step, the shape-free pose-corrected faces exhibit a number of occluded pixels which is a function of the input yaw. The more severe the input angle, the wider the occlusion band is.
- Formulated a solution to find a sparse feature vector using ℓ_1-minimization in the presence of mission dimensions (corresponding to occluded pixels). We showed this solution to have the following crucial advantages:
  - ⋆ Handles severely underdetermined systems of equations
  - ⋆ Does not suffer from numerical instability issues
  - ⋆ Is not prone to "lose" the correct solution when the number of training images increases
  - ⋆ Is stable and changes minimally when the degree of occlusion (or number of missing dimensions) changes
  - ⋆ Does not degenerate towards the mean
- Empirically showed how our pose-correction technique handles variations in both pitch and yaw.
- Demonstrated how we can use the extracted sparse feature vector to reconstruct or synthesize a frontal looking face image.
- Demonstrated how to transform an image to a shape-free domain suitable for texture analysis and synthesis, and then how to put the shape information back to generate traditional image crops.

(a) Original face image with yaw varying from -45° to $+45^\circ$ in increments of 15°.

(b) STRETCH: Pose-corrected images in shape-free representation with no occlusion detection and no reconstruction.

(c) HOLES: Pose-corrected images with occlusion detected in shape-free representation.

(d) L2-PCA reconstructed pose-corrected images in shape-free representation.

(e) L1-PCA reconstructed pose-corrected images in shape-free representation.

(f) HYBRID: L1-PCA+STRETCH pose-corrected images in shape-free representation.

Figure 3.12: MPIE subject 1 with (a) the original images at different yaw angles (b) The equivalent pose-corrected shape-free images with no occlusion detection. As a result, the texture mapping module in 3DGEM will introduce stretching artifacts due to interpolating from too few available pixels. (c) The equivalent pose-corrected shape-free images with pixels that are likely to be occluded detected in 3D. (d) The equivalent reconstructed pose-corrected images using L2-PCA in shape-free representation. (e) The equivalent reconstructed pose-corrected images using L1-PCA in shape-free representation. (f) Hybrid representation which consists of the un-occluded pixels from (c) and occluded pixels from (e). The sparse feature vector responsible for this reconstruction will be used (in the next chapter) to achieve one-to-one verification rates.

(a) Original face image with yaw varying from -45° to $+45^\circ$ in increments of 15°.

(b) STRETCH: Pose-corrected images in shape-free representation with no occlusion detection and no reconstruction.

(c) HOLES: Pose-corrected images with occlusion detected in shape-free representation.

(d) L2-PCA reconstructed pose-corrected images in shape-free representation.

(e) L1-PCA reconstructed pose-corrected images in shape-free representation.

(f) HYBRID: L1-PCA+STRETCH pose-corrected images in shape-free representation.

Figure 3.13: MPIE subject 2 with (a) the original images at different yaw angles (b) The equivalent pose-corrected shape-free images with no occlusion detection. As a result the texture mapping module in 3DGEM will introduce stretching artifacts due to interpolating from too few available pixels. (c) The equivalent pose-corrected shape-free images with pixels that are likely to be occluded detected in 3D. (d) The equivalent reconstructed pose-corrected images using L2-PCA in shape-free representation. (e) The equivalent reconstructed pose-corrected images using L1-PCA in shape-free representation. (f) Hybrid representation which consists of the un-occluded pixels from (c) and occluded pixels from (e). The sparse feature vector responsible for this reconstruction will be used (in the next chapter) to achieve one-to-one verification rates.

(a) Original face image with yaw varying from -45° to $+45^\circ$ in increments of 15°.

(b) STRETCH: Pose-corrected images in shape-free representation with no occlusion detection and no reconstruction.

(c) HOLES: Pose-corrected images with occlusion detected in shape-free representation.

(d) L2-PCA reconstructed pose-corrected images in shape-free representation.

(e) L1-PCA reconstructed pose-corrected images in shape-free representation.

(f) HYBRID: L1-PCA+STRETCH pose-corrected images in shape-free representation.

Figure 3.14: MPIE subject 3 with (a) the original images at different yaw angles (b) The equivalent pose-corrected shape-free images with no occlusion detection. As a result the texture mapping module in 3DGEM will introduce stretching artifacts due to interpolating from too few available pixels. (c) The equivalent pose-corrected shape-free images with pixels that are likely to be occluded detected in 3D. (d) The equivalent reconstructed pose-corrected images using L2-PCA in shape-free representation. (e) The equivalent reconstructed pose-corrected images using L1-PCA in shape-free representation. (f) Hybrid representation which consists of the un-occluded pixels from (c) and occluded pixels from (e). The sparse feature vector responsible for this reconstruction will be used (in the next chapter) to achieve one-to-one verification rates.

(a) Original (b) HOLES (c) L1-PCA (d) HYBRID

Figure 3.15: Pose correction for pitch angles increasing from 0° to 40° in increments of 5. (a) shows the original input images (b) shows the corresponding shape-free representation with occlusion detection (c) shows the resulting L1-PCA shape-free reconstructions (d) shows the HYBRID representation which consists of the the original pixels when the pixel is not occluded, and the L1-PCA equivalent pixel when the pixel is occluded.

(a) Original face images with yaw varying from $-45°$ to $+45°$ in increments of $15°$.

(b) 2D warped images obtained using local piecewise-linear warping.

(c) 2D warped images obtained using local weighted mean warping.

(d) Cropped L1-PCA reconstructed pose-corrected images with shape information.

Figure 3.16: Comparison with traditional cropping. MPIE subject 2 with (a) original images at different yaw angles. The equivalent pose-corrected images using traditional 2D warping methods are shown using piece-wise linear (b) and local weighted-mean (c). The equivalent crops obtained using L1-PCA after adding the shape back are shown in (d).

Chapter 4

Face Recognition with Pose Correction

In the previous chapter, we introduced a pose-correction method that consists of three main components: First, a 3D modeling step (3DGEM) that builds a 3D-model from a non-frontal face image. We can render this model at any arbitrary angle, and we choose 0° to generate a frontal-looking face. Second, we showed that this 3D model is not sufficient to achieve pose-invariance, so we introduced a shape-free representation which is ideally suited for texture analysis and synthesis. Third, we introduced a subspace modeling approach that relies on sparse ℓ_1-minimization for feature extraction in the shape-free domain. We showed this feature to be stable under a wide variation of pose and produced a smooth and artifact-free reconstruction. These three components work together to let us achieve pose-tolerance. In this chapter, we show how we can utilize this sparse pose-tolerant feature to achieve one-to-one face verification. The *Modus operandi* of most experiments in this chapter is illustrated in the flowchart of Figure 4.1. We assume we have a large database of frontal or near-frontal faces and that the test images are non-frontal and of an arbitrary pose. We run extensive and large scale experiments on the MPIE database to validate the pose-tolerance of our sparse feature extraction. We also analyze the effect of shape information in a pose-correction framework, and compare pose-correction to pose-synthesis. We also present two algorithms to perform robust face recognition across pose variations.

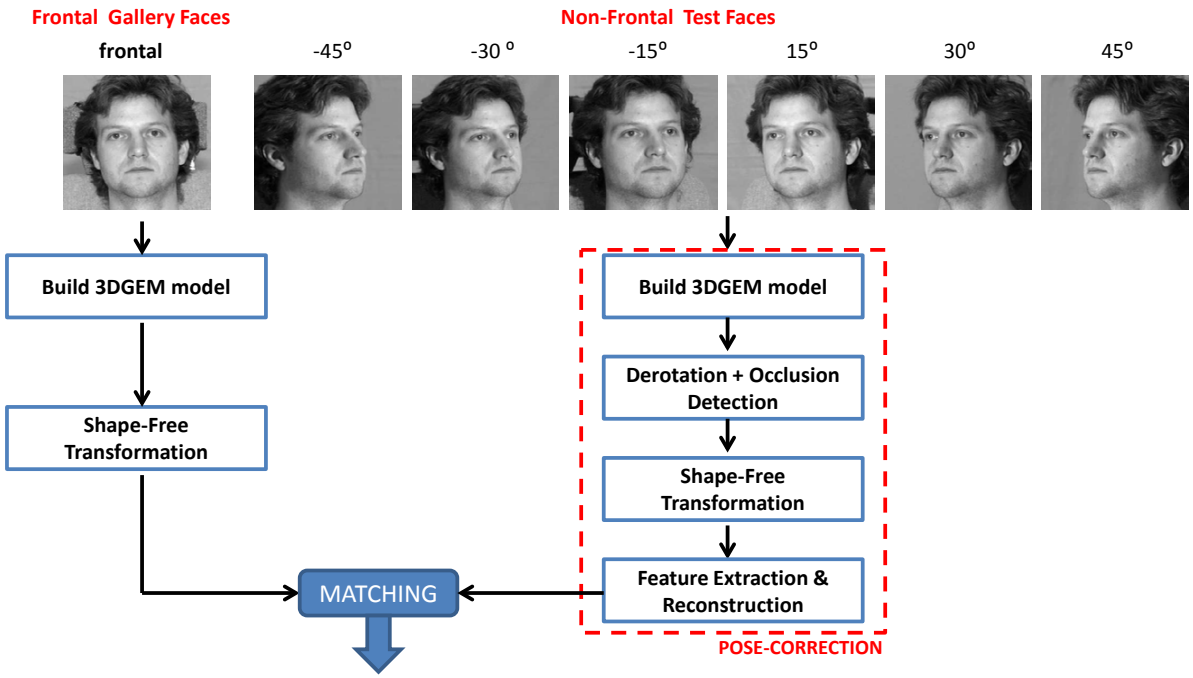


Figure 4.1: Experimental setup for face recognition with pose correction using our framework.

4.1 Analyzing Texture Information in Pose Corrected Images

The experimental setup of our first pose-correction test is depicted in Figure 4.1. Our gallery database consists of the frontal images[1] of the MPIE database. The query set consists of 6 non-frontal images [2] which roughly represent yaw angles ranging from $-45°$ to $+45°$ in increments of $15°$. Every frontal face goes through the shape-free transformation and then the feature extraction step. Every non-frontal face goes through the 3D step to fit a non-frontal 3D model, which is then rendered at $0°$ and has its occluded pixels detected. It then goes through the shape-free transformation, and then a feature extraction step just like the frontal face. All of our experiments are one-to-one, which means every pair of (frontal, non-frontal) faces is matched independently, and we have no prior information on the size of the gallery set, or the number of query images, and we do not perform any kind of score normalization. Unless we specifically say otherwise, the

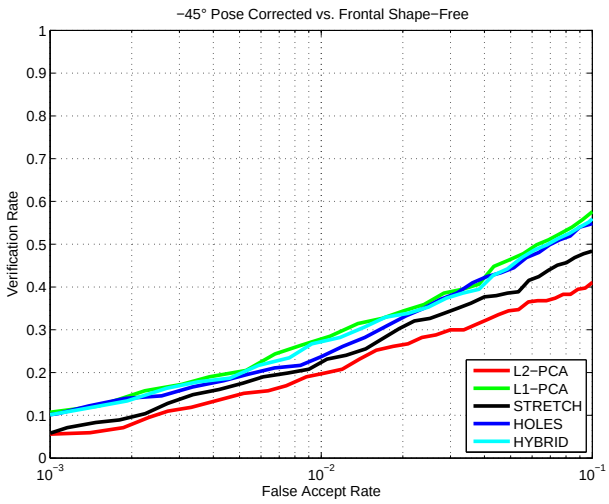
[1] angle code 05_1 in the MPIE technical description
[2] angle codes 19_0, 04_1, 05_0, 14_0, 13_0 and 08_0 in the MPIE technical description

matcher consists of a normalized cosine-distance (NCD) given by:

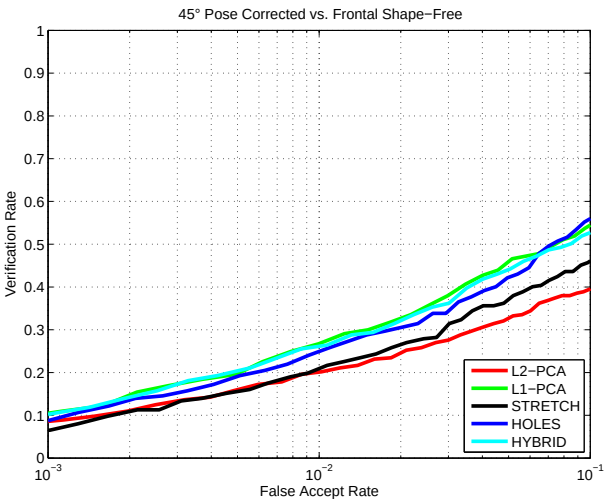
$$\mathrm{NCD}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \, \|\mathbf{b}\|} \tag{4.1}$$

where $\mathbf{a}$ and $\mathbf{b}$ are two vectors (features or vectorized images). Figure 4.2 depicts the ROCs obtained on all 6 different angles after the image synthesis step and using all the different shape-free image representations we defined in the previous chapter. These ROCs represent the normalized cosine distance performance between the reconstructed pose-corrected shape-free images and the frontal images seen in Figure 3.12. As a reminder, the black ROC corresponds to the STRETCH representation: shape-free images as rendered by the 3D model with no occlusion detection or reconstruction (such as those depicted in Figure 3.12b). The blue ROC corresponds to matching the HOLES representation: taking into account the non-occluded pixels (such as those depicted in Figure 3.12c) . The red and green ROC correspond to matching the L2-PCA and L1-PCA representations respectively (such as those depicted in Figures 3.12d and 3.12e). The cyan ROC corresponds to matching the HYBRID representation: pose-corrected images that mixes the observed un-occluded pixels from STRETCH and the occluded reconstructed pixels from L1-PCA (such as those depicted in Figure 3.12f).

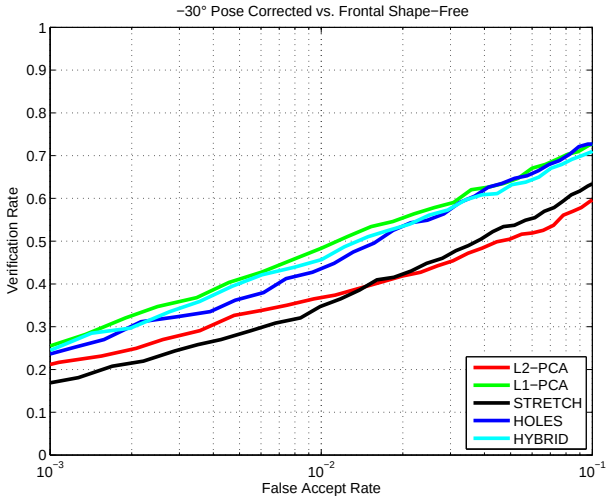
Analyzing Figure 4.2, we can conclude that no matter what the test angle is, the L2-PCA and STRETCH representations are consistently under-performing, and this is not surprising given the degree of visual distortion apparent in the reconstructed images. It is interesting to note that the L1-PCA representation very narrowly but consistently outperforms the HYBRID representation. Even though the HYBRID representation is more faithful to the observed data, its ROC shadows that of the L1-PCA representation. That is mostly because the smoothing resulting from the sparse reconstruction which suppresses the higher-order principal components is effectively denoising the images. The denoised images are performing better in a one-to-one verification framework than the original images. L1-PCA also narrowly beats the “HOLES” images except for very low FAR



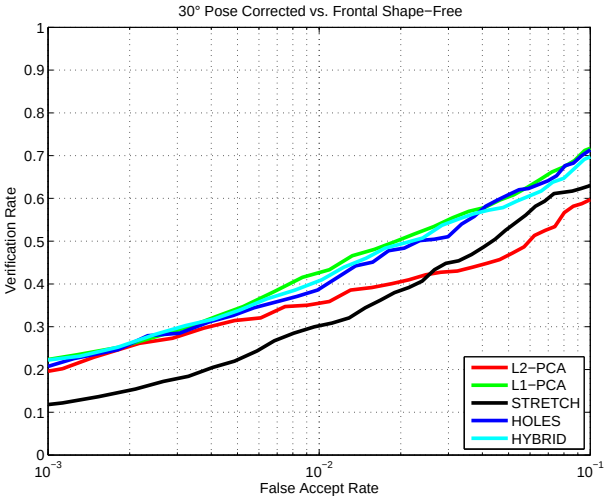
(a) Shape-free $-45°$ PoseCorrection



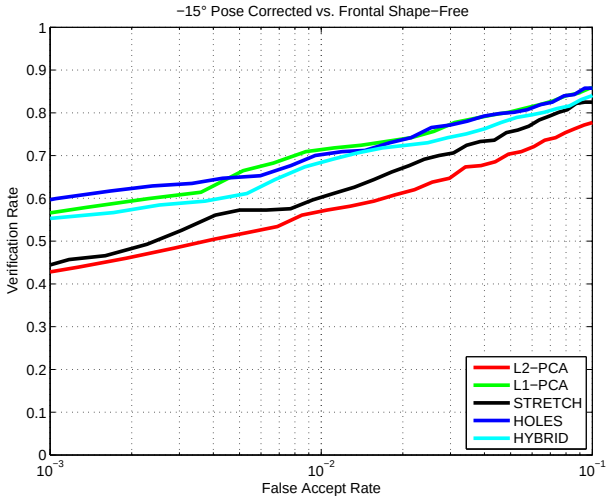
(b) Shape-free matching at $45°$



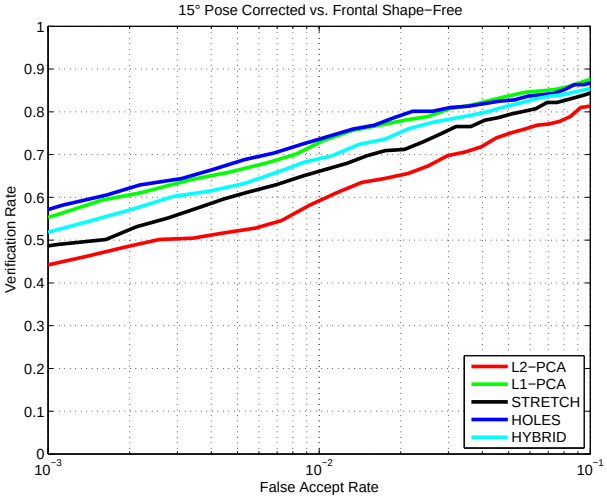
(c) Shape-free matching at $-30°$



(d) Shape-free matching at $30°$



(e) Shape-free matching at $-15°$



(f) Shape-free matching at $15°$

Figure 4.2: ROC of verification performance of pose-corrected shape-free reconstructed images on the MPIE database consisting of 337 unique subjects. The gallery images are frontal and the query images are at the angle indicated in the caption.

in the 15° case. The "HOLES" result is interesting, given that this representation is an intermediate step between L1-PCA and HYBRID. It measures the discriminative information available in the native image with no extraneous subspace reconstructions. Though the result is remarkable, this representation is not very usable, since little can be done with a shape-free image full of blank pixels. If we train an image classifier with this image representation, the algorithm will inevitably try to learn the pattern of blank pixels, which changes with every image. Therefore, for the remainder of this thesis, we will drop the comparison against the "STRETCH" and "HOLES" options and focus mainly on L1-PCA and HYBRID for feature extraction and reconstruction. Overall, the actual verification score remains poor, but the remaining sections of this chapter will provide some ideas on how to improve them.

The above result represents shape-free images, which is only a one-sided view of face information. In [59], the individual contributions of shape and texture information have been studied in a large scale analysis involving frontal faces. It was empirically shown that shape information is reasonably discriminative, and demonstrated significant improvement in performance by analyzing shape and texture information separately and fusing them to boost recognition rates. In the next section we extend the study of shape information study for non-frontal faces.

Table D.1[3] summarizes the verification rates for the different image representations measured at 1% FAR. We also include in the comparison the reconstructed images with shape information added back and the warped images using different warping schemes, such as those depicted in Figure 3.16.

4.2 Analyzing Shape Information in Non-frontal Faces

In [59], the effectiveness of shape information was demonstrated by a purely shape-based face recognition. A dense set of 79 landmark (or control) points were stacked with y coordinates fol-

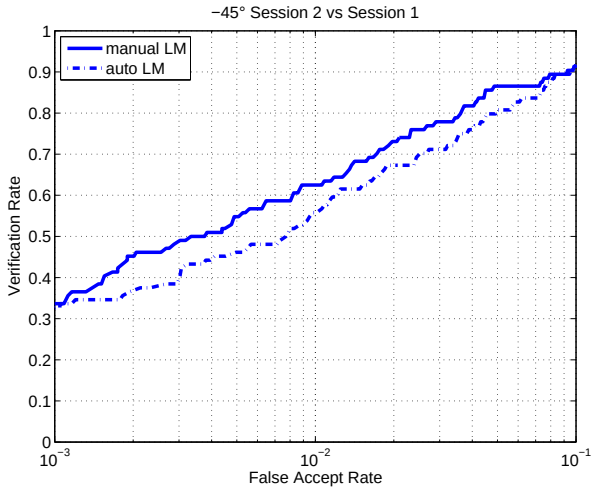
[3]See Appendix D for this result.

lowing x coordinates to form a feature vector that describes the shape of the face. Procrustes analysis [60] is used to align all shape vectors and remove all traces of rotation, scale and translation. It was shown in a large-scale analysis involving the FRGC database [61] that shape information alone can outperform the overall performance of the entire face.

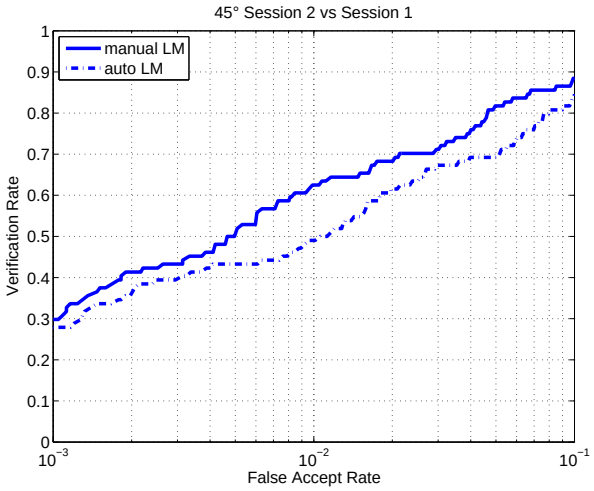
4.2.1 View-Based Shape Information

To investigate the effectiveness of shape information across pose, we recreate the same experimental setup on a per-view basis. For every test angle, we create a "generic" shape-only subspace from 88 subjects belonging to MPIE sessions 2 and 3. We then test by matching 249 MPIE session 1 subjects against 105 subjects from MPIE session 2. All those 105 subjects from session 2 are present in session 1. All 249 testing subjects are not seen in training. Figure 4.3 displays the ROCs for all 6 test angles. For every case we compare manual landmarking to automatic landmarking to validate the robustness of the findings.

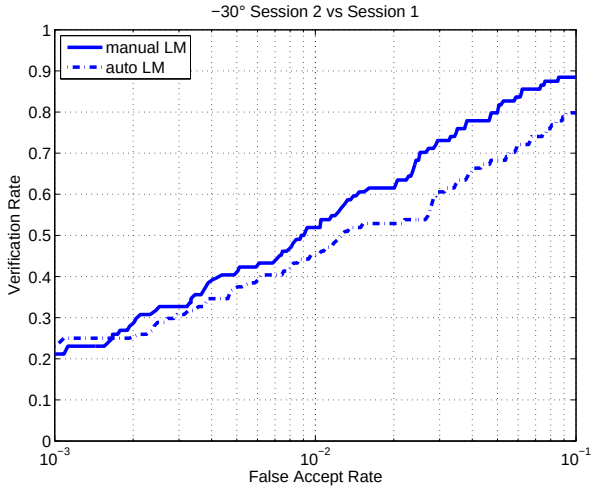
The ROCs in Figure 4.3 show the view-based verification performance using shape information only. The feature extraction corresponds to a simple projection on a shape subspace trained on unseen faces at the corresponding angle. What is remarkable to observe is that the effectiveness of the shape information does not decrease as the viewpoint deviates from frontal, for both manual and automatic landmarking. This validates the results of [59] that shape information is important, not only in frontal faces, but also off non-frontal faces. However, this finding is less useful in our pose-correction framework than it is for pose-synthesis, since we do not plan to perturb the frontal faces or their shape information. We need to find out if this discriminative shape information is carried over after the non-frontal faces go through the 3D pose-correction step.



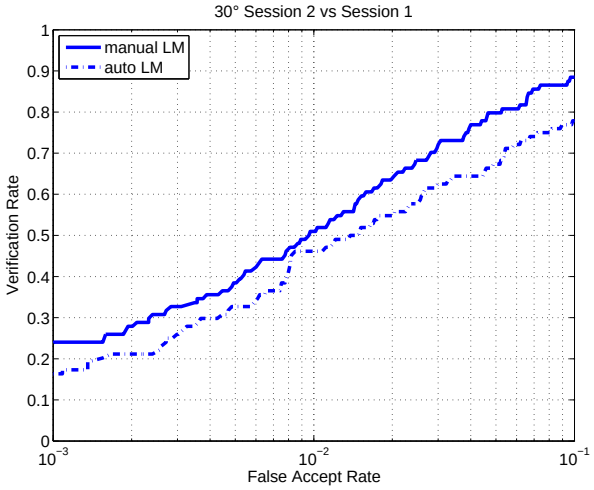
(a) Shape matching at $-45°$



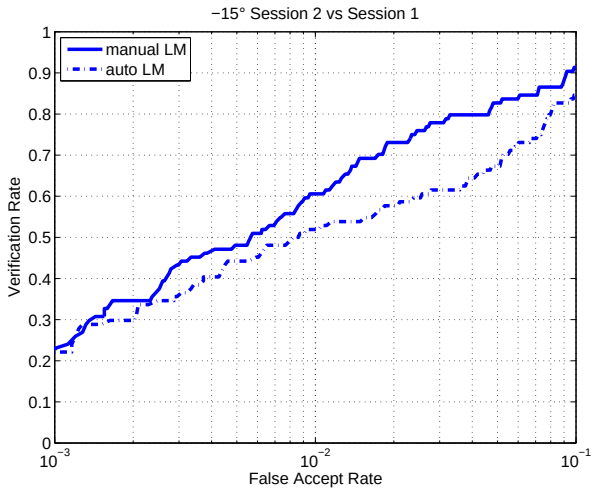
(b) Shape matching at $45°$



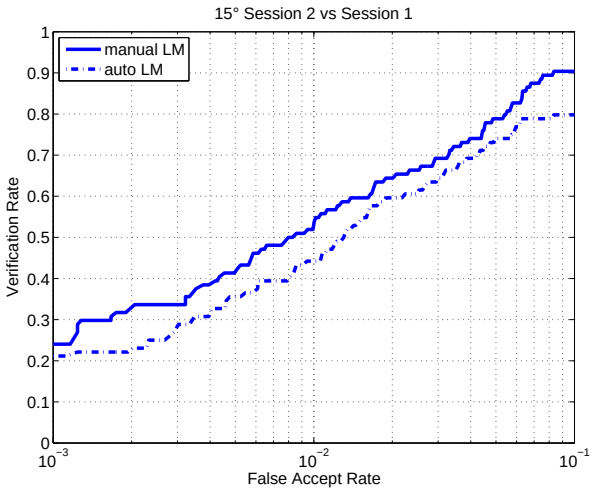
(c) Shape matching at $-30°$



(d) Shape matching at $30°$



(e) Shape matching at $-15°$



(f) Shape matching at $15°$

Figure 4.3: Verification performance of shape-only information using a subspace model trained on 88 subjects from MPIE sessions 3. ROCs represent matching the shape information of MPIE session 1 versus MPIE session 2. All subjects are unseen during training.

4.2.2 Pose-Corrected Shape Information

We assume our gallery faces are frontal, and the equivalent non-frontal shape information of the gallery faces is either not-available or too costly to obtain. We can therefore apply the techniques of the previous chapter to obtain pose-corrected shape information via 3DGEM. Figure 4.4 portrays the experimental setup used for pose-corrected shape-only matching. Non-frontal test faces go through the 3D-modeling step and the derotation step, and then we only measure the resulting x and y coordinates of the original 79 landmarks after the derotation step. Just as the previous experiment, this new shape pose-corrected vector goes through a Procustes analysis step to align all shape vectors and eliminate rotation, scale and translation variations. Since the resulting pose-corrected landmarks are not only a function of the original landmarks but also a function of the 3D-model and the derotation step, they are dependent on both the accuracy of the landmarks and the accuracy of the pose estimator. Therefore, when measuring verification rates, we consider all possible combinations of landmarking and pose estimation schemes.

For the following experiment, we train a shape-only subspace on frontal and pose-corrected non-frontal shape vectors from MPIE sessions 2 and 3. When we test, we match the pose-corrected shape information of the non-frontal images to the shape information of the frontal gallery images by projecting onto this shape-only subspace (as was done in the previous experiment). The matching method employed is a simple normalized cosine-distance metric. We compute the ROCs for all possible angles, and for all possible combinations of landmarking and pose estimation schemes. The results are depicted in Figure 4.5. The first observation to make is that there is a visible discrepancy between the performance of manual and the automatic modes (for both landmarking and pose-estimation). This suggests that shape-based matching after pose correction is highly sensitive to the accuracy of the landmarking and the pose estimation. Even though the automatic landmarking did not significantly impact the shape-based matching in a view-based mode (previous section), the errors the landmarking is making in conjunction with the pose estimator errors are getting amplified after the pose-correction step. The second observation to make is that even the

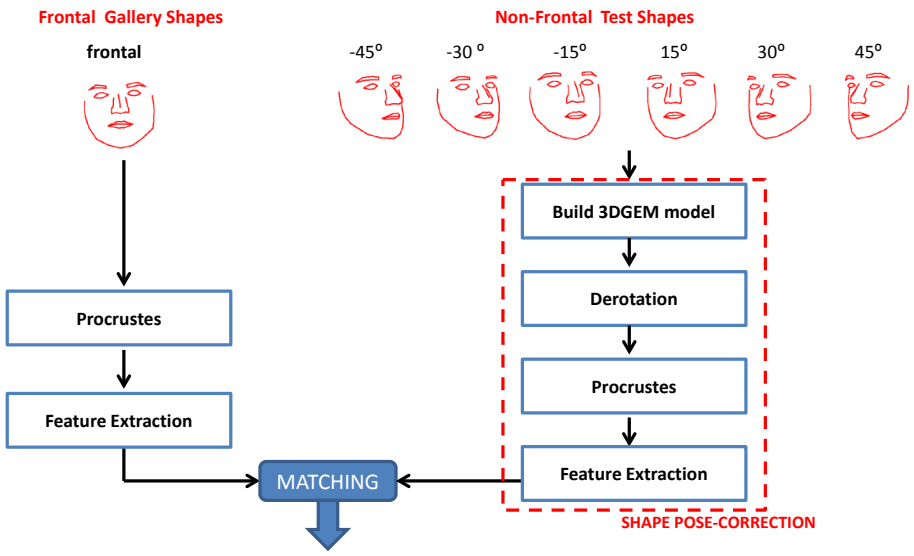
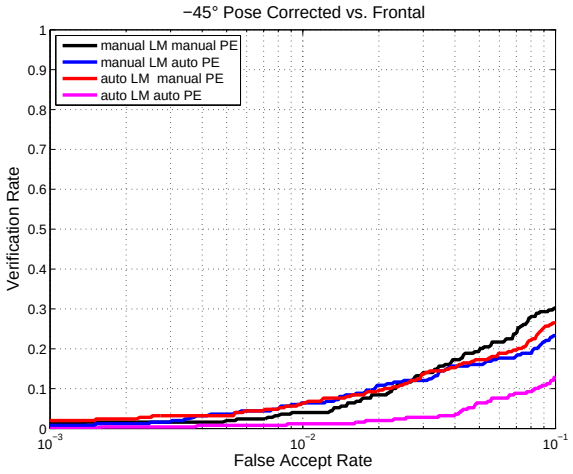


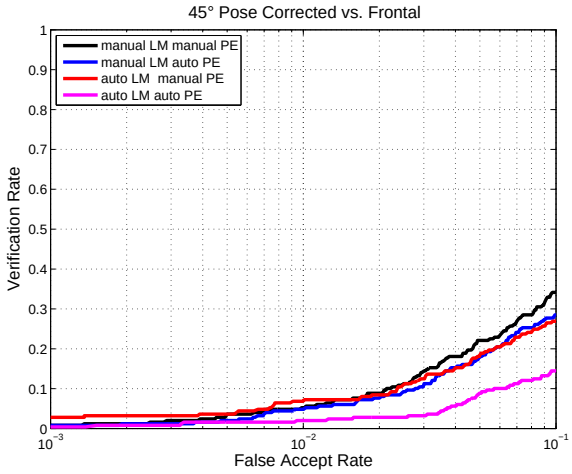
Figure 4.4: Experimental setup for face recognition with pose correction using our method on shape information only. The input data is a vector of stacked x and y coordinates. 3D modeling is used to find the corresponding frontal looking coordinates. Procrustes analysis is performed to align all shape vectors and eliminate rotation, scale and translation from the input images. The feature extraction step corresponds to projecting on a PCA shape subspace built on unseen face shapes.

fully manual results are significantly deteriorating as the input angle deviates from frontal. This suggests that the pose-corrected shape information loses a lot of its discrimination power. This problem is inherent to the landmarking scheme we are following which places landmarks on the *best available* location given the pose, rather than the true location of the landmarks. For instance, in a frontal face, the right face contour points are placed on the boundary of the face that separates the face from the background. However in a non-frontal image, the same right contour points are still placed on the boundary that separates the face from the background, but those points now fall well inside the right cheek. This inherent variance of the landmarking scheme between frontal and non-frontal gets amplified by 3DGEM, and makes shape-only based matching for pose-corrected faces practically impossible.

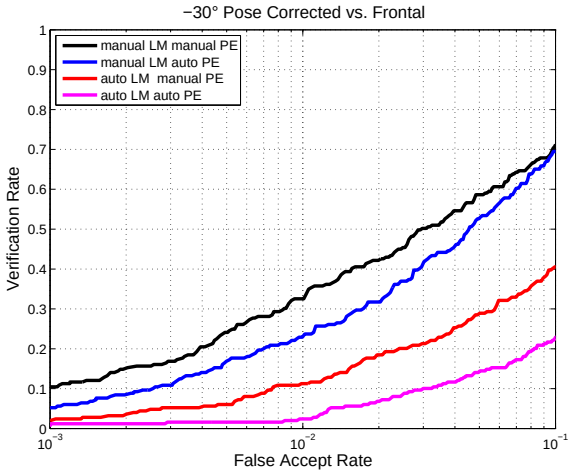
We tried to attenuate this problem by discounting the shape coordinates that fall on the far side of the face, and use L1-PCA in a missing-shape-information problem similar to the one detailed



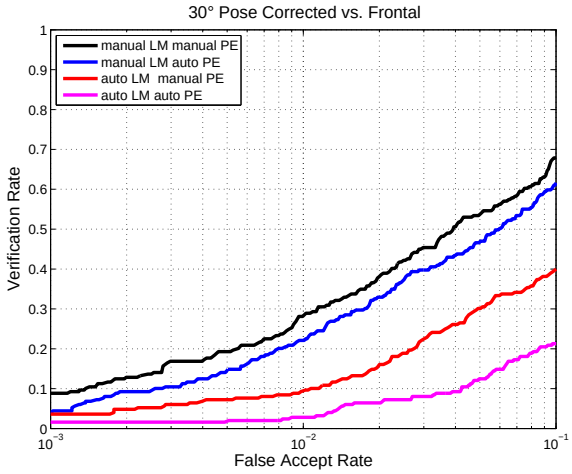
(a) -45° Shape Pose Corrected vs. Frontal



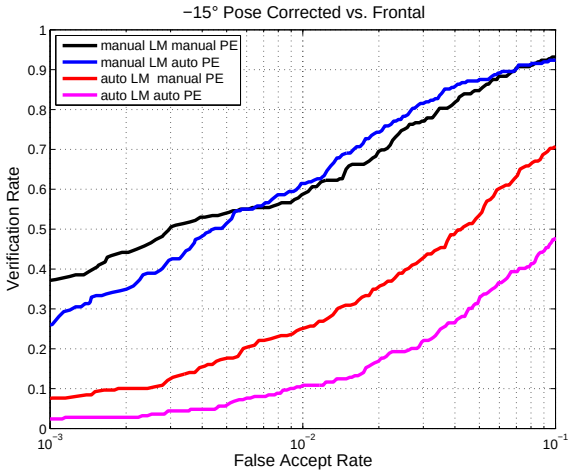
(b) 45° Shape Pose Corrected vs. Frontal



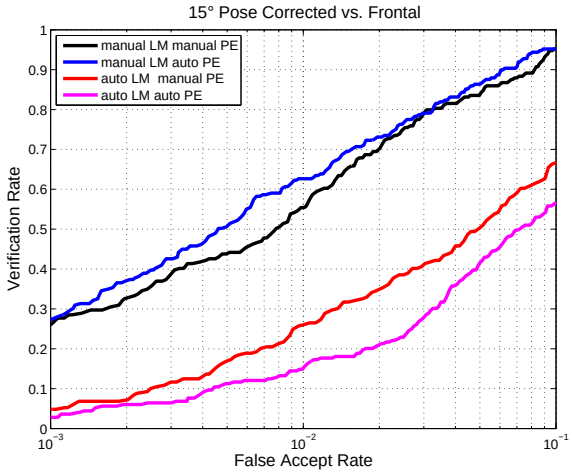
(c) -30° Shape Pose Corrected vs. Frontal



(d) 30° Shape Pose Corrected vs. Frontal



(e) -15° Shape Pose Corrected vs. Frontal



(f) 15° Shape Pose Corrected vs. Frontal

Figure 4.5: Shape-only based matching. MPIE session 1 pose corrected images versus MPIE session 1 frontal images. We show the performance for both automatic and manual landmarking and pose estimation. Shape information loses its discrimination with pose correction. The shape feature vector is a projection on a shape subspace of unseen faces, and the matcher is normalized cosine distance.

the previous chapter, but this hardly seemed to make a difference, which suggests this problem plagues all points on the face, and not just the occluded ones. That is why, until we develop or utilize a different landmarking scheme, we drop the shape information in pose-corrected images, and rely solely on the shape-free representation for feature extraction and reconstruction, which systematically alleviate those problems by discounting the shape information for all subjects.

4.3 Matching in Coefficient Space

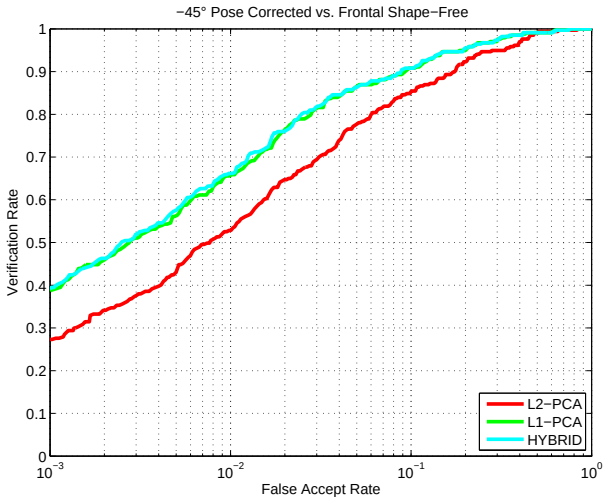
Instead of reconstructing the image and matching the reconstructed images where one could argue we are hallucinating pixel values, we could stop at the feature extracting step and use $\mathbf{c}$ in Equation 3.18 as the feature vector to match with. In our case, this could be exactly equivalent to matching with the entire image, had we not subtracted the mean when building the PCA subspace. If an image $\mathbf{x}_1$ is represented by $\mathbf{x}_1 = \mathbf{V}\mathbf{c}_1$ (instead of $\mathbf{V}\mathbf{c}_1 + \mathbf{m}$) and similarly another image $\mathbf{x}_2$ is represented by $\mathbf{x}_2 = \mathbf{V}\mathbf{c}_2$, then the normalized cosine distance between coefficients is equivalent to normalized cosine distance between images:

$$\text{NCD}\left(\mathbf{V}\mathbf{c}_1, \mathbf{V}\mathbf{c}_2\right) = 1 - \frac{\mathbf{c}_1^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{c}_2}{\|\mathbf{V}\mathbf{c}_1\| \, \|\mathbf{V}\mathbf{c}_2\|} = 1 - \frac{\mathbf{c}_1^{\mathrm{T}}\mathbf{c}_2}{\|\mathbf{c}_1\| \, \|\mathbf{c}_2\|} = \text{NCD}(\mathbf{c}_1, \mathbf{c}_2) \tag{4.2}$$

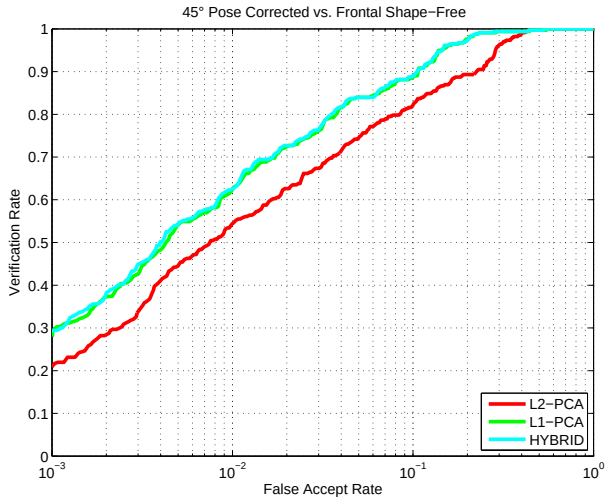
Equation 4.2 holds when $\mathbf{V}$ is an orthonormal basis, which is the case of our PCA basis. It is not uncommon to build a PCA linear subspace without subtracting the mean, and this has been shown to improve the results in certain cases. Moreover, as it is usually done with PCA representation, verification performance significantly improves when the first few coefficients, which represent the first few eigenfaces, are dropped [24]. That is because the first few eigenfaces are dominated by ambient illumination and therefore are common for all subjects in the dataset. By disregarding the first few dimensions that are common for everybody, a face matcher can focus on the dimensions that are more representative of the individual subjects. This improvement in

verification rates suggests that there is significant discriminative information in the middle band of PCA coefficients, where the lower band (representing the first eigenfaces) captures the largest variations which are common for everybody, and the higher band (representing the last eigenfaces) predominantly represents noise. Figure 4.7 depicts the resulting ROCs from matching the feature vectors extracted from the shape-free pose-corrected images using L2-PCA and L1-PCA. Moreover, we compare against the coefficients fit on the HYBRID representation. As a reminder, this representation contains the original observed pixels, and the L1-PCA occluded pixels. Therefore, extracting coefficients from the HYBRID representation represents a dual operation where we find the sparse feature from the observed data, reconstruct the occluded pixels, and then refit the coefficients on the full reconstructed image. The PCA subspace was built without subtracting the mean (for Equation 4.2 to hold), and we also disregarded the first few dimensions of every feature vector which further improved the matching performance. To gauge the statistical significance of this result, we also measure the confidence intervals using a bootstrap [62] and *threshold sweep* approach [63]. Figure D.6 shows the corresponding ROCs with a 95% confidence band centered around the mean verification rates.

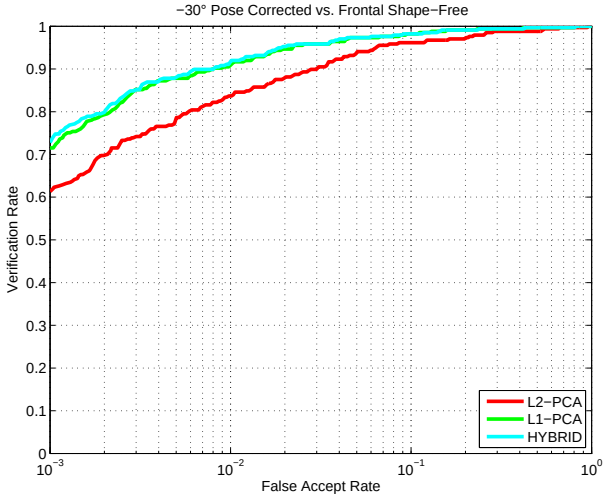
Comparing the results of Figure 4.2 to those of Figure 4.7, we observe a very visible performance improvement. For instance, at the most extreme angles, the verification rate has improved from roughly 10% and 0.1% FAR to 40% for the -45° angle and 30% for the $+45^\circ$ angle at the same FAR. This was achieved by simply dropping coefficients from the feature vector. Also note that the gap between the L1-PCA and L2-PCA ROCs has increased when matching in coefficient space, which confirms that our sparse feature extraction carries more discriminative information and has a better tolerance for pose variations. Also note that the ROCs of the HYBRID and L1-PCA coefficients are essentially identical. Very little is gained by reffiting the coefficients on the full HYBRID image compared to the sparse feature representation of L1-PCA. Therefore, for the remainder of this document, we will drop the HYBRID coefficient representation and focus on the L1-PCA coefficient representation which comes at half the computational cost of HYBRID. Table



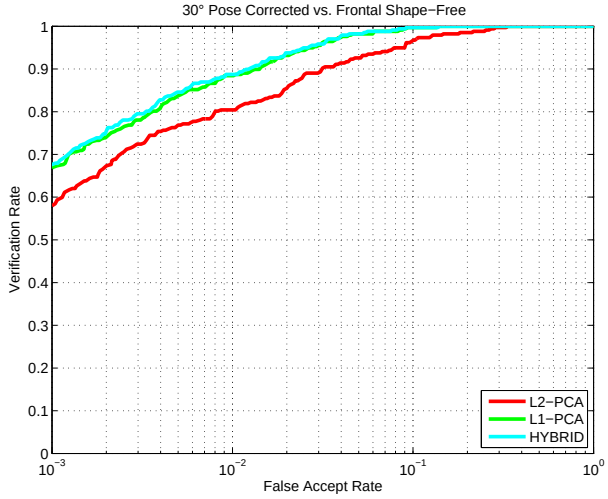
(a) Pose-corrected shape-free feature $-45°$



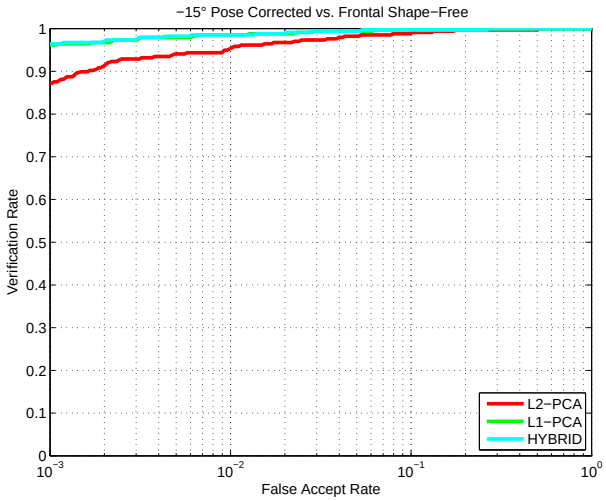
(b) Pose-corrected shape-free feature $45°$



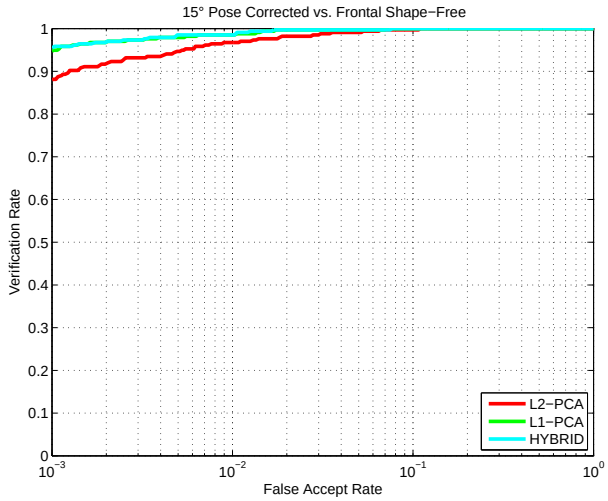
(c) Pose-corrected shape-free feature $-30°$



(d) Pose-corrected shape-free feature $30°$



(e) Pose-corrected shape-free feature $-15°$



(f) Pose-corrected shape-free feature $15°$

Figure 4.6: Verification performance of matching with the coefficients after dropping the first few dimensions. The matcher used is normalized cosine distance. The testing set contains all 337 unique MPIE subjects.

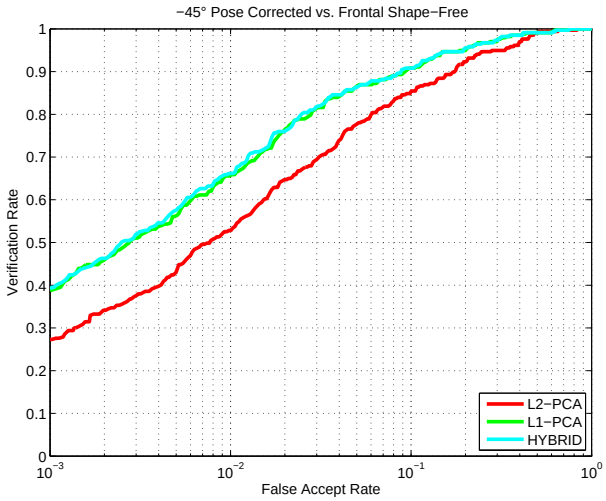
D.1 compares the verification rates obtained for matching in the coefficient space with the rest of the reconstructed image representations.

4.3.1 SimBoost for Weighted Non-Linear Coefficients Matching

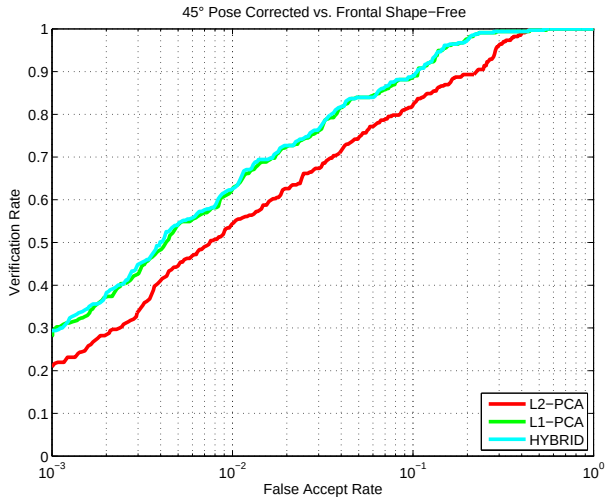
While the previous observation seems to be based more on intuition rather than rigorous mathematical analysis, it clearly shows that there is a lot of discriminative information to be obtained from the coefficient feature vector. This makes the sparse feature vector a perfect candidate for feature selection algorithms, that will methodically pick the most significant dimensions. For this reason, we present a simple matching algorithm based on boosting to select and non-linearly combine the different dimensions of our sparse feature vector. This algorithm weighs dimensions differently based on the similarity matrix the feature vector induces on a validation set, and hence we call it SimBoost. SimBoost is detailed in Appendix B. The improvement in verification performance due to SimBoost is depicted in Figure 4.19. Replacing the simple normalized-cosine distance matcher by SimBoost which corresponds to a nonlinear weighted combination of feature dimensions is allowing us compensate for errors introduced by our pose-correction technique. This in return significantly enhances the verification rates as depicted in Figure 4.19.

4.3.2 Identity Retention Across Pose Variations

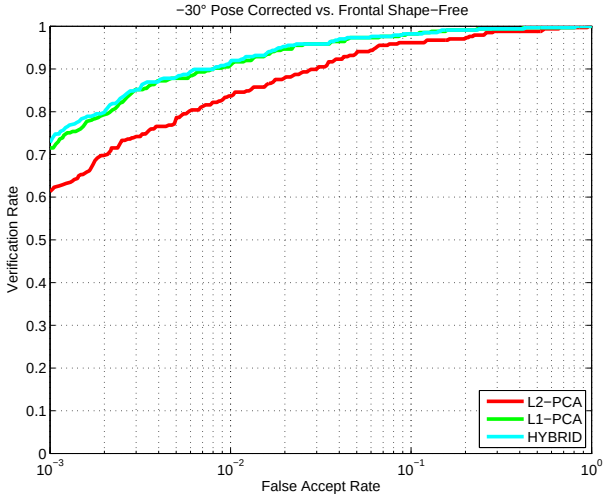
In section 3.3.3 we showed how the L1-PCA sparse feature vector minimally changed as the levels of occlusion (representing pose) increased. The minimal drift of L1-PCA in feature space is the result of the sparsity of the feature vector, as ℓ_1-minimization will abstain from assigning weights to arbitrary dimensions as the number of missing dimensions increases. It is specifically this minimal drift of the L1-PCA feature vector in coefficient space that makes it better suited for face recognition than L2-PCA, since it translates into a minimal drift in identity space. Coefficients extracted from different subjects (and different viewpoints) will be maximally separated in the



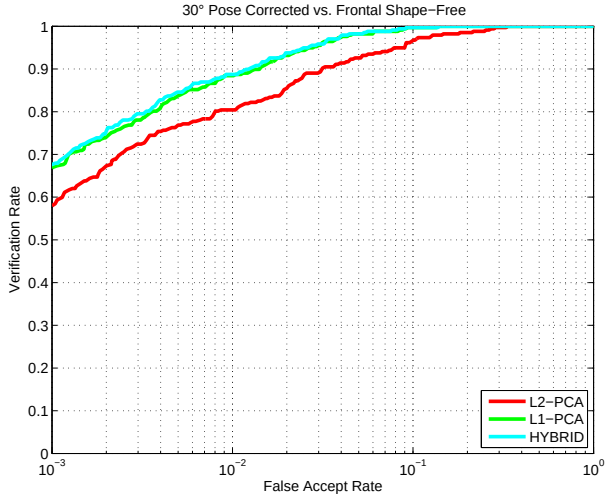
(a) Pose-corrected shape-free feature $-45°$



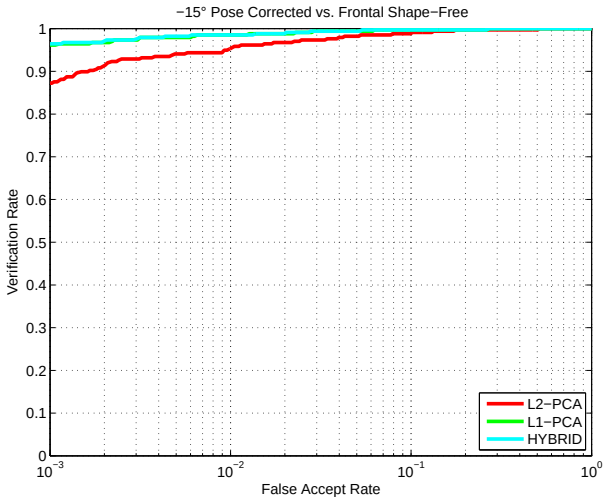
(b) Pose-corrected shape-free feature $45°$



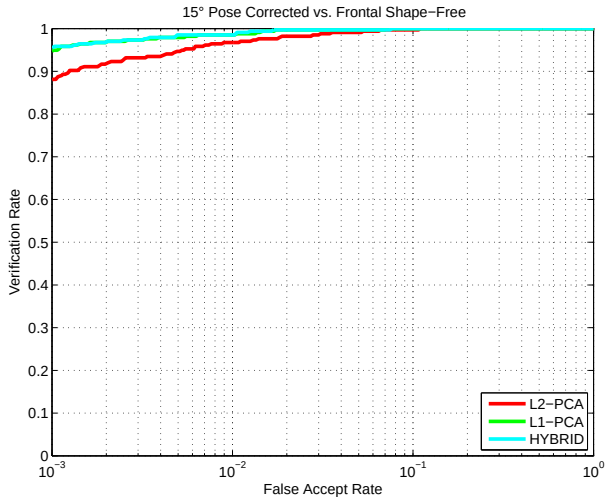
(c) Pose-corrected shape-free feature $-30°$



(d) Pose-corrected shape-free feature $30°$

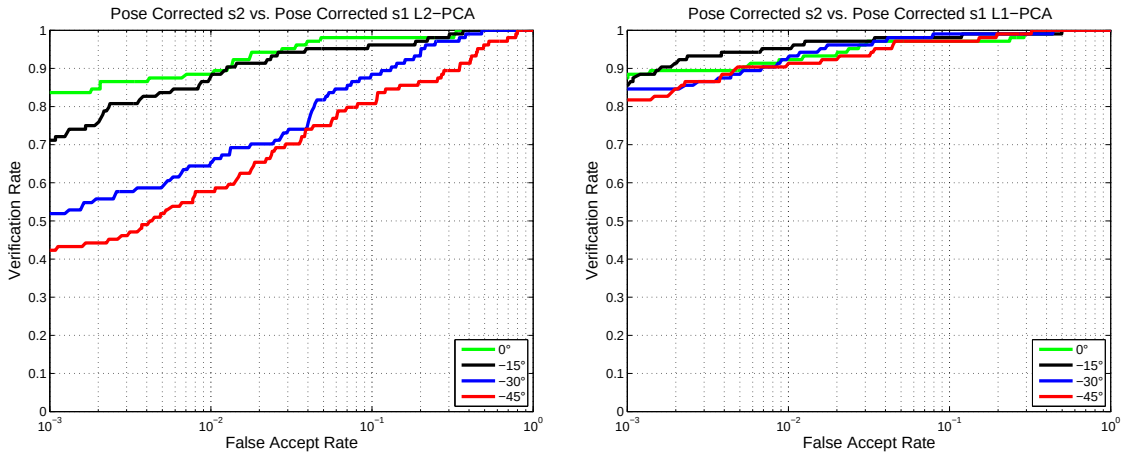


(e) Pose-corrected shape-free feature $-15°$



(f) Pose-corrected shape-free feature $15°$

Figure 4.7: Verification performance of matching with the coefficients after dropping the first few dimensions. The matcher used is normalized cosine distance. The testing set contains all 337 unique MPIE subjects.



(a) L2-PCA (b) L1-PCA

Figure 4.8: Verification performance of matching with the coefficients within each viewpoint. The gallery set consists of MPIE session 1. The probe set consists of 104 subjects from MPIE session 2 that are included in session 1. The matcher used in NCD. The rest of the angles have been omitted to make the figure more readable.

feature space, which explains why L1-PCA outperformed L2-PCA in the ROCs of Figure 4.7. To further prove this point, we set up the following experiment. The gallery set consists of the pose-corrected face images of MPIE session 1 subjects at various angles. The query set consists of the pose-corrected face images of a hundred of those same subjects, at the same corresponding angles, only from a different session (MPIE session 2). We then match the pose-corrected images within each viewpoint, to gauge the impact of pose-correction on maintaining the identities of the subjects across pose. Figure 4.8 shows the ROCs obtained for this experiment. The ROCs in Figure 4.8a exhibit a marked performance drop for L2-PCA when matching within each viewpoint, which is not the case for L1-PCA as shown in Figure 4.8b. This confirms that the intra-class variations due to the L1-PCA feature vector are stable under varying poses and will minimally impact the inter-class variations.

4.4 Pose Synthesis versus Pose Correction

In this section, we further justify the use of pose-correction over pose-synthesis. One could argue that if we reconstruct the image, pose-correction is introducing new information that is estimated

from a training set which may not be representative of the testing data. Instead of pose-correcting the query image which one could argue is introducing new information that is estimated from a training set, we could have the gallery near-frontal images match the pose of the query image, which albeit less efficient, might be more faithful to the data. This would involve building 3D models of all gallery images as shown in Figure 4.9. This approach brings its own share of challenges as well, most common of which is the masking problem. In images rendered from 3D, each face will display a different contour and traditionally masks are used to "mask-out" the uncommon regions. However, we could avoid this problem by adopting a similar approach to the previous section and render all faces to a mean non-frontal shape, which would ensure that no information is lost in a masking process. Figure 4.10 compares the shape-free images generated by the pose-correction module to those generated by the pose-synthesis module. The faces in the top row of Figure 4.10 have all been generated by the frontal image of the MPIE subject. We fit a shape-free 3DGEM model on this frontal face, and render it at different angles to synthesize different viewpoints. On the other hand, the bottom row of Figure 4.10 displays the shape-free images rendered by a non-frontal 3DGEM model built on the non-frontal image. We can observe that the discrepancies between the two approaches increase with the yaw angle, with the most visible changes occurring around the area of the cheeks.

To compare pose-synthesis against pose-correction, we match the two types of images in Figure 4.10. Specifically, we take the frontal gallery images, fit a frontal 3DGEM model on them, and rotate them to match the angle of the query image. Similarly, we fit a non-frontal 3DGEM model on the query image to generate a shape-free representation. In this shape-free domain, both gallery and query faces have exactly the same dimensions, and all features of the faces are perfectly aligned. Figure 4.11 depicts the ROCs obtained for every test angle and for every unique subject in the MPIE database. To compare, we also plot the best (L1-PCA) result obtained on the same dataset using the pose-correction approach of section 4.1 and depicted in Figure 4.2. We observe that using the naive cosine distance matcher, pose-correction narrowly outperforms pose-synthesis

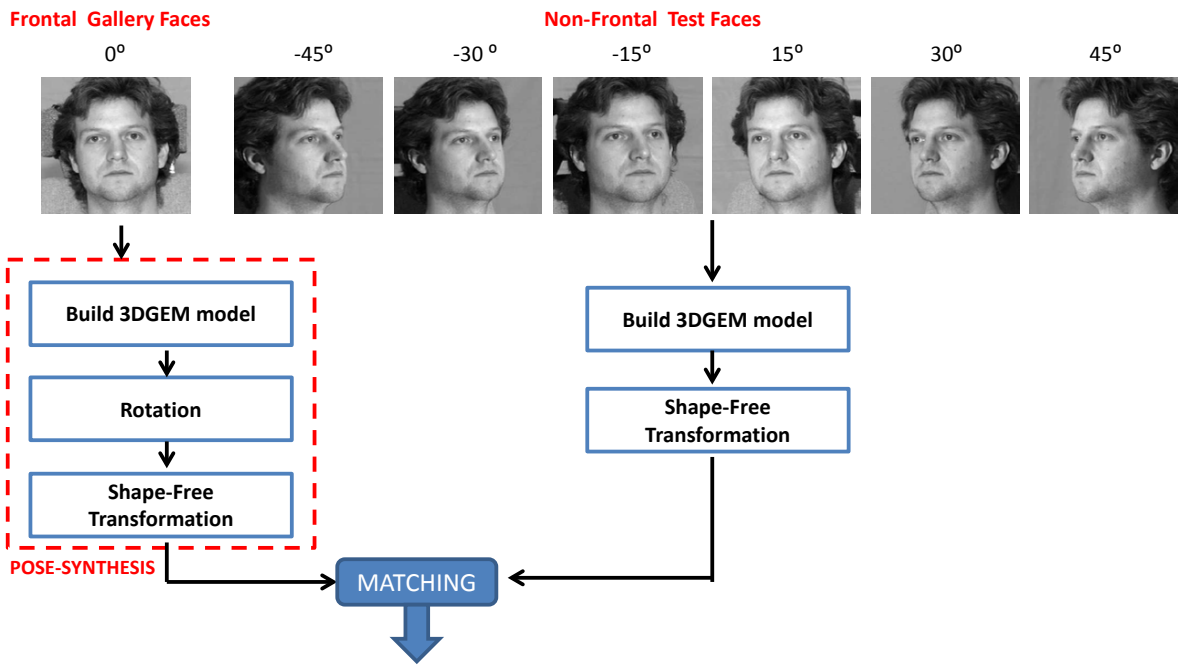


Figure 4.9: Pose synthesis experimental setup. To match in shape-free representation, we fit a non-frontal 3DGEM on the test images. We also fit a frontal 3DGEM on the gallery face and rotate it to match the angle of the test image.

in all test angles. Note that the smaller the yaw angle, the more pose-correction outperforms pose-synthesis which suggests that our current non-frontal 3D modeling is operating at its limits. With the current formulation, building non-frontal 3DGEM models from input faces beyond 45° degrees will introduce artifacts that are more prominent than those introduced by the pose-synthesis approach at these angles.

4.5 Sensitivity Analysis

Our pose-correction technique makes three important assumptions. First we assume the face has been accurately localized in the image. More importantly, we also assume that we can obtain accurate landmarks and pose estimates to generate the non-frontal 3D model. In this section, we analyze the relative impact of the accuracy of those estimates on verification performance induced

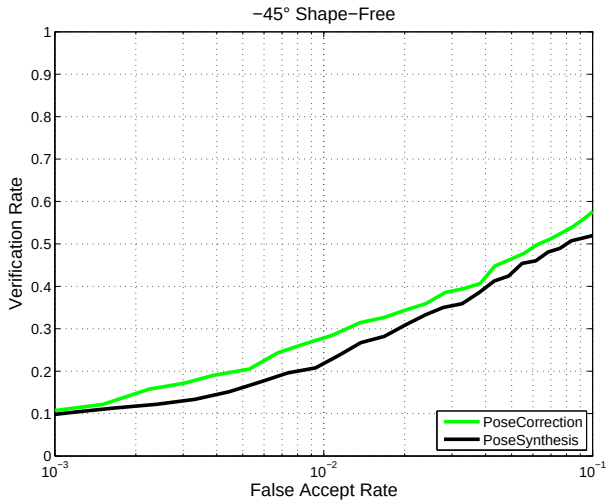
(a) $-45°$ (b) $-30°$ (c) $-15°$ (d) $0°$ (e) $15°$ (f) $30°$ (g) $45°$

Figure 4.10: Pose correction vs. pose synthesis. The top row represents the shape-free face synthetically generated using 3DGEM from a frontal image. The bottom row represents the shape-free image obtained using a non-frontal 3DGEM model. The columns (a) through (g) show different yaw angles. The differences between the top and the bottom row becomes more pronounced as the yaw angle increases.

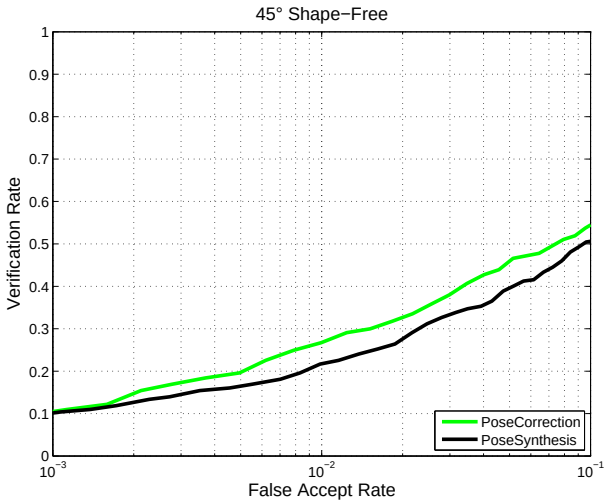
by the sparse feature vector fit in shape-free representation. (See section 4.2 for sensitivity analysis on shape-only based verification). So far all the results have been obtained on manually annotated face images. All 79 control points have been manually placed on the face for all test angles. The angles were estimated to be at fixed value following the technical description of the MPIE database acquisition protocol. We now switch to a fully automated process that requires no human operator intervention.

4.5.1 Automatic Mode of Operation for Pose Correction

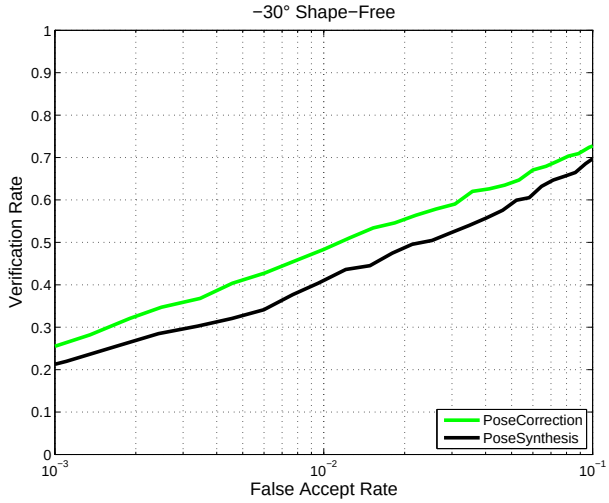
Automatic landmarking of faces is a well studied subject. Early efforts go back to the Active Shape Models (ASM) [64] and Active Appearance Models (AAM) [58]. When the aim is to accurately place landmarks on the face, ASMs have been shown to be more accurate and easier to train and fit on unseen faces [65]. That is because they rely on local texture modeling which is less susceptible to illumination variations than the global texture models used by AAMs. In this dissertation we rely on a robust variant of ASMs called Modified Active Shape Model (MASM) [32]. MASM uses 2D search profiles and relies on an image subspace of profile texture variations to update the



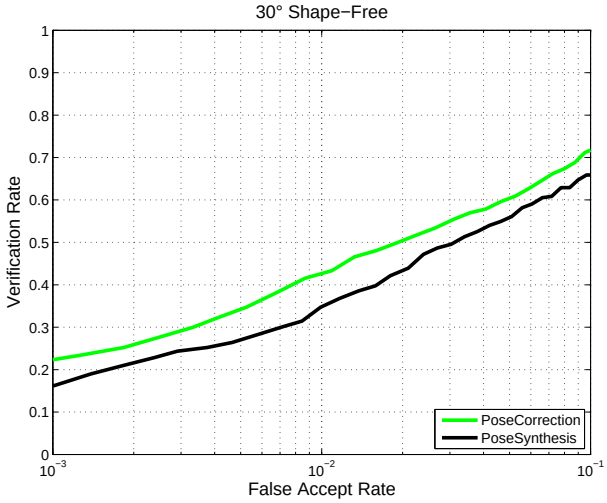
(a) Shape-free matching at $-45°$



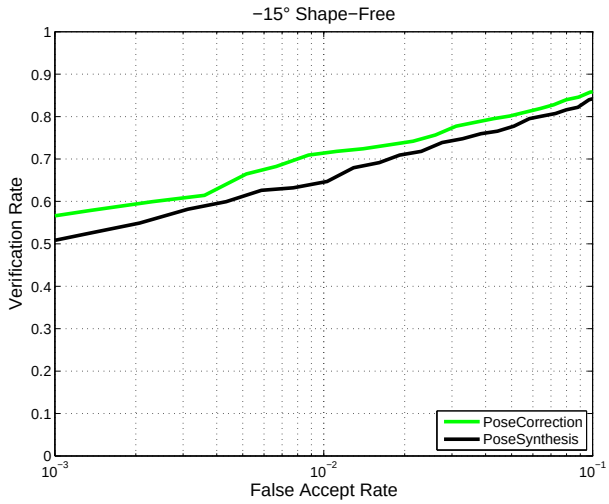
(b) Shape-free matching at $45°$



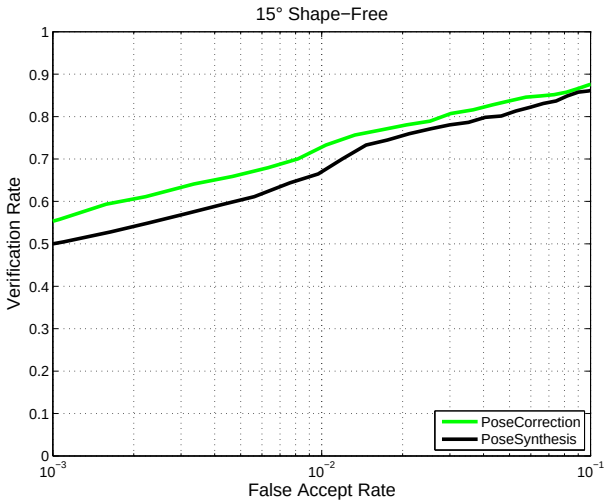
(c) Shape-free matching at $-30°$



(d) Shape-free matching at $30°$



(e) Shape-free matching at $-15°$



(f) Shape-free matching at $15°$

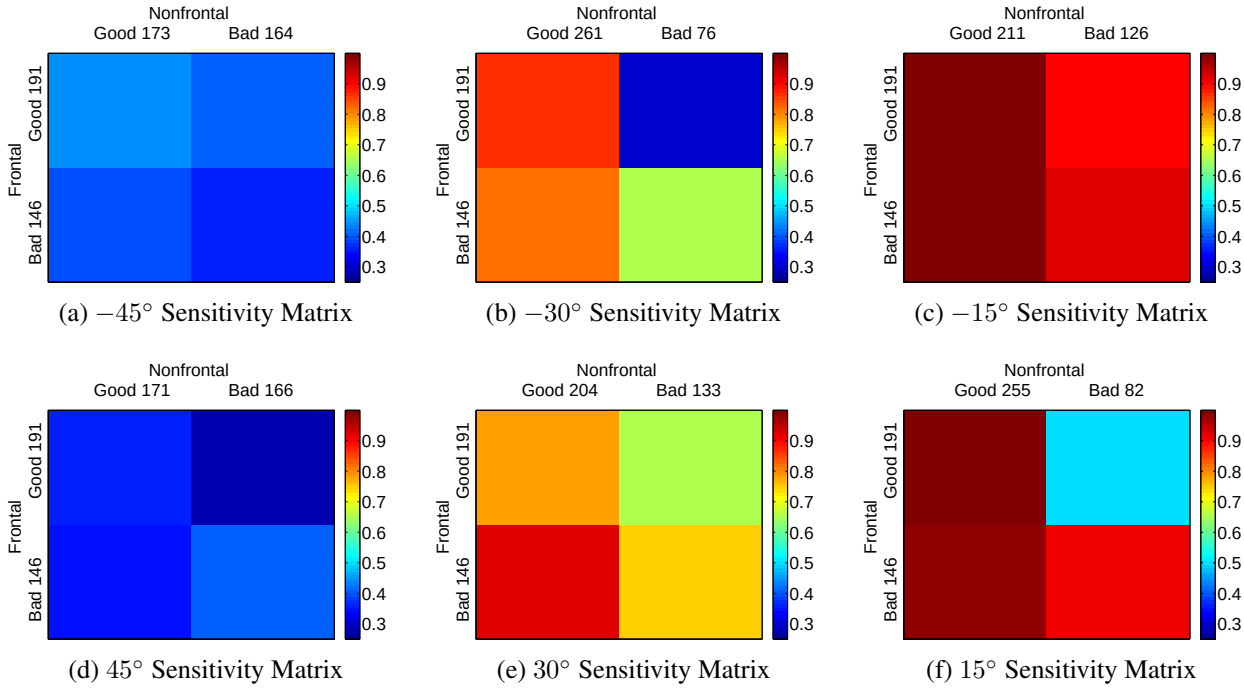
Figure 4.11: Pose correction vs. pose synthesis verification ROCs. Normalized Cosine Distance on shape-free texture images. All 337 unique MPIE subjects are included in this experiment.

location of a given landmark during the search. To handle non-frontal images, we train a view-based MASM, with a different model for every range of viewpoints, each view bin spanning 15° in yaw. As for automatic pose estimation, we rely on a commercial software that estimates the pose as a function of face detection responses. It maintains distinct view-based face detection models, and given a testing image of an unknown yaw, it measures the response triggered by each of the detection models. The yaw value is estimated by regressing on the detection model responses. When using a simple normalized cosine distance matcher on the sparse coefficients (like we did in section 4.3), we notice that the overall one-to-one verification slightly improves with using automatic landmarking and automatic pose estimation compared to manual ground-truth information. Figure D.4[4] depicts the verification performance of the sparse feature with the different combinations of landmarking and pose estimations modes. This suggests that the automatic modes are more consistent that manual ground-truth information. To investigate this further, we now analyze the impact of landmarking and pose estimation separately.

4.5.2 Landmarking Sensitivity

It is not obvious to measure the impact of landmarking accuracy. Facial landmarking is a complex multiresolution approach that is constrained by the concept of legal face shape. Therefore, adding random noise to the landmark coordinates is a not representative of real world test scenarios, since random motion of landmark points are unlikely to happen. Instead of randomly perturbing the landmarks of the subjects in the database and then measuring verification accuracies, we instead perform a more realistic study: we run the automatic facial landmarking, and then measure the drift between the manual coordinates and the automatic landmarks (a mean squared error scheme is standard in the landmarking literature). For every angle, we get a drift distribution, and then classify as "good" fits, the landmarks whose drift is below the mean drift. A "bad" fit are the images whose landmarking drift is greater than the average drift. (Figure D.5

[4]See Appendix D for this result



(a) $-45°$ Sensitivity Matrix (b) $-30°$ Sensitivity Matrix (c) $-15°$ Sensitivity Matrix

(d) $45°$ Sensitivity Matrix (e) $30°$ Sensitivity Matrix (f) $15°$ Sensitivity Matrix

Figure 4.12: Landmarking Sensitivity Matrices. To the left and top of every matrix we indicate the number of fits that we deemed to be "good" or "bad". The color-coded values indicate the verification rate for this combination measured at 0.1% FAR. For most angles, the biggest drop of performance (cool color temperature) is observed when the frontal images are well landmarked but the non-frontal images are not. Conversely, the best performance was when both frontal and non-frontal landmark fitting was good (warm color temperature)

in Appendix D depicts a few examples of bad landmarking fitting). We then observe the pose-correction performance trends between the Cartesian product of $\{\{\text{Frontal Good}\}, \{\text{Frontal Bad}\}\}$ and $\{\{\text{Nonfrontal Good}\}, \{\text{Nonfrontal Bad}\}\}$. The verification rates of these four combinations for every angle are color-coded in Figure 4.12. The cooler the color for a specific combination, the lower the verification rate (measured at 0.1% FAR) for this combination. On the other hand, the warmer the color, the higher the average verification rate. This test represents a more realistic measure of the impact of bad landmarking on the verification rates than randomly perturbing the landmarks of every subject in the dataset.

In all test cases except two, the highest verification rates occur when both the landmarking and pose estimation are accurate. Surprisingly, in all test cases except one, the worst performance

is recorded for the Frontal Good, Nonfrontal Bad combination, rather than the Frontal Bad, Nonfrontal Bad combination. And typically, if the frontal landmark fitting is "bad", it is not responsible for a worst-case scenario. What this implies is that when both fits are bad, some of the errors cancel each other out. This consistency explains why overall the ROCs displayed in Figure D.4 improved in the automatic landmarking case. Even though the fitting is sometimes off, the variance of the landmarks is reduced and this improved consistency translates into improved verification rates.

4.5.3 Pose Estimation Sensitivity

As previously mentioned, our pose-correction method requires an estimate of the face pose to compensate for it by rotating the face model in the opposite direction to give it a frontal viewpoint. This makes the results dependent on the accuracy of the pose estimator. In this section, we analyse the impact of the pose estimates on the verification performance. The standard deviation of the yaw estimates grow as the pose becomes increasingly non-frontal. Figure 4.13 shows that the standard deviations of yaw estimates for the -45 and 45 degrees is almost double that of the frontal faces.

We can compute similar sensitivity matrices to those in the previous section. We can define a good pose estimate for a given viewpoint an estimate that doesn't deviate more than one standard deviation from the mean for that viewpoint. Figure 4.14 depicts the color-coded results for all six angles of the MPIE dataset. However, no specific sensitivity pattern emerges. For $\pm 15^\circ$ the combination of "good" frontal pose-estimate and "bad" non-frontal estimates seems to result in the worst verification results, but $\pm 45^\circ$ that same combination is responsible for the top performance. This seems to suggest that the pose estimator is also significantly biased. What we deem to be bad estimates (because they deviate from the mean by more than one standard deviation) may actually be good estimates, if the mean is offset by some bias. This highlights the fact that estimating a pose is challenging because the *true* head pose is hard to evaluate.

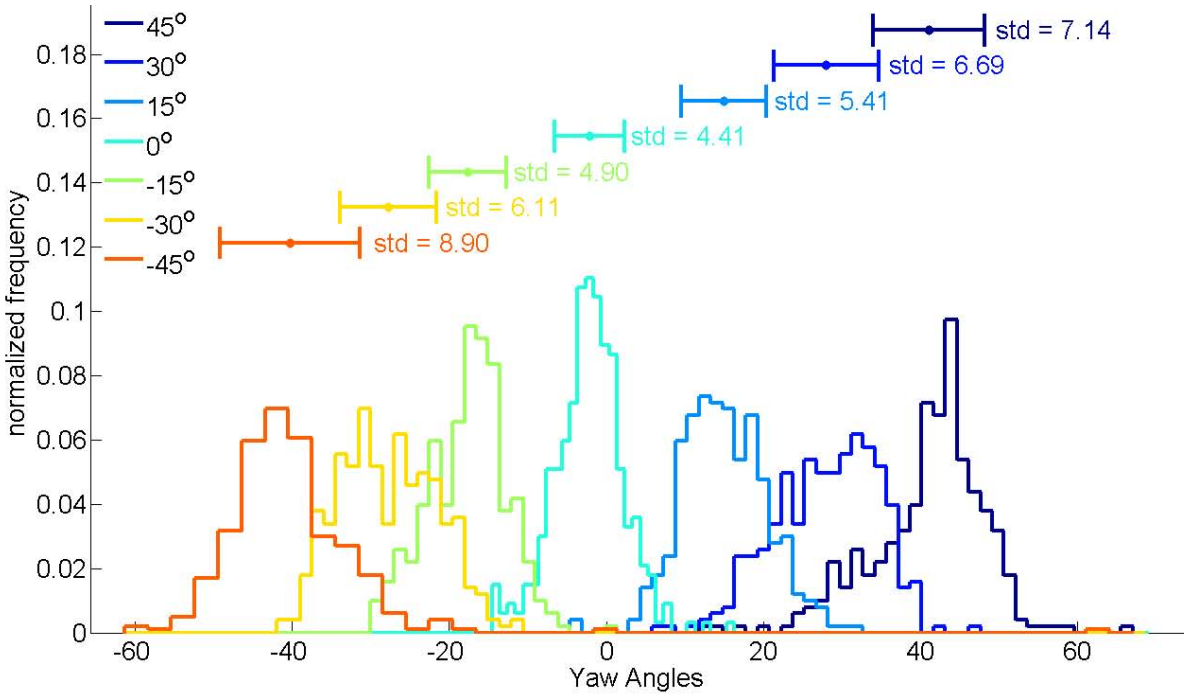
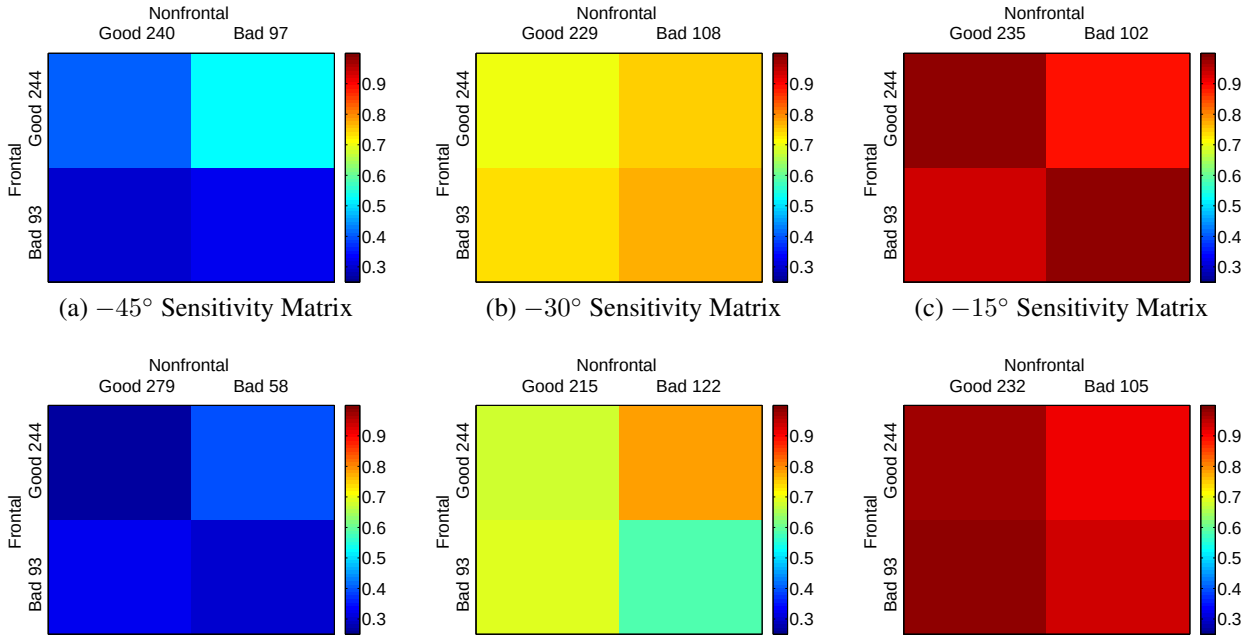


Figure 4.13: Automatic yaw estimation histograms. Note that the standard deviation at the extreme angles is almost double the standard deviation at the frontal viewpoint.

4.6 Robust Matching Algorithms

So far all face verification results have been based on simple normalized-cosine distance matchers or SimBoost which is a non-linear way of matching weighted coefficients. We now capitalize on the pose-tolerance of L1-PCA coefficients to train a more robust classifier based on advanced Correlation Filters (CF) called Class-Dependent Feature Analysis (CFA). Appendix C briefly overviews CFs and explains the CFA algorithm. Advanced correlation filters have a number of inherent advantages that make them ideal for use in 2D data, especially faces. The CFA algorithm harnesses the power of correlation filters in a multiclass problem. In a nutshell, given a training set of N classes, we build a correlation filter for each class. Each correlation filter can be designed to trigger maximum response for the elements of this class, and zero response for elements of the negative classes. Given the N filters, we can now represent an unseen test sample by a feature vector of size N, each dimension representing how much the test sample *correlates*



(a) -45° Sensitivity Matrix (b) -30° Sensitivity Matrix (c) -15° Sensitivity Matrix

(d) 45° Sensitivity Matrix (e) 30° Sensitivity Matrix (f) 15° Sensitivity Matrix

Figure 4.14: Pose Sensitivity Matrices. To the left and top of every matrix we indicate the number of pose estimates that we deemed to be "good" or "bad". The color codes indicate the verification rates for a specific combination at 0.1% FAR.

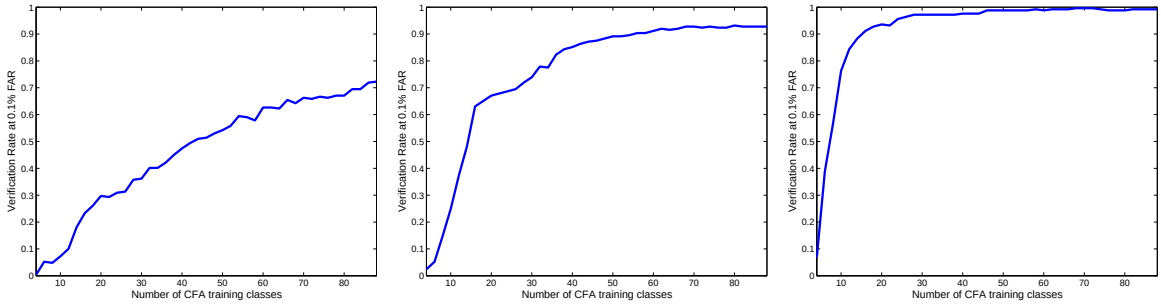
Matching Mode	Feature Mode	Yaw Test Angles					
		$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$
Verification Rate at 0.1% FAR	L1-PCA CFA	72.29	92.77	99.19	99.6	97.19	77.92
	L1-PCA SimBoost	62.65	91.57	97.19	97.99	89.56	57.03
Rank-1 Identification	L1-PCA CFA	85.95	97.19	100	100	98.78	89.16
	L1-PCA SimBoost	51.00	85.94	97.18	97.18	87.95	53.41

Table 4.1: Matching performance on MPIE session 1 with L1-PCA CFA used for feature extraction. The verification rate is measured at 0.1% FAR. For CFA, the matching metric is normalized cosine distance. The experiment represent a genuine one-to-one scenario as we do not perform score normalization of any kind.

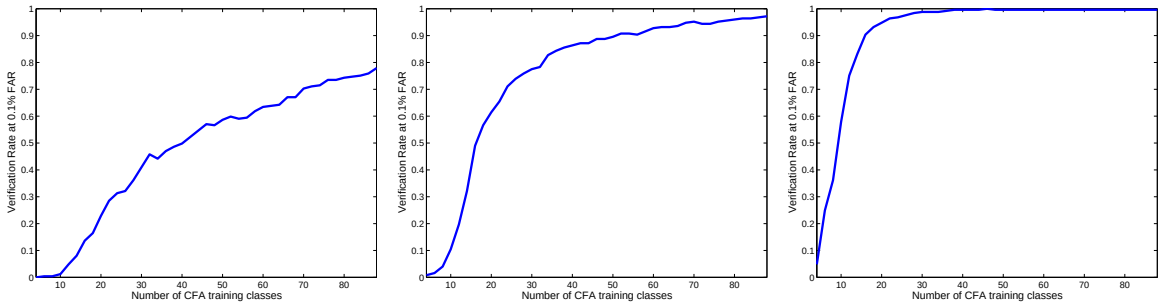
with a given class filter. This sequence of *one-against-all* responses creates a robust representation for any arbitrary test sample. We apply the CFA algorithm on the sparse feature vector extracted from shape-free pose-corrected images, and the verification improvement is very significant. Figure 4.19 depicts the ROC for different angles when using CFA with the L1-PCA features. The training set for CFA consists of 88 subjects from MPIE sessions 2 and 3. The testing set consists of 249 different subjects from MPIE session 1. There is no overlap between testing and training subjects. Table 4.1 summarizes the performance of our CFA algorithm.

4.6.1 CFA Experimental Analysis

The performance of CFA depends on the number of training classes available. With too few training classes, the algorithm will find it hard to discriminate between unseen testing faces, and with too many classes, the algorithm will saturate and essentially starts learning noise because it will be challenging to enforce the correlation constraints for a very large number of training faces and training classes. In this experiment, we analyze the impact of the number of available training classes on the pose-corrected verification rates. As a reminder, the training faces used with CFA are all frontal (pose-corrected frontal and originally frontal faces), and therefore our implementation of CFA is not a view-based approach. We could probably significantly improve our results by



(a) L1-PCA CFA at $-45°$ (b) L1-PCA CFA at $-30°$ (c) L1-PCA CFA at $-15°$



(d) L1-PCA CFA at $45°$ (e) L1-PCA CFA at $30°$ (f) L1-PCA CFA at $15°$

Figure 4.15: Impact of available number of training classes on Verification Rate. The y axis represents the verification rate measured at 0.1 % False Accept Rate. The x axis represents the number of available training classes.

adopting a view-based CFA approach, but for reasons that will become evident in section 4.8, we will stick to training the CFA on pose-corrected face images from multiple angles. Figure 4.15 depicts the progression of verification rates as a function of the number of available training classes. The verification rates are measured at 0.1% FAR.

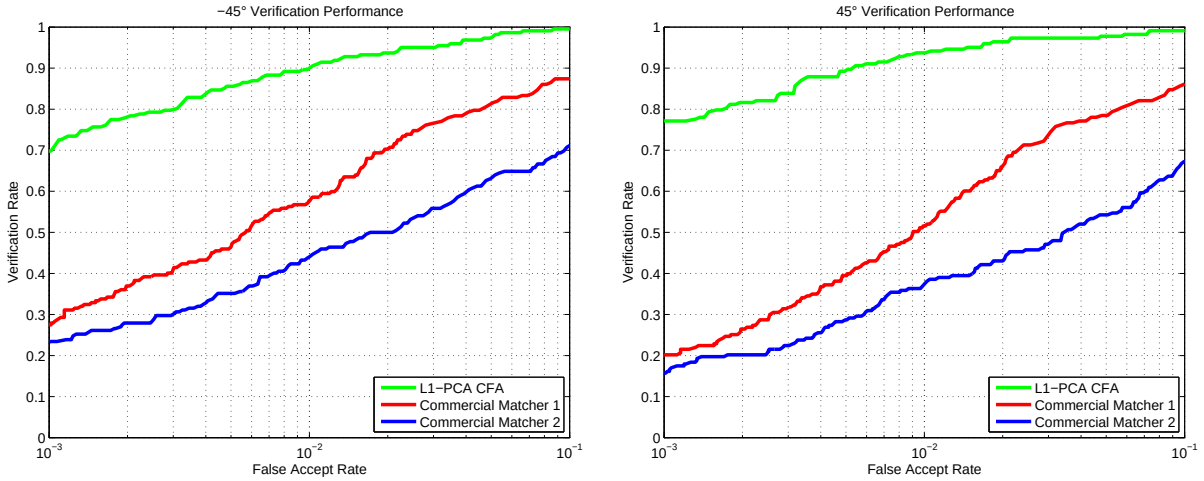
The test set consists of MPIE session 1. The training set consists of the 88 subjects in sessions 2 and 3 that are not present in session 1. Figure 4.15 shows that for the near-frontal angles, as few as 30 classes are enough for CFA to get very close to its maximum verification rate and then saturates. In the other cases, as the yaw angles increase, the dependence on the number of training classes becomes more linear, since the task becomes more challenging, and suggests that with more training data the verification rates could have kept improving.

4.7 Evaluation Against Commercial Face Recognition Engines

Most commercial face matchers do not explicitly correct for pose beyond the classical 2D warping-based methods. In this section, we benchmark our best result against commercial matchers. Note that this comparison cannot be fair for a number of reasons. First, commercial face recognition software developers have access to unlimited amount of training data, including all the database that are available to us and that we experiment with (MPIE, FERET, etc.). Therefore, it is to be expected that the matching accuracies they obtained on the MPIE frontal and near-frontal faces are very high. (The fact that they score perfect scores on MPIE but perform poorly on a small set of random images captured in the lab with similar viewpoints made us suspect that they are training on our testing dataset). Second, to avoid this problem, we had to push the matcher beyond their advertised pose operating range, and that's when we saw their performance drop sharply.

Commercial Matcher 1 is a popular commercial matcher that is only a few years old, and it explicitly mentions that it can handle pose up to $\pm 40^{\circ}$ in yaw. The gallery dataset consists of MPIE's frontal images from session 1 (249 subjects) and the query dataset consists of MPIE's $\pm 45^{\circ}$ images. For 23 of these query images, it rejected the comparison saying that these input faces were"invalid". Therefore, to be as fair as possible, we discarded the same 23 images from the query set for all algorithms involved in this experiment. Commercial Matcher 2 is another popular commercial matcher, and while it does not explicitly advertise that it handles poses, it did not reject any images due to the severely non-frontal viewpoint. Moreover, to guarantee a fair comparison, we bypass the commercial matchers' face detection and manually provide the location of the face and eye coordinates, so that the results are not biased by errors due to face detection. We also disabled all score normalization schemes which typically affect results significantly. Matcher 1 did not permit us to override the normalization engine, so we ran the software on pairwise matches and populated the similarity matrix one entry at a time.

Figure 4.16 compares the performance of the commercial matchers with our best matcher L1-



(a) Matching $-45°$ to frontal

(b) Matching $45°$ to frontal

Figure 4.16: One-to-one verification performance comparison against commercial face matchers.

PCA CFA at the most extreme viewpoint. Commercial matchers 1 and 2 obtained a perfect score on the non-frontal MPIE images, but their performance severely dropped at the extreme viewpoint while our method exhibited a more graceful performance drop off.

4.8 Matching Non-frontal Images to Other Non-frontal Images

So far we have assumed a standard *mugshot* scenario where the gallery images are frontal and test images are assumed to be of an arbitrary viewpoint. However, the methodology we presented in the previous chapter can be easily extended to handle the case where both *gallery and query* images are non-frontal. This could be a more realistic assumption in surveillance footage or in crime scene investigations where several non-frontal snapshots of a presumed suspect are available, and we would like to find a match between them. Figure 4.17 depicts the flowchart of our technique modified to handle non-frontal gallery images. The remarkable aspect of our method is that the flowchart in Figure 4.17 requires little or no extra effort compared to the one depicted in Figure 4.1. That is because our algorithms for pose-correction have all been trained on a frontal and

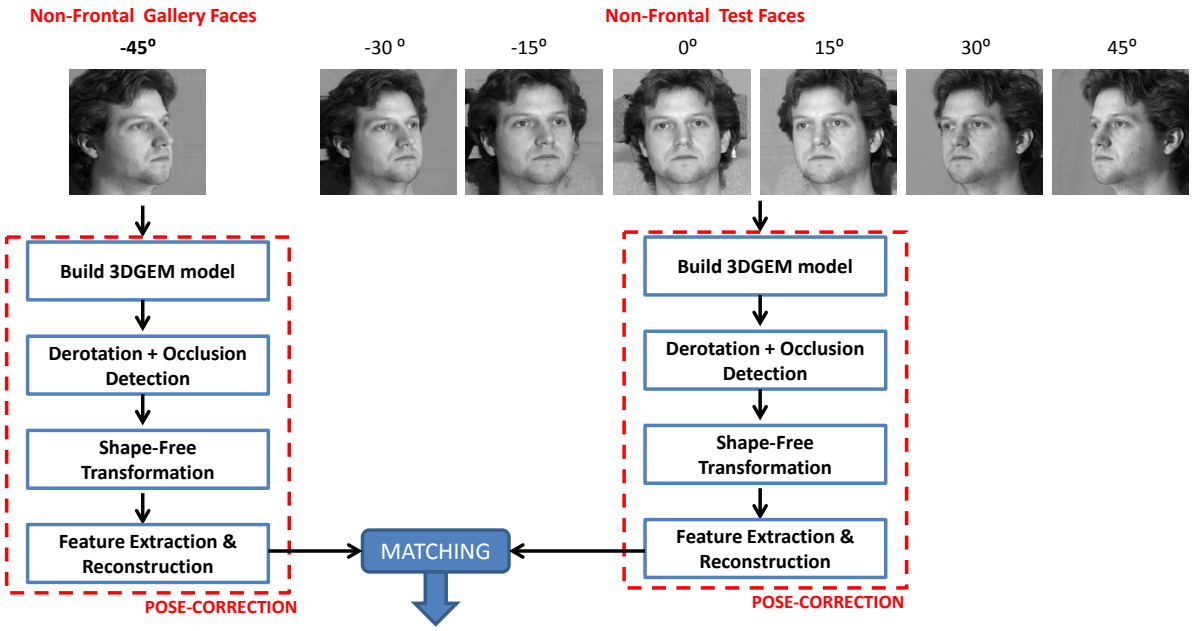
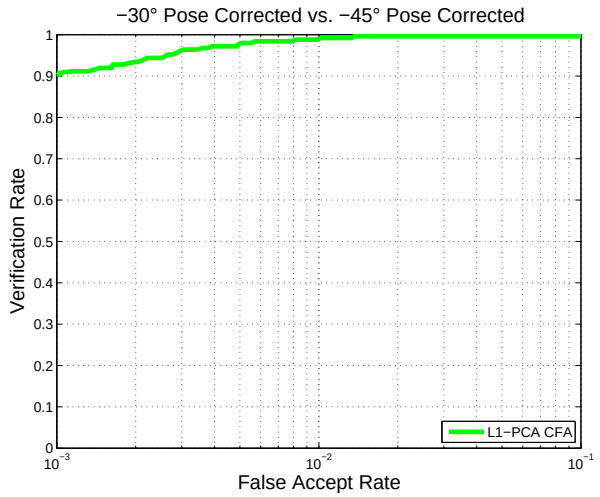


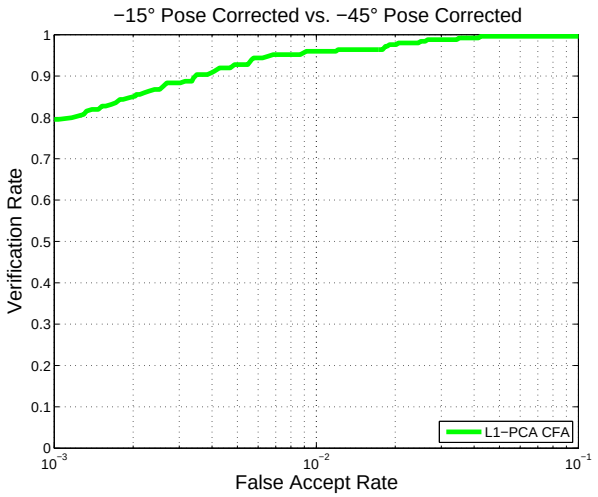
Figure 4.17: Experimental setup for face recognition with pose correction when the gallery image is non-frontal and the test image is of an arbitrary viewpoint.

pose-corrected images. For instance, our CFA and SimBoost algorithms have been trained on the pose-corrected unseen subjects of MPIE sessions 2 and 3 (88 subjects total unseen in session 1). These images represent all available angles (including frontal) all pooled together in a unified generic training set. In testing, our algorithm is completely blind as to which angle constitutes the gallery image, and which angles represent the query images. Figure 4.18 depicts the ROC for matching gallery images at -45° against query images of random yaw angles.

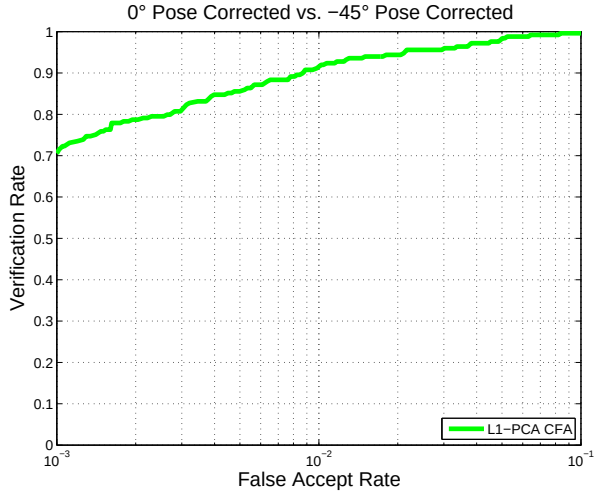
It is remarkable to note that when the gallery images are at -45°, the performance does not significantly drop when the test angles vary. In fact, when the query images are at 45° which is the opposite angle of the gallery images, the verification performance at 0.1% FAR remains around 80%. In fact, with our method, we can *mix and match* angles in both gallery and query sets. In this next experiment, we let the gallery faces be of a given angle, and have the query set be a mix of all the remaining angles. Table 4.2 summarizes the resulting verification rates measured at 0.1% FAR when using our L1-PCA CFA algorithm for feature extraction with cosine distance as the matcher.



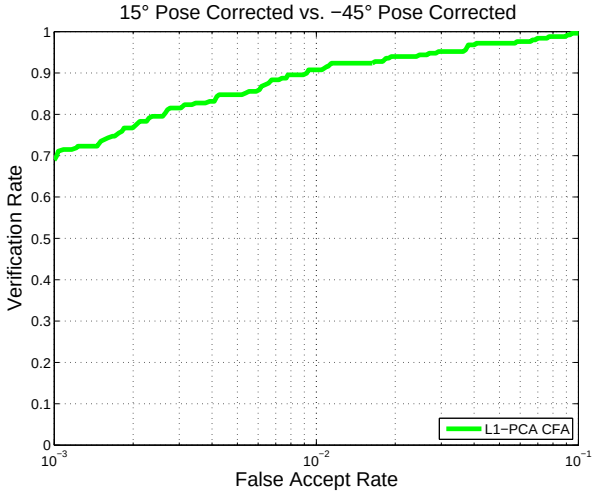
(a) $-30°$ Pose-Corrected vs. $-45°$ Pose-Corrected



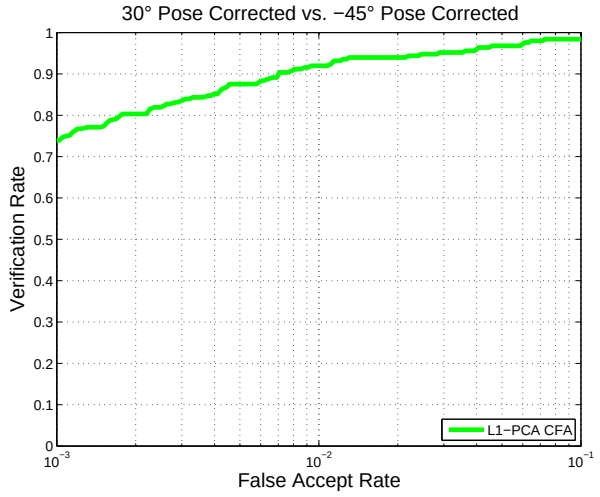
(b) $-15°$ Pose-Corrected vs. $-45°$ Pose-Corrected



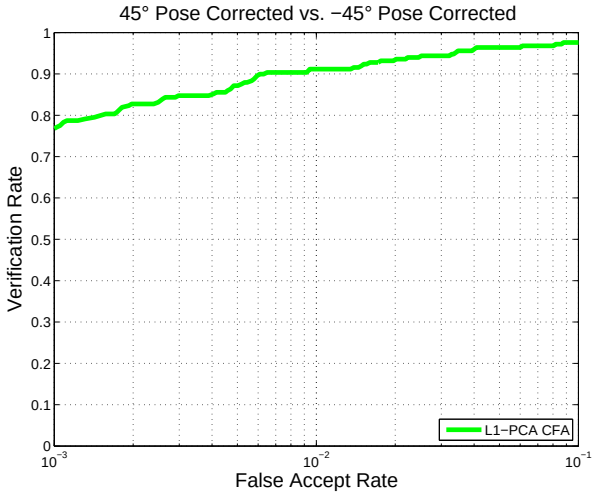
(c) $0°$ Pose-Corrected vs. $-45°$ Pose-Corrected



(d) $15°$ Pose-Corrected vs. $-45°$ Pose-Corrected



(e) $30°$ Pose-Corrected vs. $-45°$ Pose-Corrected



(f) $45°$ Pose-Corrected vs. $-45°$ Pose-Corrected

Figure 4.18: Verification performance on MPIE session1 with the non-frontal gallery images at $-45°$.

Feature Mode	Matching Mode	Angle of Gallery Images						
		$-45°$	$-30°$	$-15°$	$\mathbf{0°}$	$15°$	$30°$	$45°$
L1-PCA CFA	Verification	77.04	89.49	88.62	**86.01**	90.29	91.76	78.64
L1-PCA CFA	Rank-1 Identification	88.82	96.45	96.25	**95.18**	96.45	96.39	90.43

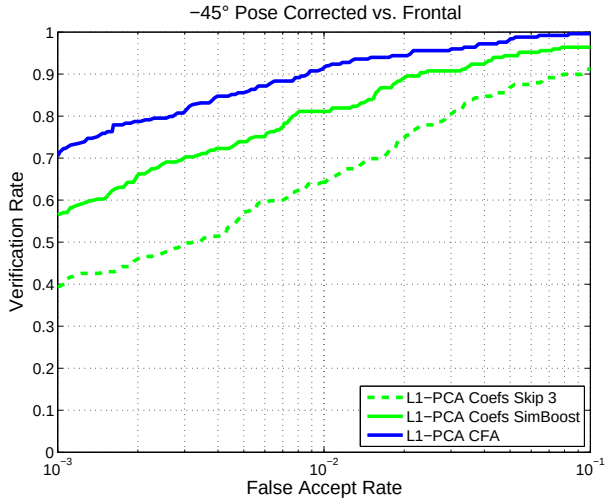
Table 4.2: Matching performance with a mixed-angle testing set. The verification rate is measured at 0.1% FAR using normalized cosine distance on L1-PCA CFA features. We indicate the angle of the gallery images. The corresponding test angles are all the remaining angles. For example, if the Gallery angle is $0°$, the test angles are $[-45°, -30°, -15°, 15°, 30°, 45°]$ combined together.

4.9 Summary of Findings and Results

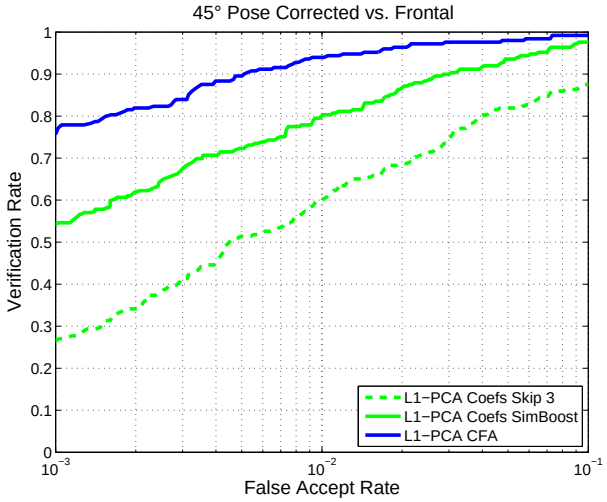
In this chapter we capitalized on the sparse feature extraction and reconstruction to achieve high one-to-one verification rates. Figure 4.19 summarizes the verification performance improvement using the different methods presented in this chapter.

Here are the main contributions of this chapter:

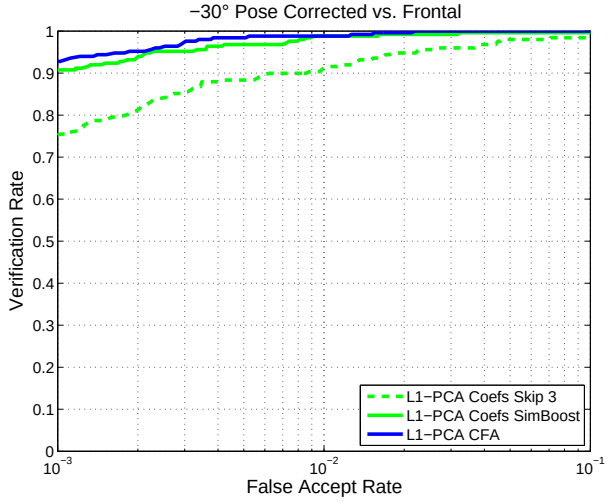
- Benchmarked the one-to-one verification results of all the different image representations presented in Chapter 3 and showed that the L1-PCA representation is the best.
- Empirically showed that pose-corrected shape information carries too much variance. Therefore, all pose-correction results took place in the shape-free representation.
- Showed that verification performance in coefficient domain can outperform that of reconstructed pixel domain. This proves that there is a lot of discriminative information available in the sparse feature extracted from the pose-corrected images.
- Developed a feature selection technique called SimBoost based on a non-linear weighted combination of similarities between feature vector components. This technique improved pose-correction verification rates significantly.
- Presented a CFA-based technique to perform robust face recognition.
- Benchmarked our technique against commercial face matchers and comprehensively outper-



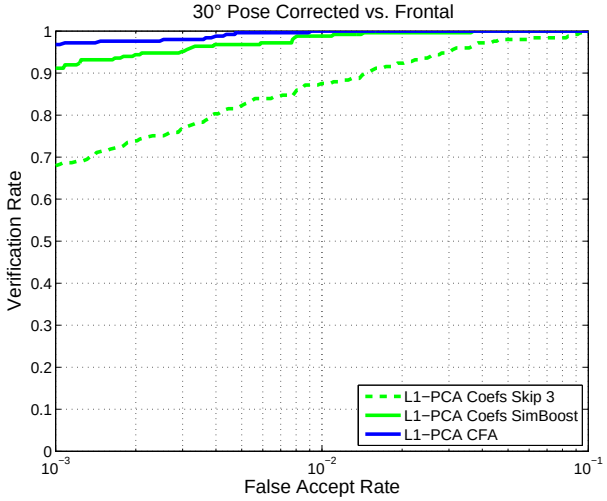
(a) $-45°$ PoseCorrection Comparison



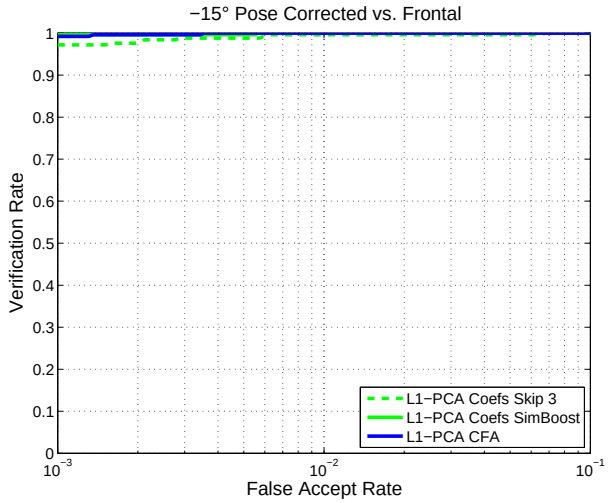
(b) $+45°$ PoseCorrection Comparison



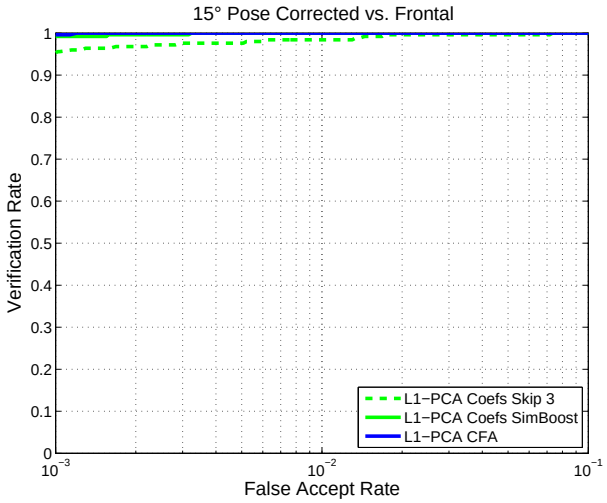
(c) $-30°$ PoseCorrection Comparison



(d) $+30°$ PoseCorrection Comparison



(e) $-15°$ PoseCorrection Comparison



(f) $+15°$ PoseCorrection Comparison

Figure 4.19: Verification performance on MPIE session1. The training set consists of 88 unseen subjects from MPIE sessions 2 and 3. All test subjects are unseen in the training. For each angle we depict the performance progression for the different methods presented in this chapter.

formed them.

- Since our matchers are not view-based, and have been trained on pose-corrected images, the gallery images are not restricted to being frontal. We showed that when matching gallery images at -45°, the verification rates do not significantly drop with test angles as extreme as $+45^{\circ}$.

Chapter 5

Application in Single Face Image Super-resolution

Super-resolution refers to the enhancement of the visual qualities of a low resolution image. Low resolution images are the result of how far the object of interest is from the camera, the specifications of the imaging sensor (and its tolerance to noise and low-light capabilities), and the quality of the optical lenses attached to the sensor. The widespread availability of affordable digital imaging devices, such as cell phone cameras, surveillance cameras, etc., comes as a result of the recent mass production and the shrinking in size of these devices. However, this does not always translate into increased image quality, and in general the visual quality of footage obtained by these devices remains poor. The problem is exacerbated by the ever-shrinking size of the CCD sensor which increases the amount of noise in the image, and the use of wide-angle lenses to increase the view angle but in return introduces barrel distortion. Moreover, the popular need for wireless communications with the devices has forced the manufacturers to implement aggressive compression algorithms to increase throughput, but in return, further introduce image artifacts and blurring to the footage. Finally, in surveillance applications, most cameras are placed high up on a wall or on a ceiling to offer them a vantage viewpoint. However, this makes the object of interest too distant

from the camera and decreases the number of pixels apportioned to the face.

The holy grail of face recognition is to build a system that can handle true real world data. This data emerges from surveillance footage, or mobile devices, where it is not only the viewpoint that poses a challenge, but also the low resolution image. Even though face recognition algorithms do not require megapixel images, having high-resolution texture information certainly helps. Most algorithms can produce respectable results with image resolutions as low as 50×50 pixels [66]. The problem is that most faces in the wild do not even meet these minimum size requirements. For instance, most surveillance cameras currently deployed consist of a wide-angle lens (to capture as much of the scene as possible) coupled to a 640×480 pixel sensor. A subject standing a few meters away will not reproduce a face with a significant number of pixels between the eyes, rendering it impossible even for a human operator to identify. In one of the earlier studies on this same topic, Bachmann [67] observes that there is an abrupt fall in identification efficiency by human operators when the faces become smaller than 24×18 pixels.

Super-resolution, or reversing this last problem, is a severely ill-posed problem, where the solution does not always exist, and when it exists, it is not unique and is very sensitive to perturbation in the input. Despite years of active research, it still remains an open challenge, particularly when trying to break the 4x magnification barrier in resolution. In this chapter, we apply the same technique we developed in Chapter 3 to attempt to solve the low resolution problem. We rely on sparse feature extraction and a multiresolution face model to develop a *single-image* super-resolution (or face hallucination) technique. If we can achieve great results using single image resolution, we can enhance this solution using a sequence of images.

This chapter is structured as follows: we first briefly overview some of the classical approaches to face super-resolution. We then adapt the method we developed earlier for pose-correction to handle the resolution problem. This involves learning a different face model based on a multiresolution pyramid of faces. We then evaluate our method on synthetic low-resolution images (obtained from downsampling the high resolution images). We also show that our method is inher-

ently robust to noise by reformulating it as a Bayesian approach. Finally, we analyze the case of non-frontal faces and show that theoretically our method represents a unified framework to solve the pose-correction and super-resolution problems simultaneously.

5.1 Background

Most single-image super-resolution techniques are limited because they are constrained by the number of available pixels. Early methods mostly relied on interpolation techniques, such as nearest-neighbor, bilinear or cubic B-spline interpolation [68] or convolution kernels [69]. Interpolation-based methods assume global continuity and maintains smoothness constraints that often produce results with blurred edges and textures which are essentially unusable in a face recognition framework. Edge-preserving interpolation techniques have been proposed, such as adaptive splines [70], and POCS (Projection Onto Convex Sets) interpolation [71]. Interpolation techniques aware of edge information were also proposed. Nonlinear interpolation with edge fitting [72] incorporated local edge fitting to avoid interpolation across edges. Similarly, edge-directed interpolation [73] makes use of edge orientation information to perform directional interpolation, and the interpolation occurs along an edge rather than across one. While interpolation gives satisfactory results when the input image size is reasonable, its performance abruptly drops when the input resolution is low. More powerful techniques are therefore needed to generate higher resolution images.

Another class of methods make the assumption that multiple low-resolution images available. Typical approaches of this kind relied on a registration method (such as cross-correlation) to correctly align the low resolution images, followed by the inversion of the resolution-reduction function (assumed to be a low-pass filter followed by a decimation operation), and use regularization to resolve the severely ill-posed nature of the problem [74]. A common way of regularizing the inverse problem was to cast the problem in a Bayesian framework (find the high-resolution estimate that maximizes the likelihood of observing the low-resolution image conditioned on the true high

resolution image). This is typically done by a assuming a prior on the image model and a noise-model [75, 76, 77]. Different Bayesian approaches use different priors (such as a Gaussian [78] distribution or a Gibbs [79] distribution). These priors express assumptions about the local relationship between the pixels values in the high-resolution image, and as a result act as a smoothing agent in the MAP optimization to encourage each pixel to take on the average of its neighbours in the reconstructed image. Another method [80] also used a Bayesian approach to estimate the unknown point spread function, by marginalizing the likelihood of the image registration parameters over the unknown high-resolution image using Gaussian process priors.

A large part of the earlier super-resolution algorithms are based on homogeneous (stationary) Markov Random Fields (MRFs) assumptions, since the Markovianity assumption lends itself very elegantly to super-resolution applications. This implies that a pixel value depends solely on the neighborhood of that pixel. The homogeneous part of their assumption implies that different neighborhoods have equal size. De Bonet *et al.* [81] proposed a non-parametric sampling method to infer high-frequency features from available low-frequency features in the context of multiresolution texture synthesis.

Super-resolution of face images is slightly different than super-resolution of general images, because we can make more assumptions about the structure of the human face, and have face-specific image priors. In their seminal work on super-resolution, Baker and Kanade [82] abandoned the MRF framework for a more general Bayesian MAP formulation which is more suitable for synthesizing global textures as with face images. The authors used a large number of training images to compute a multiresolution pyramid of features (such as Laplacian and Gradients features). Each level of the pyramid corresponds to a different resolution that is obtained by reducing the original native full resolution. In testing, given an input low-resolution image, they populate the top of the pyramid of features (that they call "the Parent Structure"), and for every pixel location, they exhaustively search (using the nearest-neighbor approach or Gradient descent) the training set for a pixel value that generates a similar feature vector. Moreover, they use a Bayesian framework to

incorporate more than one input low-resolution images. Assuming the input images have been accurately aligned together (using control points on the face or a registration method such as optical flow), the nearest-neighbor search is replaced by a maximum *a posteriori* computation that reflects the likelihood of observing several low-resolution pixel values and a prior on the high-resolution values.

Liu *et al.* [83] combined a global parametric model and local non-parametric approach in a two-step statistical approach. They assumed a high and low frequency mixture model, and for super-resolution, the reconstructed low-frequency information is inferred by a global linear model that learns the relationship between high and low resolution images, while the reconstructed high-frequency information is captured by learning the residual between the original high-resolution and the super-resolution image using a patch-based nonparametric Markov network.

In [84], a simple PCA-based global approach was attempted. Given high-resolution training images, they downsample them and build a PCA subspace of reduced-resolution images, where they project the input low-resolution image to obtain the PCA coefficients. The authors use those low-resolution induced coefficients to reconstruct the high-resolution equivalent face using the full-resolution eigenfaces. To make sure that the final reconstruction is "face-like", they introduce artificial constraints on the coefficient values. These constraints are a function of the eigenvalues learned from training for every principal component.

A more recent work [1] exploited the properties of sparse image representation for single-image super resolution. Their patch-based local approach simultaneously learns two distinct over-complete dictionaries. One for high-resolution patches, and one for low-resolution ones. Given an input low-resolution patch, its sparse representation in the low-resolution dictionary is used to recover a high-resolution patch from the high-resolution dictionary. Local consistency is achieved by requiring the patches to overlap, and requiring the high-resolution patches to agree on the overlapped area. In the specific case of face hallucination, a two-step approach was adopted, similar to previous methods. The first step is a global approach to reconstruct what they call a "medium

high-resolution" smooth face using NMF [38], which is then enhanced with high-frequency information using the local patch-based approach that makes use of the sparse feature extraction using the coupled dictionaries learned at the training stage.

Another recent work [85] exploited sparsity. The basic idea in this work is to use kernel ridge regression to learn the mapping between low and high resolution images. To avoid blurring artifacts, a post-processing step that relies on a prior model is employed.

5.2 Sparse Feature Extraction for Hallucinating Faces

Global face super-resolution approaches are criticized for not being able to render sharp edges and high-frequency information, or that PCA-based methods often degenerate towards the mean face. However, we have empirically shown in Chapter 3 that ℓ_1-minimization did not exhibit any of these problems, and the sparse PCA feature extraction preserved sharp edges in the reconstruction.

In this section we borrow several concepts from Chapter 3 to achieve face super-resolution. Since the problem at hand is a texture reconstruction problem, we first rely on the shape-free representation that decouples the shape information from the text information.

One of L1-PCA's main strengths are its ability to produce a "good" solution in the PCA space in the presence of severe occlusions (missing data dimensions) leading to a highly underdetermined system. The ℓ_1-minimisation solution produced a sparse feature vector that we showed to be largely tolerant to pose when properly utilized. To export the same benefits to the hallucinating face problem and we utilize a high-dimensional multiresolution PCA subspace (represented by a matrix $\mathbf{V}$ of vectorized eigenfaces and a mean vector $\mathbf{m}$) trained on a Gaussian pyramid [86] of training images (the same images used for training the pose-correction subspace). This subspace will represent the correlation between the equivalent pixels in high and low resolution. Similar to [82], this pyramid will enable us to avoid explicitly modeling the resolution-reduction function or parameters of a point-spread-function.

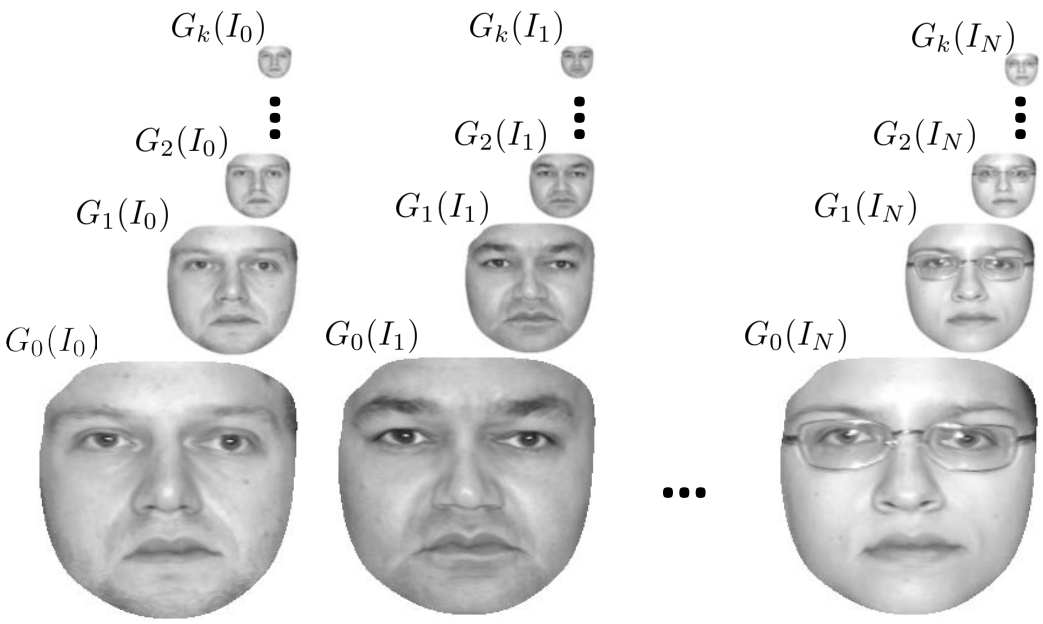


Figure 5.1: Depiction of a Gaussian Pyramid of k levels for N face images in the shape-free representation taken from the MPIE dataset.

Given a training image I_i, the Gaussian pyramid $G_0(I_i), \ldots, G_k(I_i)$ for such an image is depicted in Figure 5.1. Following [86], the bottom level of the pyramid is the image itself, and every subsequent level is obtained by $G_{i+1}(I) = \text{REDUCE}\,(G_i(I))$ where the REDUCE operator is defined by the following equation:

$$\text{REDUCE}(I)\,[i,j] = \sum_{m=1}^{5}\sum_{n=1}^{5} w\,[m,n]\,I\,[2i+m, 2j+n] \tag{5.1}$$

w in this case is a 5×5 low-pass Gaussian filter kernel. We build the training subspace by concatenating the vectorized images of all k levels into a single column. For testing, we assume we only have access to the k^{th} level of the pyramid of an unseen face image. The active dimensions in this current "missing-data" problem are the dimensions corresponding to the pixels of the input low-resolution image. The missing dimensions are the dimensions corresponding to the pixels of the images in the lower levels of the pyramid. In the case where a middle resolution image is available and as is common in most super-resolution algorithms that rely on Gaussian pyramids,

the smaller levels of the pyramid can be populated, and the active dimensions become the top of the pyramid, while the missing dimensions are at the bottom of the pyramid.

For notational simplicity, let $\mathbf{x}'$ be the vector of active pixels of $\mathbf{x}$ provided by the l^{th} level of the Gaussian pyramid and let $\mathbf{x}'$ be of size d' and $d' < d$ pixels. Similarly, let $\mathbf{m}'$ be the mean of the active mean pixels. For notational simplicity, we can also introduce $\mathbf{x}'_{\mathbf{c}} = \mathbf{x}' - \mathbf{m}'$, which represents the centered version of $\mathbf{x}'$ and is of size d'. $\mathbf{V}'$ is the matrix of active rows that are in $\mathbf{V}$. We need to solve for the coefficient vector $\mathbf{c}$. As in L1-PCA, we can solve for the following cost function:

$$\min \|\mathbf{c}\|_1 \text{ subject to } \|\mathbf{V}'\mathbf{c} - \mathbf{x}'_{\mathbf{c}}\|_2 \leq \epsilon \tag{5.2}$$

Given the solution coefficient vector $\mathbf{c}_{\text{L1-PCA}}$, we can simply reconstruct the entire pyramid column $\mathbf{x}_{\text{L1-PCA}} = \mathbf{V}\mathbf{c}_{\text{L1-PCA}} + \mathbf{m}$, and extract whichever level of the pyramid we want. Figure 5.3 depicts the reconstruction results for different levels magnification levels (2^k).

Our low-resolution parent-structure feature extraction step is similar to the one described in [82], but unlike in [82], our method is global and relies on the ℓ_1-minimization to avoid degenerating towards a mean face. Unlike [84] where the coefficient vector $\mathbf{c}$ is extracted by projecting on low-resolution eigenfaces and reconstructing using high-resolution eigenfaces (and which does not provide any theory as to what ties the two subspaces together), our subspace learns the coupled high-and-low frequency from the training Gaussian pyramids (or equivalent features) simultaneously and hence our eigenfaces are multiresolution so no artificial constraints need to be introduced to guarantee a smooth face-like reconstruction.

More importantly, the simplicity of our method keeps it modular and allows for further improvements. For instance, it could replace the global reconstruction part of any two-step approach (such as [1, 83]) which first compute a rough global reconstruction and further enhance it with a slower and more specialized local patch-based algorithm. Similarly, the Gaussian pyramid-based

subspace could be augmented with any other features such as a Laplacian pyramid or gradient-based features.

5.3 Evaluation on Experimental Data

We evaluate L1-PCA for global face hallucination by downsizing the original MPIE session 1 images in the shape-free domain. The original shape free images provide an interocular distance of 100 pixels, and when we reconstruct a lower resolution image, we recover that original size. We evaluate the quality of our reconstruction for different magnification factors (different k-levels of the Gaussian pyramid).

We benchmark results against the best interpolation techniques, such as bicubic polynomial interpolation [69] and Lanczos resampling which uses a Lanczos kernel (a windowed sinc function) to smoothly interpolate the value between samples. This latter method is usually employed by most commercial photo displaying software. We also benchmark against cubic B-spline interpolation [70] which marginally outperforms traditional bicubic polynomial interpolation. Table 1 summarizes the average PSNR obtained for 249 reconstructions using the different methods. The PSNR between two images I and J, given by Equation 5.3, is a common objective image reconstruction quality metric typically used in denoising and image/video compression applications. As the MSE approaches zero, the PSNR goes to infinity. In lossy image/video compression, the typical PSNR values range from 30 to 50 dB, and anything below 20 dB is deemed unacceptable [87]. As is also common, since the PSNR is a logarithmic scale, the average PSNR reported was computed by first measuring the average MSE and then converting the average MSE to PSNR rather than averaging PSNR values directly.

$$\text{PSNR}(I,J) = 20 \cdot \log_{10}\left(\frac{255}{\sqrt{\text{MSE}(I,J)}}\right) \tag{5.3}$$

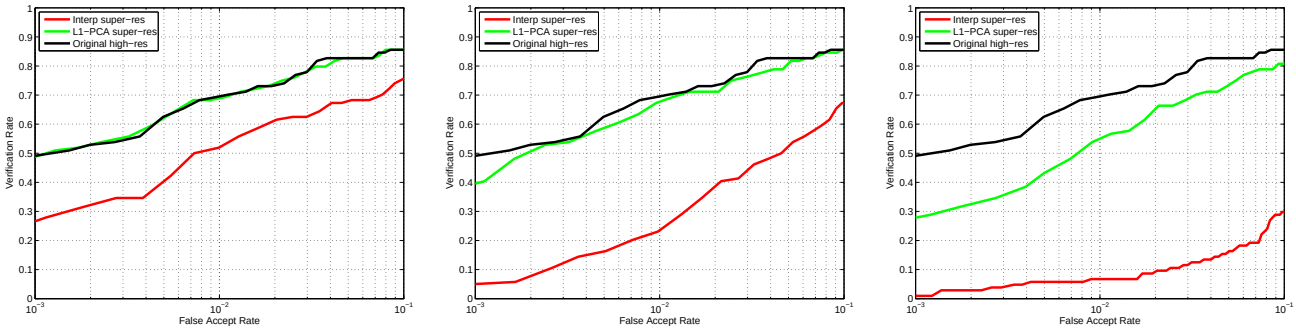
Interocular distance	Magnif. factor	Average PSNR (in dB)					
		Interpolation					
		Bicubic	Cub B-spline	Lanczos3	Yang *et al.*	Kim *et al.*	**L1-PCA**
25 pixels	4x	24.21	24.23	24.32	24.53	24.37	**32.80**
12.5 pixels	8x	20.59	21.04	20.66	20.77	20.57	**28.97**
6.25 pixels	16x	8.47	19.14	8.46	17.84	17.66	**26.33**

Table 5.1: Summary of the average PSNR (in dB) for different super-resolution techniques.

The results tabulated in Table 5.1 represent the average PSNR in dB produced by different magnification rates (starting from different interocular distances) and for different techniques.

Whereas the MSE and PSNR metrics quantify the faithfulness of the reconstruction to the original face, our next experiment measures how much of the discriminating information of the original face the reconstruction retains. For that we set up a simple face matching experiment, and measure the verification rate at different magnification levels, summarized by the ROCs in Figure 5.2. The gallery set consists of the original high-resolution MPIE session 1 shape-free faces, while the query set contains 341 subjects with 104 of them seen in the gallery set (same subjects, but different images as they originate from MPIE session 2) while the rest of them are unseen in the gallery set (a mix of MPIE session 2, session 3, and FERET faces). The input query face images have been downsized to provide 25, 12.5 and 6.25 pixels between the eyes and then reconstructed to the original 100 pixels between the eyes, to provide the magnification ratios of 4, 8 and 16 respectively. Normalized cosine distance is the metric used for the simple matcher. Figure 5.2 shows that L1-PCA hallucinated faces with 4x magnification offer the same discrimination between faces as the original high resolution images. Naturally the ROCs drop for high magnification ratios, but the relative drop for L1-PCA is significantly less than for naive interpolation.

We also benchmark our L1-PCA method for super-resolution against the method in [1] (an im-



(a) 4x magnification (b) 8x magnification (c) 16x magnification

Figure 5.2: Effect of super resolution on cross-session matching. The gallery images are 249 full-resolution face images from MPIE session 1. The query images are 341 reconstructed super-resolution images from MPIE sessions 2, 3 and FERET. 104 of test subjects are seen in the gallery set. Our face hallucination method handles the drop in resolution much more gracefully than bicubic interpolation. With an input resolution of 25 pixels between the eyes, our super-resolution technique fares as well as the original high-resolution in this face verification experiment.

plementation of the approach is available [88]) since theoretically this approach is closest to ours. The differences are that our approach is global, and our dictionary is given by Gaussian-pyramid face model, while theirs is twofold, one global based on NMF to generate a smooth intermediate face, and then an expensive local patch-based approach to infer high-frequency information. The intermediate global face provides patches whose sparse representation (from a low-resolution dictionary) is used to generate a high-resolution patch (using the high-resolution dictionary, and both dictionaries were learned jointly)[1]. Since their method is local patch-based, their storage requirement is much lower than our method, since they only need to keep the two compact dictionaries. On the other hand, their method is much slower because of the number of overlapping patches it needs to process, while our method extracts the sparse representation once from the entire low-resolution face. Even though their method marginally beats interpolation methods at 4x

[1]We retrained the method of [1] using the software made publicly available by the first author using different crops and different number of faces, and different dictionary sizes. The results we show use the following parameters: $\lambda = 0.1$, 100000 patches, dictionary size $= 1024$, upscaling factor 2, patch size $= 5$, overlap $= 4$, 1000 shape-free tightly-cropped and registered training faces. The results reported were obtained without the use of their backprojection global method which seemed to hurt the MSE of the reconstructed high-resolution face image.

magnification, it breaks down for lower input resolutions and returns a very blurry reconstruction[2].

We also benchmark against the method of [85]. An open-source implementation by the author is available here [89]. However, we could not retrain their algorithm on face images only so we used it as is. It is important to note that their magnification factor was limited to 4x, so for higher magnification rates we ran the algorithm twice on the same image.

5.4 Sensitivity to Noise

So far we have assumed that input signal is clean. The motivation for super resolution was low resolution faces that come from poor quality surveillance cameras or video footage. In these cases, the signal will most likely be contaminated with noise. One could first denoise the image and then apply super-resolution. However, we can demonstrate that our approach can intrinsically handle noise without an explicit denoising step. Let's assume that our signal $\mathbf{x} = \mathbf{V}\mathbf{c} + n$ is corrupted with additive white Gaussian noise n (white noise with a constant spectral density and a Gaussian distribution of amplitude). Let the noise be normally distributed with zero mean and variance σ^2.

Equation 5.2 can be reformulated using Lagrange multipliers

$$\arg\min_{\mathbf{c}} \|\mathbf{V}'\mathbf{c} - \mathbf{x}'_{\mathbf{c}}\|_2^2 + \lambda \|\mathbf{c}\|_1 \tag{5.4}$$

This in return corresponds to the traditional Bayesian formulation that seeks to maximise the posterior probability:

$$P(\mathbf{c}|\mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{c})P(\mathbf{c})}{P(\mathbf{x})} \tag{5.5}$$

which corresponds to solving the following MAP problem:

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} P(\mathbf{c})P(\mathbf{x}|\mathbf{c}) \tag{5.6}$$

[2]We tried different product of upscale ratios, for example to achieve 8x, upscale with a factor of 2 three times, or with a factor of 4 followed by a factor of 2.

where the prior $P(\mathbf{c})$ is assumed to be Laplacian with location parameter 0 and scale parameter b given by:

$$P(\mathbf{c}) = \frac{1}{2b}\exp\left(-\frac{\|\mathbf{c}\|_1}{b}\right) \tag{5.7}$$

Since $n \sim N(0, \sigma^2)$ the likelihood $P(\mathbf{x}|\mathbf{c})$ is also normally distributed with mean $\mathbf{Vc}$ and variance σ^2:

$$P(\mathbf{x}|\mathbf{c}) = \frac{1}{2\sigma^2}\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Vc}-\mathbf{x}\|_2^2\right) \tag{5.8}$$

Combining Equations 5.8 and 5.7 back in Equation 5.5 translates into maximizing the sum of the exponents, or minimizing the negative of the sum of the exponents:

$$\arg\min_{\mathbf{c}} \frac{1}{2\sigma^2}\|\mathbf{Vc}-\mathbf{x}\|_2^2 + \frac{\|\mathbf{c}\|_1}{b} \tag{5.9}$$

Equations 5.9 and 5.4 are identical for $\lambda = \sigma^2/b$. This shows that our technique is inherently a Bayesian approach that is designed to handle noise by modeling the prior of the coefficient vector $\mathbf{c}$, assuming white Gaussian noise, and looking for the solution vector that is optimal is the MAP-sense. λ controls the sparsity of the solution, so the sparser the model, the bigger σ^2 can be (assuming b to be constant) which means the more noise in the data our model assumes. Figure 5.6 depicts noisy image super resolution reconstruction. The images have been corrupted by the AWGN channel with variance increasing by a factor of 10 in each case ($\sigma^2 = 0.0001$, 0.001 and 0.001 respectively). The parameter ν that controls the sparsity of the L1-solver had to be increased accordingly. We selected the optimal ν by finding the optimal value with respect to reconstruction error on a validation set. As predicted by our above calculation, those optimal values were greater by a factor of 10 each time, confirming our model.

This last analysis highlights the similarities between our method and standard Bayesian super-resolution approaches [75, 77, 78, 79, 82]. The most significant contrasts is that their methods operate in pixel space (or a related space, such as Laplacian or Gradient features), while we merely

operate in the PCA-space of a shape-free Gaussian pyramid representation. Moreover, our smoothness prior is an ℓ_1-based measure which prevents the reconstruction from degenerating towards the mean, which is often the case in traditional Bayesian super-resolution techniques.

5.5 Pose-Correction with Low Resolution Images

The remarkable reconstruction results our method achieved with magnification rates of 16x so far assumed a frontal low-resolution input face. However, the technique of section 5.2 is essentially the same as the one we followed for pose-correction, and this has great implications: we could potentially achieve pose-correction *and* super-resolution in the same unified framework provided by the sparse feature extraction. This section investigates this idea further.

5.5.1 Pose-Correction Sensitivity to Low Resolution

We first measure the impact of resolution on our pose-correction method outlined in Chapter 3. In that chapter, the 3D model that 3DGEM built for the shape-free representation set the distance from the camera to the face to a certain value that guarantees an interocular distance of 100 pixels. All training and testing shape-free images were preprocessed in this fashion to generate images with 100 pixels between the eyes. In this experiment, we use the same training data (and hence the same face model), and the gallery images are still assumed to be of full resolution. What changes now is that the non-frontal images have been downsampled to offer a lower resolution and fewer pixels between the eyes (measured in the frontal viewpoint). This, in return, reduces the amount of non-frontal shape and texture information available in the testing images. In this experiment, 3DGEM still performs the de-rotation of the face at the full resolution, and it is 3DGEM's texture mapping module that interpolates the texture information up to 100 pixels between the eyes. This essentially corresponds to extracting features from the upsampled non-frontal faces.

We first measure the impact of low-resolution on the L1-PCA coefficients. Figure 5.7 shows the

ROCs obtained for the experiment described earlier using a simple Normalized Cosine Distance (NCD) measure on the coefficients directly, without passing them through the CFA algorithm, which will compensate for a drop in discriminative power and bias the results positively. The presence of pose makes the task harder on super-resolution, and it is not surprising that we see a drop in performance when matching across low-resolution *and* non-frontal faces. In Figure 5.2, it can be seen that with 25 pixels between the eyes, the verification rates remains largely unaffected, but in the presence of pose, we need as many as 50 pixels between the eyes to leave the overall performance unaffected. Figure 5.7 shows the ROCs obtained for verification using the L1-PCA coefficients extracted from the low-resolution non-frontal test images. We can see that degradation is controlled, and with 25 pixels between the eyes, the performance does not deviate too much from the original resolution. However the loss of discriminative information in smaller resolutions drops the verification rates significantly as we start to lose features. Note that we can even detect a minor improvement at the $\pm 45^\circ$ angles with 50 interocular pixels, since the downsampling cancels out some landmarking human errors.

We now apply the CFA algorithm on the L1-PCA coefficients extracted from the low-resolution face images, to measure how much variation in resolution can the CFA algorithm handle and correct for. Figure 5.8 shows the ROCs obtained for the low-resolution images. We also report the rank-1 identification rates as the testing resolution drops. Table D.4[3] summarizes all the matching results and Figure 5.9 presents a bar graph of the identification rates of the L1-PCA CFA algorithm. We see virtually no drop at half the resolution (50 pixels) and an average drop of 15% at the quarter of the resolution (25 pixels). Figure 5.8 shows the corresponding verification rates.

It must be noted that for the first three resolutions the identification rates remain somewhat flat across the testing angles, which indicates that our algorithm behaves predictably when dealing with low-resolutions. It is only at the lowest resolution that the performance finally breaks down. With only 13 pixels between the eyes for a frontal viewpoint, the pose image at $\pm 45^\circ$ offers very

[3] see Appendix D for this result

little information for our algorithm to work with without an explicit super-resolution step. This hypothesis is validated by the observation that the verification ROCs for these extreme angles at that resolution are essentially the same whether we use the CFA algorithm or the simple NCD matcher. Most of the discriminative information has been lost. Hence the need to extract super-resolution features from the low-resolution non-frontal image, which we briefly introduce in the following section and leave as future work.

5.5.2 Sparse Feature Extraction from Non-frontal Low Resolution Faces

Previously, we were just measuring the impact of low-resolution images on the pose-correction technique as presented and described in Chapter 3, without actively correcting for low-resolution. The sparse feature was extracted from the input images interpolated up to the face model's size. However, for magnification factors greater than 3x, the performance visibly starts to degrade. We now investigate whether actively correcting for low-resolution by extracting features in the multiresolution pyramid and performing pose correction using the sparse feature vector can further improve our results.

Matching With Low-Resolution Non-frontal Faces

When matching across different resolutions, several approaches have been researched over the years. The most common approach is to enhance the low-resolution image using super-resolution techniques, and then perform the matching. Alternatively, one could downsample the high-resolution gallery faces and match at low resolutions. As a compromise, several approaches that combine super-resolution and face matching in the same framework have been studied. In [90], a method called S^2R^2 that simultaneously performs super-resolution and recognition was proposed. The authors of [90] augment the super-resolution formulation by incorporating terms that reflect the classifier's metric. They repeat the optimization for every gallery subject to obtain impressive results. In [91], a generative model that separates the illumination and downsampling effects was

proposed. The authors show that the downsampling effects are person-specific and propose a statistical method to learn the parameters of a subject model. In [92], the authors use Multidimensional Scaling (MDS) to map low and high-resolution faces to a Euclidean space where the separation between classes is maintained. In [93], they presented a Bayesian framework to perform super-resolution in tensor space. Given a low resolution face, they directly compute a maximum likelihood identity parameter vector in the high-resolution tensor space. This can be used for either reconstruction or recognition.

Our pose-correction framework lends itself very easily to extracting the feature vector from the low-resolution input faces, just as we did to achieve super-resolution. Unifying our super-resolution and pose-correction frameworks, and similar to Equation 3.18, we can now extract the following coefficient feature-vector:

$$\min \|\mathbf{c}\|_1 \text{ subject to } \|\mathbf{V}'_{\mathbf{MR}}\mathbf{c} - \mathbf{x}'_{\mathbf{MR}_c}\|_2 \leq \epsilon \tag{5.10}$$

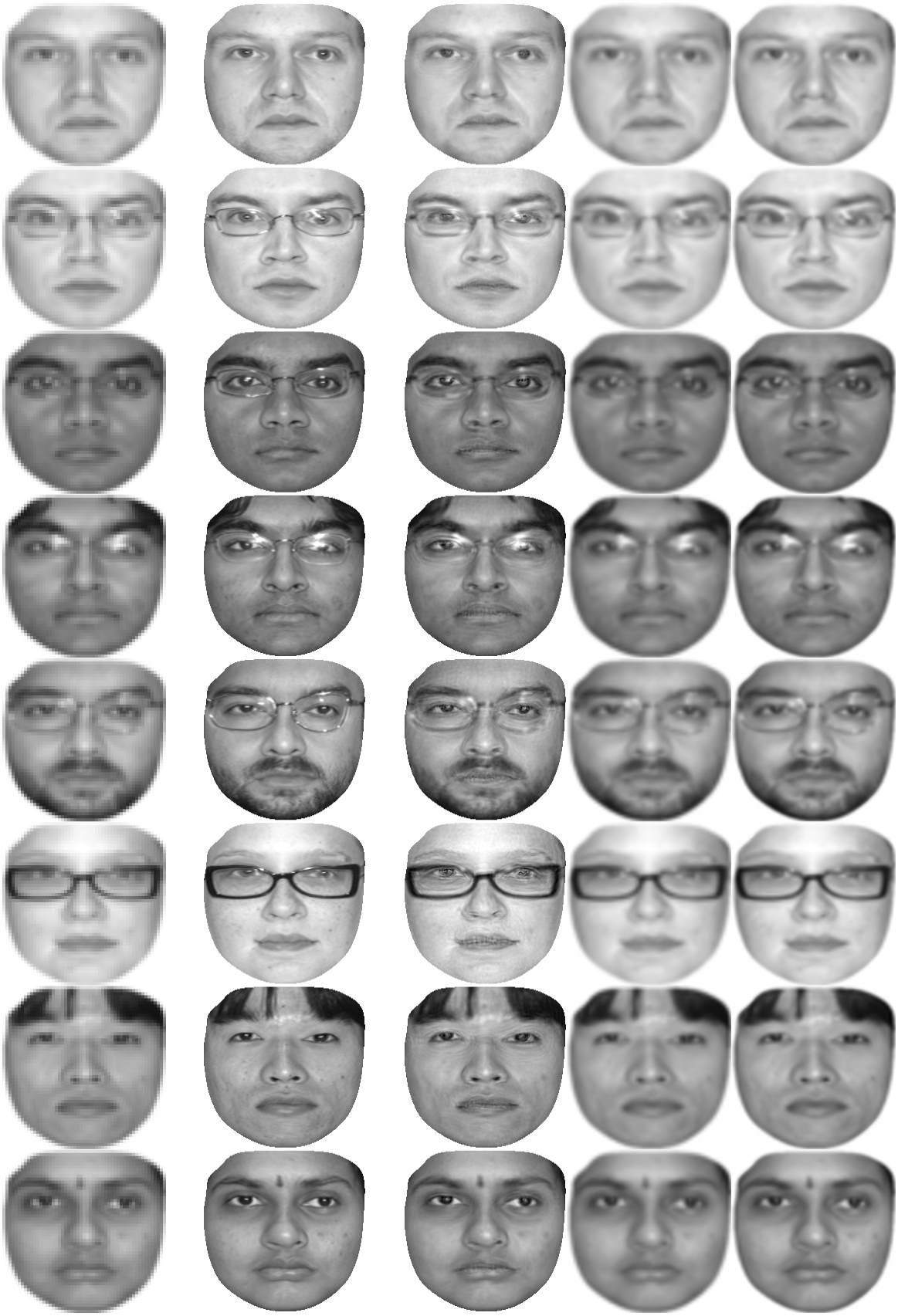
where $\mathbf{V}'_{\mathbf{MR}}$ represents the matrix of active rows of $\mathbf{V}_{\mathbf{MR}}$ which is the matrix of vectorized *multiresolution* eigenfaces. Similarly, $\mathbf{x}'_{\mathbf{MR}}$ is the vector of observed *low-resolution and pose-corrected* pixels in the *multiresolution* vector $\mathbf{x}$. In other words, the same methodology we presented in Chapter 3, can be applied in a low-resolution setting, by swapping the subspace in which we extract features with a multiresolutional subspace. We leave this last method to be pursued as future work.

5.6 Summary of Results

In this chapter we achieved the following:

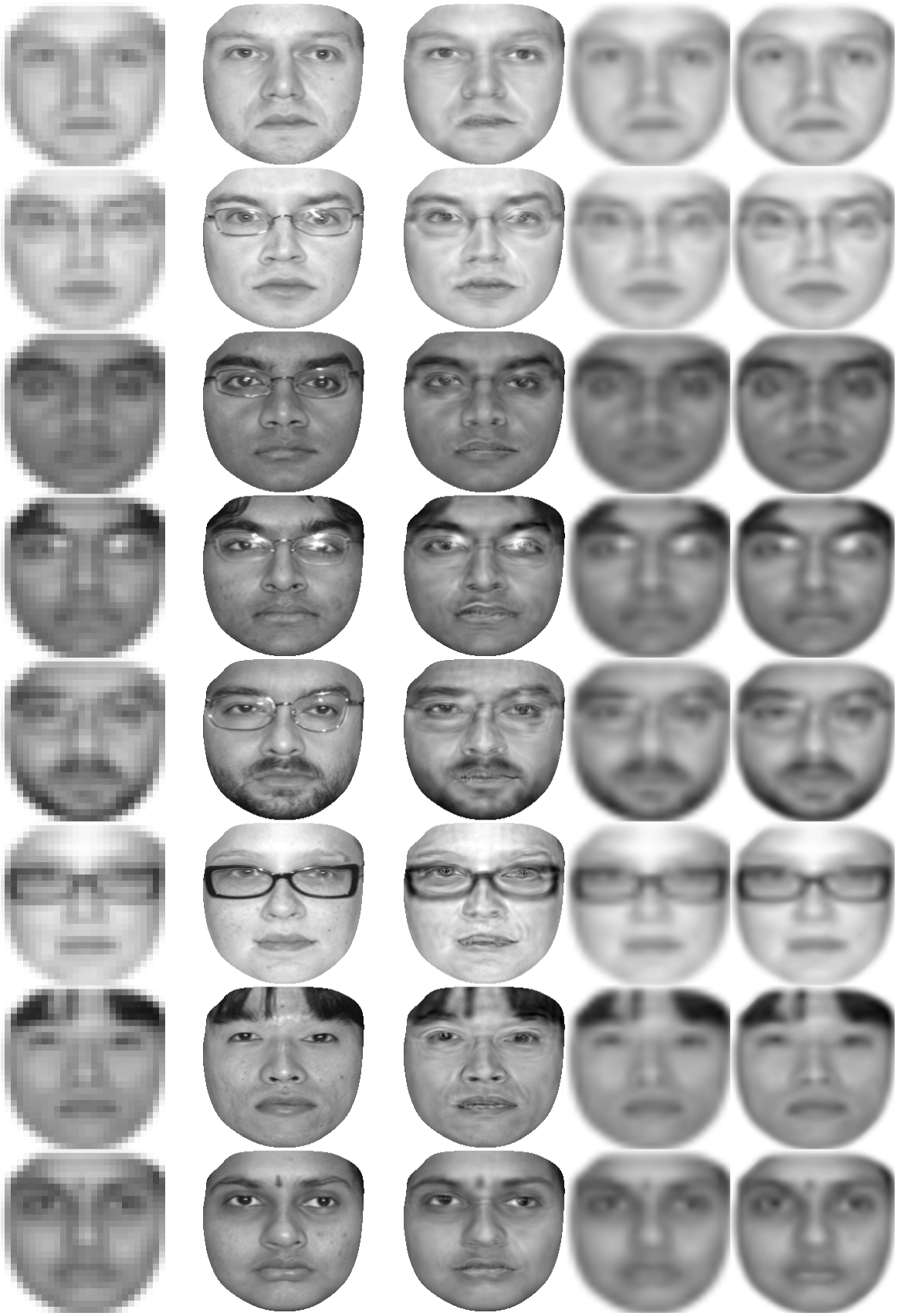
- Showed that our L1-PCA method can be interpreted as a Bayesian approach with a Laplacian prior on the coefficient vector $\mathbf{c}$.

- Showed that L1-PCA has an inherent robustness to noise as a result of the previously described bullet point.
- Showed that L1-PCA can be used as a single-image face super-resolution technique to reconstruct a global face.
- Showed that this super-resolution application of L1-PCA fits the same framework that we used for pose-correction and therefore can be used to achieve both.
- Showed that our pose-correction results from Chapter 3 are robust to low resolution input images down to 25 pixels between the eyes (for the equivalent frontal viewpoint).



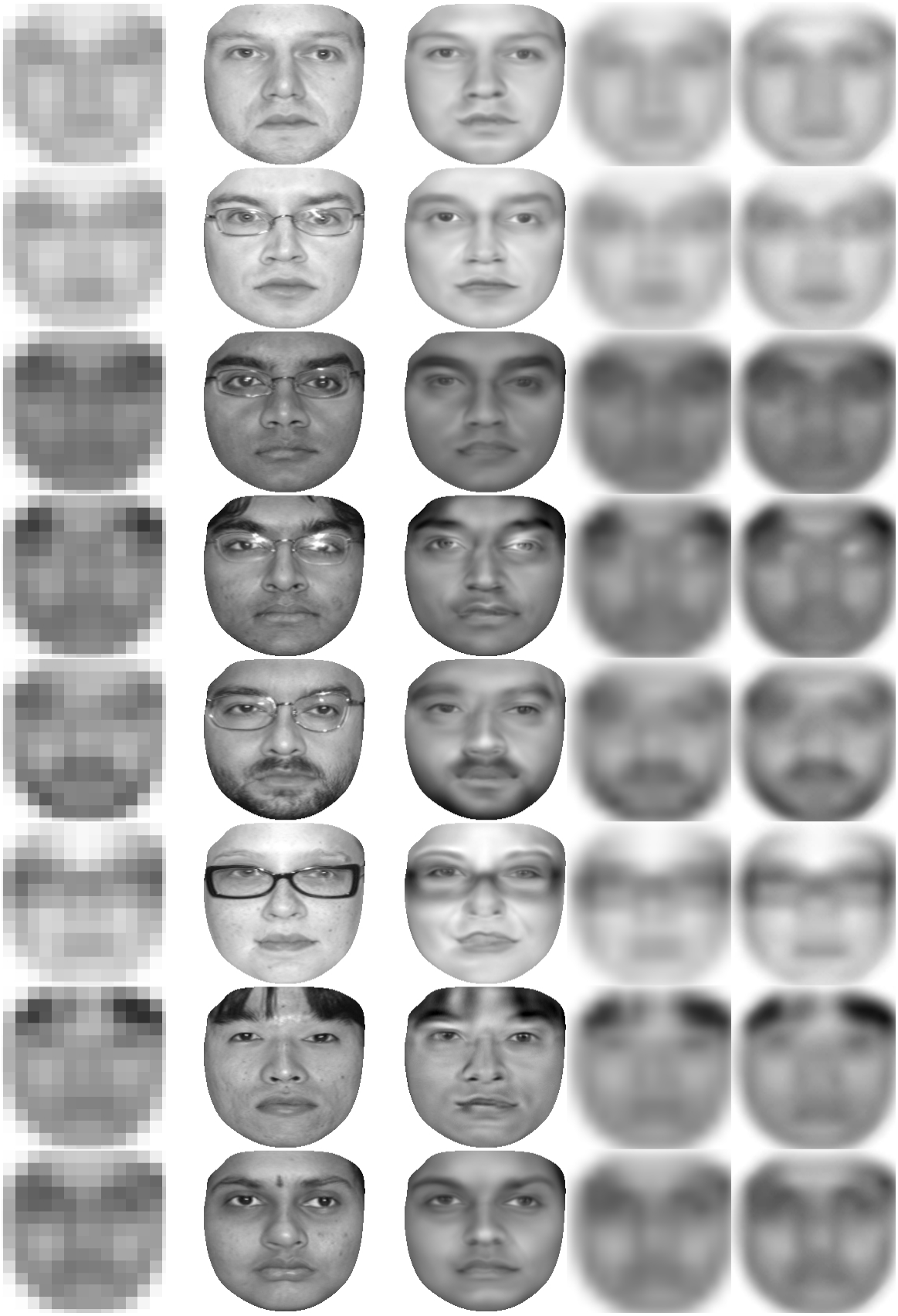
(a) Input (b) Original (c) L1-PCA (d) Bspline (e) Yang *et al.*

Figure 5.3: 4x magnification results. Starting with 25 pixels between the eyes (a) input image (b) original (c) L1-PCA (d) cubic B-spline (e) the method of [1].



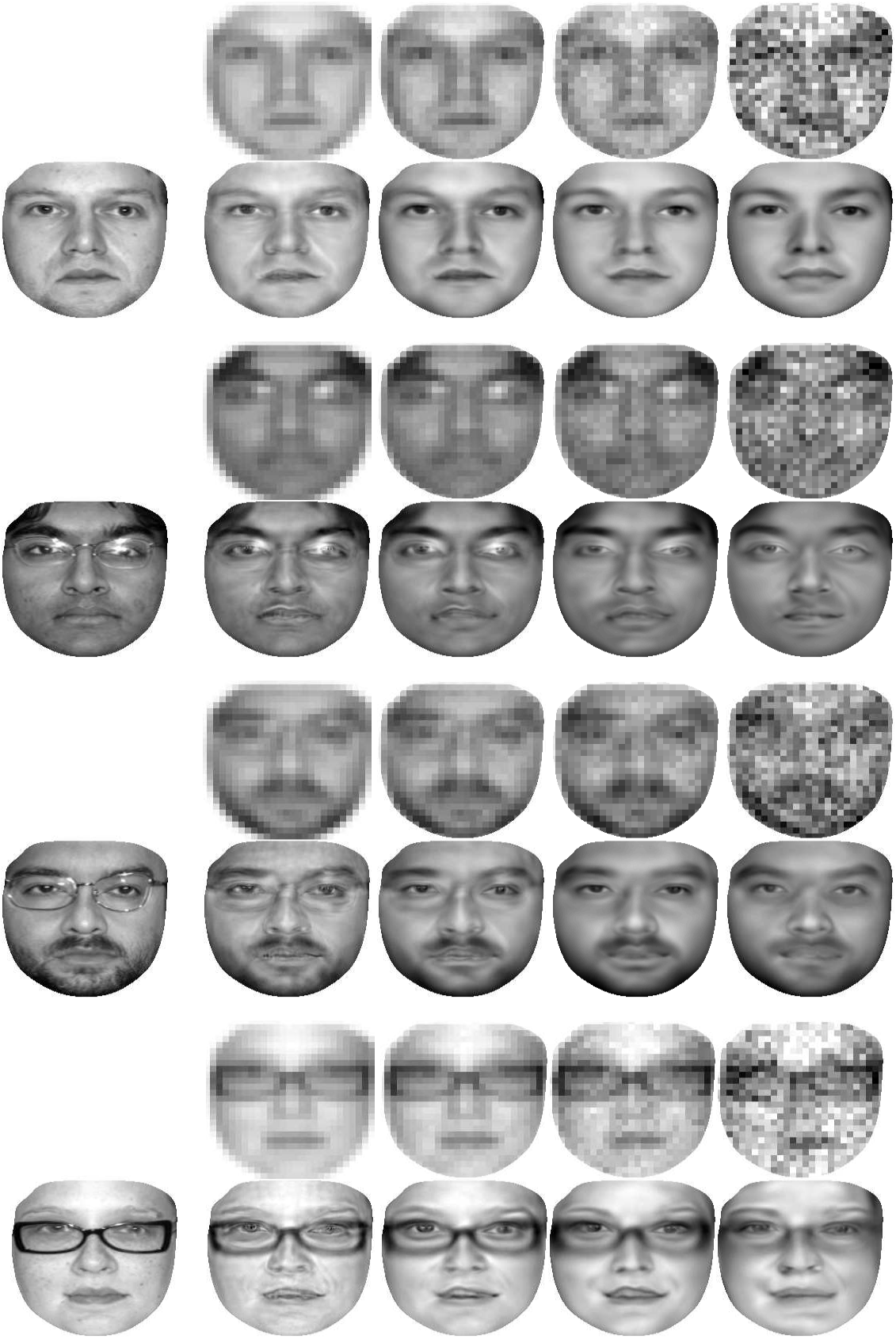
(a) Input (b) Original (c) L1-PCA (d) Bspline (e) Yang *et al.*

Figure 5.4: 8x magnification results. Starting with 12.5 pixels between the eyes (a) input image (b) original (c) L1-PCA (d) cubic B-spline (e) the method of [1].



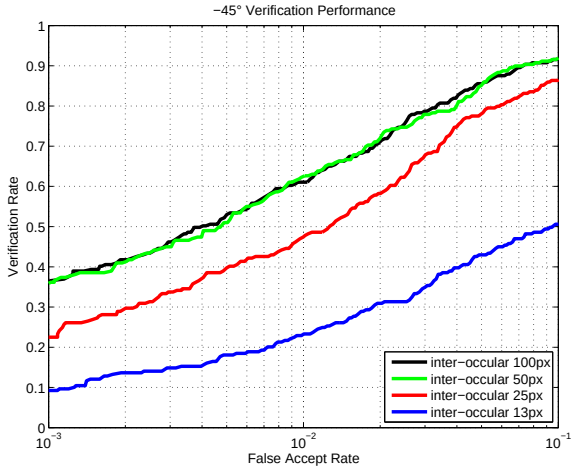
(a) Input (b) Original (c) L1-PCA (d) Bspline (e) Yang *et al.*

Figure 5.5: 16x magnification results. Starting with 6.75 pixels between the eyes (a) input image (b) original (c) L1-PCA (d) cubic B-spline (e) the method of [1].

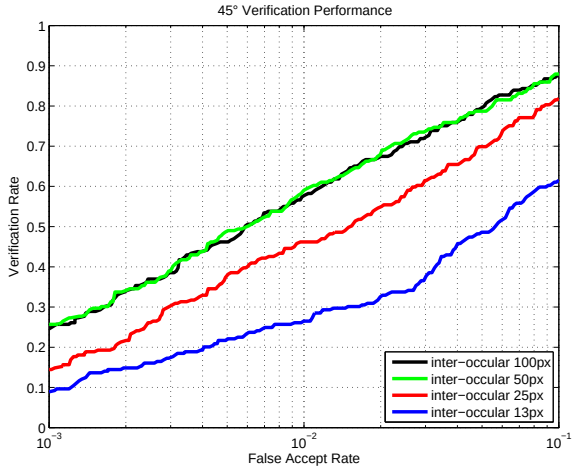


(a) Original (b) No Noise (c) $\sigma^2 = 0.0001$ (d) $\sigma^2 = 0.001$ (e) $\sigma^2 = 0.01$

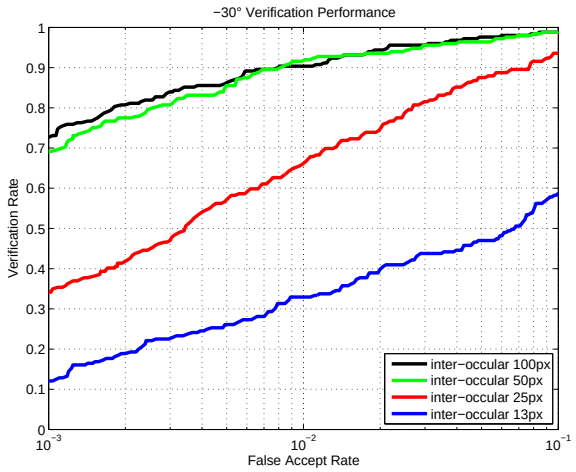
Figure 5.6: Noise Tolerance. Reconstruction with 8x magnification, starting from an interocular distance of 12.5 pixels. The low resolution images were corrupted with AWGN of increasing variance.



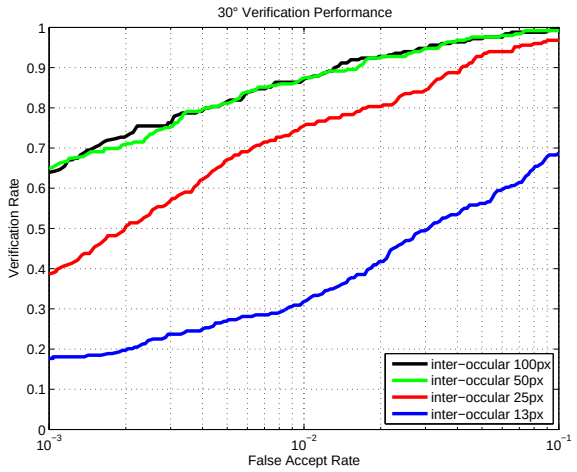
(a) Sensitivity at $-45°$



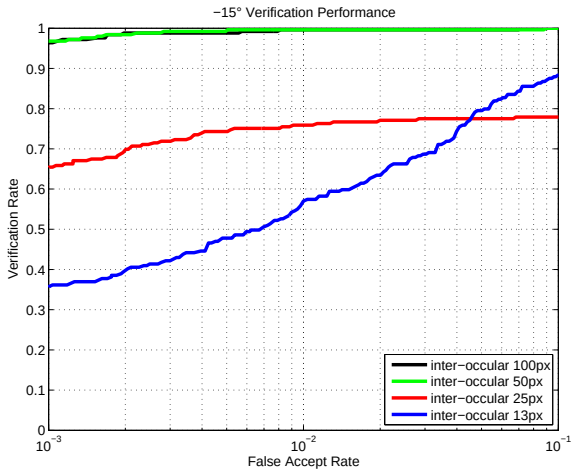
(b) Sensitivity at $45°$



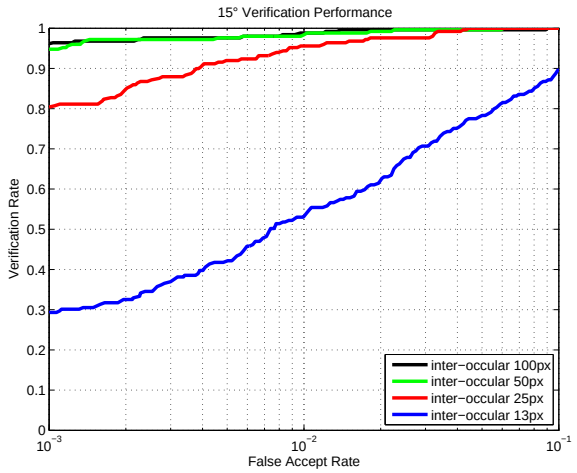
(c) Sensitivity at $-30°$



(d) Sensitivity at $30°$

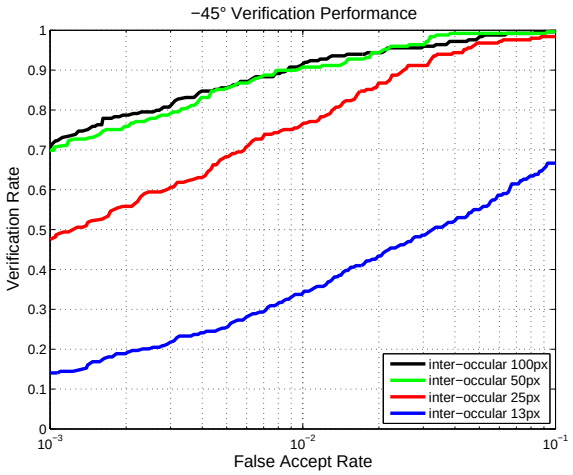


(e) Sensitivity at $-15°$

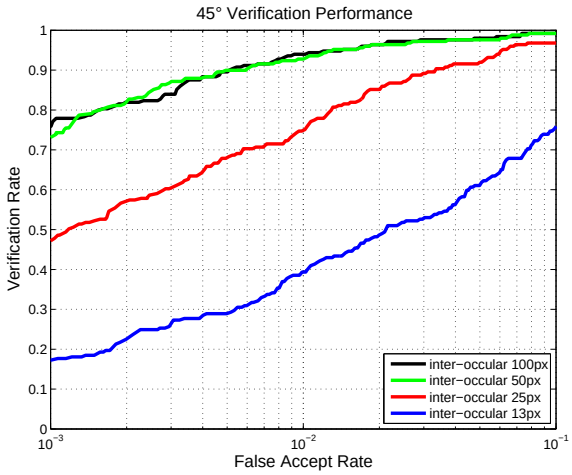


(f) Sensitivity at $15°$

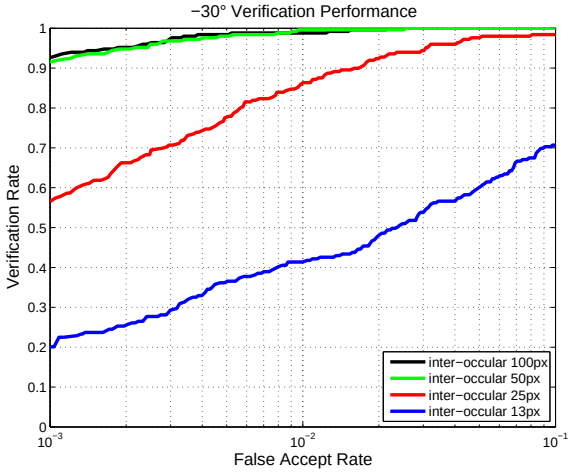
Figure 5.7: Verification rate of pose-corrected images as a function of query image interocular distance. The matcher used is normalized cosine-distance in coefficient space. The gallery images are full resolution with 100 pixels between the eyes. Note that this result represents the inherent tolerance of our method to low-resolution without an explicit super-resolution reconstruction or feature extraction which is the topic of this chapter.



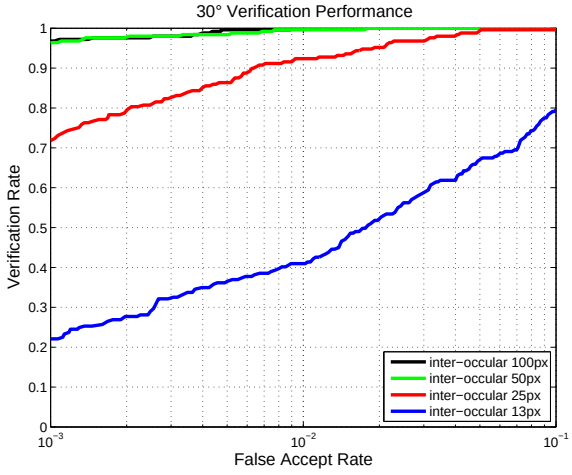
(a) Sensitivity at $-45°$



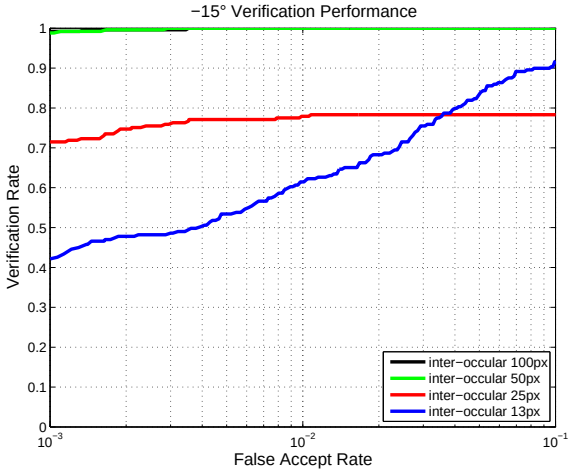
(b) Sensitivity at $45°$



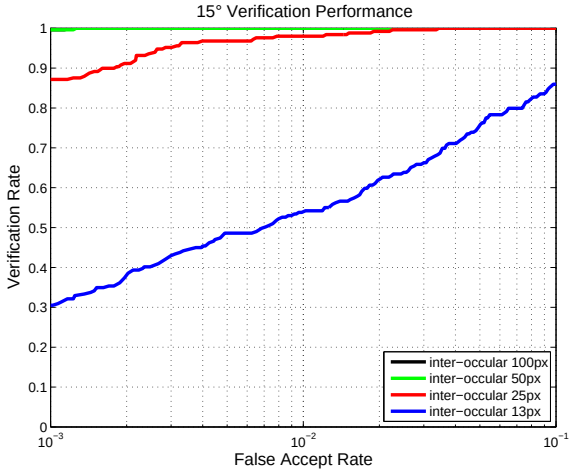
(c) Sensitivity at $-30°$



(d) Sensitivity at $30°$



(e) Sensitivity at $-15°$



(f) Sensitivity at $15°$

Figure 5.8: Verification rate of pose-corrected images as a function of query image interocular distance. The matcher used is CFA on L1-PCA coefficients. The gallery images are full resolution with 100 pixels between the eyes. Note that this result represents the inherent tolerance of our method to low-resolution without an explicit super-resolution reconstruction or feature extraction which is the topic of this chapter.

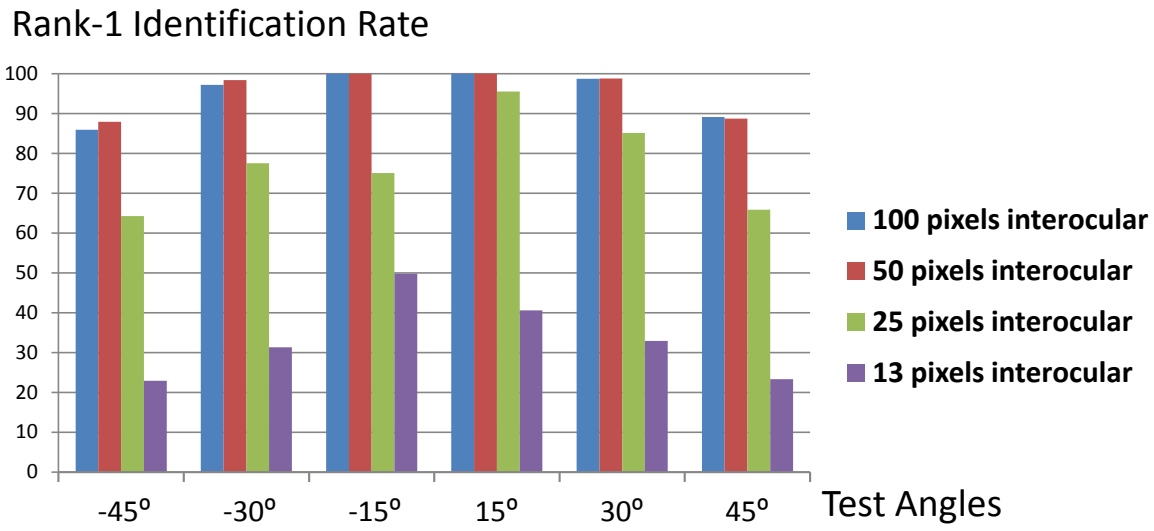


Figure 5.9: Tolerance of the L1-PCA CFA features to low-resolution when identifying across pose. The gallery images are the original resolution MPIE session 1 frontal faces. The query images are the non-frontal MPIE session 1 images downsampled to different sizes. We measure the rank-1 identification rates for different resolutions denoted by the available pixels between the eyes (measured for a frontal viewpoint). Note that this represents the inherent tolerance of our method to low-resolution without explicit super-resolution reconstruction or feature extraction which is the topic of this chapter.

Chapter 6

Conclusion

In this thesis, we presented a novel sparse feature extraction technique that we showed was tolerant to pose variations and low-resolution. We built a face model with a large number of training images that have been preprocessed to adopt the same shape (represented by x and y coordinates of landmarks) and represented by a PCA basis of eigenvectors and mean face. We then fit a 3D model on the non-frontal face and rotated the 3D model to render a frontal-looking shape-free face. We used ℓ_1-minimization to extract a feature vector from the pose-corrected face discounting the occluded pixels which fall on the parts of the face we do not observe. We showed that our ℓ_1-based feature vector carries a great deal of discriminative information that we could use to perform face recognition. We presented two different algorithms that make use of this discriminative information to achieve a strong verification one-to-one performance that outperforms commercial matchers. Furthermore, we showed that we could reconstruct a frontal-looking face that corresponds to the input non-frontal face in case one requires a reconstructed face to feed into commercial matchers. We also showed that our method is tolerant to changes in yaw, pitch, and resolution. Our verification performance was barely affected when input faces were reduced to offer no more than 25 pixels between the eyes. We also applied the same methodology to enhance the resolution of low-resolution faces. We showed that we could obtain magnification factors of 16x and starting

from images with as few as seven pixels between the eyes. Despite the multitude of components

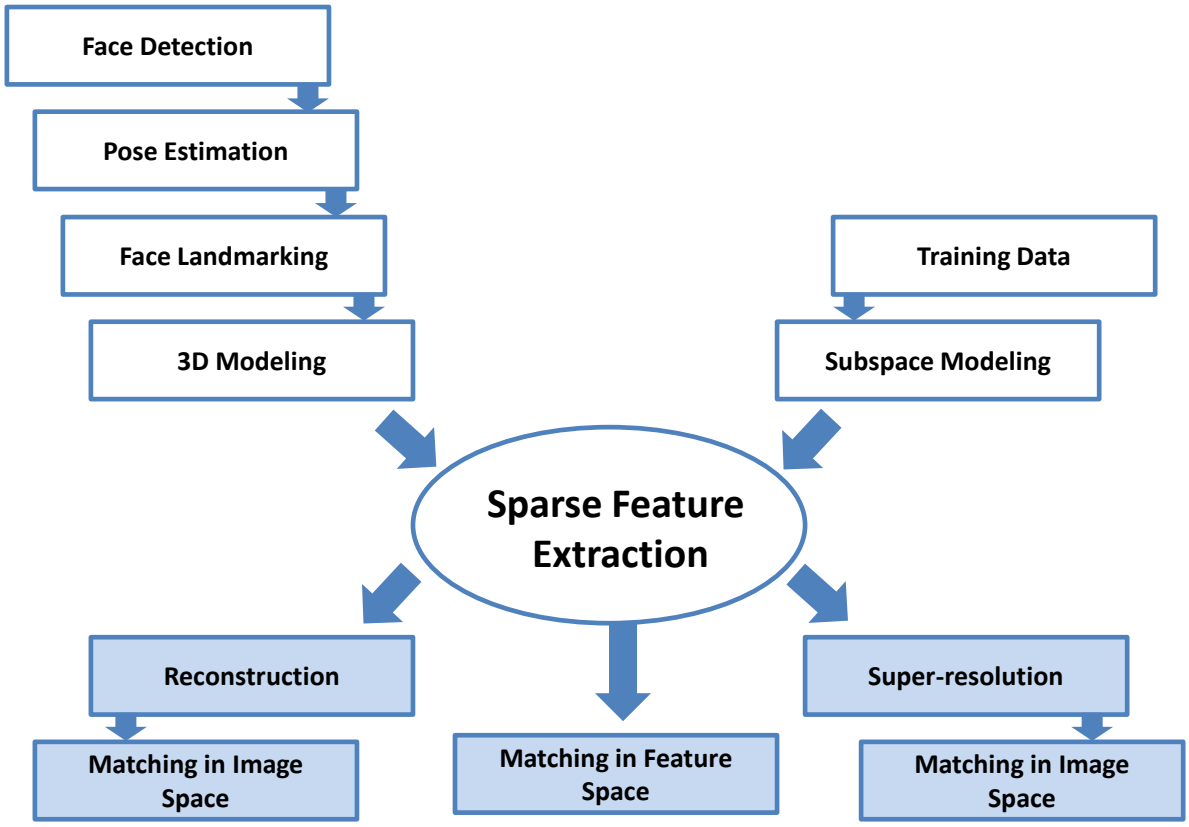


Figure 6.1: The big picture which highlights the modularity of our approach. At the heart of our method lies a sparse feature extraction step (due to the missing dimensions problem). The rest of the modules are interchangeable. It is possible to upgrade to any face detector, pose estimator, landmarker, 3D modeling technique and subspace modeling training technique.

involved in our feature extraction technique, our framework remains highly modular and flexible. The flowchart depicted in Figure 6.1 highlights the modularity of our approach. Each component remains independent to a very large extent from the preceding and succeeding modules and can be easily interchanged with another equivalent module. This *plug-and-play* flexibility keeps our technique easily upgradable. The sparse feature extraction step remains at the heart of our framework, and this step too can be modified. This layered architecture in our framework holds a clear advantage over most other statistical 2D modeling approaches which build a vertical architecture comprising of complicated and interconnected steps. All of our components are very active fields for research (for example face landmarking, pose estimation, 3D modeling, etc.), and upgrades to

any individual component will undoubtedly translate into improved overall performance.

6.1 Summary of Contributions

This thesis aimed to isolate and solve the lack the pose-tolerance in current face recognition algorithms. We presented a framework to perform a pose-correction feature extraction step. The main contributions of this thesis are listed below:

- We invented a new 3D modeling technique to generate a 3D model from a *single non-frontal* image.
- We described a new shape-free representation for 2D and 3D faces.
- We created a sparse *global* pose-tolerant feature extraction method based on ℓ_1-minimization.
- We showed that with current facial landmarking scheme, the shape information for non-frontal images carry too much variance and can hinder face recognition.
- We showed that the L1-PCA feature can be used for pose and resolution tolerant feature extraction.
- We showed how we can use the L1-PCA feature to synthesize pose-corrected and resolution-enhanced frontal faces.
- We showed drastic improvement of verification results over commercial matchers when matching across different viewpoints (an average 50% verfication rate improvement at 0.1% FAR for the $\pm 45^\circ$ yaw angles).
- We applied the framework to achieve single-image face hallucination (with up to 16x magnification ratio).
- We presented a new algorithm based on boosting to non-linearly combine dimensions of a feature vector.

6.2 Relation to Previous Work

Most of the serious efforts that seek pose-tolerance in face recognition seem to be divided between two groups. One that explicitly relies on 3D face modeling to geometrically correct the posture of the face, and one that explicitly avoids the 3D modeling step and rely on statistical methods to model the relationship between frontal and non-frontal faces. Methods belonging to the latter group justify their approach by citing that the 3D step is too slow or computationally expensive. However, 3DGEM provided a simple and easily implementable framework to rapidly obtain a fairly accurate 3D model that approximates the true 3D face. Our method capitalizes on rapid 3D prototyping to avoid complex statistical modeling approaches that make rigorous assumptions that can be hard to justify or verify. Most 2D modeling statistical approaches resort to local patch-based approaches in an effort to improve their modest global-based results. This however adds a layer of complexity that our method avoids by extracting a global feature. We have reasons to believe our results could improve if we were to adopt a patch-based approach, but the high verification rates we already achieve with a simple global approach is a testament to the soundness of our approach. Moreover, certain patch-based approaches (such as [20]) place patches on the ears and hairlines of the subjects. Our 3D modeling technique so far only models a tight mask around the face, and we believe that by incorporating extra information available on the head (such as ears and hair information) our results could further improve.

Our method bears some similarity to the approach of [17] in that we both treat the problem as a missing-data problem. While they consider the missing-data as entire viewpoints of the face, our missing data corresponds to a few self-occluded pixels that fall on the far-side of the face. We also borrow ideas from [24] by dropping the first few PCA coefficients.

The 3DGEM modeling technique we use is a direct competitor to the 3DMM algorithm of [27]. While the speed advantage of 3DGEM over 3DMM is obvious, the former also holds another advantage: it is more *forensically* accurate, as it does not not iteratively reconstruct the texture like

3DMM does. Instead, 3DGEM just works with whatever observable pixel information is there, and relies on texture mapping to cast the texture on top of the 3D model. Should the face have a distinctive textural feature (such as a mole or scar), 3DGEM will reproduce this feature unaffected, while 3DMM might smooth it out.

Although it is hard to compare our 3D geometric approach to statistical approaches that eschew 3D modeling and learn the relationship between frontal and non-frontal images via pattern recognition methods, it is useful to contrast the end-result performance. For this purpose, we compare our method to the recent work of [22] which adopts a unified approach to pose and resolution tolerance similar to ours. As a reminder, in this approach they rely on Tensorfaces to estimate the pose and the fiducial points on which to extract SIFT features, reduce the dimensionality of the combined SIFT features using PCA, and resort to an MDS-based mapping to learn the relationships between frontal and non-frontal faces (and in this case low and high resolution faces). They mostly rely on the robustness of the SIFT features to handle the illumination variations. However, they reported pose-tolerance performance for the $\pm 30^\circ$ range only, and resolution-tolerance results with only 3x magnification rates and at the -30° and -15° angles only. For higher magnification rates they restrict the pose angle further to 15°. Curiously, their SIFT+MDS peak performance happens at the extreme illuminations[1] which cast significant shadows on the face. This seems to suggest either a bias in their training or a dependency on the presence of sharp shadows on the face. More importantly, their rank-1 identification rates remain on average 8% *lower* than the identification rates we obtain using a simple *normalized-cosine* matcher on the reconstructed *pixels* of our L1-PCA reconstruction. As a reminder, this specific image representation is easily outperformed by the coefficient representation, and by our L1-PCA CFA features. Moreover, our results were computed on all 337 unique subjects of the MPIE dataset (their method uses 100 subjects for training and the remaining 237 for testing). Furthermore, our method handles a wider range of angles ($\pm 45^\circ$). Even though we have not analyzed the effect of illumination on our methods, we firmly believe

[1] illuminations number 14, 15 and 18 in the MPIE technical description

that our framework is more than capable of handling the presence of illumination variations. As for the resolution-tolerance performance, their 3x decrease-in-resolution results are comparable to our 4x decrease-in-resolution tolerance. This is without our method *explictly* correcting for low-resolution and simply working with whatever resolution is available, while their method is *designed and optimized* to handle low-resolution images.

6.3 Future Research Directions

The pose-correction framework we presented represents a baseline approach that has vast room for improvement in every module. Next we offer some insight on how different modules in Figure 6.1 could be upgraded to improve the overall performance. We can try:

Performing super-resolution on non-frontal images: As explained in section 5.5.2, our pose-correction framework lends itself very easily to super-resolution, and vice versa. It is therefore relatively easy to extract features from super-resolution coefficients from non-frontal faces for matching or reconstruction.

Different subspace modeling techniques to learn features on training data: Our method relied a simple PCA basis to represent our training data. There are more advanced subspace learning methods that could potentially improve our feature extraction and reconstruction accuracies. Our framework is general and lends itself to any subspace modeling method, and we used PCA as a simple example to demonstrate our approach.

Incorporating soft-biometric information: Our training data included a large mix of different ethnicities, genders and other extraneous face information such as beards, glasses, etc... Arguably, we could have gender and ethnic specific subspaces, and rely on the soft-biometric information of the test subject to select the adequate model. This soft-biometric information can be automatically obtained (see [94]) with reasonable accuracy. As importantly, the 3D modeling step can make very good use of the soft-biometric prior information, as we could use a gender and ethnic specific 3D

mean shape model to generate the 3D model.

Different landmarking scheme that is better suited for pose changes: In the current landmarking scheme we used, contour points falling on the far side of the face get occluded beyond 25°, and as a result get placed on the best available boundary of the face which usually falls on the cheekbone far away from the actual contour. This discrepancy made the shape information inconsistent in a pose-correction framework, and made us discount the shape information and rely solely on the shape-free representation. With a more comprehensive landmarking scheme, we have every reason to believe that our pose-correction face matching results will improve when we incorporate the pose-corrected shape information.

Reformulating the feature extraction step to incorporate more than one input face (whether doing pose-correction or super-resolution). This would involve re-engineering the ℓ_1-minimizing solver and modifying it to incorporate more constraints.

Extending the 3D modeling step to build half-face models: Currently our 3D face modeling requires observing both eye sockets in the test face. By considering only the half face, we could potentially handle poses beyond the half-profile and all the way to full profile.

Optimizing the matching in a watch-list $1 : N$ identification scenario: In this case, one could use clever score normalization techniques [95] to boost matching rates significantly.

Incorporating an illumination-tolerant component in the pose-correction framework: This could either be a separate module in the sequence of modules depicted in Figure 6.1 that will explicitly perform a pre-processing step to normalize illumination (such as [96, 97]), or could be part of the feature extraction module. We believe that our method with its sparse feature extraction scheme on L1-PCA coefficients that disregard the first few dimensions is well equipped to handle the adverse effects of illumination variations.

Appendix A

Brief Overview of ℓ_1-minimization

In the nineteenth century ℓ_2-minimization became more mainstream due to the simplicity and elegance of deriving closed-form solutions using linear algebra, while ℓ_1-minimization often relied on expensive iterative numerical approaches. But with the ubiquity of affordable and powerful computing which mitigated the disadvantages of iterative solutions, the advantages ℓ_1-minimization started to become more obvious and ℓ_1-minimization underwent a huge resurgence.

The main conceptual differences between ℓ_1 and ℓ_2 minimization can be observed in the fundamental problem of solving for $\mathbf{x}$ in $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$. Let's first assume an overdetermined problem where $m > n$, the ℓ_2-minimization or least-squares operates on a sum-of-square-differences concept and will put a small weight on small residuals, and a large weight on large residuals. This in return makes it sensitive to outliers. Alternatively, we could minimize the ℓ_1 distance or sum of the absolute values as the residuals between $\mathbf{Ax}$ and $\mathbf{b}$, which comparatively will put more weight on the small residuals, and less weight on larger residuals, since they are not squared anymore. In the underdetermined case with $m < n$, the problem becomes ill-posed due to the null-space of $\mathbf{A}$, and there are infinitely many solutions. Throughout literature, people have chosen the minimum-energy solution (such as the one given in Equation 3.16). By analogy, we could also instead pick the solution that has minimum ℓ_1-norm.

The underdetermined case, where we have considerably more unknowns than equations, comes up very frequently in endless applications and domains, such as geophysics, biomedical imaging and biomedical engineering, signal processing, error correction, etc.... The sparsity-inducing properties of ℓ_1-minimization have been empirically known since the 1970s [98]. In [99], it was already demonstrated that if a bandlimited signal is corrupted on an interval of less than a certain fraction of its bandlimit, then the original signal can be recovered simply by finding the closest signal in ℓ_1. Solving for the minimum ℓ_1-norm of a vector with an ℓ_1-norm regularization term as a goodness-of-fit criterion was suggested in [100] and a ℓ_2-norm regularization term in [101]. (Those ideas were later more thoroughly analyzed and reintroduced more recently in the statistics community as Lasso [48] and in the signal processing community as Basis Pursuit [46, 102]). In [103], the matrix $\mathbf{A}$ represented a combined "dictionary" of different representations or bases, each ideally suited to model a different phenomenon (for example a sinusoidal basis for periodic signals, wavelet, or curvelet-based representations ideal to represent discontinuous and signals with edges). Given a signal, the authors proposed a greedy algorithm (called *Matching Pursuit*) to extract features that takes advantage of the strengths of each representation, with numerous applications in broad areas such as transform coding, deconvolution and deblurring applications, statistical estimation, etc. Many variants of matching pursuit have been proposed since then (OMP [104], ROMP [105], StOMP [106], CoSaMP [107], etc.). Theoretical foundations of exact recovery conditions for greedy algorithms were also studied (see [108]). For a review and analysis of greedy methods see [109].

Sparsity in the ℓ_0-sense (which represents the number of nonzero elements) is generally an NP-complete problem [110]. The current best algorithms are of exponential complexity and look for a best fit via an exhaustive search of the subsets of columns of the dictionary matrix. Non-exhaustive algorithms that directly operate on the ℓ_0-regularized cost function (and the S-sparse constrained optimization) have been proposed (such as in [111]). However, since the optimization problem is non-convex, the strategies that ℓ_0-minimization solvers follow only guarantee to find local solu-

tions, which make them sensible to good initialization. In [102] it was empirically demonstrated that if the signal has a sparse expansion in the ℓ_0-sense than ℓ_1-minimization can be used instead to recover it.

The concept of sparsity steadily regained more prominence and started to significantly impact how people viewed the traditional data acquisition paradigms, especially in the vastly undersampled cases where measurements are very expensive (the classical example is that of IR-sensitive CCDs), or the acquisition process is slow (scan time in magnetic resonance imaging), or few sensors are available, etc. This gave rise to the field of Compressive Sensing (CS) [44], which aimed to non-adaptively acquire information efficiently with as few measurements as possible to reconstruct the signal, instead of the traditional approach of measuring a very large amount of data, then compressing it (minimizing some reconstruction error) by throwing away redundant information.

A.1 Compressed Sensing Primer

Some of the earlier theoretical foundations of CS were laid out by [56] where the concept of incoherence was reintroduced[1]. They make the distinction between two systems, the sparsity basis, and the measurement basis. To illustrate this dual concept, let $\mathbf{f}$ be the image of size n and we assume that $\mathbf{f}$ has a sparse or nearly sparse expansion in the basis Ψ such that $\mathbf{f} = \mathbf{\Psi x}$. Let S represent the cardinality of $\mathbf{x}$ or sparsity degree such that $\|\mathbf{x}\|_0 := |supp(\mathbf{x})| \leq S$. Another basis $\mathbf{\Phi}$, called the sensing or measurement basis is introduced. It represents the domain in which we are observing the measurements $\mathbf{y}_k = \langle \boldsymbol{\phi}_k, \mathbf{f} \rangle$, where k = 1 through m, where m is the number of measurements and $m < n$. For instance, if we are sensing in the Fourier domain, $\mathbf{\Phi}$ could be complex exponentials and Ψ could be a wavelet representation or a canonical basis (where the

[1] A related concept called mutual-coherence and defined equivalently was already used in [103].

signal is assumed to be sparse) [2]. The coherence between the two bases is given by:

$$\mu(\mathbf{\Psi}, \mathbf{\Phi}) = n. \max_{j,k} \left| \langle \boldsymbol{\phi}_j, \boldsymbol{\psi}_k \rangle \right| \tag{A.1}$$

Roughly speaking, it represents how far apart are the bases. For instance, a time basis, represented by shifted delta functions and a frequency basis represented by complex exponentials, are maximally incoherent. Note that if $\boldsymbol{\psi}_k$ and $\boldsymbol{\phi}_j$ are unit-norm, the coherence obeys $1 < \mu(\Psi, \Phi) \leq n$. One practical way to achieve a high incoherence is to have the sensing basis be constructed with Gaussian White Noise, or sequences of random binary entries (± 1), or random projections, etc. This guaranteed global measurements that are incoherent with the original signal. In [112] it was showed that if $\mathbf{f}$ is S-sparse in $\mathbf{\Psi}$, and we uniformly select m measurements at random in $\mathbf{\Phi}$, then ℓ_1-minimization will let us exactly recover the original image with very high probability if

$$m \geqslant \mu\left(\mathbf{\Psi}, \mathbf{\Phi}\right) .S. \log(n) \tag{A.2}$$

Note that this process is non-adaptive, i.e, we can fix the sensing basis (any random basis incoherent with the original basis in which the original signal expansion is nearly-sparse) and keep it unmodified regardless of the signal or the number of measurements. More importantly, observe that if S is small, then the number of measurements m needed can be significantly smaller than the Shannon-Nyquist rate. This breakthrough later gave rise to a large number of ideas and papers on how to better design this random "dictionary" or sensing matrices, such as [113, 114] and many more.

A more rigorous and more general foundational result was later introduced in [47]. The two separate *sparsity* and *sensing* systems were unified under a *Restricted Isometry Property* (RIP) concept, which dropped the probabilistic guarantee for an exact reconstruction for a deterministic

[2] In our framework of Chapter 3, $\mathbf{\Psi}$ is the eigenfaces basis $\mathbf{V}$, and $\mathbf{\Phi}$ is the canonical basis, where 1 represents a pixel we observe and 0 the rest of the pixels.

one. Now we need to recover $\mathbf{x} \in \Re^n$ from $\mathbf{y} = \mathbf{\Phi x}$. From [115], The RIP of a sensing matrix $\mathbf{\Phi}$, which is a function of the sparsity S and a scalar δ is defined by:

$$(1 - \delta) \|\mathbf{x}\|_2^2 \leq \left\|\mathbf{\Phi x}_2^2\right\| \leq (1 + \delta) \|\mathbf{x}\|_2^2 \tag{A.3}$$

Roughly speaking, this means that the sensing matrix $\mathbf{\Phi}$ acts as an approximate isometry on the set of vectors that are S-sparse in the basis. This also implies that the columns of $\mathbf{\Phi}$ have to be approximately orthogonal. The RIP is a necessary condition if we wish to be recover all sparse signals x from all the measurements in y. If $\|\mathbf{x}\|_0 = S$ then $\mathbf{\Phi}$ must satisfy the lower bound of the RIP with $\delta < 1$. Moreover, the RIP also suffices to ensure that several practical algorithms can recover any nearly-sparse or sparse signal from noisy measurements. It was also shown that for a small enough δ, ℓ_1-minimization can be used in a linear program to relax the ℓ_0 problem and recover the sparse signal. It was later proved ([116, 117]), that for random Gaussian noise sensing matrices, the RIP holds if $m \geqslant S \log(n/S)$. Furthermore, in [118], it was shown that another way of designing the sensing basis $\mathbf{\Phi}$ that meets the RIP was to start with a unitary matrix basis (like the Fourier Basis or in our case an eigenfaces basis) and to randomly select rows.

Independently and around the same time, the same findings about the recoverability of sparse signals using ℓ_1-minimization was demonstrated by [119]. Instead of incoherence/RIP, they introduced a related concept called the *spark* of a matrix A, given by the smallest number of columns from A that are linearly dependent.

A.2 Brief Overview of ℓ_1 Solvers

Going back to the original $\mathbf{Ax} = \mathbf{b}$ problem, we have argued in the previous section that if $\mathbf{x}$ is sufficiently sparse and the sensing matrix $\mathbf{A}$ is incoherent with the basis under which $\mathbf{x}$ is sparse, then it can be recovered by ℓ_1-minimization using Equation A.4. In the absence of noise, the

general optimization problem is given by:

$$\min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b} \tag{A.4}$$

In the presence of noise, the constraints are relaxed to instead minimize the reconstruction error:

$$\min \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \epsilon \tag{A.5}$$

where ϵ is a small positive constant to represent the anticipated noise level. The unconstrained equivalent of Equation A.5 is given by:

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{A.6}$$

A.2.1 Second-Order Methods

While Equation A.4 is a linear program (LP) which can be solved by classical algorithms, Equation A.6 can be cast a second order cone programming problem and thus can be solved via interior point methods [53]. Although it is very hard to review the vast literature on this subject, see [55] for an in-depth review of modern ℓ_1-minimization techniques. The classical approach is to iteratively recast the inequality constrained problem into an equality constrained problem, and then solve it using Newton's method [120] with the barrier method. This is what *Primal-Dual Interior-Point Algorithms* (PDIPA) (such as [121]) attempt to do. Homotopy methods [50, 122] take advantage of piecewise-linear properties of the ℓ_1-regularization path. They identify the next breakpoints along the solution path by examining the optimality conditions. However these path-following methods (which compute the entire solution path) become slow for large-scale problems. To alleviate these problems, an alternative method called the *Truncated Newton Interior-Point Method* (TNIPM) [123] was suggested. It uses a specialized interior-point method that uses preconditioned conjugate gradient [124] algorithms to find search direction and step size for the following

optimization problem:

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \sum_{i=1}^{n} v_i \quad \text{such that} - v_i \leq x_i < v_i \quad \text{with } i = 1, ..., n \tag{A.7}$$

In general, most of the second-order methods become intractable in modern large-scale and high-dimensional datasets and this motivated the need for simpler gradient-based (first-order) algorithms for solving Equation A.6 based on simple matrix/vector multiplications.

A.2.2 First-Order Methods

First-order methods, rely on renewed interest in iterative thresholding ideas (for the original ℓ_0-optimization [111, 125]) and the subdifferential of the ℓ_1-norm to reduce the per-step complexity, at the cost of increasing the total number of iterations. Some of these methods roughly fall under the following families of methods: *Proximal Gradient* (PG) (such as [49, 123]) and the related *Iterative Shrinkage-Thresholding Algorithms* (ISTA) (such as [51, 52]). They operate on the classical method of minimizing an unknown function by going in the negative direction of the gradient. For instance let $f : \Re^n \to \Re$ be a continuously differentiable convex function:

$$\min_{x} f(\mathbf{x}) \tag{A.8}$$

We can iteratively solve Equation A.8 by starting from an initial $\mathbf{x}_0 \in \Re^n$ and updating it according to:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_{k+1} \nabla f(\mathbf{x}_k) \tag{A.9}$$

where t_{k+1} is a suitable step-size and k is the iteration number.

Furthermore, we can also assume that $\nabla f(\mathbf{x})$ is Lipschitz continuous with constant $1/t_k$. Roughly speaking this means the gradient does not vary too fast (for twice differential functions it implies that $\nabla^2 f(\mathbf{x}) \leq (1/t_k)\mathbf{I}$ and the eigenvalues of the Hessian are smaller than $1/t_k$). The

Lipschitz property allows us to fix a quadratic upper bound on $f(\mathbf{x})$ (see [126]). So at a given iteration k, we can solve the following optimization:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left(f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^{\mathrm{T}} \nabla f(\mathbf{x}_k) + \frac{1}{2t_{k+1}} \|\mathbf{x} - \mathbf{x}_k\|_2^2 \right) \tag{A.10}$$

By completing the squares, we can rewrite Equation A.10 into the following quadratic optimization problem:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \frac{1}{2t_{k+1}} \|\mathbf{x} - (\mathbf{x}_k - t_{k+1} \nabla f(\mathbf{x}_k))\|^2 \tag{A.11}$$

Taking the gradient of the function to minimize in Equations A.10 or A.11 with respect to $\mathbf{x}$ and setting it to zero will yield the well known gradient descent update of Equation A.9.

Let us now aim to solve a composite optimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}) \tag{A.12}$$

where $f(\mathbf{x})$ is a convex smooth function but $g(\mathbf{x})$ is a convex non-smooth function. Proximal-Gradient methods maintain the quadratic upper bound on $f(\mathbf{x})$ but simply add $g(\mathbf{x})$ to it to obtain an upper bound on the global function:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left(f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^{\mathrm{T}} \nabla f(\mathbf{x}_k) + \frac{1}{2t_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2 + g(\mathbf{x}) \right) \tag{A.13}$$

The same manipulation yields the following *proximal* optimization:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \frac{1}{2t_{k+1}} \|\mathbf{x} - (\mathbf{x}_k - t_{k+1} \nabla f(\mathbf{x}_k))\|^2 + g(\mathbf{x}) \tag{A.14}$$

The solution of Equation A.14 is the *proximal-gradient* (PG) algorithms update rule:

$$\mathbf{x}_{k+1} = \mathrm{prox}_{t_k} (\mathbf{x}_k - t_{k+1} \nabla f(\mathbf{x}_k)) \tag{A.15}$$

where prox is the proximal operator. Intuitively, this operator corrects for the addition of $g(\mathbf{x})$ which made us deviate from the global unique solution that minimized $f(\mathbf{x})$. In the general case, this operator is expensive to obtain, however in the special case where $g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$, it can be done efficiently. As done earlier, we can take the gradient of the function to minimize with respect to $\mathbf{x}$ and set it to 0. More importantly, we can make use of the separatibility of the ℓ_1-norm ($\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$) to make of the minimization of $\mathbf{x}$ a one-dimensional element-wise minimization problem and obtain (see [127]) the corresponding update rule to Equations A.9 and A.15:

$$\mathbf{x}_{k+1} = \tau_{\lambda t_{k+1}} \left(\mathbf{x}_k - t_{k+1} \nabla f(\mathbf{x}_k)\right) \tag{A.16}$$

with τ_α the shrinkage operator (also called soft-thresholding) defined by the following function and depicted in Figure A.1:

$$\tau_\alpha(\mathbf{x})_i = max\left(|x_i| - \alpha, 0\right) sign(x_i) \tag{A.17}$$

Note the similarities between Equations A.9 and A.16. This method is referred to as (ISTA). To recap, in this case where the gradient of $\nabla f(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{x} - \mathbf{b})$, the general update step of IST methods look like the following:

$$\mathbf{x}_{k+1} = \tau_{\lambda t}\left(\mathbf{x}_k - 2t\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{x}_k - \mathbf{b})\right) \tag{A.18}$$

The convergence of PG and ISTA (which is a function of α) has been analyzed in literature and it although it is a simple algorithm, it is often criticized for its slow convergence speed (which is dominated by the gradient computation above). To alleviate this, more recent methods (such as NESTA [128] and FISTA [127]) make use of improved and fast-converging first-order gradient methods first discovered in ([129]) called the *Accelerated*-Gradient method and whose update rule

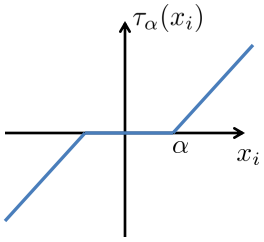


Figure A.1: Shrinkage-thresholding function.

is given by:

$$\mathbf{x}_{k+1} = \mathbf{y}_k - t_{k+1}\nabla f(\mathbf{y}_k) \tag{A.19}$$

$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \alpha_{k+1}\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) \tag{A.20}$$

This variant of the gradient algorithm incorporates a momentum parameter (α) that takes into account the previous two iterations to accelerate convergence. The corresponding accelerated version of the IST method above is called Fast-ISTA [127] and whose update rule is roughly given by:

$$\mathbf{x}_{k+1} = \tau_{\lambda t}\left(\mathbf{y}_k - t_{k+1}\nabla f(\mathbf{y}_k)\right) \tag{A.21}$$

$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \alpha_{k+1}\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) \tag{A.22}$$

We next focus on a variation of the FISTA algorithm which we used in our study, which was deemed in [54] to be good compromise between accuracy and speed.

A.2.3 Augmented Lagrangian Method

The Augmented Lagrangian method (ALM) [130, 131] has re-emerged lately in the Compressed Sensing community, is included in most solvers and packages, and has seen many new variants and development. The main idea is to eliminate the equality constraints by adding a penalty term to the original cost function that assigns a very heavy weight to points that fall outside the feasible

set.

Augmented Lagrangian methods aim to solve constrained optimization problem below:

$$\min h(\mathbf{x}) \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{A.23}$$

by minimizing the equivalent (unconstrained) cost function:

$$J_\rho(\mathbf{x}, \lambda) = h(\mathbf{x}) + (\rho/2) \left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|_2^2 + \lambda^{\mathrm{T}} (\mathbf{A}\mathbf{x} - \mathbf{b}) \tag{A.24}$$

The first two terms of Equation A.24 correspond to the penalty method (and ρ is the penalty parameter), which is then *augmented* by the additional penalty mimicking the method of Lagrange multipliers (λ is the Lagrangian). In other words, J_0 is the traditional Lagrangian for problem (A.23)). Equation A.24 can be solved the same way as a traditional Lagrangian method, obtaining the gradient of the augmented function (or its dual) and then evaluating the resulting equality constraint residual. From [130] we have the following update rules:

$$\mathbf{x}^{k+1} := \arg\min_{\mathbf{x}} J_\rho(\mathbf{x}, \lambda^k) \tag{A.25}$$

$$\lambda^{k+1} := \lambda^k + \rho \left(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\right) \tag{A.26}$$

ALM has been shown [126] to converge much faster than with the quadratic penalty method alone since ρ does not need to go to infinity and there is little extra computational overhead from adding the extra penalty term.

Observe that when $h(\mathbf{x}) = c \left\|\mathbf{x}\right\|_1$ in Equation A.23 (which then becomes identical to Equation 3.17), we can use any of the solvers previously mentioned to solve Equation A.25. The solver used in our pose-correction study utilized an ISTA-type approach for the dual form of the Augmented Lagrangian Method.

Appendix B

SimBoost: A Meta-Algorithm to Measure Similarity Between Multidimensional Feature Vectors

In this appendix, we present a novel yet simple way of obtaining a similarity measure between two high-dimensionality vectors motivated by AdaBoost [132] and based on a non-linear weighted combination of distance measures between the two vectors. The intuition behind this algorithm emerged from the need to select the dimensions that are the most discriminative for a verification purpose. For that reason, it is based on AdaBoost. When applied to a two-class classification problem, AdaBoost has proved very successful and was dubbed "best off-the-shelf classifier in the world" [133]. However, AdaBoost struggles in the case of high-dimensional multiclass problems. Several multiclass extensions to AdaBoost have been proposed over the years, but in this appendix we present a simple way of extending AdaBoost for a multiclass verification problem without the explicit need for a multiclass classification scheme.

AdaBoost [132] is an ensemble-learning method designed to select and combine weak classifiers into stronger classifier. AdaBoost operates by maintaining a distribution of weights over all

the training samples. The weight of a given training sample implies how "important" is this sample at a given iteration of the algorithm.

There exists several variants of the AdaBoost algorithm. We will review two of the more common ones, the traditional Discrete AdaBoost and the more advanced Real AdaBoost, and then present the algorithm which we call SimBoost, since it is obtained by boosting classifiers trained on different similarity matrices that represent the same data.

B.1 Discrete AdaBoost

AdaBoost is an iterative algorithm that can construct a "strong" classifier as a linear combination:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{B.1}$$

of "weak" high-bias classifiers $h_t(x) : \chi \rightarrow \{-1, +1\}$, where T is the total number of classifiers. The final classifier or final hypothesis is given by:

$$H(x) = \textit{sign}\left(f(x)\right) \tag{B.2}$$

Algorithm 1 Discrete AdaBoost

Given samples $(x_1, y_1), \ldots, (x_N, y_N)$ where $x_i \in \chi$ and $y_i \in \{-1, +1\}$ and N is the total number of samples. Initialize the weight distribution $W_1(i) = 1/N$
For $t = 1, \ldots, T$

- $h_t = \underset{h_j \in \mathcal{H}}{\arg\min}\, \epsilon_j = \sum_{i=1}^{N} W_t(i)\mathbf{1}\{y_i \neq h_j(x_i)\}$
- Set $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
- Update:
  $W_{t+1}(i) = \frac{W_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t}$
  where Z_t is a normalization factor such that W_t remains a distribution

The final output if the final strong hypothesis:
$H\left(x\right) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

If we have N total samples x_i for $i \in 1, \ldots, N$, samples x_i's classified as to belong to the negative class will have $f(x_i)$ values that are negative. Samples classified to belong to the positive class will have positive $f(x_i)$ values. Therefore for misclassified samples (where the sign of y_i does not agree with the sign of $f(x_i)$) will have the product $f(x_i)y_i$ negative. Crucially, AdaBoost relies on the exponential function that enables us to write that

$$y_i h_t(x_i) \leq 0 \tag{B.3}$$

$$e^{-y_i h_t(x_i)} \geqslant 1 \tag{B.4}$$

and therefore we can now find an upper bound on the training error given by:

$$\text{training error}\,(H_{final}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{H(x_i \neq y_i)\} \tag{B.5}$$

$$= \frac{1}{N}\begin{cases} 1 & \text{if } y_i H(x_i) \leq 0 \\ 0 & other \end{cases} \tag{B.6}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N} exp\left(-y_i f(x_i)\right) \tag{B.7}$$

By unwrapping the recursion of the weight update rule and using induction, we can prove that

$$W_{T+1}(i) = \frac{1}{N}\frac{exp\left(-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)\right)}{\prod_{t=1}^{T} Z_t} \tag{B.8}$$

$$= \frac{1}{N}\frac{exp\left(-y_i f(x_i)\right)}{\prod_{t=1}^{T} Z_t} \tag{B.9}$$

$$exp\left(-y_i f(x_i)\right) = N W_{T+1}(i) \prod_{t=1}^{T} Z_t \tag{B.10}$$

Plugging the last equation in the inequality of (B.7) we get

$$\text{training error}\,(H_{final}) \leq \frac{1}{N}\sum_{i=1}^{N} NW_{T+1}(i)\prod_{t=1}^{T} Z_t \tag{B.11}$$

$$\leq \prod_{t=1}^{T} Z_t \tag{B.12}$$

We detailed the proof that the training error is bounded by the product of the normalizing Z_t. Now we will express Z_t as a function of the error of every weak classifier. For the $0/1$ loss function, the training error ϵ_t of a weak classifier h_t at iteration t is the sum of the weights of the misclassified samples:

$$\epsilon_t = \sum_{i=1}^{N} W_t(i)\mathbf{1}\{y_i \neq h_t(x_i)\} \tag{B.13}$$

At a given iteration t, to make sure the sum of all weights is always equal to one, Z_t is defined as the sum of the current weights, which are the weights of the previous iteration exponentially modified as specified by the update rule:

$$Z_t = \sum_{i=1}^{N} W_t(i)exp\left(-\alpha_t y_i h_t(x_i)\right) \tag{B.14}$$

$$= \sum_{i\text{ s.t }h(x_i)=y_i} W_t exp\left(-\alpha_t y_i h_t(x_i)\right) + \sum_{i\text{ s.t }h(x_i)\neq y_i} W_t exp\left(-\alpha_t y_i h_t(x_i)\right) \tag{B.15}$$

$$= \sum_{i\text{ s.t }h(x_i)=y_i} W_t e^{-\alpha_t} + \sum_{i\text{ s.t }h(x_i)\neq y_i} W_t e^{-\alpha_t} \tag{B.16}$$

$$= e^{-\alpha_t}\left(1-\epsilon_t\right) + e^{\alpha_t}\epsilon_t \tag{B.17}$$

To find the minimum Z_t with respect to α_t set the derivative to zero:

$$\frac{\mathrm{d}Z_t}{\mathrm{d}\alpha_t} = e^{-\alpha_t}\left(1-\epsilon_t\right) + e^{\alpha_t}\epsilon_t = 0 \tag{B.18}$$

$$\epsilon_t e^{\alpha_t} = \left(1-\epsilon_t\right)e^{-\alpha_t} \tag{B.19}$$

$$e^{2\alpha_t} = \frac{\left(1-\epsilon_t\right)}{\epsilon_t} \tag{B.20}$$

$$\alpha_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t} \tag{B.21}$$

This is the penalty that the missclassified samples exponentially "pay", and at the same time the reward that the positive samples exponentially "earn". We can also also express Z_t in terms of ϵ_t as follows:

$$Z_t = 2\sqrt{\epsilon_t\left(1-\epsilon_t\right)} \tag{B.22}$$

Let $\epsilon_t = 1/2 - \gamma_t$ where γ_t represents how much better than chance is the classifier doing at iteration t, we can then write

$$\text{training error}\left(H_T\right) \leqslant \prod_{t=1}^{T} Z_t \tag{B.23}$$

$$= \prod_{t=1}^{T}\left[2\sqrt{\epsilon_t\left(1-\epsilon_t\right)}\right] \tag{B.24}$$

$$= \prod_{t=1}^{T}\sqrt{1-4\gamma_t^2} \tag{B.25}$$

$$\leqslant exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right) \tag{B.26}$$

So as long as γ_t is positive, the training error is bounded and decreasing. AdaBoost is adaptive, and we do not need to have an *a priori* knowledge of γ or T. As long as we make sure $\gamma_t >> \gamma$, then the training error is guaranteed to decrease.

B.2 Real AdaBoost

Real AdaBoost [134] improves upon Discrete AdaBoost by using confidence-rated predictions at every iteration of the algorithm. The weak hypotheses in Real AdaBoost make their predictions by partitioning the domain into disjoint blocks. There is no single weight for the vote of a specific classifier, but different weights for every domain block. The alpha term now is absorbed into the weak classifier whose response depends on where in the domain space the sample falls.

B.3 SimBoost: Using boosting to combine feature vector coefficients

We now adapt AdaBoost, which is a feature selection and classification algorithm, to handle a multiclass matching problem where every sample is represented by a multidimensional feature vector.

As in Algorithm 1, at every iteration t if we maintain a positive γ_t where $\epsilon_t = 1/2 - \gamma_t$, we guarantee that the training error is decreasing. This means that for a given similarity matrix $\mathrm{M}^{(\mathrm{i})}$ based on a subset of the total dimensions, the corresponding ROC should be better than the diagonal. By controlling how big is d' with respect to d, we can control the area under the curve (AUC) of the ROC.

B.4 Evaluation

To evaluate the efficacy of Algorithm 2, we setup the following experiment. We use the FRGC [61] generic dataset to train SimBoost. We test on the FRGC gallery dataset of 16028 face images for a total number of comparisons that exceeds 128 million.

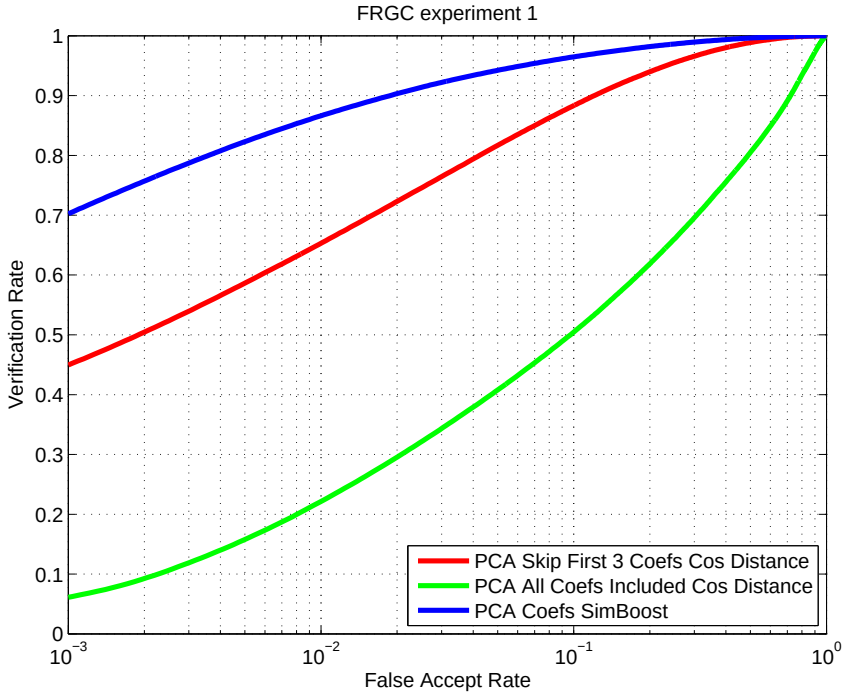


Figure B.1: ROC of FRGC Experiment 1 which contains over 128 million matches. The feature vector is a standard PCA coefficient vector obtained by projecting onto the PCA basis trained on the FRGC generic dataset.

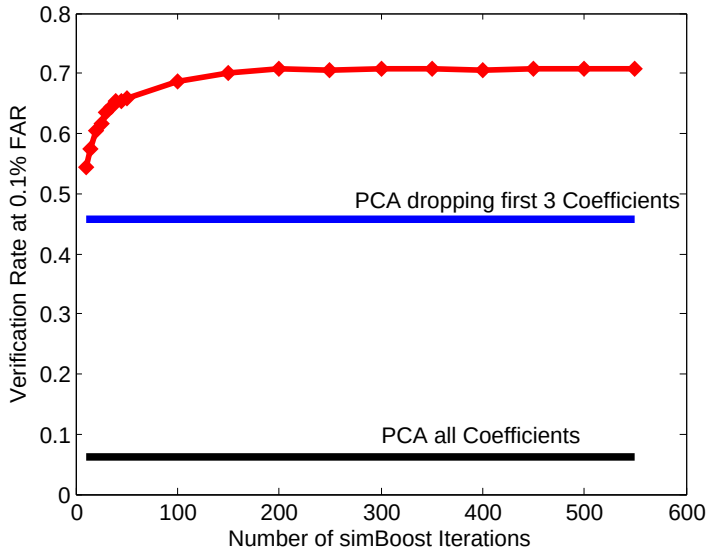


Figure B.2: This plot depicts the Verification Rate increase, reported at 0.1% FAR, as a function of SimBoost iterations. Also depicted on the same graph the performance of baseline PCA, with and without skipping any coefficients.

Algorithm 2 Discrete SimBoost

Given a training set of samples, partition the set into s_1 gallery samples, and s_2 query samples. Let $(\mathbf{v_1}^{gallery}, c_1), \ldots, (\mathbf{v_{s1}}^{gallery}, c_{s_1})$ be the gallery samples and $(\mathbf{v_1}^{query}, c_1), \ldots, (\mathbf{v_{s2}}^{query}, c_{s_2})$ be the query samples where $v_i \in \Re^d$ and $c_i \in \{C_1, C_2, \ldots, C_k\}$. k is the total number of training classes and C_i represents the class label.

Create a set $\mathcal{S} = \{\text{dimset}_1, \text{dimset}_2, \cdots, \text{dimset}_m\}$ of random and overlapping sets of dimensions. Each dimset_i is a set of d' random dimensions sampled from a discrete uniform distribution $\mathcal{U}(1, d)$, with $d' < d$ and $m > d$. These sets account for every dimension $\bigcup_{i=1}^{m} \text{dimset}_i = \{1, ..., d\}$ to make S an over-complete set of random dimensions.

Generate a set $\mathcal{M} = \left\{\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \cdots, \mathbf{M}^{(m)}\right\}$ of similarity matrices. Every $\mathbf{M}^{(i)}$ is an $s_1 \times s_2$ matrix where every $\mathbf{M}^{(i)}_{k,l}$ entry represents the similarity given a distance metric between samples $\mathbf{v}_k^{gallery}$ and $\mathbf{v}_l^{query}$ considering dimensions dimset_i only.

Each $\mathbf{M^{(i)}}$ yields a distribution of authentic scores and impostor scores. $(x_1^{(i)}, y_1), \ldots, (x_N^{(i)}, y_N)$ represent the scores extracted from $\mathbf{M^{(i)}}$ and $y_i \in \{-1, +1\}$ the labels where -1 represents an impostor score and $+1$ an authentic score. Let $N = s_1 \times s_2$ be the total number of scores in the similarity matrix.

Let $\mathcal{H}$ be the set of all possible weak classifiers h_i given score distributions generated by all similarity matrices in $\mathcal{M}$

Initialize the weight distribution $W_1(i) = 1/N$

For $t = 1, \ldots, T$

- $h_t = \underset{h_j \in \mathcal{H}}{\arg\min}\, \epsilon_j = \sum_{i=1}^{N} W_t(i)\mathbf{1}\{y_i \neq h_j(x_i)\}$
- Set $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
- Update:
  $W_{t+1}(i) = \frac{W_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t}$
  where Z_t is a normalization factor such that W_t remains a distribution

The final output if the final strong hypothesis:

$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Appendix C

Class-Dependent Feature Analysis

In this appendix we review a classifier based on advanced correlation filter theory, called Class-Dependent Feature Analysis (CFA). The kernel version of this classifier is called KCFA. In this dissertation, we use CFA on the pose-corrected shape-free images or their feature vector representation (see Chapters 3 and 4).

C.1 Advanced Correlation Filters Overview

Correlation filters evolved from the simple Matched Filter [135] into complex multi-image multi-class filters that exhibit very advantageous properties that are ideal for challenging tasks such as unconstrained face recognition. We briefly review two of the filters we use in this dissertation.

C.1.1 Equal Correlation Peak Synthetic Discriminant Function Filters

The basic correlation filter is the Equal Correlation Peak Synthetic Discriminant Function Filter (ECP-SDF)[136]. Unlike the Matched Filter, which matches one template at a time to a given scene, the ECP-SDF can handle multiple templates built into one filter. The desired correlation value at the origin is specified in a constraint vector $\mathbf{u}$ which is the size of the number of images

we are training the filter on. Typically, one would set a desired value of 1 for positive training images (enforcing maximum correlation) and a value of 0 or -1 for the negative samples (enforcing no correlation or maximum de-correlation respectively). Let $\mathbf{X}$ denote the matrix containing the training images vectorized and places along its columns. We need to design the filter $\mathbf{h}_{\text{ECP-SDF}}$ such that:

$$\mathbf{X}^{\mathrm{T}}\mathbf{h}_{\text{ECP-SDF}} = \mathbf{u} \tag{C.1}$$

By assuming that the filter is a linear combination of the training images or $\mathbf{h}_{\text{ECP-SDF}} = \mathbf{X}\alpha$ and substitution into the first equation, we can solve for α to obtain:

$$\alpha = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{u} \tag{C.2}$$

We then derive the following filter:

$$\mathbf{h}_{\text{ECP-SDF}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{u} \tag{C.3}$$

THe ECP-SDF essentially corresponds to the minimum-ℓ_2-norm filter $\mathbf{h}$ such that Equation C.1 holds. The ECP-SDF controls the correlation values at the origin (as specified by $\mathbf{u}$) and has no control over the rest of the correlation plane, and therefore yields many false peaks located elsewhere in the correlation plane. The MACE filter was developed to fix this problem.

C.1.2 Minimum Average Correlation Energy Filters

Minimum Average Correlation Energy (MACE) filters [137] solve the false peaks problem by suppressing the rest of the correlation plane. The average correlation energy (ACE) of the cross-correlation outputs are minimized while maintaining the constraints at the origin. Therefore typical correlation outputs from MACE filters typically exhibit very sharp peaks which makes detection and localization easy. The formulation of MACE happens in the frequency domain where corre-

lations can be very efficiently computed. Applying Parseval's theorem (which relates the spatial domain energy to the frequency domain energy), we can efficiently obtain a "prewhitening" matrix $\mathbf{D}$ which contains the average power spectrum of all training images along its diagonal.

$$\text{ACE} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}^{+}\mathbf{D_i}\mathbf{h} = \mathbf{h}^{+}\mathbf{D}\mathbf{h} \tag{C.4}$$

where N is the number of training images, and $\mathbf{D} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{D}_i$. We can now minimize the following cost function:

$$\min J(\mathbf{h}) = \mathbf{h}^{+}\mathbf{D}\mathbf{h} \quad \text{subject to} \quad \mathbf{X}^{+}\mathbf{h} = \mathbf{u} \tag{C.5}$$

where now $\mathbf{X}$ is a complex matrix containing the 2D Fourier transform of the training images vectorized and stacked along its columns, and $\mathbf{h}$ is the vectorized 2D Fourier transform of the filter. The $+$ operator denotes the conjugate transposition of a complex matrix. Following the same derivation as Equation 3.11 in section 3.2.4, we obtain the following filter $\mathbf{h}_{\text{MACE}}$:

$$\mathbf{h}_{\text{MACE}} = \mathbf{D}^{-1}\mathbf{X}\left(\mathbf{X}^{+}\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{u} \tag{C.6}$$

Note the similarities between Equations C.3 and C.6. However, the former is in spatial domain while the latter is in frequency domain. Filtering with a MACE will generally return a very sharp peak with reduced side-lobes, as the overall correlation place energy is minimized by including $\mathbf{D}$ in the formulation. The correlation output of a ECP-SDF filter $\mathbf{h}_{\text{ECP-SDF}}$ given an input $\mathbf{y}$ is given by $\mathbf{y}^{\text{T}}\mathbf{h}_{\text{ECP-SDF}}$. In the case of MACE, we can further exploit the diagonal structure of $\mathbf{D}$ in equation

C.6 as shown below:

$$\mathbf{y}^{+}\mathbf{h}_{\text{MACE}} = \mathbf{y}^{+}\mathbf{D}^{-1}\mathbf{X}\left(\mathbf{X}^{+}\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{u} \tag{C.7}$$

$$= \left(\mathbf{D}^{-0.5}\mathbf{y}\right)^{\mathrm{T}}\left(\left(\mathbf{D}^{-0.5}\mathbf{X}\right)^{\mathrm{T}}\left(\mathbf{D}^{-0.5}\mathbf{X}\right)\right)^{-1}\mathbf{u} \tag{C.8}$$

$$= \mathbf{y}'^{\mathrm{T}}\mathbf{X}'\left(\mathbf{X}'^{\mathrm{T}}\mathbf{X}'\right)^{-1}\mathbf{u} \tag{C.9}$$

where $y' = \mathbf{D}^{-0.5}\mathbf{y}$ and $X' = \mathbf{D}^{-0.5}\mathbf{X}$ indicates pre-whitened versions of $\mathbf{y}$ and $\mathbf{X}$ respectively. What we have achieved in Equation C.9 is converting the MACE formulation back into the spatial domain by incorporating a pre-whitening step of the training data. This can be efficiently done by taking the 2D Fourier transform of the training matrix $\mathbf{X}$, pre-multiplying it by $\mathbf{D}^{-0.5}$ and then taking the inverse 2D Fourier transform, such that the resulting $\mathbf{y}'$ and $\mathbf{X}'$ are in spatial domain.

The MACE formulation can be extended to incorporate some noise tolerance[138]. Equation C.6 becomes:

$$\mathbf{h}_{\text{OTSDF}} = \mathbf{T}^{-1}\mathbf{X}\left(\mathbf{X}^{+}\mathbf{T}^{-1}\mathbf{X}\right)^{-1}\mathbf{u} \tag{C.10}$$

where $\mathbf{T} = \left(\alpha\mathbf{D} + \sqrt{1-\alpha^2}\mathbf{C}\right)$. $\mathbf{C}$ presents a diagonal matrix whose elements represent the noise power spectral density. For a white noise model, $\mathbf{C} = \mathbf{I}$. α represents the blending parameter (any positive scalar ≤ 1) which controls the trade-off between noise tolerance and peak sharpness. Therefore the filter in Equation C.10 is called the Optimal Tradeoff Synthetic Discriminant Filter (OTSDF). Note that when $\alpha = 1$, the OTSDF degenerates into a MACE filter.

C.1.3 Kernel Extension

The MACE and OTSDF formulations can be easily be extended to kernel space since the data appears in the form of inner products. Suppose we map the data to some other feature space by the non-linear mapping Φ. The kernel "trick" lets us map the data into an inner product space without having to compute the mapping Φ explicitly. For two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, $K(\mathbf{x}_i, \mathbf{x}_j) =$

$\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$. The kernel function defined here can be used as long as it forms an inner-product and satisfies Mercer's theorem[139] to ensure that we are still working in a Hilbert inner-product space. The kernel extension of Equation C.9 becomes:

$$\Phi(\mathbf{y})^{\mathrm{T}} \Phi(\mathbf{h}_{\mathrm{MACE}}) = K(\mathbf{y}', \mathbf{X}') \left(K\left(\mathbf{X}', \mathbf{X}'\right) \right)^{-1} \mathbf{u} \tag{C.11}$$

Examples of kernel functions are polynomial, Radial Basis Function (RBF), and neural net sigmoidal kernels.

C.2 Class-Dependent Feature Analysis

Correlation filters have been applied with great success to a number of 2D-based classification problems in biometrics (face [140, 141], fingerprints [142], iris [143]) and Automatic Target Recognition (ATR) [144]. They have shown to be highly modular and robust to occlusion, misalignment, and exhibit a graceful degradation. The CFA algorithm aims at harnessing the power of correlation filters in a classification framework. We first assume that all input images are correctly centered, and only measure the correlation peak at the center. Given a training set of C classes, one could train C filters (for instance MACE or OTSDF), each designed to provide a peak of 1 (maximum correlation) for images belonging to the given class, and no peak at all (zero correlation) for images belonging to all other classes. This way one could represent unseen testing images by a feature vector of length C each dimension representing the correlation peak by the corresponding filter. This feature vector represents how much the unseen testing sample correlates with each one of the training classes. Figure C.1 shows a diagram that depicts how an unseen face image is tested with a MACE filter built for class 2, in an N-class problem, each class containing 4 training images. For this specific case depicted in the Figure C.1, the constraints vector $\mathbf{u}$ will be given by:

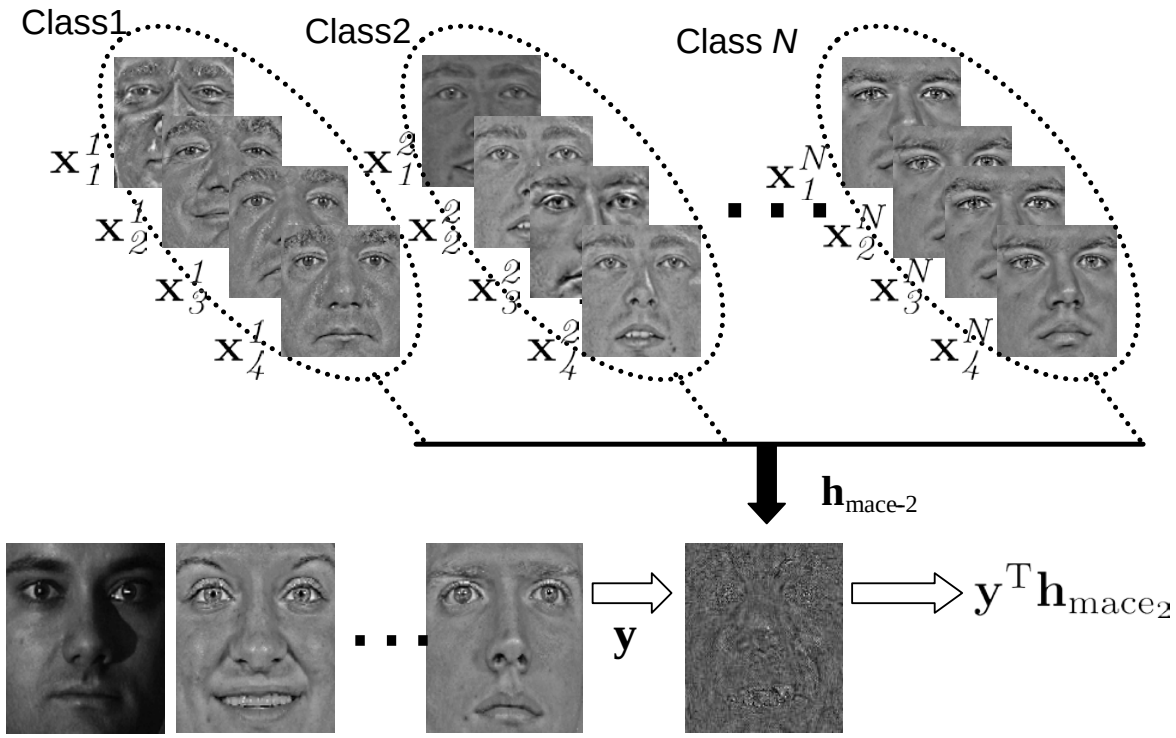


Figure C.1: Schematic to illustrate the building of a MACE filter $\mathbf{h}_{\text{mace}_2}$ for class 2 out of N classes each with 4 images. The unseen testing image $\mathbf{y}$ is filtered with the resulting MACE filter. The filter response in given by the inner product between $\mathbf{h}_{\text{mace}_2}$ and $\mathbf{y}$. This scalar value will represent how much the test face $\mathbf{y}$ triggers the filter built for class 2. By repeating the process for N filters, we obtain a feature vector of size N, each entry representing how much the test face *correlates* with all training classes.

$$\mathbf{u}_1 = [0 \quad 0 \quad 0 \quad 0]^{\mathrm{T}} \tag{C.12}$$

$$\mathbf{u}_2 = [1 \quad 1 \quad 1 \quad 1]^{\mathrm{T}} \tag{C.13}$$

$$\mathbf{u}_N = [0 \quad 0 \quad 0 \quad 0]^{\mathrm{T}} \tag{C.14}$$

$$\mathbf{u}_{\mathrm{mace}_2} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \ldots \quad \mathbf{u}_N]^{\mathrm{T}} \tag{C.15}$$

And the $\mathbf{X}$ matrix will contain all the vectorized training images $\mathbf{X} = [\mathbf{x}^1_{1\ldots4}\mathbf{x}^2_{1\ldots4}\ldots\mathbf{x}^N_{1\ldots4}]$. The filter response for an input face image $\mathbf{y}$ is given by Equation C.9.

Appendix D

Additional Experimental Results

D.1 Results on MPIE

The following is a list of additional results and figures for the MPIE dataset:

- Figure D.1: An additional example of different image representations after reconstructing pose-corrected images.
- Figure D.2: An additional example of different image representations after reconstructing pose-corrected images.
- Figure D.3: ROCs of traditional 2D warping techniques versus pose-correction using L1-PCA.
- Figure D.4: ROC of L1-PCA coefficient matching for different combination of automatic and manual landmarking and pose estimation.
- Figure D.5: Examples of bad landmarking fit.
- Table D.1: Verification results on MPIE for different image representations.
- Table D.2: Rank-1 identification rates on MPIE for different image representations.
- Table D.3: Rank-1 identification rates for different combinations of automatic and manual

landmarking and pose estimation.

- Table D.4: Rank-1 identification rates when the resolution (represented by the number of pixels between the eyes in a frontal viewpoint) of the query images drops.
- Figure D.6: ROCs of matching in coefficient space using the L1-PCA, L2-PCA and HYBRID representations using NCD as the matcher. 95% confidence intervals are plotted on the ROCs to illustrate the statistical significance of the result.

		Yaw Test Angles					
Verification Mode	Image Representation	$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$
128×128 crop with shape added back	2D warp plin	29.37	48.96	64.68	64.68	38.57	31.45
	2D warp lwm	23.73	43.91	62.9	63.79	40.65	25.22
	L1-PCA crop	35.31	60.53	85.75	85.45	57.56	32.34
Shape-Free Reconstruction	L1-PCA	29.37	48.96	64.68	64.68	38.57	31.45
	L2-PCA	23.73	43.91	62.9	63.79	40.65	25.22
	HYBRID	26.7	45.69	71.81	73.5	40.94	26.40
	STRETCH	23.14	34.71	59.64	66.46	29.97	21.66
	HOLES	23.73	42.73	70.02	74.18	38.57	25.81
Shape-Free Coefficients	L1-PCA	66.46	91.69	98.51	98.81	88.72	63.5
	L2-PCA	52.81	84.56	95.25	96.73	81.00	54.3
	HYBRID	66.76	91.69	98.51	98.81	88.72	63.2

Table D.1: Verification Rates measured at 1% FAR using normalized cosine distance. The Mode denotes whether the shape information is incorporate in the image or suppressed and whether we are matching in the reconstructed images or in the coefficient space. The Image Representation lists all the different representations we defined in Section 3.4. All 337 unique subjects from the combined MPIE sessions are included in this experiment.

Identification Mode	Image Representation	Yaw Test Angles					
		$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$
128×128 crop	2D warp plin	58.75	85.45	95.54	96.73	86.35	53.70
with shape added back	2D warp lwm	51.92	86.05	97.03	97.03	87.83	52.81
	L1-PCA crop	48.07	90.20	98.22	99.12	87.54	45.42
Shape-Free Reconstruction	L1-PCA	55.49	89.32	97.63	98.22	89.62	50.45
	L2-PCA	43.62	78.64	96.14	97.33	85.46	46.89
	HYBRID	53.12	88.43	97.33	97.92	89.02	48.96
	STRETCH	53.11	85.16	97.03	98.51	87.24	51.63
	HOLES	53.11	84.27	98.22	98.51	89.61	53.70
Shape-Free Coefficients	L1-PCA	61.72	93.18	98.81	97.92	82.49	45.70
	L2-PCA	41.25	81.00	96.74	98.22	78.93	38.58
	HYBRID	62.02	93.18	98.81	98.22	83.98	45.99

Table D.2: Rank-1 Identification rates using normalized cosine distance. The Mode denotes whether the shape information is incorporate in the image or suppressed and whether we are matching in the reconstructed images or in the coefficient space. The Image Representation lists all the different representations we defined in Section 3.4. All 337 unique subjects from the combined MPIE sessions are included in this experiment.

(a) Original face images with yaw varying from $-45°$ to $+45°$ in increments of $15°$.

(b) STRETCH: Pose-corrected images in the shape-free representation with no occlusion detection and no reconstruction.

(c) HOLES: Pose-corrected images with occlusion detected in the shape-free representation.

(d) L2-PCA reconstructed pose-corrected images in the shape-free representation.

(e) L1-PCA reconstructed pose-corrected images in the shape-free representation.

(f) HYBRID: L1-PCA+STRETCH pose-corrected images in the shape-free representation.

Figure D.1: MPIE subject 14 with (a) the original image from different angles (b) The equivalent pose-corrected shape-free images with no occlusion detection. As a result the texture mapping module in 3DGEM will introduce stretching artifacts due to interpolating from too few available pixels. (c) The equivalent pose-corrected shape-free images with pixels that are likely to be occluded detected in 3D. (d) The equivalent reconstructed pose-corrected images using L2-PCA in shape-free representation. (e) The equivalent reconstructed pose-corrected images using L1-PCA in shape-free representation. (f) Hybrid representation which consists of the unoccluded pixels from (c) and occluded pixels from (e). The sparse feature vector responsible for this reconstruction is used in Chapter 4 to achieve one-to-one verification rates.

(a) Original face image with yaw varying from -45° to $+45^\circ$ in increments of 15°.

(b) STRETCH: Pose-corrected images in the shape-free representation with no occlusion detection and no reconstruction.

(c) HOLES: Pose-corrected images with occlusion detected in the shape-free representation.

(d) L2-PCA reconstructed pose-corrected images in the shape-free representation.

(e) L1-PCA reconstructed pose-corrected images in the shape-free representation.

(f) HYBRID: L1-PCA+STRETCH pose-corrected images in shape-free representation.

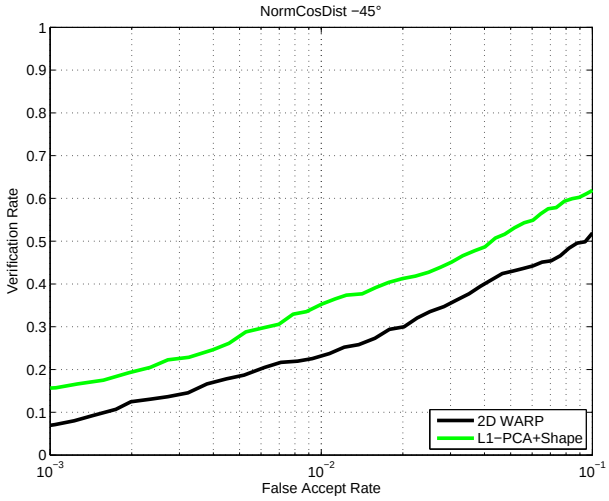
Figure D.2: MPIE subject 23 with (a) the original image from different angles (b) The equivalent pose-corrected shape-free images with no occlusion detection. As a result the texture mapping module in 3DGEM will introduce stretching artifacts due to interpolating from too few available pixels. (c) The equivalent pose-corrected shape-free images with pixels that are likely to be occluded detected in 3D. (d) The equivalent reconstructed pose-corrected images using L2-PCA in shape-free representation. (e) The equivalent reconstructed pose-corrected images using L1-PCA in shape-free representation. (f) Hybrid representation which consists of the unoccluded pixels from (c) and occluded pixels from (e). The sparse feature vector responsible for this reconstruction is used in Chapter 4 to achieve one-to-one verification rates.

Image Representation	Landmarking	Pose Estim.	Yaw Test Angles					
			$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$
L1-PCA Reconstruct.	Manual	Manual	55.49	89.32	97.63	98.22	89.62	50.45
	Manual	Auto	56.68	89.02	97.33	97.92	89.91	54.89
	Auto	Manual	48.96	84.27	96.44	97.33	83.09	44.80
	Auto	Auto	51.63	84.27	96.44	97.03	85.76	48.37
L1-PCA Coefficients	Manual	Manual	61.72	93.18	98.81	97.92	82.49	45.70
	Manual	Auto	60.53	92.29	98.52	97.92	85.16	44.21
	Auto	Manual	56.68	89.02	97.32	98.22	89.61	49.55
	Auto	Auto	58.16	88.72	97.33	98.22	90.50	50.15

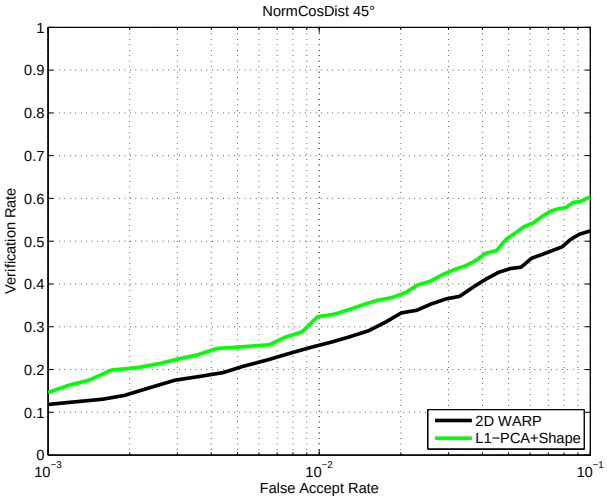
Table D.3: Rank-1 Identification rates using normalized cosine distance for the two L1-PCA representations. We tabulate the effect of automatic operation on both landmarking and pose estimation. All 337 unique subjects from the combined MPIE sessions are included in this experiment.

Identification Mode	Interocular Dist	Low Resolution Yaw Test Angles					
		$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$
L1-PCA CFA	100 pixels	85.95	97.19	100	100	98.78	89.16
	50 pixels	87.95	98.39	100	100	98.79	88.75
	25 pixels	64.25	77.51	75.10	95.58	85.14	65.86
	13 pixels	22.89	31.32	49.80	40.57	32.93	23.29
L1-PCA Coefs	100 pixels	65.06	93.57	99.59	97.99	81.12	45.78
	50 pixels	62.66	90.77	99.19	97.59	80.32	42.16
	25 pixels	42.16	59.84	75.90	91.57	59.84	30.12
	13 pixels	14.45	23.69	48.59	39.35	23.69	16.90

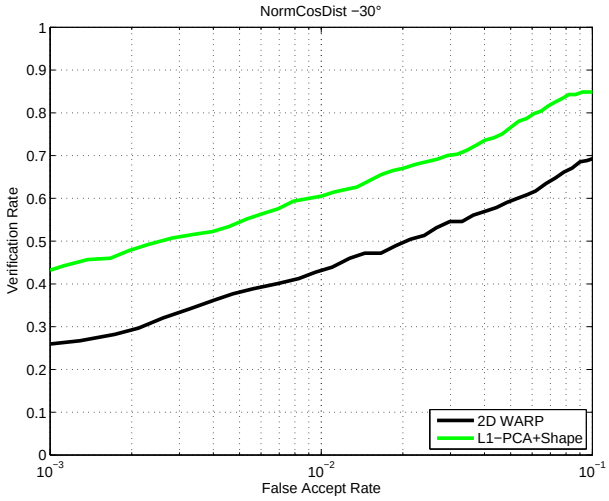
Table D.4: Rank-1 Identification rates using normalized cosine distance on MPIE session 1 when the gallery set is of high-resolution and the query set is of decreasing resolution.



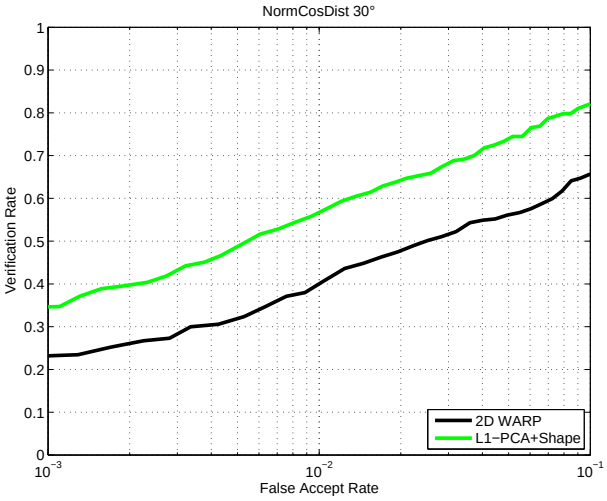
(a) 2D warping at $-45°$



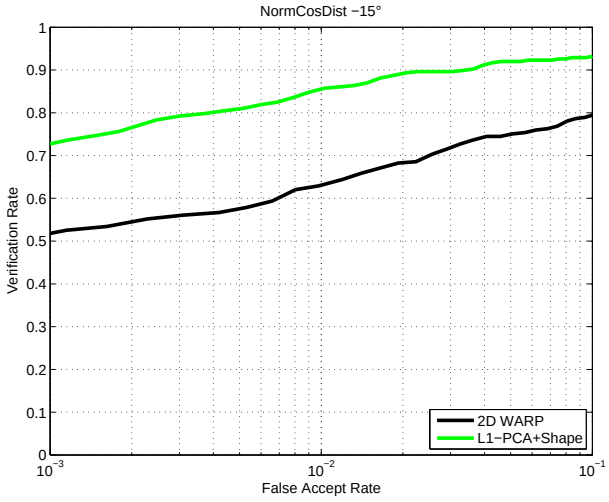
(b) 2D warping at $45°$



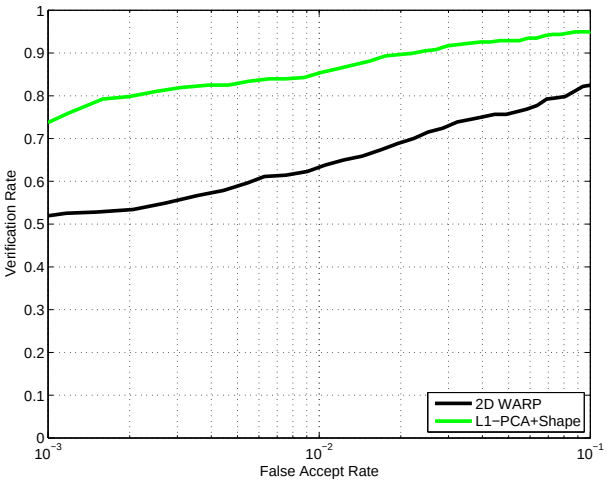
(c) 2D warping at $-30°$



(d) 2D warping at $30°$

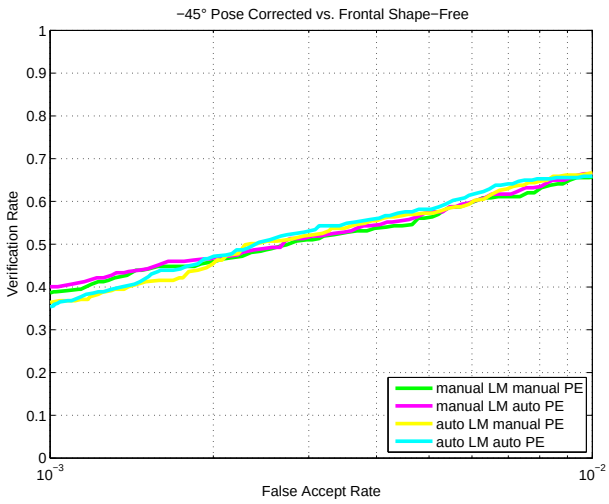


(e) 2D warping at $-15°$

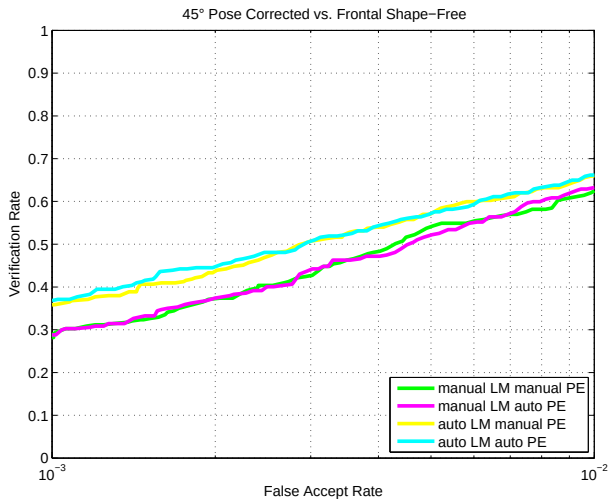


(f) 2D warping at $15°$

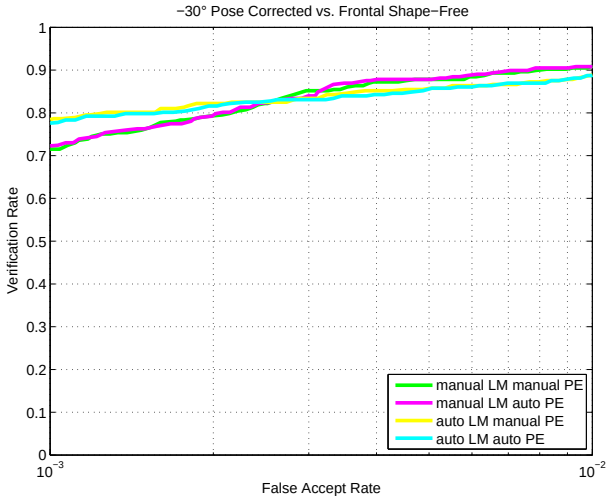
Figure D.3: ROC to compare the verification advantage of our pose-correction method (having added the shape back such that those depicted in Figure 3.16d) to traditional 2D-warping methods such as those depicted in Figure 3.16c. All unique MPIE 337 subjects using Normalized Cosine Distance as matcher.



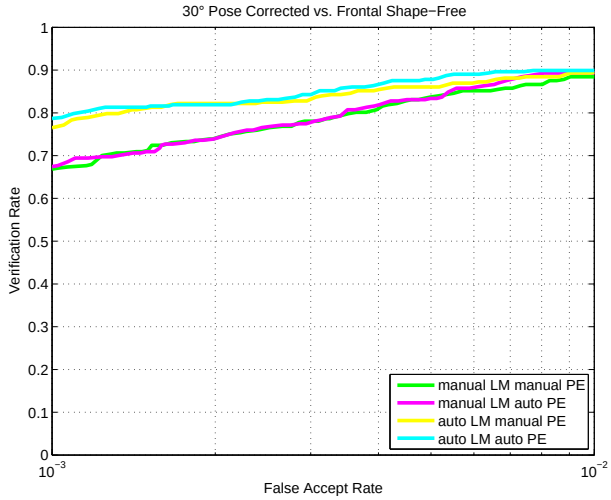
(a) L1-PCA coefficients $-45°$



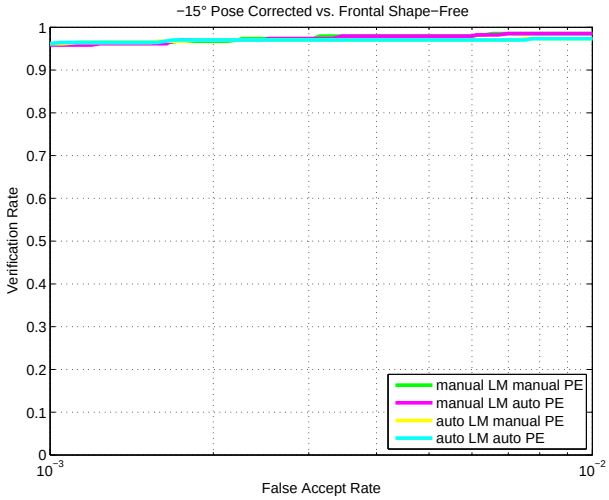
(b) L1-PCA coefficients at $45°$



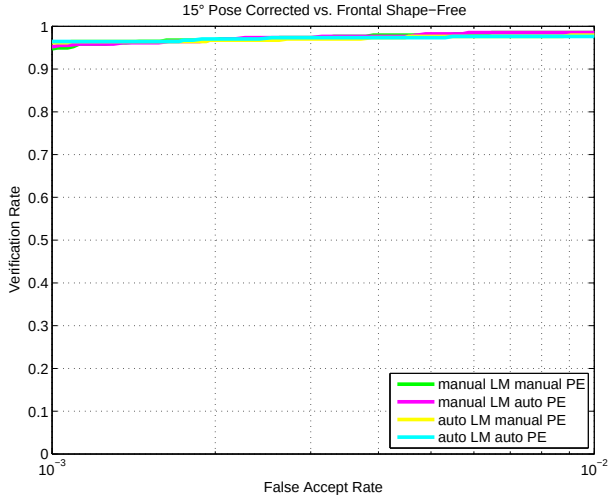
(c) L1-PCA coefficients at $-30°$



(d) L1-PCA coefficients at $30°$



(e) L1-PCA coefficients at $-15°$

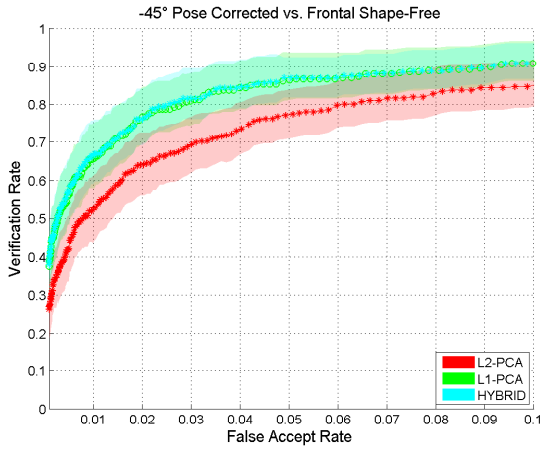


(f) L1-PCA coefficients at $15°$

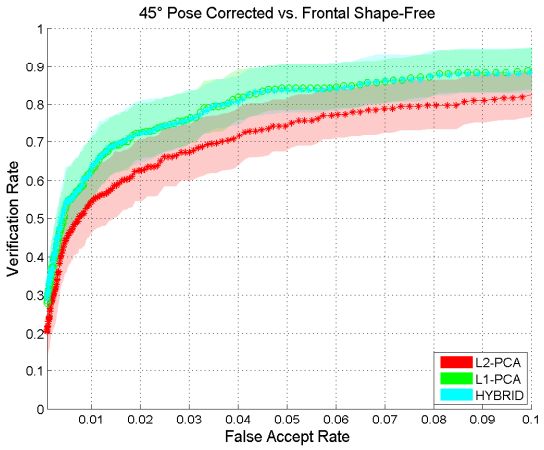
Figure D.4: MPIE sensitivity analysis on all unique 337 subjects. Normalized Cosine Distance on L1-PCA shape-free coefficients. LM denotes landmarking. PE denotes Pose Estimation.

(a) $-45°$ (b) $-30°$ (c) $-15°$ (d) $0°$ (e) $15°$ (f) $30°$ (g) $45°$

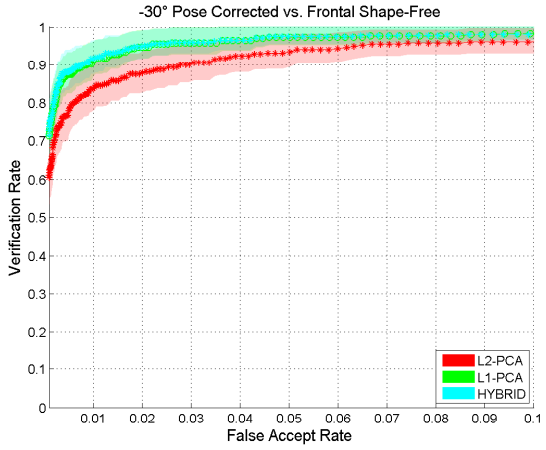
Figure D.5: Example of bad landmarking automatic fitting for all angles in the MPIE dataset. The top row represents the worst fit (defined as the highest drift between manual landmarked points and automatically landmarked points in a mean squared error sense) for the given angle. The second and third rows represents the second and third worst drift respectively. The columns (a) through (g) denotes the different yaw angles.



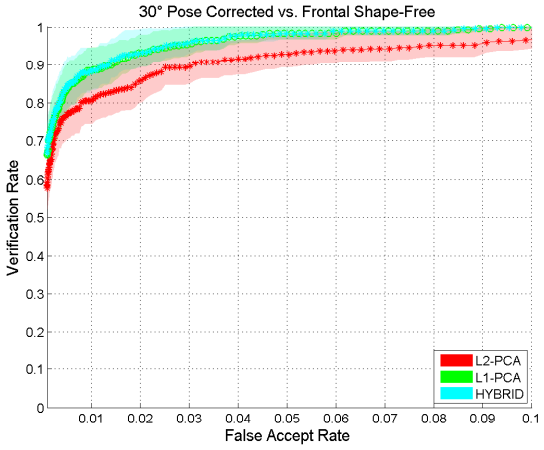
(a) 2D warping at $-45°$



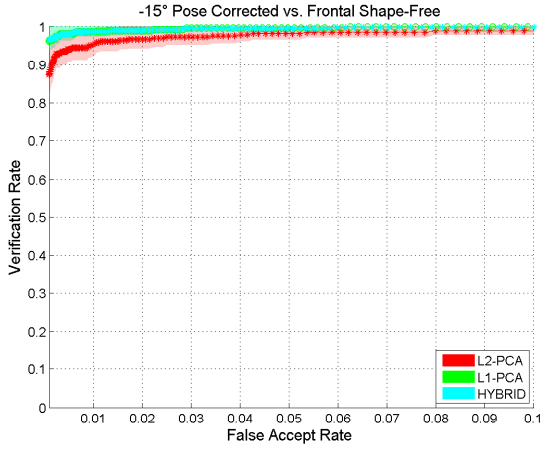
(b) 2D warping at $45°$



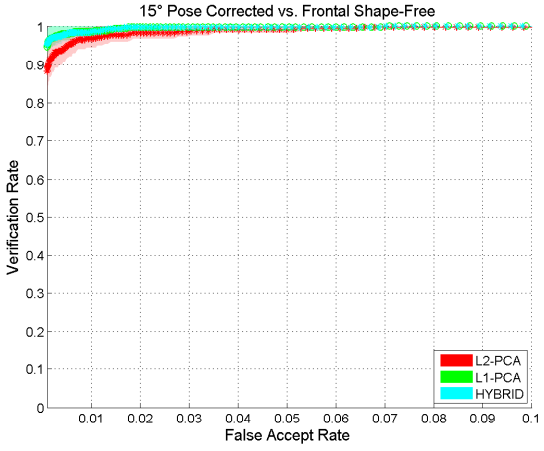
(c) 2D warping at $-30°$



(d) 2D warping at $30°$



(e) 2D warping at $-15°$



(f) 2D warping at $15°$

Figure D.6: Verification performance of matching with the coefficients after dropping the first few dimensions. The matcher used is normalized cosine distance. The testing set contains all 337 unique MPIE subjects. 95% confidence bands are centered on the ROCs to depict the statistical significance of the result.

D.2 Results on FERET Database

The following is a list of additional results and figures for the FERET pose dataset. See [57] for a description of the construction and use of the database.

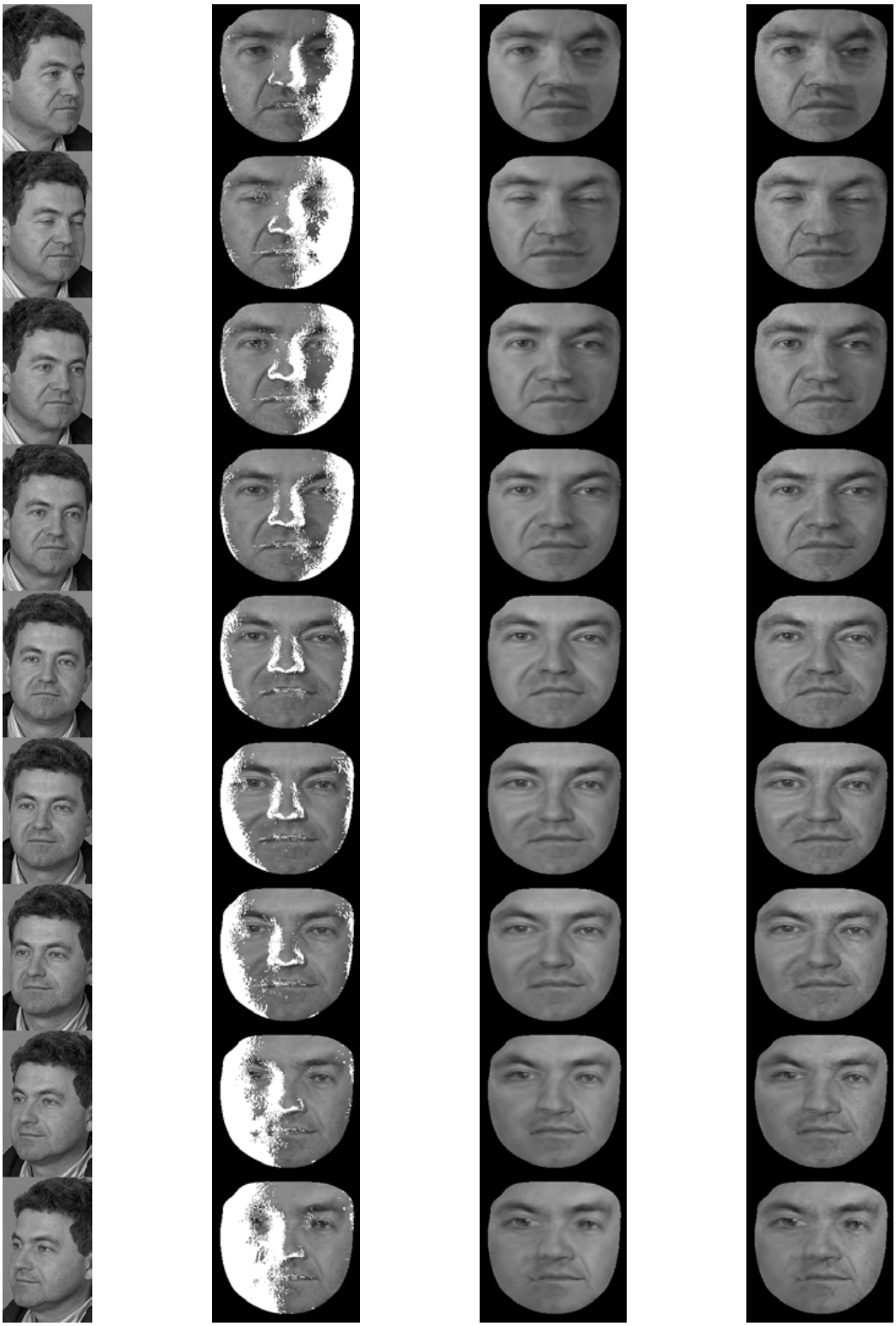
- Figure D.7: Example of different image representations after reconstructing pose-corrected images.
- Figure D.8: An additional example of different image representations after reconstructing pose-corrected images.
- Figure D.9: An additional example of different image representations after reconstructing pose-corrected images.
- Figure D.10: ROC of verification rates for matching pose-corrected images to frontal in coefficient space.
- Figure D.11: ROC of verification rates for matching pose-corrected images to frontal after applying the CFA algorithm on the L1-PCA features.
- Table D.5: Rank-1 identification rates on the FERET dataset which includes 8 different non-frontal viewpoints and 200 subjects. MPIE session 1 is used to train the CFA algorithm.
- Table D.6: Rank-1 identification rates on the FERET dataset when the gallery images are at $-40°$. MPIE session 1 is used to train the CFA algorithm.

	Yaw Test Angles							
Rank-1 Identification	$-40°$	$-30°$	$-20°$	$-10°$	$10°$	$20°$	$30°$	$40°$
L1-PCA CFA	63.50	95.00	99.50	99.50	98.00	96.50	90.00	75.00

Table D.5: Rank-1 Identification rates using normalized cosine distance on L1-PCA CFA coefficients. The size of the gallery and test set is 200 subjects. The training set consists of MPIE session 1 subjects.

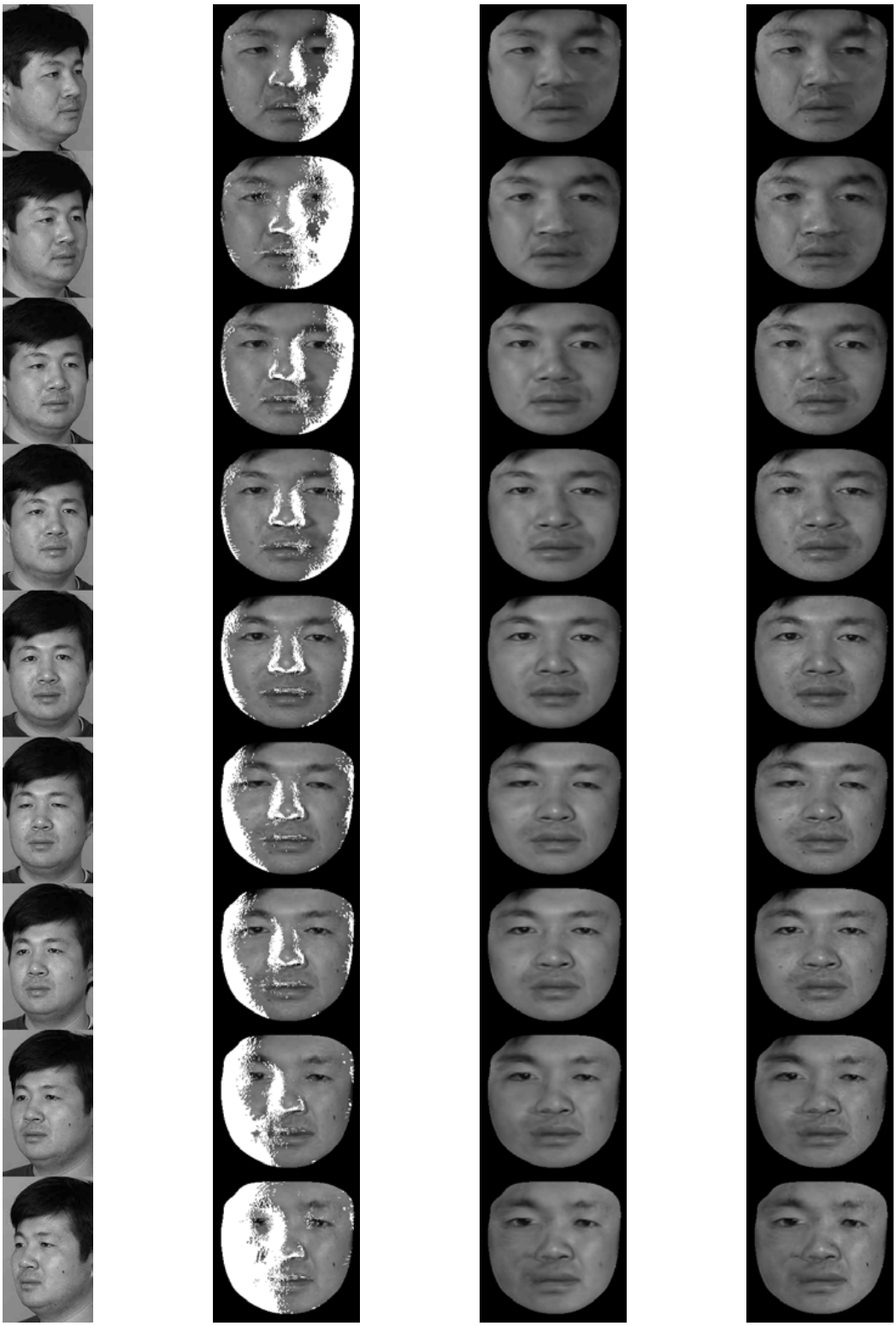
	Yaw Test Angles							
Rank-1 Identification	$-30°$	$-20°$	$-10°$	$0°$	$10°$	$20°$	$30°$	$40°$
L1-PCA CFA	90.00	86.50	78.50	75.50	68.50	66.00	59.50	54.00

Table D.6: Rank-1 Identification rates using normalized cosine distance on L1-PCA CFA coefficients when the gallery faces are at $-40°$. The size of the test set is 200 subjects. The training set consists of MPIE session 1 subjects.



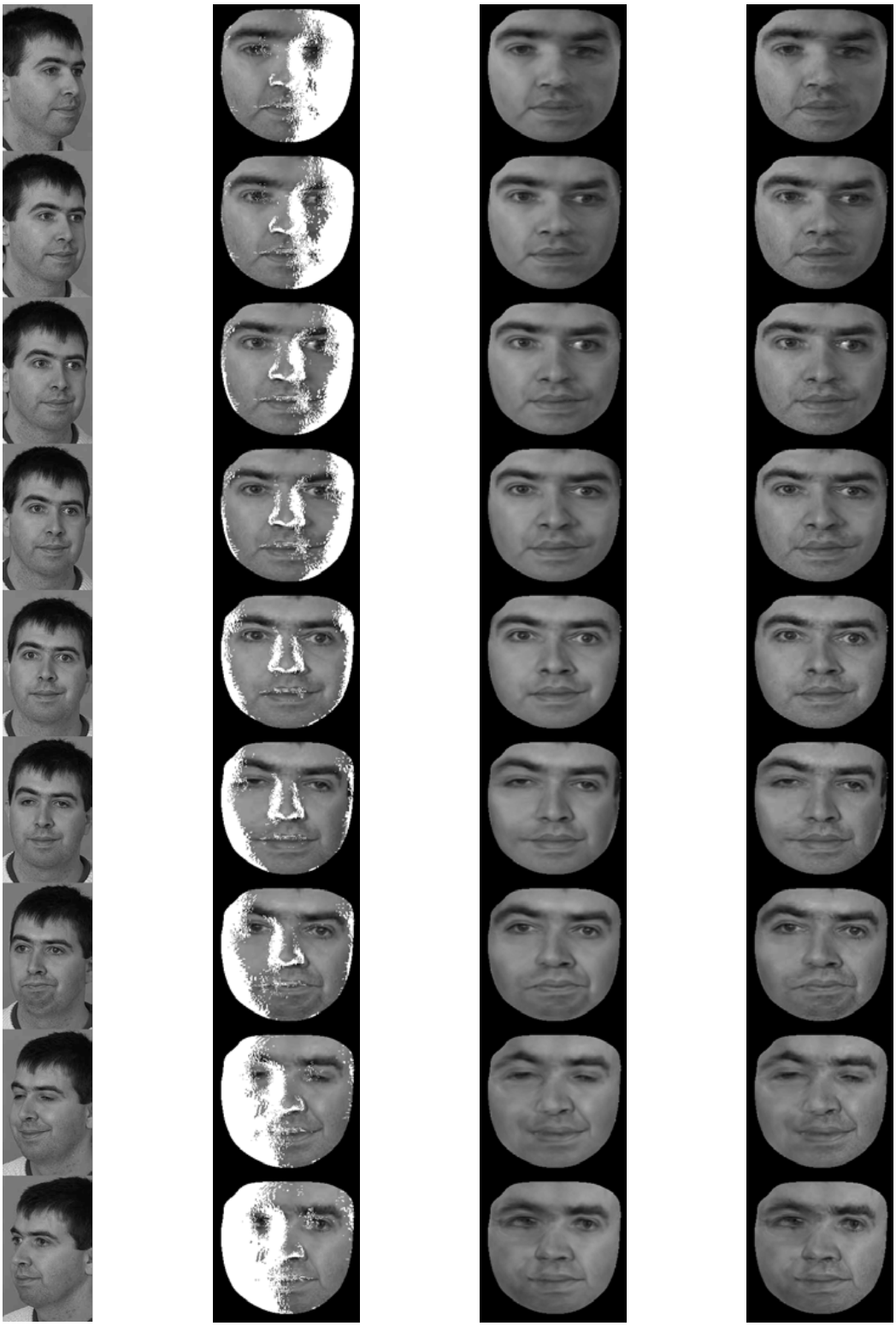
(a) Original (b) HOLES (c) L1-PCA (d) HYBRID

Figure D.7: Pose-correction for yaw angles increasing from -40° to 40° in increments of 10°. (a) input images (b) the corresponding shape-free representation with occlusion detection (c) corresponding L1-PCA shape-free reconstructions (d) HYBRID representation which consists of the the original pixels when the pixel is not occluded, and the L1-PCA equivalent pixel when occlusion is detected.



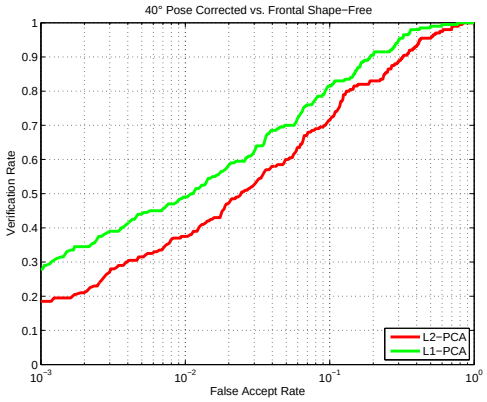
(a) Original (b) HOLES (c) L1-PCA (d) HYBRID

Figure D.8: Pose-correction for yaw angles increasing from -40° to 40° in increments of 10°. (a) input images (b) the corresponding shape-free representation with occlusion detection (c) corresponding L1-PCA shape-free reconstructions (d) HYBRID representation which consists of the the original pixels when the pixel is not occluded, and the L1-PCA equivalent pixel when occlusion is detected.

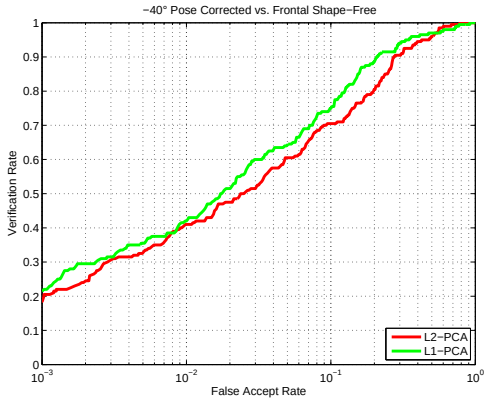


(a) Original (b) HOLES (c) L1-PCA (d) HYBRID

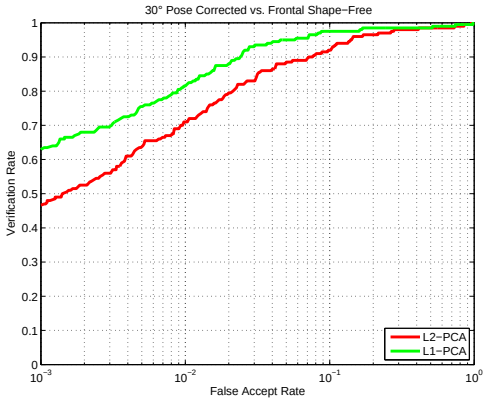
Figure D.9: Pose-correction for yaw angles increasing from -40° to 40° in increments of 10°. (a) input images (b) the corresponding shape-free representation with occlusion detection (c) corresponding L1-PCA shape-free reconstructions (d) HYBRID representation which consists of the the original pixels when the pixel is not occluded, and the L1-PCA equivalent pixel when occlusion is detected.



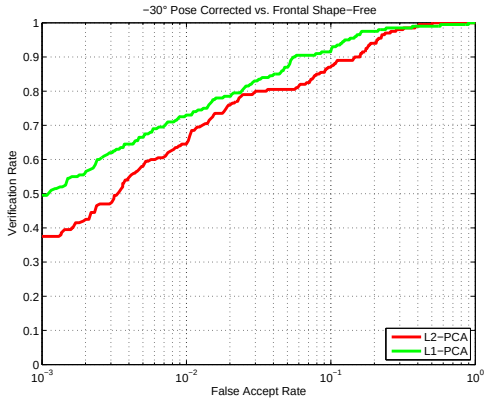
(a) Shape-free matching at $-45°$ PoseCorrection



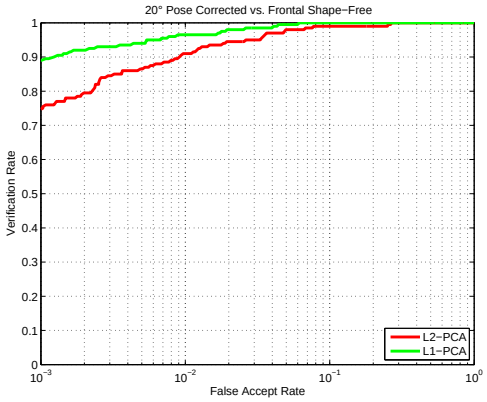
(b) Shape-free matching at $45°$



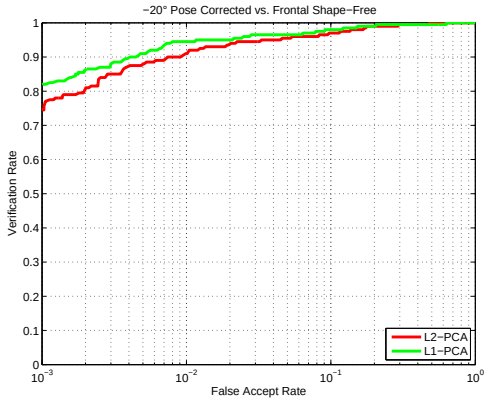
(c) Shape-free matching at $-30°$



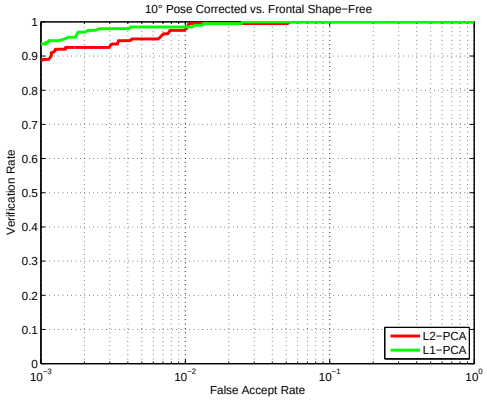
(d) Shape-free matching at $30°$



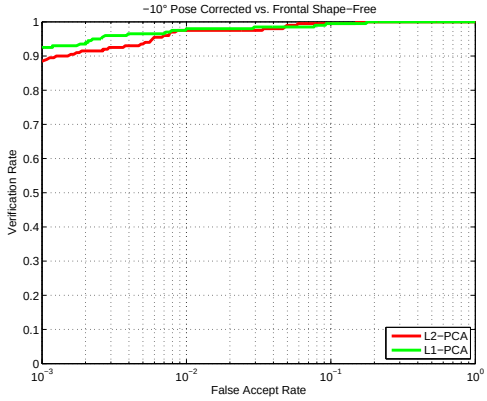
(e) Shape-free matching at $-15°$



(f) Shape-free matching at $15°$

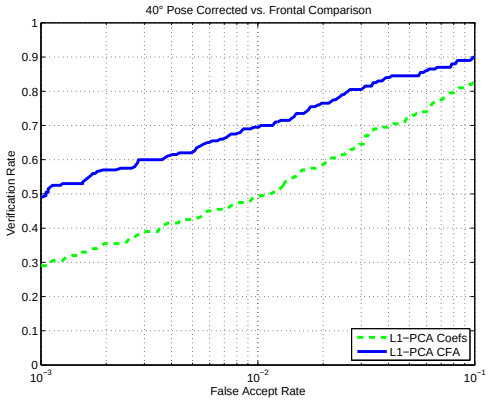


(g) Shape-free matching at $-15°$

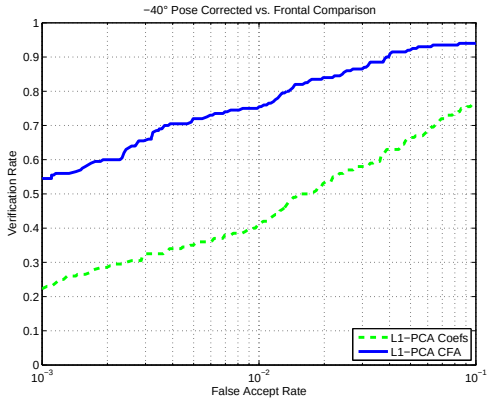


(h) Shape-free matching at $15°$

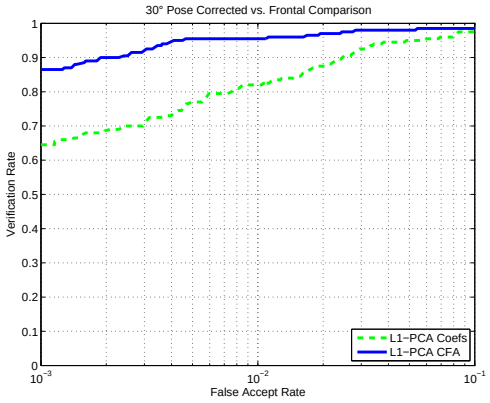
Figure D.10: Verification rates using NCD in coefficient space for the L1-PCA and L2-PCA representations. The size of the FERET dataset is 256 subjects.



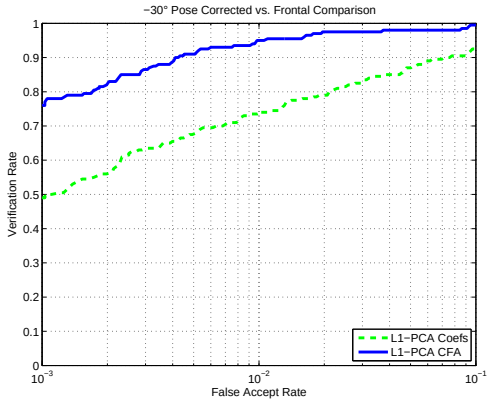
(a) Shape-free matching at $-40°$



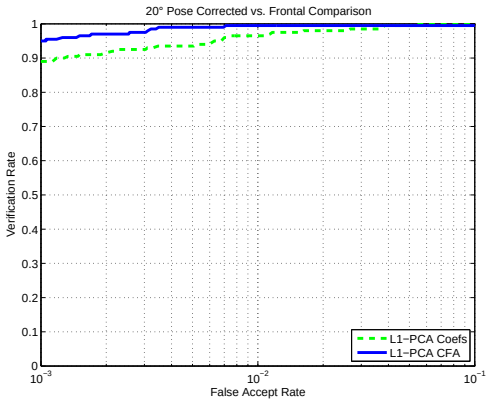
(b) Shape-free matching at $40°$



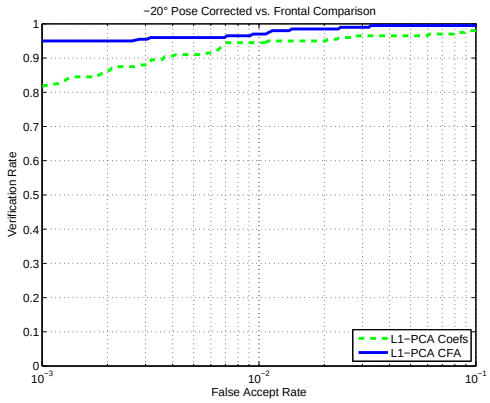
(c) Shape-free matching at $-30°$



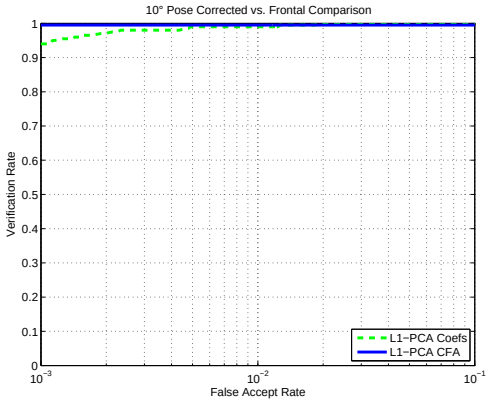
(d) Shape-free matching at $30°$



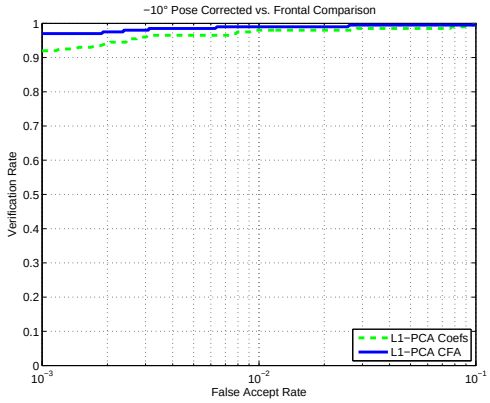
(e) Shape-free matching at $-20°$



(f) Shape-free matching at $20°$



(g) Shape-free matching at $-10°$



(h) Shape-free matching at $10°$

Figure D.11: Verification improvement due to CFA algorithm on FERET database. The size of the dataset is 200 subjects.

Bibliography

[1] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma, "Image super-resolution via sparse representation," *Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010. xxiv, 89, 92, 94, 95, 103, 104, 105

[2] Albert Mehrabian, *Silent Messages*, 1st ed. Wadsworth Publishing Company, 1971. 2

[3] Julius Fast, *Body Language*, 1st ed. Pocket Books, 1988. 2

[4] Ray L. Birdwhistell, *Kinesics and Context: Essays on Body Motion Communication (Conduct and Communication)*. University of Pennsylvania Press, 1970. 2

[5] P. Jonathon Phillips, Patrick Grother, Ross J. Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone, "Face recognition vendor test 2002: Evaluation report," Mar. 2003. 7, 11

[6] Xiaozheng Zhang and Yongsheng Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009. 7

[7] Alex Pentland, Baback Moghaddam, and Thad Starner, "View-based and modular eigenspaces for face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994, pp. 84–91. 7

[8] David Beymer, "Face recognition under varying pose," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994, pp. 756–761. 7

[9] M. Alex O. Vasilescu and Demetri Terzopoulos, "Multilinear image analysis for facial recognition," in *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 2, 2002, pp. 511–514. 7

[10] David Beymer and Tomaso Poggio, "Face recognition from one example view," in *Proceedings of the 5th International Conference on Computer Vision*, 1995, pp. 500–507. 8

[11] Thomas Maurer, Christoph Von Der Malsburg, and Ruhr universitat Bochum, "Single-view based recognition of faces rotated in depth," in *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, 1995, pp. 176–181. 8

[12] Laurenz Wiskott, Jean-Marc Fellous, Norbert Kruger, and Christoph Von der Malsburg, "Face recognition by elastic bunch graph matching," in *Proceedings of the International Conference on Image Processing*, vol. 1, 1997, pp. 129–132. 8

[13] Kin-Wang Cheung, Jiansheng Chen, and Yiu-Sang Moon, "Pose-tolerant non-frontal face recognition using ebgm," in *Proceedings of the 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2008, pp. 1–6. 8

[14] Frank Wallhoff, Stefan Muller, and Gerhard Rigoll, "Hybrid face recognition systems for profile views using the mugshot database," in *Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 2001, pp. 149–156. 8

[15] Conrad Sanderson, Samy Bengio, and Yongsheng Gao, "On transforming statistical models for non-frontal face verification," *Pattern Recognition*, vol. 39, no. 2, pp. 288–302, Feb. 2006. 8

[16] Tae-Kyun Kim and Joseph Kittler, "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 318–327, 2005. 8

[17] Ralph Gross, Iain Matthews, and Simon Baker, "Eigen light-fields and face recognition

across pose," in *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, ser. FGR '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 3–. 9, 10, 114

[18] Simon Lucey and Tsuhan Chen, "Learning patch dependencies for improved pose mismatched face verification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Jun. 2006. 9

[19] Takeo Kanade and Akihiko Yamada, "Multi-subregion based probabilistic approach toward pose-invariant face recognition," in *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 2003, pp. 954–959. 9

[20] Simon J. D. Prince, James H. Elder, Jonathan Warrell, and Fatima M. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 970–984, Jun. 2008. 9, 114

[21] Volker Blanz, Patrick Grother, P. Jonathon Phillips, and Thomas Vetter, "Face recognition based on frontal views generated from non-frontal images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, Jun. 2005, pp. 454–461. 10, 11

[22] Soma Biswas, Gaurav Aggarwal, Patrick J. Flynn, and Kevin W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, p. 1, 2013. 10, 115

[23] Wenyi Zhao and Rama Chellappa, "SFS based view synthesis for robust face recognition," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 285–292. 11

[24] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–

660, Jun. 2001. 11, 61, 114

[25] Kazuhiro Fukui and Osamu Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," in *Robotics Research*, ser. Springer Tracts in Advanced Robotics, Paolo Dario and Raja Chatila, Eds. Springer Berlin / Heidelberg, 2005, vol. 15, pp. 192–201. 11

[26] Xiaoming Liu and Tsuhan Chen, "Pose-robust face recognition using geometry assisted probabilistic modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 502–509. 11

[27] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194. 11, 12, 14, 114

[28] Lei Zhang and Dimitris Samaras, "Pose invariant face recognition under arbitrary unknown lighting using spherical harmonics," in *Biometric Authentication*, ser. Lecture Notes in Computer Science, David Maltoni and Anil K. Jain, Eds. Springer Berlin Heidelberg, 2004, vol. 3087, pp. 10–23. 11

[29] M. Judith Leo and D. Manimegalai, "3D modeling of human faces - a survey," in *Proceedings of the 3rd International Conference on Trends in Information Sciences and Computing (TISC)*, Dec. 2011, pp. 40–45. 11

[30] Jingu Heo, "Generic elastic models for 2D pose synthesis and face recognition," Ph.D. dissertation, Carnegie Mellon University, Pennsylvania, USA, 2009. 13

[31] Utsav Prabhu, Jingu Heo, and Marios Savvides, "Unconstrained pose-invariant face recognition using 3D generic elastic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1952 –1961, Oct. 2011. 13

[32] Keshav Seshadri and Marios Savvides, "Robust modified Active Shape Model for automatic facial landmark annotation of frontal faces," in *Proceedings of the IEEE 3rd International*

Conference on Biometrics: Theory, Applications, and Systems, Sep. 2009, pp. 1–8. 14, 69

[33] Charles T. Loop, "Smooth subdivision surfaces based on triangles," Department of Mathematics, University of Utah, Utah, USA, Aug. 1987. 14

[34] Jingu Heo and Marios Savvides, "3D generic elastic models for fast and texture preserving 2-d novel pose synthesis," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 563–576, 2012. 14

[35] A. Ardeshir Goshtasby, "Piecewise linear mapping functions for image registration." *Pattern Recognition*, vol. 19, no. 6, pp. 459–466, 1986. 21, 44

[36] Ralph Gross, Iain Matthews, John Cohn, Takeo Kanade, and Simon Baker, "Multi-PIE," in *Proceedings of the 8th IEEE International Conference on Automatic Face Gesture Recognition*, Sep. 2008, pp. 1–8. 22, 41

[37] Matthew Turk and Alex Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991. 22, 25

[38] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999. 22, 90

[39] Sinjini Mitra, Marios Savvides, and B. V. K. Vijaya Kumar, "Face identification using novel frequency-domain representation of facial asymmetry," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 3, pp. 350–359, 2006. 25

[40] Harold A. Sackeim, Ruben C. Gur, and Marcel C. Saucy, "Emotions are expressed more intensely on the left side of the face," *Science*, vol. 202, pp. 434–436, 1978. 25

[41] Harold A. Sackeim and Ruben C. Gur, "Lateral asymmetry in intensity of emotional expression," *Neuropsychologia*, vol. 16, pp. 473–481, 1978. 25

[42] R. Campbell, "Asymmetries in interpreting and expressing a posed facial expression," *Cortex*, vol. 14, pp. 327–342, 1978. 25

[43] Morris Moscovitch and Janet Olds, "Asymmetries in emotional, facial expressions and their possible relation to hemispheric specialization," *International Neuropsychology Society*, Jun. 1979. 25

[44] David L. Donoho, "Compressed Sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006. 30, 121

[45] David L Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Communications on Pure Applied Mathematics*, vol. 59, pp. 797–829, 2004. 30, 31

[46] Scott Shaobing Chen, "Basis pursuit," Ph.D. dissertation, Stanford University, Nov. 1995. 31, 120

[47] Emmanuel J. Candes, Justin K. Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006. 31, 122

[48] Robert J. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996. 31, 120

[49] Mário A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," IEEE Journal of Selected Topics in Signal Processing, Tech. Rep., 2007. 31, 125

[50] David L. Donoho and Yaakov Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789 –4812, Nov. 2008. 31, 124

[51] Patrick L. Combettes and Valérie R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005. 31, 125

[52] Elaine T. Hale, Wotao Yin, and Yin Zhang, "A fixed-point continuation method for ℓ_1-

regularized minimization with applications to compressed sensing," Rice University, Tech. Rep., Jul. 2007. 31, 125

[53] Yurii Nesterov and Arkadii Nemirovski, *Interior point polynomial methods in convex programming*. SIAM Studies in Applied Mathematics, 1994, vol. 13. 31, 124

[54] Junfeng Yang and Yin Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM Journal on Scientific Computing*, vol. 33, no. 1, pp. 250–278, Feb. 2011. 31, 128

[55] Allen Y. Yang, Arvind Ganesh, Zihan Zhou, Shankar Sastry, and Yi Ma, "A review of fast ℓ_1-minimization algorithms for robust face recognition," *Computing Research Repository*, vol. abs/1007.3753, Jul. 2010. 31, 124

[56] Emmanuel J. Candes, Justin K. Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006. 38, 121

[57] P. Jonathon Phillips, Sandor Z. Der, Patrick J. Rauss, and Or Z. Der, "FERET (face recognition technology) recognition algorithm development and test results," Army Research Laboratory, Tech. Rep. ARL-TR-995, Oct. 1996. 41, 157

[58] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, Jun. 2001. 44, 69

[59] Ramzi Abiantun, Utsav Prabhu, Keshav Seshadri, Jingu Heo, and Marios Savvides, "An analysis of facial shape and texture for recognition: A large scale evaluation on FRGC ver2.0," in *Proceedings of the IEEE Workshop on Applications of Computer Vision*, Jan. 2011, pp. 212–219. 55, 56

[60] J. C. Gower, "Generalized Procrustes Analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, Mar. 1975. 56

[61] P. Jonathon Phillips, Patrick J. Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek, "Overview of the Face Recognition Grand Challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2005, pp. 947–954. 56, 136

[62] Bradley Efron and Robert Tibshirani, *An introduction to the bootstrap*. Chapman & Hall/CRC, 1993, vol. 57. 62

[63] Sofus A. Macskassy, Foster Provost, and Saharon Rosset, "Roc confidence bands: an empirical evaluation," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 537–544. 62

[64] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995. 69

[65] T. F. Cootes, G. Edwards, and C.J. Taylor, "Comparing Active Shape Models with Active Appearance Models," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 1999, pp. 173–182. 69

[66] Ramzi Abiantun, Marios Savvides, and B. V. K. Vijaya Kumar, "How low can you go? Low resolution face recognition study using kernel correlation feature analysis on the FRGCv2 dataset," in *Proceedings of the Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, 2006, pp. 1–6. 86

[67] Talis Bachmann, "Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity?" *European Journal of Cognitive Psychology*, vol. 3, no. 1, pp. 87–103, 1991. 86

[68] Philippe Thevenaz, Thierry Blu, and Michael Unser, "Interpolation revisited," *IEEE Transactions on Medical Imaging*, vol. 19, no. 7, pp. 739–758, Jul. 2000. 87

[69] Robert G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE*

Transactions on Acoustics, Speech, and Signal Processing, vol. 29, pp. 1153–1160, 1981. 87, 93

[70] Seong Won Lee and Joon Ki Paik, "Image interpolation using adaptive fast B-spline filtering," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 1993. 87, 93

[71] Krishna Ratakonda and Narendra Ahuja, "POCS based adaptive image magnification," in *Proceedings of the International Conference on Image Processing*, vol. 3, 1998, pp. 203–207. 87

[72] Kris Jensen and Dimitris Anastassiou, "Subpixel edge localization and the interpolation of still images," *Transactions on Image Processing*, vol. 4, no. 3, pp. 285–295, Mar. 1995. 87

[73] Jan Allebach and Ping Wah Wongt, "Edge-directed interpolation," in *Proceedings of the IEEE International Conference on Image Processing*, 1996. 87

[74] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar, "Fast and robust multiframe super-resolution," *IEEE Transactions on Image Processing*, vol. 13, pp. 1327–1344, 2003. 87

[75] Richard R. Schultz and Robert L. Stevenson, "A Bayesian approach to image expansion for improved definition," *IEEE Transactions on Image Processing*, vol. 3, no. 3, pp. 233–242, May 1994. 88, 97

[76] Richard R. Schultz and Richard R. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, pp. 996–1011, 1996. 88

[77] Peter Cheeseman, Bob Kanefsky, Richard Kraft, John Stutz, and Robin Hanson, "Super-resolved surface reconstruction from multiple images," in *Maximum Entropy and Bayesian Methods*, ser. Fundamental Theories of Physics, Glenn R. Heidbreder, Ed. Springer Netherlands, 1996, vol. 62, pp. 293–308. 88, 97

[78] Michael Elad and Arie Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1646–1658, 1997. 88, 97

[79] Russell C. Hardie, Kenneth J. Barnard, and Ernest E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997. 88, 97

[80] Michael E. Tipping and Christopher M. Bishop, "Bayesian image super-resolution," in *Advances in Neural Information Processing Systems*. MIT Press, 2003, pp. 1303–1310. 88

[81] Jeremy S. De Bonet, "Multiresolution sampling procedure for analysis and synthesis of texture images," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 361–368. 88

[82] Simon Baker and Takeo Kanade, "Hallucinating faces," Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-99-32, Sep. 1999. 88, 90, 92, 97

[83] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang, "A two-step approach to hallucinating faces: global parametric model and local nonparametric model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. I–192–I–198 vol.1. 89, 92

[84] Xiaogang Wang and Xiaoou Tang, "Hallucinating face by eigentransformation with distortion reduction," in *Biometric Authentication*, ser. Lecture Notes in Computer Science, David Zhang and Anil K. Jain, Eds. Springer Berlin Heidelberg, 2004, vol. 3072, pp. 88–94. 89, 92

[85] Kwang In Kim and Younghee Kwon, "Single-image super-resolution using sparse regression and natural image prior," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 6, pp. 1127–1133, 2010. 90, 96

[86] Peter J. Burt, "Fast filter transform for image processing," *Computer Graphics and Image Processing*, vol. 16, no. 1, pp. 20 – 51, 1981. 90, 91

[87] Nikolaos Thomos, Nikolaos V. Boulgouris, and Michael G. Strintzis, "Optimized transmission of JPEG2000 streams over wireless channels," *IEEE Transactions on Image Processing*, vol. 15, no. 1, pp. 54–67, Jan. 2006. 93

[88] http://www.ifp.illinois.edu/~jyang29/resources.html. 95

[89] http://www.mpi-inf.mpg.de/~kkim/supres/supres.htm. 96

[90] Pablo H. Hennings-Yeomans, Simon Baker, and B. V. K. Vijaya Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8. 100

[91] Ognjen Arandjelović and Roberto Cipolla, "A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution." *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2007. 100

[92] Soma Biswas, Estefan Ortiz, and Kevin W. Bowyer, "Multidimensional scaling for matching low-resolution facial images," in *Proceedings of the 4th IEEE conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–6. 101

[93] Kui Jia and Shaogang Gong, "Multi-modal tensor face for simultaneous super-resolution and recognition," in *Proceedings of the 10th IEEE International Conference on Computer Vision*, vol. 2, 2005, pp. 1683–1690 Vol. 2. 101

[94] Jameson Merkow, Brendan Jou, and Marios Savvides, "An exploration of gender identification using only the periocular region," in *Proceedings of the 4th IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–5. 116

[95] Walter J. Scheirer, Anderson Rocha, Ross Michaels, and Terrance E. Boult, "Robust fu-

sion: Extreme value theory for recognition score normalization," in *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, Sep. 2010. 117

[96] Ralph Gross and Vladimir Brajovic, "An image preprocessing algorithm for illumination invariant face recognition," in *Proceedings of the 4th International Conference on Audio and Video-Based Biometric Person Authentication*, Jul. 2003, pp. 10–18. 117

[97] Xiaoyang Tan and Bill 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, Jun. 2010. 117

[98] Jon F. Claerbout and Francis Muir, "Robust modeling with erratic data," *Geophysics*, vol. 38, no. 5, pp. 826–844, 1973. 120

[99] Benjamin F. Logan, "Properties of high-pass signals," Ph.D. dissertation, Columbia University, 1965. 120

[100] H. L Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1 norm," *Geophysics*, vol. 44, no. 1, pp. 39–52, 1979. 120

[101] Fadil Santosa and William W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 4, pp. 1307–1330, 1986. 120

[102] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998. 120, 121

[103] Stephane G. Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397 –3415, Dec. 1993. 120, 121

[104] Y.C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive

function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40–44 vol.1. 120

[105] Deanna Needell and Roman Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of Computational Mathematics*, vol. 9, no. 3, pp. 317–334, Apr. 2009. 120

[106] David L. Donoho, Yaakov Tsaig, Iddo Drori, and Jean-Luc Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1094–1121, 2012. 120

[107] Deanna Needell and Joel A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Communications of the ACM*, vol. 53, no. 12, pp. 93–100, Dec. 2010. 120

[108] Joel A. Tropp, "Topics in sparse approximation," Ph.D. dissertation, University of Texas at Austin, Aug. 2004. 120

[109] Joel A Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004. 120

[110] Geoffrey Davis, "Adaptive nonlinear approximations," Ph.D. dissertation, New York University, Sep. 1994. 120

[111] Thomas Blumensath and Mike E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008. 120, 125

[112] Emmanuel J. Candes and Justin K. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, no. 3, pp. 969–985, Jun. 2007. 122

[113] Louise Benoit, Julien Mairal, Francis Bach, and Jean Ponce, "Sparse image representation with epitomes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition, 2011, pp. 2913–2920. 122

[114] Jianping Shi, Xiang Ren, Guang Dai, Jingdong Wang, and Zhihua Zhang, "A non-convex relaxation approach to sparse dictionary learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1809–1816. 122

[115] Emmanuel J. Candes and Terence Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005. 123

[116] Richard Baraniuk, Mark Davenport, Ronald Devore, and Michael Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, pp. 253–263, Dec. 2007. 123

[117] Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and subgaussian ensembles," *Constructive Approximation*, vol. 28, no. 3, pp. 277–289, 2008. 123

[118] Mark Rudelson and Roman Vershynin, "Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements," in *Proceedings of the 40th Annual Conference on Information Sciences and Systems*, 2006, pp. 207–212. 123

[119] David L. Donoho and Michael Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2197–2202 (electronic), 2003. 123

[120] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004. 124

[121] N. Karmarkar, "A new polynomial-time algorithm for linear programming," in *Proceedings of the 16th Annual ACM Symposium on Theory of Computing*, ser. STOC '84. New York, NY, USA: ACM, 1984, pp. 302–311. 124

[122] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert J. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004. 124

[123] Seung-jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dimitry Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007. 124, 125

[124] Magnus R. Hestenes and Eduard Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, pp. 409–436, Dec. 1952. 124

[125] Ingrid Daubechies, Michel Defrise, and Christine De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004. 125

[126] Dimitri P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999. 126, 129

[127] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Mar. 2009. 127, 128

[128] Stephen Becker, Jérôme Bobin, and Emmanuel J. Candès, "NESTA: A fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011. 127

[129] Yurii Nesterov, "A method of solving a convex programming problem with convergence rate $\mathrm{O}(1/k^2)$," *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983. 127

[130] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, Jan. 2011. 128, 129

[131] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," in *Opti-*

mization, R. Fletcher, Ed. New York: Academic Press, 1969, pp. 283–298. 128

[132] Yoav Freund and Robert E. Schapire, "Experiments with a new boosting algorithm," in *13th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 1996, pp. 148–156. 131

[133] Leo Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996. 131

[134] Robert E. Schapire and Yoram Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, Dec. 1999. 136

[135] A.Vander Lugt, "Signal detection by complex spatial filtering," *IEEE Transactions on Information Theory*, vol. 10, no. 2, pp. 139–145, 1964. 139

[136] Charles F. Hester and David Casasent, "Multivariant technique for multiclass pattern recognition," *Applied Optics*, vol. 19, pp. 1758–1761, 1980. 139

[137] Abhijit Mahalanobis, B. V. K. Vijaya Kumar, and David Casasent, "Minimum average correlation energy filters," *Applied Optics*, vol. 26, no. 17, pp. 3633–3640, Sep. 1987. 140

[138] B. V. K. Vijaya Kumar, "Minimum-variance synthetic discriminant functions," *Journal of the Optical Society of America A*, vol. 3, no. 10, pp. 1579–1584, Oct. 1986. 142

[139] Vladimir Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995. 143

[140] Ramzi Abiantun, Marios Savvides, and B. V. K. Vijaya Kumar, "Generalized low dimensional feature subspace for robust face recognition on unseen datasets using kernel correlation feature analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2007, pp. I–1257–I–1260. 143

[141] Jingu Heo, Marios Savvides, Ramzi Abiantun, Chunyan Xie, and B. V. K. Vijaya Kumar, "Face recognition with kernel correlation filters on a large scale database," in *Proceedings*

of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, 2006, pp. II–II. 143

[142] Jason Thornton, Pablo Hennings, Jelena Kovacevic, and B. V. K. Vijaya Kumar, "Wavelet packet correlation methods in biometrics," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2005, pp. 81–84. 143

[143] Ryan Kerekes, Balakrishnan Narayanaswamy, Jason Thornton, Marios Savvides, and B. V. K. Vijaya Kumar, "Graphical model approach to iris matching under deformation and occlusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6. 143

[144] Abhijit Mahalanobis, B. V. K. Vijaya Kumar, and S.R.F. Sims, "Distance-classifier correlationfilters for multiclass target recognition," *Applied Optics*, vol. 35, no. 17, pp. 3127–3133, Jun. 1996. 143