Heterogeneous Face Recognition

By

Brendan F. Klare

A Dissertation

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Computer Science and Engineering

2012

Abstract

Heterogeneous Face Recognition

By

Brendan F. Klare

One of the most difficult challenges in automated face recognition is computing facial similarities between face images acquired in alternate modalities. Called heterogeneous face recognition (HFR), successful solutions to this recognition paradigm would allow the vast collection of face photographs (acquired from driver's licenses, passports, mug shots, and other sources of frontal face images) to be matched against face images from alternate modalities (e.g. forensic sketches, infrared images, aged face images). This dissertation offers several contributions to heterogeneous face recognition algorithms. The first contribution is a framework for matching forensic sketches to mug shot photographs. In developing a technique called Local Feature-based Discriminant Analysis (LFDA), we are able to significantly improve sketch recognition accuracies with respect to a state of the art commercial face recognition engine. The improved accuracy of LFDA allows for facial searches of criminal offenders using a hand drawn sketch based on a verbal description of the subject's appearance, called a forensic sketch. The second contribution of this dissertation is a generic framework for heterogeneous face recognition. By representing images from alternate modalities with their non-linear similarity to a set of prototype subjects who provide images from each corresponding modalities, the need to directly compare face images from alternate modality is eliminated. This property generalizes the algorithm, called Prototype Random Subspaces (P-RS), to any HFR scenario. The viability of this algorithm is demonstrated on four separate HFR databases (near infrared, thermal infrared, forensic sketch, and viewed sketch). The third contribu-

tion of this dissertation is a large scale examination of face recognition algorithms in the presence of aging. We study whether or not aging-invariant face recognition algorithms generalize to non-aging scenarios. By demonstrating that they do not generalize, we conclude that the heterogeneous appearances between faces that have aged casts aging-invariant face recognition problem in the same category as heterogeneous face recognition. That is, much like images acquired in alternate modalities, aged face images should be matched using specially trained algorithms. The fourth contribution of this dissertation is an examination of how heterogeneous demographics (i.e. gender, race, and age) affect the recognition accuracy of face recognition systems. Using six different face recognition systems (including commercial systems, non-trainable systems, and a trainable face recognition system), the experiments conclude that all systems have a consistently lower recognition accuracy on the following demographic cohorts: (i) females, (ii) black subjects, and (iii) young subjects. This study also examined whether or not recognition accuracy could be improved for a specific demographic cohort by training a system exclusively on that cohort. The fifth contribution of this dissertation is an examination of the problem of identifying a subject from a caricature. A caricature is a facial sketch of a subject's face that exaggerates identifiable facial features beyond realism, yet humans still have a profound ability to identify subjects from their caricature sketch. Automated caricature recognition with the intent of discovering improved facial feature representations with respect to face recognition as a whole. To enable this task, we propose a set of qualitative facial features that encodes the appearance of both caricatures and photographs. We utilized crowdsourcing, to assist in the labeling of the qualitative features. Using these features, we combine logistic regression, multiple kernel learning, and support vector machines to generate a similarity score between a caricature and a facial photograph. Experiments are conducted on a dataset of 196 pairs of caricatures and photographs, which we have made publicly available.

To my loving wife, Christina

Acknowledgments

My efforts towards this thesis would have been in vein had it not been for the contributions of a wide range of people.

Foremost on the list of those to whom I am indebted is my advisor, Anil Jain (or as I refer to him, "Professor Jain"). I am not a superstitious person, but instead I believe that randomness (and hence, luck) shapes our life experiences. With this perhaps being the case, I consider myself infinitely lucky to have crossed paths with Professor Jain. While the more celebrated achievements in his career are clearly evident when reading his CV or doing any Google Scholar search on the topic of biometrics or pattern recognition, I believe that his passion and innate talent in mentoring and developing his students is his greatest quality. The causality is not dubious; the success of Professor Jain's students is the result of his hard work and passion towards developing them into greater researchers, engineers, and scientists.

Professor Jain has taught me the value of collaborating with fellow researchers and those those who support our academic fields from an administrative capacity, such as program managers and government employees. I have learned that great research often involves looking ahead of your peers, and trying to find solutions to a problem before people realize the problem even exists. I have learned it does not matter how great a particular algorithm or body of research is if its efficacy has been demonstrated with a flawed experimental design. I have implicitly been taught by Professor Jain to adhere to Occam's razor when developing a solution to a given problem. Finally, despite often being privy to catered lunches and Mrs. Jain's sweets in the lab, I have painfully learned the persistence of the "no free lunch" theorem as the joys of research are constantly balanced by trying times.

The most important thing I have learned from Professor Jain is that nothing

trumps passion for your profession. Professor Jain is arguably the smartest person I know, however his intelligence alone did not make him the person we all look up to. It is his day in, day out passion for research. I hope to never forget from him to approach my profession with great passion.

Success in one's career is all for naught without someone to share it with. To this end, my hard work is always with the knowledge that at the end of the day I get to come home to my truly better half: my wife Christina. My having met Christina is the only thing that prevents me from completely denouncing fate because I do not feel anyone is this lucky. No matter what I achieve in field of pattern recognition, biometrics, and beyond, I will first and foremost be defined by Christina and our amazing relationship. I would not trade anything in the world for Christina, including this dissertation. The only joy greater joy than reflecting on our nearly the near six years we have spent together is thinking of the of the rest of our lives we have ahead of us.

While Professor Jain had a profound influence on the last four years of my life, my father, Fred Klare (or "Dad"), has had the largest single influence in my life. In life, one does well to follow the proverb "you should not count your eggs until they are hatched". However, if eggs were in fact my father, this sage advice would not be necessary. You can always count on my Dad. Always. He is the most consistent and integrable person I will ever know in my life.

My father is also endlessly selfless. He lives for his four children, and never showed it more then with the sacrifices he made during the untimely passing of our mother, Anna Klare, when I was around 13 years old. My mother was a truly amazing person herself: no one who met her ever seemed to forget her. I owe my (arguable) talents in math, and intense personality (for better or worse) to her. However, after her passing, my father was left to raise four children while maintain his then career as a Special Agent in the Secret Service. No matter how much I would veer from the path

of being a hard working citizen, my father would consistently (and often painfully) remind me of this path. I have no doubt that I would not be completing a PhD had it not been for my father sacrificing so much time and effort in helping me understand the difference between right and wrong.

I am fortunate to have an amazing family. My younger sisters Kelly and Kristy are the best sisters anyone could ask for. My brother Kevin has always been there to give me advise and opinions when I need them. My Aunt Gracie has meant a great deal to our family, and I am very grateful for how much she has done to help get my mind off of tough times over the years. My step-mother Connie and my step-brother Erik make our family so much better. Christina's mother Diane Pearson has been so good to me over the years - I truly feel comfortable calling her "Mom". It has been wonderful calling the rest of Christina's family my own as well, including Rick, Andrew, Tony, Gloria, Nanny, and Poppop.

When I was 19 years old I dropped out of school and joined the Army Rangers, mostly at the advice of my father. Four years later I left a changed man that was ready for any challenge in life. My time in the 75th Ranger Regiment was profound. The rangers I served with taught me how lead by example. They taught me self sacrifice; many by paying the ultimate sacrifice to the airborne ranger in the sky. They forged in me the mindset to always strive to be better than anyone else, yet always be humble in any such endeavors. In 3rd Ranger Battalion I met some of my best friends in life (e.g. Chris K., Tony V., Andrew S., Paul O.). Finally, I learned that RLTW: Rangers Lead the Way. As I will always consider myself a ranger, my subsequent academic studies have all been with the mind set of RLTW.

When I left the Army, I applied to eight different Universities. While I was not a serious student before I joined the Army, I was indeed quite focused and determined afterwards. Such was the message of my application letter. Only one institution believed me: the University of South Florida. I am fully indebted to USF for taking

a chance on me. I tried to reward them by getting an A in ever course I toke, but I feel short with a lone B+.

My original plan upon graduating with by B.S. in Computer Science from USF was to get a position in industry. However, Christina still had a year remaining in her studies so I decided to pursue an M.S. degree at USF. During this time I was able to work with Sudeep Sarkar. Working with Professor Sarkar allowed me to discover my love for research. I am very thankful to Professor Sarkar, and I learned a tremendous amount from him. Had I not taking the undergraduate computation geometry course from him my senior year (to which he did one of the finest teaching jobs I have ever seen), then it is unlikely I would have ended up in the field of pattern recognition.

Upon arrival at the Pattern Recognition and Image Processing (PRIP) Lab at Michigan State, I quickly realized that my fellow lab mates were well trained in Professor Jain's ideals. There is a phrase for such scholars: pripies. I am privileged to be able to call myself a pripie.

The two pripies that had the great influence on my studies are Dr. Unsang Park and Serhat Bucak. Unsang help guide my through the vast field of face recognition when I arrived. I could always count on Unsang's help with any problem I had during my studies. When Unsang left for greener pastures, I felt like I lost my safety net in the lab.

Serhat joined the lab as a PhD student when I arrived in Fall 2008. I will likely look back at our time here and think of our common love of Spartan basketball (which was kindly rewarded with two trips to the Final Four), his help guiding Christina and I through the streets of Istanbul during our honeymoon, and the enjoyment of being his friend. However, Serhat also played an invaluable role in my research. I consider Serhat's knowledge of pattern recognition to be second to none. Because of this, I have consistently relied on Serhat's feedback and advice in my studies.

I am very grateful the many other pripies in the lab, which includes (but is not

limited to): Alessandra Paulino, Soweon Yoon, Abhishek Nagar, Pavan Mallapragada, Radha Chitta, Kate Bonnen, Scott Klum, Shengcai Liao, Zhifeng Li, Hu Han, Qijun Zhao, Tim Havens, Koichiro Niinuma, and Jianjiang Feng. I need give an extra mention to Alessandra who (thankfully) always seemed to catch the smallest of my mistakes. I am very thankful have been able to work with Kate as well. While Kate is my peer, I learned a lot through my fortunate role of helping to advise her studies. Kate also spent considerable time reviewing this thesis.

I am very grateful to have such an capable and cooperative thesis committee. Professor Rong Jin, Professor Yiying Tong, and Professor Selin Aviyente have been a great help during my studies. I am truly proud to have them serving as committee members for this dissertation. Professor Jin and Professor Tong have each given me specific advice on different papers that helped comprise this thesis. Professor Emeritus George Stockman was also quite important in the early stages of this thesis.

Many outside collaborators have made this thesis possible. No one has been a bigger asset and colleague than Scott McCallum. His fingerprints are on nearly every project in this thesis. Unfortunately, all I have done in return was teach him the joy of an Oberon. Others who have helped greatly along the way include: Scott Swann, Kelly Faddis, Richard Vorder Bruegge, John Manzo, Greg Michaud, Lois Gibson, Tayfun Akgul, Professor Arun Ross, Sheila Meese, Catyana Sawyer, Paul Moody, Mark Burge, and Josh Klontz. Mark Burge in particular has been a tremendous mentor and colleague during the past three years, and I am very thankful for the time I have spent working with him.

I want to thank all my friends. Sometimes my better ideas would come after enjoying a IPA from any of the fine craft breweries in the great state of Michigan. Christian Weeder, Adam Dorr, Jim Marr, Tom Robinson, Joe Sanchez, Jake Flynn, Srijana Pradhan, Kate Flynn, and many others have given me the proper balance I have needed between studies and a social life. Kent and Michaelle Rehmann are the

greatest neighbors anyone could ask for. Michael Horton and his tribe at Spartan Crossfit have also been one of the best things that have ever happened to Christina and I, and I am not sure if I could have endured the stress from this last year without blowing off steam at "the box".

Thank you everyone!

RLTW.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Automated face recognition is a rapidly growing field that uses computer algorithms to determine the similarity between two face images [73]. Automating this process of facial identification has enormous implications towards improving public safety and security, and increasing the ubiquitous nature with which we interact with intelligent machines.

The progress of face recognition technology over the past two decades has been substantial, as benchmarked by the National Institute of Standards and Technology (NIST) [34] (see Figure 1.1). Because error rates are shown to have dropped at an exponential rate, one would justifiably assume that face recognition is becoming a largely solved problem. Unfortunately, this is far from the case as many challenges in face recognition still remain (see Figure 1.2). The reduction in error rates shown in Figure 1.1 is for face images captured in a controlled environment with cooperative subjects. However, face recognition performance significantly deteriorates when variations in facial pose, facial expression, and illumination (collectively known as PIE) are introduced [108]. Examples of such variations can be found in Figure 1.3. Other factors such as image quality (e.g., resolution, compression, blur), time lapse or facial aging (see Figure 1.4), and occlusion also contribute to face recognition errors [43,44].

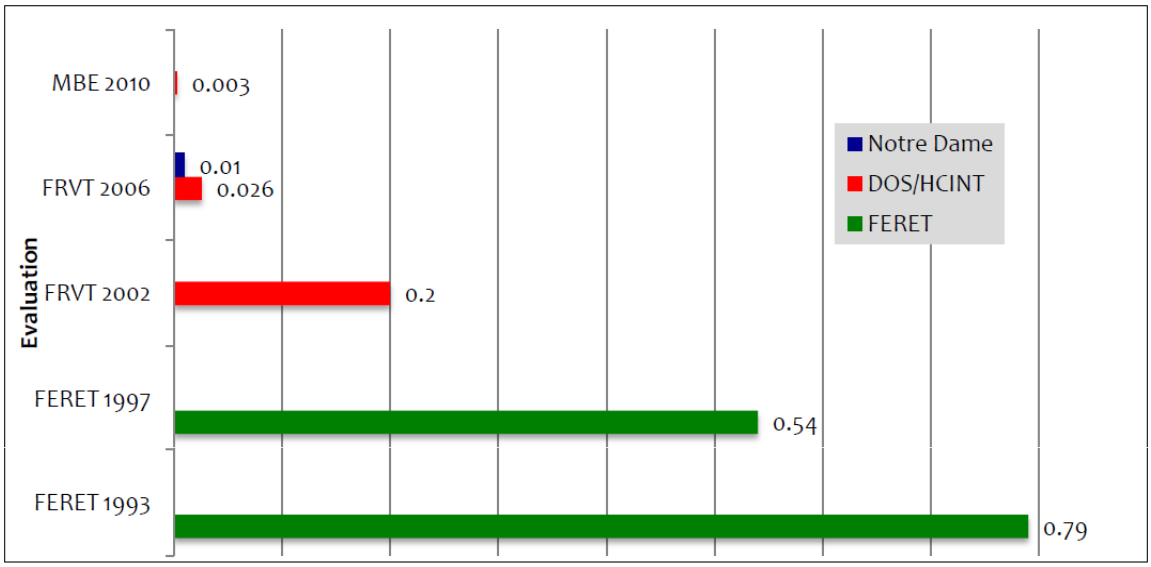
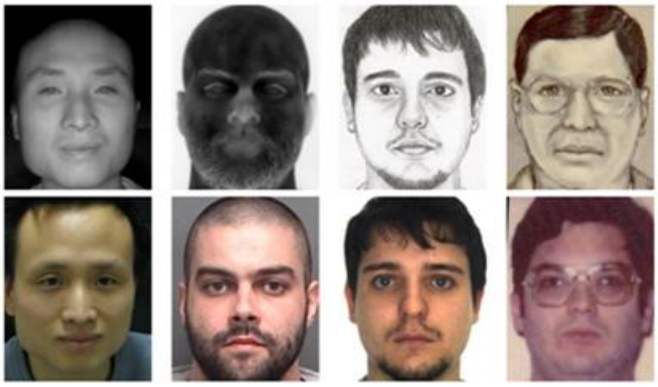


Figure 1.1: The reduction in error rates over the past 20 years for state of the art face recognition systems, as benchmarked by the National Institute of Standards and Technology [34]. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

When considering face recognition in videos, issues such as segmenting the face in varying illuminations [63] and compression artifacts [61] must be considered as well.

One of the most challenging tasks in automated face recognition is matching between two face images that have been sensed in either alternate imaging modalities (e.g. infrared images, hand drawn sketches, or depth images) or in different sensing environments and time (e.g. face images of the same person taken 10 years apart). Called *heterogeneous face recognition*, successful solutions to this face recognition paradigm extend the capabilities of face recognition to covert capture scenarios (e.g face recognition at a distance and face recognition in nighttime environments), situations where no face image even exists (forensic sketch recognition), or in situations where face images exhibit changes through the effects of aging (aging-invariant face recognition). Thus, while the majority of face recognition research seeks to mimic the capabilities of humans, heterogeneous face recognition offers the prospect of recognition capabilities beyond that of humans. *The goals and objectives of the research*

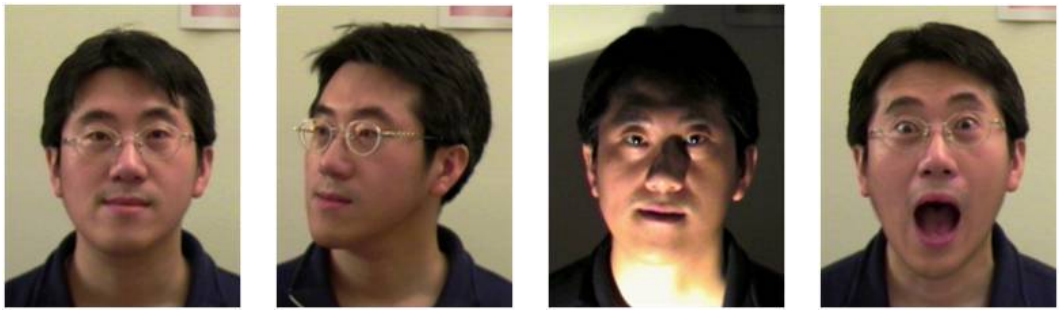


(a)

(b)

(c)

Figure 1.2: Some of the major challenges in automated face recognition. These challenges include (a) heterogeneous face recognition (top row shows face images of subjects in non-visible modalities and the bottom row shows corresponding faces in visible light), (b) unconstrained face recognition (images from [39]), and (c) aging-invariant face recognition (same subject at four different ages).



(a) (b) (c)

Figure 1.3: Pose, illumination and expression (PIE) challenges. (a) A face image with controlled pose, expression, and illumination. Face images with variations in (b) facial pose, (c) illumination, and (d) facial expression. These factors are common sources of error in automated face recognition systems.

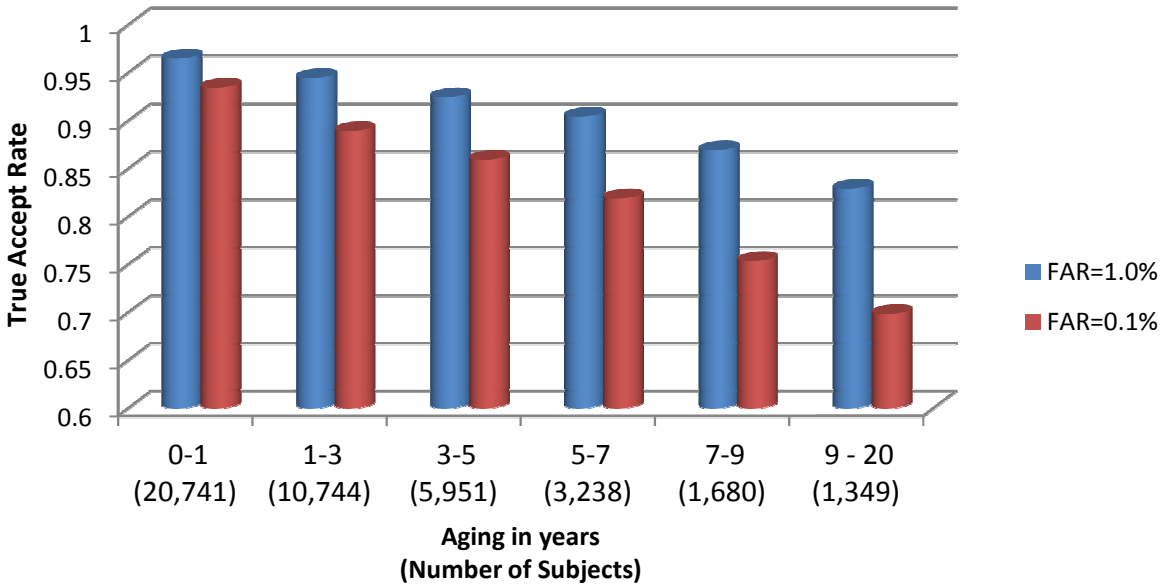


Figure 1.4: The decrease in accuracy of a leading commercial face recognition algorithm as a function of the time lapse between the probe and gallery images. Measured on a mug shot database containing 94,631 face images of 28,031 subjects, this performance degradation illustrates one of several challenges in face recognition research.

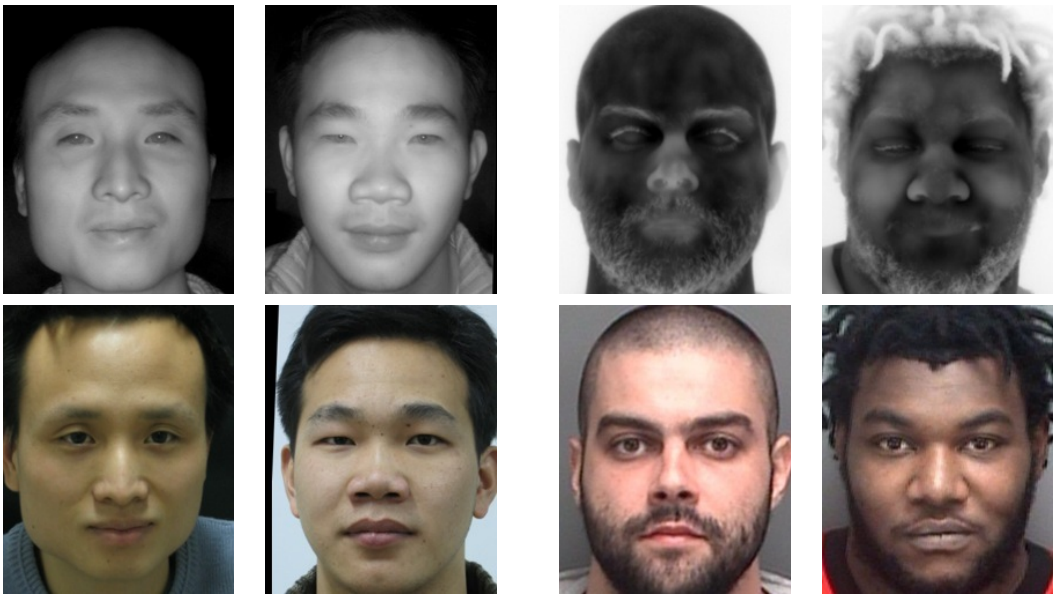
presented in this thesis are to develop representations and matching schemes that will improve the state of the art in heterogeneous face recognition.

The remainder of the introduction to this thesis on heterogeneous face recognition is organized as follows. In Section 1.1 we will trace some of the lineage of face recognition research. In Section 1.2 we will provide an overview of how automated face recognition algorithms are designed. An overview of heterogeneous face recognition will be presented in Section 1.3. Section 1.4 will outline the contributions of this thesis. Finally, Section 1.5 will discuss the organization of the remaining chapters of this thesis.

1.1 The Lineage of Face Recognition

Dating as far back as the invention of the abacus, we have sought for machines to replicate intelligent tasks performed by the human brain. The current state of intelligent automation is such that we are able to design machines that perform some of the most complicated human tasks, such as piloting vehicles, natural language processing, and face recognition. The key advancements that have allowed us to realize these technologies are the progression of computing capacities at a rate predicted by Moore's Law [106], and advancements in computer algorithms.

The progression in computing algorithms that has enabled today's intelligent machines may be attributed to research in the broad field known as artificial intelligence. With an aim to automate intelligent tasks otherwise performed by biological lifeforms, artificial intelligence spans a host of engineering applications and has attempted to leverage mathematical developments from almost every academic field.



(a) (b)

(c)

Figure 1.5: Examples images from three different heterogeneous face recognition scenarios. The top row contains probe images from (a) near-infrared, (b) thermal infrared, and (c) forensic sketch modalities. The bottom row contains the corresponding gallery (visible band face image) photographs.

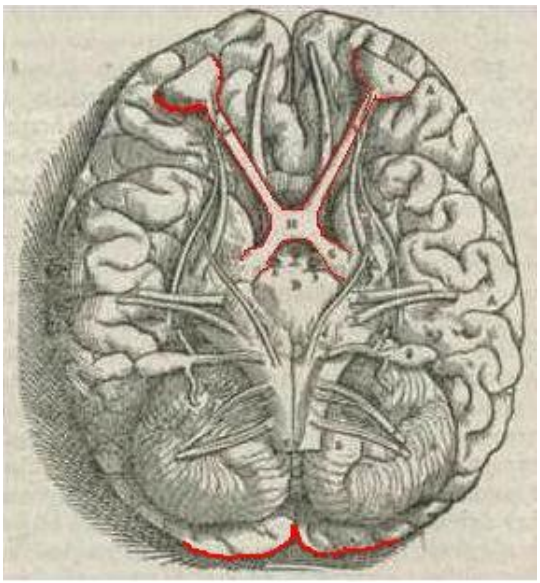


Figure 1.6: The human visual system starts with the eyes (top of the image in red) which senses visible light waves. The path of information detected by the eyes ends with the visual cortex of the brain (bottom of the image in red). The visual cortex is the largest component in the brain, and is responsible for such intelligent tasks as object recognition, motion detection, and face recognition.

The academic discipline of pattern recognition is a field within artificial intelligence that (broadly) seeks to infer high level information from low level data. Many similarities lie between pattern recognition and the closely related field of machine learning. In theory, pattern recognition algorithms are not concerned with the particular source of the data (e.g. digital images, object measurements, depth fields), but instead on the structure of the data (e.g. real-valued, nominal, ordinal, unstructured) and what information needs to be inferred from the data (e.g. classification, clustering, regression) [24]. Typically, inference from data is achieved after implementing a learning stage, which uses empirical samples of data exemplar to the inference task at hand in order to develop decision boundaries that best generalize to the aggregate of the training data to unseen test samples.

When applying pattern recognition algorithms to digital images we enter the field of computer vision [126]. Computer vision faces the bold task of replicating the largest processing system in the human brain: the visual system. The human visual system is the primary method of sensing (and hence responding to) the environment for most humans. The visual system operates by transmitting visible light waves detected by the eye to the visual cortex region of the brain (see Figure 1.6). When the signals arrive to the visual cortex, the information is transmitted through the dorsal and ventral streams of the brain. Studies have indicated that the ventral stream is used to process information regarding a person's location in relation to his environment, and the ventral stream has been shown to perform object recognition related tasks from the visible light waves sensed [32].

Strong evidence shows that when the appearance of a face enters the ventral stream, the task of associating an identity with that face is performed by a dedicated region of the brain, called the fusiform face area [47, 91]. The Thatcher effect [140], illustrated in Figure 1.7, allows us to readily observe evidence supporting the suggestion that we have a dedicated area of the brain for face processing.

Shifting focus to automated face recognition algorithms, we realize that research in the field of automated face recognition goes beyond finding solutions to a typical application of pattern recognition and computer vision. Face recognition research seeks to replicate an entire region of the brain that is predominately dedicated to this one task: extracting information from human faces. That the human brain has evolved to weight this identification task with such resources demostrates the importance and the benefit of designing computer algorithms capable of replicating this task.

1.2 Automated Face Recognition Algorithms

The challenges in designing automated face recognition algorithms are numerous. Charged with the task of outputting a measure of similarity between a given pair of face images, such challenges manifest in the following stages performed by most face recognition algorithms: (i) face detection, (ii) face alignment, (iii) appearance normalization, (iv) feature description, (v) feature extraction, and (vi) matching.

This section provides an overview of each of the above mentioned stages in automated face recognition, and follows the same ordering illustrated in Figure 1.8. The predominant focus will be on the face representation and feature extraction stages. This is because our research on heterogeneous face recognition has generally relied on improvements in these two stages to increase recognition accuracies between heterogeneous face images.

1.2.1 Detection, Alignment, and Normalization

The first step in automated face recognition is the detection and alignment of face images. Often viewed as a preprocessing step, this stage is critical to both detecting the presence of a face in a digital image, and alignment of the face with the spatial

(a)

(b)

Figure 1.7: The Thatcher effect. (a) Most people will notice only minor differences between the two inverted face images shown. (b) However, when turned upright, the differences between the same two face images are noticed to be far more severe. Our reduced ability to see the strong differences between the images in (a) is believed to be because inverted faces do not trigger the fusiform face area of the brain.

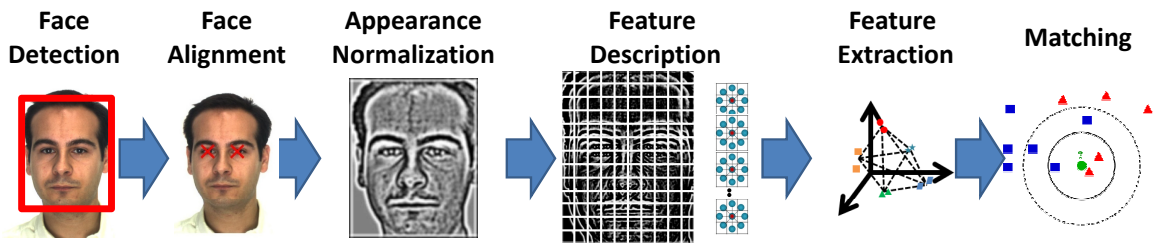


Figure 1.8: The common steps utilized by most face recognition algorithms.

(a) (b) (c)

Figure 1.9: Different methods for face alignment. (a) Face images before (left column) and after (right column) alignment through planar rotation and scaling. (b) Face images aligned using a morphable model (images from [13]), and (c) a video sequence aligned using a 3D model whose parameters were solved from a structure from motion algorithm (images from [102])

coordinate system used in the succeeding face description.

The face detector proposed by Viola and Jones [146], which uses a cascaded classifier in conjunction with images represented using a verbose set of Haar-like features, set the precedent for all modern detectors with it's robust accuracy and scalable computational complexity. While many methods have been proposed to improve upon Viola and Jones detector, it still serves as an optimistic baseline of state of the art performance [105].

Face alignment is typically performed by first detecting the location of some fixed set of anthropometric landmarks on the face. In its simplest form, these landmarks are the centers of the two eyes. Using the two eye locations, a 2D affine transformation is performed to fix the angle and distance between the two eyes. More advanced methods use 3D affine transformations, or Procustes alignment, on a more verbose set of landmarks (such as a set of landmarks outlining the locations on the mouth, nose, and face outline). The landmarks are generally detected by Active Shape Models (ASM) [21] or active appearance appearance models (AAM) [22]. Additional techniques include the use of 3D morphable models [13] and structure from motion [102]. Examples of face image alignment are shown in Figure 1.9.

Appearance normalization seeks to compensate for variations in illumination. A variety of methods have been proposed to perform such compensation, including the contrast equalization and difference of Gaussian filters proposed by Tan and Triggs [136], cones models by Georghiades et al. [30], and light field modeling by Gross et al. [33].

1.2.2 Feature Representation

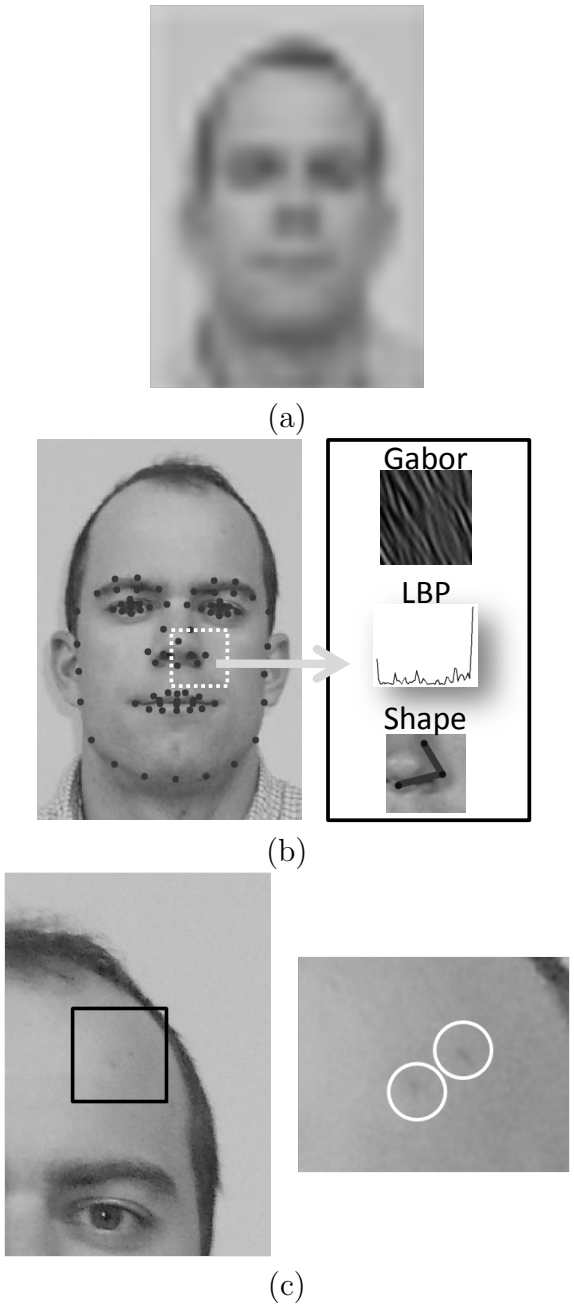
The feature representation stage encodes different facial characteristics (often implicitly) in a feature descriptor vector. Such descriptive information can range from a vector of ordered image pixel values, to distance measurements between facial com-

ponents (e.g. the distance from the nose to the mouth), or to even more complex features such as convolutions of a face image with a set of Gabor filters.

The range of representations used in face recognition is quite. Klare and Jain developed an organization of such features to facilitate studies of facial individuality and help standardize the face recognition process [56]. Below we introduce this taxonomy to provide a better understanding of the different methods by which a face images can be represented.

Klare and Jain's taxonomy organized the vast gamut of facial feature representations leveraged in automated and manual face recognition into three levels: Level 1, Level 2, and Level 3. Level 1 features consist of gross facial characteristics that are easily observable in a face, such as skin color, gender, and the general appearance of the face. Level 2 features consist of localized face information that requires specialized cortex processing, such as the structure of the face, the relationship among facial components, and the precise shape of the face. Level 3 features consist of certain irregularities in the facial skin, which includes micro features such as facial marks, skin discoloration, and moles. An example of this proposed feature grouping can be found in Figure 1.10.

This categorization of facial features is intended to provide a better understanding and standardization of both manual and automated face recognition processes. The benefit of this categorization is two fold: (i) facilitating an individuality measure for face images that can be used in legal testimony, and (ii) improving the accuracy of commercial matchers through a more careful selection of facial features. The current fingerprint feature categorization [87], accepted by both forensic scientists as well as fingerprint vendors, served as a guiding principle for our categorization of facial features. Compared to face recognition, fingerprint matching has over 100 years of history and success. Furthermore, features used in automatic fingerprint matchers (AFIS) are compact and have a physical interpretation in terms of the ridge flow



(a)

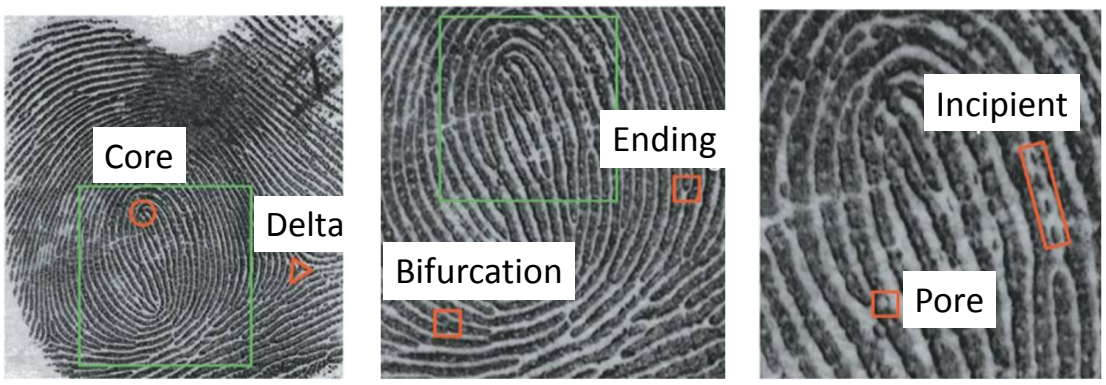
(b)

(c)

Figure 1.10: Examples of the three levels of facial features [56]. (a) Level 1 features contain low dimensional appearance information that is useful for determining highly level identifying information such as ethnicity, gender, and the general shape of a face. (b) Level 2 features require detailed processing for face recognition and captures information regarding the structure and specific shape and texture of the face. (c) Level 3 features include marks, moles, scars, and other irregular micro features of the face.

patterns in the fingerprint. Indeed state-of-the-art AFIS utilize essentially the same features that are utilized by human fingerprint examiners. This is not necessarily true for face recognition; features extracted by humans are not easy to precisely describe, and thus cannot be utilized in automatic face recognition systems. Salient features in fingerprints are categorized into three levels: Level 1 features encompass the global structure or ridge pattern (e.g. arch, loop, whorl). Level 2 features consist of minutiae location and orientation, and are primarily used for matching. Level 3 features consist of information available at higher spatial resolutions, such as dots, incipients and ridge width. An example of these fingerprint features can be found in Figure 1.11. The analogy between these widely accepted fingerprint feature levels and the proposed face feature levels will be established below.

A major benefit of Klare and Jain's facial feature taxonomy is that the same feature levels can be defined for both face recognition engines as well as human face examiners. The lack of a well defined and accepted method used in human face identification is being noticed as automated face recognition systems continue to mature [132]. The rapid growth in the use of face images captured from surveillance cameras in legal proceedings in courts has also drawn into question the methods by which human face examiners determine a person's identity using typically low quality video frames [25]. The absence of a defined set of face features prevents: (i) a generally well accepted method of human face examination, and (ii) an understanding of the statistical uniqueness of face features derivable by humans [132], and (iii) a likelihood of a false association occurring in automated face recognition systems. Ongoing studies on the individuality of fingerprints [101] are also motivated by challenges to fingerprint evidence in court cases. A report from the National Academy of Sciences on forensics [23] highlights the need for such individuality studies not only for fingerprints but for other biometric traits as well. A recent volume on forensic facial comparison [26] also mentions this report among other motivating factors for



(a) (b) (c)

Figure 1.11: A fingerprint image and its (a) Level 1, (b) Level 2, and (c) Level 3 features. The organization of facial feature is analogous to such feature levels [56].

developing face individuality models. The organization of facial features assists in conducting a study on the individuality of facial features.

Face Feature Levels

Level 1 Level 1 facial features encompass the global nature of the face, and can be extracted from low resolution face images (< 30 interpupilary pixel distance (IPD)). In automated face recognition, Level 1 features include appearance-based methods such as PCA (Eigenfaces [143]) and LDA (Fisherfaces [10]). For example, these features can generally discriminate between: (i) a short round face and an elongated thin face; (ii) faces possessing predominantly male and female characteristics; or (iii) faces from members of different ethnicities. Level 1 features cannot, however, accurately identify an individual over a large population of candidates. This is illustrated in Figure 1.12, where a query image can easily be differentiated from a subject that has a very different appearance, but cannot be distinguished from a more similar looking subject.

Level 1 facial features derivable by humans and machines are the gender, race, and general age. The postulated feedfoward nature of human face recognition also uses Level 1 features, where the initial layers can quickly discard a match candidate

Table 1.1: Example features from each of the three different levels of facial features that are used to represent face images by (a) humans, and (b) machines.

	Source: Humans and machine
Level 1	gender, race, age
Level 2	anthropometric features
Level 3	moles, scars, freckles, birth marks

(a)

	Source: Machine Only
Level 1	appearance-based methods (PCA, LDA, etc.)
Level 2	distribution-based feature descriptors (LBP, SIFT, etc.), shape distribution models, texture descriptors
Level 3	high spatial frequency

(b)

if they have a largely different facial appearance [18].

Level 1 face features are quite analogous to Level 1 fingerprint features. In each of these two traits, Level 1 features are simple to compute even in low resolution images. However, Level 1 features alone are generally only useful for indexing or reducing the search space. Level 1 features should be explicitly leveraged to improve the matching speed by using them in early stages of a cascaded face recognition system.

Level 2 Level 2 features are representations that are explicit to face recognition, and require more detailed face observations. These features are locally derived and describe structures in the face that are only relevant to face recognition (as opposed to general object recognition) due to their spatial uniqueness. Examples of such face features in automated face recognition include the use of Gabor wavelets in elastic bunch graph matching (EBGM) [151], local binary patterns (LBP) [3], SIFT feature descriptors [59, 93], point distribution models [22], texture appearance models [22], biologically inspired features proposed by Riesenhuber and Poggio (R&P) [93, 122], and explicit face geometry [124] (which includes the Bertillon system [11]).

Level 2 features are essential for face recognition. Given the strong evidence that suggests face recognition activity in humans takes place in the fusiform face area [47, 142], which is a cortical region that appears to be dedicated to face recognition. In an attempt to replicate human visual processing for face recognition, the use of Level 2 biologically inspired features in the form of Gabor wavelets have been successfully utilized in machine face recognition [93]. Along with other features such as the local binary patterns and gradient-based methods, these features are face specific, provided they are defined with respect to their spatial coordinates on the face. For example, EBGM extracts Gabor descriptors at specified locations on the face [151], and LBP and SIFT-descriptor methods extract these descriptors at uniformly distributed locations on a face that has been normalized using the eye coordinates [3, 59].

Level 2 face features are analogous to minutiae location and orientation in fingerprint recognition. In both face and fingerprint, the Level 2 features are defined with respect to a particular spatial coordinate reference, and in each case the local features can generally be computed independently of one another.

While Level 2 features are the most discriminative face features, and are predominantly used for face recognition, certain matching scenarios exist in which they alone are not sufficient. One example is face recognition in monozygotic twins [60, 134] (i.e. identical twins). Because the facial appearance of monozygotic twins is nearly identical at medium resolutions (roughly 20 to 100 IPD), Level 2 features alone are generally not sufficient for such a task. Another example where Level 2 features alone may be insufficient is age-invariant face recognition [57, 104]. As humans age, the bone structure (in early aging) and cartilage (in late aging) of the face expands and the skin wrinkles, causing both the facial shape and texture to change.

While humans extract "biological features" (i.e. neuron encodings of facial features) to recognize faces, we are limited in our knowledge of how to precisely describe these features. As a result, expert testimony for face recognition in the legal system is generally restricted to the geometric Level 2 features, such as face measurements and ratios (e.g. the ratio of the distance between the eyes and the nose width). These anthropomorphic methods were applied to systematic face recognition, prior to the advent of fingerprint identification, in the Bertillon system [11, 119]. While the uniqueness of such anthropometric features is not currently leveraged in face examination, anthropometric features: (i) have gained informal acceptance in the legal system, and (ii) are computable by both humans and machines [22]. Thus, despite the fact that anthropometric-based face recognition (i) is not a typical approach to automated face recognition, and (ii) is currently used without a consistent and proven methodology in court cases [25], *a thorough examination of their uniqueness must be undertaken*. Such a study could be guided by similar statistical studies on the unique-

Figure 1.12: An example of how Level 1 features can easily filter out faces that exhibit large differences, but cannot distinguish faces that possess many similarities. The probe image in (a) was matched to each gallery image using a Level 1 image pixel representation (difference in PCA features using the euclidean distance). Note that a larger PCA distance indicates that the faces are less similar. Using this Level 1 representation, the face in (a) matched well to an image of a similar looking subject (c) than its true mate (b), but was easily differentiated from other subjects that looked largely different (d). The information in Level 1 features is sufficient for quickly discarding some subjects (d), but more detailed Level 2 features are needed to discriminate between similar looking subjects (c). These images are from the AR face database [89].

ness of fingerprints [101] that are critical for the acceptance of fingerprint evidence in the legal system.

Level 3 Level 3 features contain unstructured, micro level features on the face, which includes scars and facial marks. Only recently has this identifiable information been explicitly considered for face recognition [60, 103]. One challenging face recognition problem where Level 3 features are critical is the discrimination of monozygotic (i.e. identical) twins [60]. Because identical twins are extremely difficult for even humans to distinguish, the presence of any small identifying mark on a face could be the difference between successful and mistaken identification. Research in the medical community has shown that while the number of moles (or nevus) in monozygotic twins is correlated, the locations of these moles are not [159] (see Figure 1.13). Level 3 features have been shown to also improve the matching accuracy in standard face

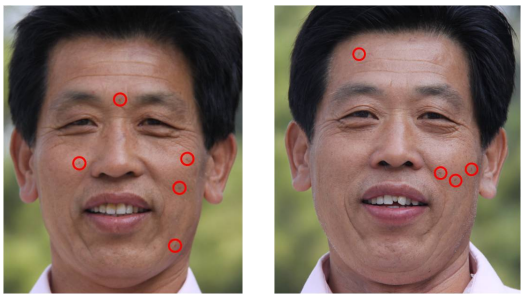


Figure 1.13: Face images of two identical twins. While their Level 1 and Level 2 features are the same, the facial mark information contained in the Level 3 features (shown in red circles) offers discriminating information [60].

recognition scenarios [103].

Level 3 features in the form of marks should be relatively easy to extract by both humans and computers. Given a good quality face image, the presence of freckles, moles, marks, and scars can be manually marked. An automated approach to mark extraction is also viable [103], though more attention is needed to develop robust solutions. For high resolution images ($>$ 100 IPD) machines are also able to extract micro texture information, though very few studies have been conducted to explicitly understand how micro texture analysis can improve face recognition. Results from the 2006 Face Recognition Vendor Test (FRVT) [110] demonstrated that high resolution face images are able to improve the matching accuracy of most commercial matchers, supporting the usefulness of micro texture information.

In fingerprints, Level 3 features include micro information such as incipient ridges and pores, and irregular information such as scars, creases and other permanent details [41]. This information is typically used by latent fingerprint examiners. In the case of AFIS matching, higher resolution fingerprint images (1000 ppi) are necessary to extract pore and ridge information to improve the matching accuracy, which is generally consistent with the proposed Level 3 face features: many moles and facial marks are not detectable at lower image resolutions. In the context of latent examination, the partial fingerprints available may require the use of Level 3 features to make

a reliable determination of a subject's identity since there may not be a sufficient number of Level 2 features (minutiae) available. Similarly, forensic examination of face images may need to leverage face mark information to make a successful identity determination [133].

It is clear now that no optimal feature representation or encoding exists for face images. However, the feature description stage is consolidated across all such representations in that each representation outputs some feature vector x that describes the face. It is from this feature vector representation that automated algorithms ultimately measure how similar two face images are.

1.2.3 Feature Extraction

With a face image I now represented by a vector x, where x is the feature vector from the above mentioned feature descriptors encodings (LBP descriptors, pixel values, etc.), a host of subspace manifold methods exist for leveraging training data (i.e. exemplar face images) to extract feature combinations which project the original features into a feature space with improved face class separation.

Principal Component Analysis Dating back to the original Eigenfaces method [143], principal component analysis (PCA) has played a vital role in the field of face recognition. PCA seeks to find an orthogonal subspace Ψ that reduces the dimensionality of the original feature space while preserving the majority of the data variance. This is achieved by performing an eigendecomposition on the covariance matrix computed from samples in the feature space. Given n samples $x_i \in \mathbb{R}^{d,1}, i = 1 \ldots n$, where x_i can be any of the feature representations discussed previously (LBP, pixels, etc.), the first step in PCA is to compute the sample mean $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$. Letting $X \in \mathbb{R}^{d,n}$ be a matrix containing each data instance centered at the mean (i.e. $X = [x_1 - \mu, x_2 - \mu, \ldots, x_n - \mu]$), we compute the scatter matrix S as $S = XX^{\mathrm{T}}$.

Finally, we seek the subspace Ψ, where

$$\Psi = \arg\max_{\Psi'} \Psi'^{\mathrm{T}} S \Psi' \tag{1.1}$$

Ψ can be solved by performing an eigendecomposition on S, which yields the matrices of eigenvectors Ψ and eigenvalues Λ, where the eigenvector in the i-th column of Ψ corresponds to the i-th diagonal entry in Λ. That is,

$$S\Psi = \Lambda\Psi \tag{1.2}$$

Generally, $d' < d$ eigenvectors are retained, such that

$$d' = \left(\arg\min_{\tilde{d}} \frac{\sum_{i=1}^{\tilde{d}'} \Lambda(i,i)}{\sum_{i=1}^{d} \Lambda(i,i)} > V_e \right) \tag{1.3}$$

where $V_e \in (0,1)$ is the fraction of data variation to be retained (e.g. 0.98).

A computational burden in solving for Ψ lies in the computation of S. This is due to the fact that often $d >> n$. That is, the dimensionality d of the feature vectors x is much greater than the number of images n. For example, if we had 1,000 images to learn the PCA space Ψ, and x was an image pixel representation for 128 x 128 sized images, then $n = 10^3$ and $d \approx 1.6 \cdot 10^4$. Thus, d is an order of magnitude larger than n. This means that the computational complexity for computing S is $O(d^2 n)$.

Though S is a d x d dimensional matrix, the rank of S will only be n. Turk and Pentland [143] showed that Ψ could instead be solved by

$$X^{\mathrm{T}} X \Psi = \Lambda' \Psi' \tag{1.4}$$

because

$$X X^{\mathrm{T}} X \Psi = \Lambda' \Psi' \tag{1.5}$$

which means $\Psi = X\Psi'$. Solving Ψ in this manner reduces the computational complexity to $O(dn^2)$.

The chief benefit of PCA lies in reducing the feature dimensionality from d to d' ($d' < d$). Typically the majority of the data variation is captured in the eigenvectors associated with large eigenvalues, and the eigenvectors associated with small eigenvalues correspond to noisy measurements. By discarding the eigenvectors associated with small eigenvalues, the feature dimensionality is greatly reduced without losing data variance information.

Linear Discriminant Analysis While PCA is effective in reducing the feature dimensionality to a more tractable size, it is not able to leverage the category information (class labels) in the training data to improve recognition accuracy. Belhumeur et al. [10] adapted linear dicrminant analysis (LDA) as a face recognition subspace technique that seeks a linear subspace projection Ψ that maximizes the discriminablity of the feature space with respect to the Fisher criterion

$$\Psi = \arg\max_{\Psi'} \frac{\Psi^{\mathrm{T}} S_B \Psi}{\Psi^{\mathrm{T}} S_W \Psi} \tag{1.6}$$

where S_B is the between-class scatter matrix and S_W is the within-class scatter matrix. That is, S_B is the scaled covariance between images of different subjects, and S_W is the scaled covariance between images of the same subject. By solving Eq. 4.1, a subspace projection is learned where (ideally) the images of the same subjects form compact groups, and images of different subjects are not well separated.

An LDA subspace projection is learned from a training set of face images of n_S different subjects, with at least two images per subject. For each subject i, the n_i feature vectors $x_i^j, j = 1 \ldots n_i$, for the i-th subject are used to compute the mean vector $\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_i^j$. From this, we compute the between-class scatter matrix S_B and the within-class matrix S_W as

$$S_B = \sum_{i=1}^{n_S} n_i(\mu_i - \mu)(\mu_i - \mu)^{\mathbb{T}} \tag{1.7}$$

$$S_W = \sum_{i=1}^{n_S} \sum_{j=1}^{n_i} (x_i^j - \mu_i)(x_i^j - \mu_i)^{\mathrm{T}} \tag{1.8}$$

where μ is the mean of all feature vectors x_i^j. Finally, Ψ is detertmined from the solution of the generalized eigenvalue problem

$$S_B\Psi = \Lambda S_W\Psi \tag{1.9}$$

Generally, this is equivalent to performing an eigendecomposition on $S_W^{-1}S_B$. However, the rank r_W of S_W will be $r_W = (\sum_{j=1}^{n_S} n_j) - n_S$. If $r_W < d$ (as is typically the case), then S_W will be singular and non-invertible. The common solution to this problem is to first perform PCA on the feature space to reduce the dimensionality to d' where $d' \leq r_W$. After this dimensionality reduction, LDA can be applied on the PCA reduced feature space.

Random Sampling Linear Discriminant Analysis Ideally, the aforementioned LDA algorithm will learn subspace projections that offer improved recognition accuracies over a PCA subspace. However, in practice, this is often not the case due to the small sample size (SSS) problem [115]. Specifically, the problem being solved by LDA is often ill-posed because the number of training samples per subject (i.e. face images for each subject) is too small with respect to the number of feature dimensions. Because of this, subspaces learned through LDA may have high generalization errors.

A solution to the SSS problem is the random sample linear discrminant analysis (RSLDA). This approach was first introduced by Wang and Tang [148] using image

pixel representations. Li et al. later extended this method in order to make it applicable to Level 2 features (SIFT and LBP) [76]. Klare and Jain showed the effectiveness of the RSLDA framework on a large aging dataset [57].

Random sampling LDA mitigates the small sample size problem by decomposing the feature space into more compact and solvable subsets. This approach follows the concept of an ensemble classifier, where Schapire combined multiple weak classifiers into a single strong classifier [125].

RSLDA learns B different subspaces. For each subspace, the d-dimensional feature space spanned by x is randomly sampled so that a subset of size $d_r < d$ of the original d features is retained. LDA is then performed using this reduced feature space. However, in addition to sampling the feature space, the training subjects are also randomly sampled. Thus, only a portion of the original subjects are used in each of the B stages.

Once the set of subspaces has been learned, a face feature vector is projected into each of the B subspaces (resulting in B feature vectors describing the face). These feature vectors may then be concatenated into a single vector for matching.

1.2.4 Matching

The matching stage outputs a measure of similarity (or dissimilarity) between two face images, where the feature vectors used to compute such (dis)similarities are the outputs from the feature extraction stage discussed above. Most simply, matching is performed using the nearest neighbor classification algorithm [24]. That is, a probe (or query) image is matched against a gallery (or database) by finding the face image in the gallery with the minimum distance (such as the Euclidean or cosine distance) or maximum similarity.

Often the matching stage can be augmented by an additional stage of statistical learning (that is, in addition to the learning that occurred in the feature extraction

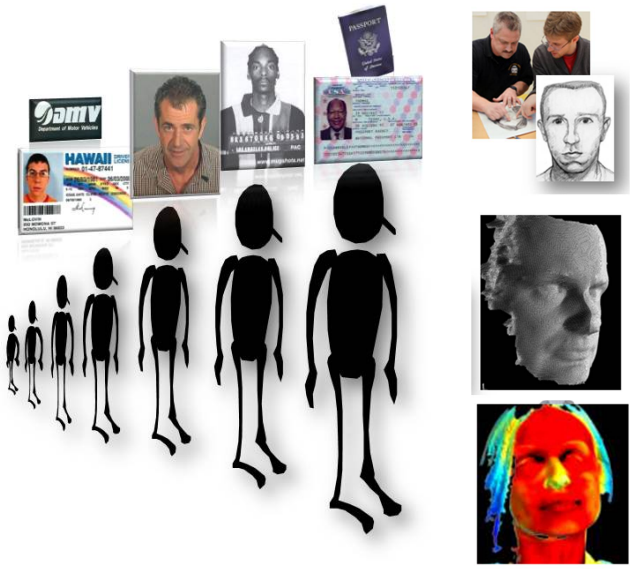
stage). A common notion here is to map the task of generating a measure of similarity between two faces images to a binary classification problem that determines whether or not two face images are of the 'same subject' or a 'different subject'. This notion can easily leverage a host of binary classification algorithms from machine learning and pattern recognition literature by creating new feature vectors that are the difference between feature vectors extracted from two face images.

The method by Moghaddam et al. was seminal in using this binary classification approach by modeling the difference vectors with Bayesian maximum a posteriori density estimation [96] to generate a probabilistic measure of similarity between two face images. This technique has also been applied using support vector machines [36].

Fusion techniques, as discussed by Ross and Jain [123], may also by exploited to improve face matching. While typically applied to multi-biometric scenarios, the same approach is viable for face recognition scenarios where, for example, the use of multiple face representations (such as LBP and Gabor), multiple views of a face, or multiple RSLDA subspaces can be consolidated to achieve better discrimination.

1.3 Heterogeneous Face Recognition

Now that we better understand the face recognition process, we can shift our attention to the topic of heterogeneous face recognition (HFR). The key difficulty in matching face images from alternate modalities is that face images of the same subject may differ in appearance due to the change in image modality. Heterogeneous face recognition algorithms must develop representation schemes to be invariant to such intra-class variations. Two of the most intuitive methods for achieving this invariance are the selection of feature descriptor encodings that are stable between heterogeneous modalities, and learning feature extractions from such encodings that further compensate for such undesired differences.



(a) (b)

Figure 1.14: Most heterogeneous face recognition scenarios leverage large visible light face image databases to determine a subject's identity from their face image acquired in some non-visible modality. (a) Between driver's licenses, passports, and mug shots, a visible light face image exists for a majority of the population. (b) Many forensic and law enforcement scenarios only have face images available from alternate imaging sources such as infrared, LIDAR, or forensic sketches.

The most frequent heterogeneous face recognition scenario involves gallery databases with visible light face images, and probe images from some alternate modality such as infrared, sketch, or depth images (see Figure 1.14). The motivation behind solutions to these scenarios is that through sources such as state DMV driver license photos, law enforcement mug shot records, the FBI's Next Generation Identification initiative, and the US-VISIT program, visible photograph databases cover the majority of the U.S. population. In fact, visible face images databases cover the population of most other developed nations as well.

While the standard face recognition paradigm is to match against the above mentioned face databases with homogeneous face images (e.g. surveillance images, mug shots, images from social networking sites), heterogeneous face recognition seeks to query these databases with images captured from imaging devices of an alternate modality.

Many scenarios exist in which the only available probe images are not visible light face images. For example, when no face image exists of a subject (suspect), a forensic sketch may be developed through a verbal description of a subject's appearance. In nighttime environments infrared imaging must be used to capture a subject's face biometric. In order to identify subjects in these scenarios, specialized algorithms for heterogeneous face recognition must be employed.

The collection of solutions to heterogeneous face recognition can be organized into three approaches:

- **Synthesis methods**: Synthesis methods seek to generate a synthetic visible light photograph from the alternate modality face image. Once a synthetic visible face image has been generated, it can be matched using standard face recognition algorithms. Synthesis solutions to heterogeneous face recognition are generative methods, and have been solved using local linear embedding [82] or Markov random fields [149]. Park et al. handled the heterogeneity in facial

aging by synthesizing the aging process [104].

- **Feature-based methods**: Feature-based methods encode face images from both modalities using feature descriptors that are largely invariant to changes between the two domains. For example, local binary patterns [99] and SIFT feature descriptors [84] have been shown to be stable between sketch and visible photographs [59], near-infrared face images and visible photographs [77], and time-separated (aged) face images [57, 76]. Once face images from both modalities are represented using feature descriptors, feature extraction methods such as LDA can be used to improve the discriminative abilities of the representation. The matching stage of the feature-based methods is performed by measuring the distance or similarity between the feature vector representation of two face images.

- **Prototype similarity methods**: As we will discuss in Chapter 3, prototype similarity methods represent a face image as a vector of similarities to a collection of prototype face images [54]. The prototypes are a collection of subjects that each contain a face image from both the probe and gallery modalities. The prototypes are analogous to a training set - in this case they help approximate the distribution of faces. Because each prototype has a face image from each modality, the vector of similarities for a face image is measured against the images from the corresponding modality. Similarities are measured between feature-based representations (e.g. LBP, SIFT) of the face images. The use of similarities naturally extends to kernel similarities, with the kernel space offering a non-linear feature space. Linear discriminant analysis may also be applied on the vectors of prototype similarities to improve the recognition accuracy. A chief benefit of prototype similarity algorithms is that the feature representation in which the similarities are computed may be different for the probe and gallery

modalities. This property is important for scenarios such as 3D to 2D matching, where common feature descriptors do not exist between the modalities.

1.4 Contributions

The contributions of this dissertation are as follows:

1. A framework for feature-based heterogeneous face recognition is developed. This framework, called Local Feature-based Discriminant Analysis (LFDA), achieves state of the art accuracy when applied to the heterogeneous face recognition task of matching sketches and photographs.

2. An approach to heterogeneous face recognition is developed which uses prototype similarities to eliminate the need to directly measure the similarity between images from heterogeneous modalities. In requiring that face similarities be computed within each modality only, the prototype heterogeneous face recognition method generalizes to any HFR scenario.

3. In showing that aging-invariant face recognition systems do not generalize to face images that have not aged, it is demonstrated that face recognition in the presence of time lapse can be viewed as a heterogeneous face recognition problem.

4. Different sources of demographic information (race, gender, and age) are exploited to perform dynamic face matcher selection. This paradigm for face recognition uses the available demographics of the probe image to improve face recognition accuracies.

5. A set of qualitative facial features is developed to enable matching caricature sketches to photographs. These features are: (i) able to encode key facial characteristics that are used in caricatures to convey a subject's identity, and (ii)

robust to facial variations such as the unconstrained exaggerations performed by a caricaturists.

1.5 Thesis Organization

In Chapter 2 we present a solution to the problem of heterogeneous face recognition between forensic sketches and mug shot photographs. In Chapter 3, a framework for heterogeneous face recognition is introduced which uses prototype similarity features to generalize the heterogeneous face recognition task to any scenario. Chapter 4 presents a study on the generalization of aging-invariant face recognition to non-aging scenarios. Chapter 5 presents a study on how heterogeneous demographics may be exploited to improve face recognition performance. Chapter 6 studies the task of a matching a caricature sketch to a facial photograph. Finally, we conclude with the findings of this dissertation and suggestions for future research in Chapter 7.

Chapter 2

Forensic Sketch Recognition

2.1 Introduction

Progress in biometric technology has provided law enforcement agencies additional tools to help determine the identity of criminals. In addition to DNA and circumstantial evidence, if a latent fingerprint is found at an investigative scene, or a surveillance camera captures an image of a suspect's face, then these cues may be used to help determine the culprit's identity using automated biometric identification. However, many crimes occur where none of this information is present, but instead an eyewitness account of the crime is available. In these circumstances a forensic artist is often used to work with the witness in order to draw a sketch that depicts the facial appearance of the culprit according to the verbal description. Once the sketch image of the transgressor is complete, it is then disseminated to law enforcement officers and media outlets with the hope that someone will come forward who knows the suspect. These sketches are known as forensic sketches and this chapter describes a robust method for matching forensic sketches to large mughshot (image) databases maintained by law enforcement agencies.

Improving forensic sketch recognition performance is perhaps the most impactful

area of heterogeneous face recognition research [55]. This is because enabling a search of a large mug shot or driver license database using a forensic sketch is equivalent to a face search using a verbal description. That is, we are able to search digital face image databases without even having a face image as a query. As such, the research presented in this chapter offers a strong contribution to the goals of this dissertation.

Two different types of face sketches are discussed in this chapter: *viewed sketches* and *forensic sketches* (see Figure 2.1). *Viewed sketches* are sketches that are drawn while viewing a photograph of the person or the person himself. *Forensic sketches*, on the other hand, are drawn by interviewing a witness to gain a description of the suspect. Published research on sketch to photo matching to this point has primarily focused on matching viewed sketches [53] [138] [82] [158] [149], despite the fact that real world scenarios only involve forensic sketches. Both forensic sketches and viewed sketches pose challenges to face recognition due to the fact that probe sketch images contain different textures compared to the gallery photographs they are being matched against. However, forensic sketches pose additional challenges due to a witness's inability to exactly remember the appearance of a suspect and her subjective account of the description, which often results in inaccurate and incomplete forensic sketches. Experimental results on viewed sketches[1] are included primarily for historical reasons since all available research to date on sketch recognition has focused on viewed sketches.

We highlight two key difficulties in matching forensic sketches: (1) Matching across image modalities, and (2) performing face recognition despite possibly inaccurate depictions of the face. In order to solve the first problem we use *local feature-based discriminant analysis (LFDA)* to perform minimum distance matching between sketches

[1]A viewed sketch is a facial sketch drawn while viewing a photograph of the subject. The scenario is not particularly interesting because the photograph itself could be queried in the FR system.

(a) (b)

(c)

Figure 2.1: The difference between viewed sketches and forensic sketches. (a) Viewed sketches and their corresponding photographs. (b) Two pairs of good quality forensic sketches and the corresponding photographs. (c) Two pairs of poor quality forensic sketches and the corresponding photographs. Sketches were labeled as "good" if they (subjectively) exhibited a mostly accurate portrayal of a subject. Otherwise, if a sketch did not strongly resemble the subject, it was labeled as "poor".

and photos, which is described in Section 2.3 and summarized in Figure 2.2 and Figure 2.3. The second problem is considered in Section 2.5, where analysis and improvements are offered for matching forensic sketches against large mugshot galleries.

The contributions of the chapter are summarized as follows: (i) We observe a substantial improvement in matching viewed sketches to photos over published algorithms using the proposed local feature-based discriminant analysis; (ii) we present the first large-scale published experiment on matching operational forensic sketches; (iii) using a mugshot gallery of 10,159 images, we perform race and gender filtering to improve the matching results; (iv) all experiments are validated by comparing the proposed method against a leading commercial face recognition engine. The last point is significant since earlier studies on viewed sketches used PCA (eigenface) matcher as the baseline. It is now well known that the performance of a PCA matcher can be easily surpassed by other face matchers.

2.2 Related Work

Most research on sketch matching has dealt with viewed sketches. Much of the early work in matching viewed sketches was performed by Tang et al. [137] [138] [82] [149] [78]. These studies share a common approach in that a synthetic photograph is generated from a sketch (or vice-versa), and standard face recognition algorithms are then used to match the synthetic photographs to gallery photographs. The different synthesis methods used include an eigentransformation method (Tang and Wang [137] [138]), Local Linear Embedding (Liu et al. [82]), and belief propagation on a Markov random field (Wang and Tang [149]). Other synthesis methods have been proposed as well [158] [29] [152] [83] [74]. The impact of matching sketches drawn by different artists was studied by Al Nizami et al. [98].

We also proposed a method of sketch matching that uses the same feature-based

approach that has been successful in other heterogeneous face recognition scenarios (specifically matching near infrared face images to visible light) [53]. In using SIFT feature descriptors [84], the intrapersonal variations between the sketch and photo modality was diminished while still maintaining sufficient information for interclass discrimination. Such an approach is similar to other methods proposed in the literature [77] [68] [52] of matching near infrared images (NIR) to visible light images (VIS), where local binary pattern (LBP) [99] feature descriptors are used to describe both NIR and VIS images.

In this chapter we extend our previous feature-based approach to sketch matching [53]. This is achieved by using local binary patterns (LBP) in addition to the SIFT feature descriptor, which is motivated by LBP's success in a similar heterogeneous matching application by Liao et al. [77]. Additionally, we extend our feature-based matching to learn discriminant projections on "slices" of feature patches, which is similar to the method proposed by Lei and Li [68].

2.3 Feature-based Sketch Matching

Feature descriptors describe an image or image region using a feature vector that captures the distinct characteristics of the image [95]. Image-based features have been shown to be successful in face recognition, most notably with the use of local binary patterns [3].

2.3.1 Feature-based Representation

We will now describe how to represent a face with image descriptors. Because most image descriptors are not sufficiently verbose to fully describe a face image, the descriptors are computed over a set of uniformly distributed subregions of the face. The feature vectors at sampled regions are then concatenated together to describe the en-

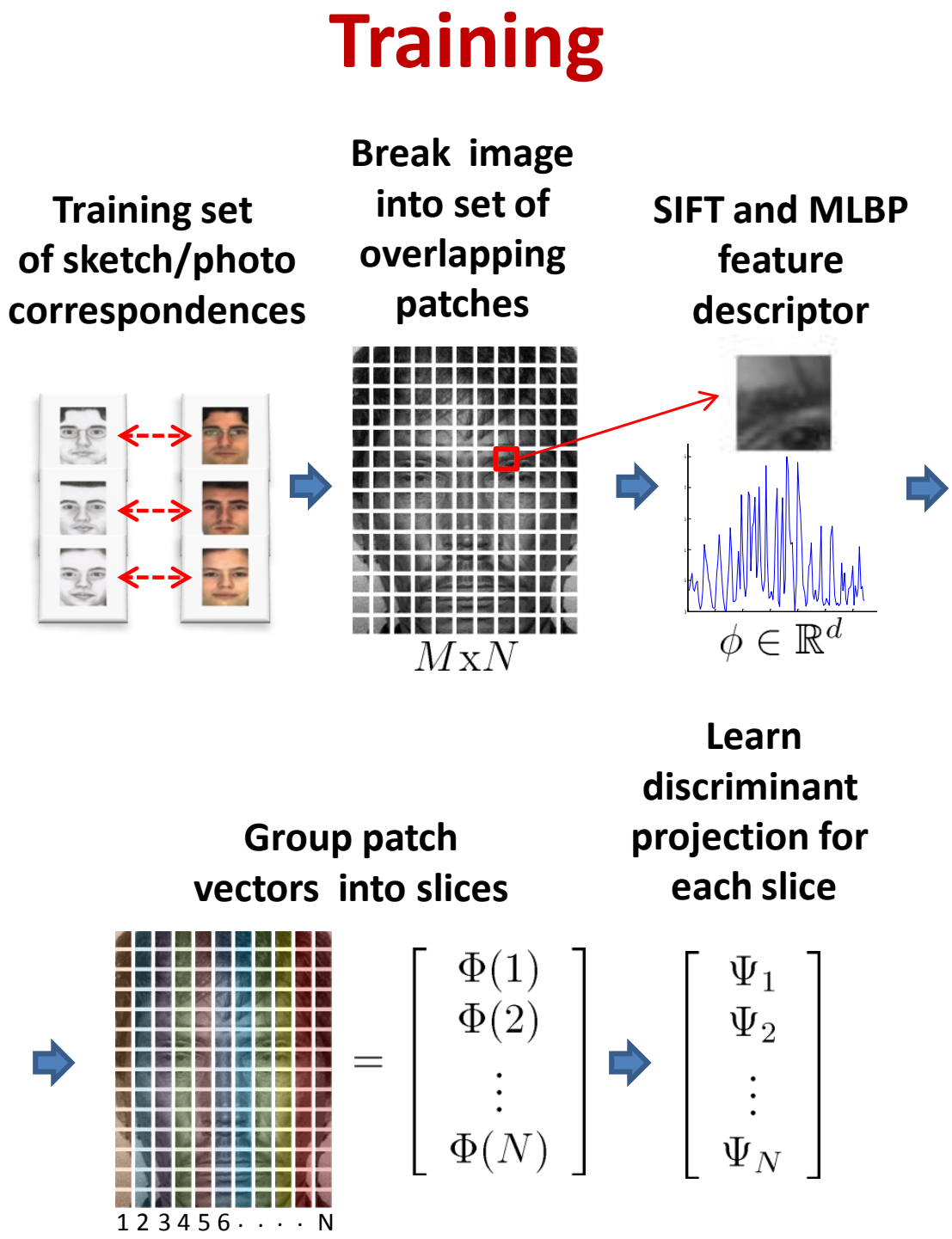


Figure 2.2: An overview of training using the *LFDA* framework. Each sketch and photo is represented by SIFT and MLBP feature descriptors extracted from overlapping patches. After grouping "slices" of patches together into feature vectors $\Phi(k)$ $(k = 1 \cdots N)$, we learn a discriminant projection Ψ_k for each slice.

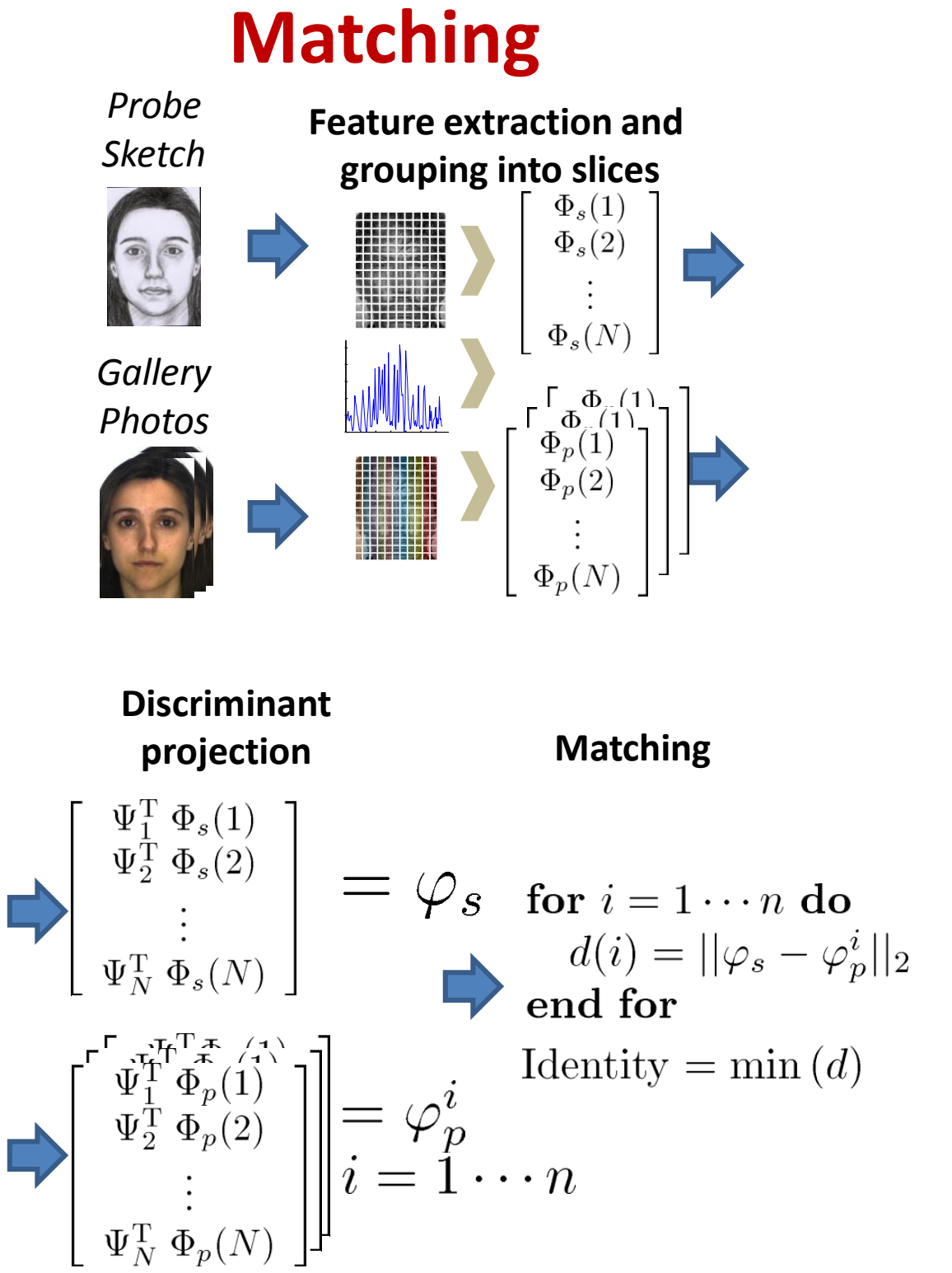


Figure 2.3: An overview of matching using the *LFDA* framework. Recognition is performed after combining each projected vector slice into a single vector φ and measuring the normed distance between a probe sketch and gallery photo.

tire face. The feature sampling points are chosen by setting two parameters: a region (or patch) size s, and a displacement size δ. The region size s defines the size of the square window over which the image feature is computed. The displacement size δ states the number of pixels the patch is displaced for each sample, thus $(s-\delta)$ is the number of overlapping pixels in two adjacent patches. This is analogous to sliding a window of size sxs across the face image in a raster scan fashion. For a HxW image the number of horizontal (N) and vertical (M) sampling locations are given by $N = (W-s)/\delta+1$, and $M = (H-s)/\delta+1$. At each of the $M \cdot N$ patches, we compute the d-dimensional image feature vector ϕ. These image feature vectors are concatenated into one single $(M \cdot N \cdot d)$-dimensional image vector Φ. Whereas $f(I) : I \rightarrow \phi$ denotes the extraction of a single feature descriptor from an image, sampling multiple features using overlapping patches is denoted as $F(I) : I \rightarrow \Phi$. Minimum distance sketch matching can be performed directly using this feature-based representation of subjects i and j by computing the normed vector distance $||F(I^i) - F(I^j)||$ [53].

In our sketch matching framework, two feature descriptors are used: SIFT and LBP. The SIFT feature descriptor quantizes both the spatial location and gradient orientations within a sxs sized image patch, and computes a histogram in which each bin corresponds to a combination of a particular spatial location and orientation. For each image pixel, the histogram bin corresponding to its quantized orientation and location are incremented by the product of: (i) the magnitude of the image gradient at that pixel, and (ii) the value of a Gaussian function centered on the patch with a standard deviation of $s/2$. Tri-linear interpolation is used on the quantized location of the pixel, which addresses image translation noise. The final vector of histogram values is normalized to sum to one. The reader is referred to [84] for a more detailed description of how the SIFT feature descriptor is designed. It is important to reiterate that because we are sampling SIFT feature descriptors from a fixed grid, we do not use SIFT keypoint detection; the SIFT feature descriptor is computed at predetermined

locations.

For the local binary pattern feature descriptor [99], we extended the LBP to describe the face at multiple scales, by combining the LBP descriptors computed with radii $r \in \{1, 3, 5, 7\}$. We refer to this as the multi-scale local binary pattern (MLBP). MLBP is similar to other variants of the LBP, such as MB-LBP [77], but we obtained slightly improved face recognition accuracy using MLBP.

The choice of the MLBP and SIFT feature descriptors was based on reported success in heterogeneous face recognition and through a quantitative evaluation of their ability to discriminate between subjects in sketches and photos [58]. Though variants of LBPs have lead to substantial success in previous heterogeneous face recognition scenarios, the use of SIFT feature descriptors for this application is quite novel. However, recent work [53] clearly demonstrates the success of SIFT feature descriptors for viewed sketch recognition. SIFT feature descriptors have also been shown to perform comparatively with LBP feature descriptors in a standard face recognition scenario [93]. These feature descriptors are well suited for sketch recognition because they describe the distribution of the direction of edges in the face; this information is contained in both sketches and photos. By densely sampling these descriptors, sufficient discriminatory information is retained to more accurately determine a subject's identity over previously used synthesis methods [53].

The feature-based representation requires each sketch and photo image to be normalized by rotating the angle between the two eyes to 0°, scaling the images to a 75 interocular pixel distance, and cropping the image size to 200 by 250 pixels. The experimental results reported in Sections 2.4 and 2.6 for each of the two descriptors are based on a sum of score fusion of the match scores generated from computing descriptors with patch sizes of $s = 16$ and $s = 32$. This also holds for the global discriminant described in Section 2.3.2; we fuse the matching scores computed using two separate patch sizes of 16 and 32. When combining the SIFT and MLBP features,

sum of score fusion is used as well.

2.3.2 Local Feature-based Discriminant Analysis

With both sketches and photos characterized using SIFT and MLBP image descriptors, we further refine this feature space using discriminant analysis. This is done to reduce the large dimensionality of the feature vector Φ. A straightforward approach would be to apply classical subspace analysis (such as LDA) directly on Φ, and to extract discriminant features for classification. However, there are several problems with this approach. First, the feature dimensionality is too high for direct subspace analysis. In our experiments each image is divided into either 154 overlapping patches (for $s = 32$) or 720 overlapping patches (for $s = 16$), with each patch producing a 128-dimensional SIFT descriptor or a 236-dimensional MLBP descriptor. The second problem is the possibility of overfitting due to the small sample size (SSS) problem [115].

In order to handle the combination of a large dimensionality (feature size) and small sample size, an ensemble of linear discriminant classifiers, called *local feature-based discriminant analysis (LFDA)*, is proposed. Other discriminant analysis methods have been proposed to handle the SSS problem, such as random sampling LDA [148], regularized LDA [85], and direct LDA [45]. However, we choose the proposed LFDA method because it is designed to work with a feature descriptor representation (as opposed to an image pixel representation), and it resulted in high recognition accuracy.

In LFDA framework each image feature vector Φ is first divided into "slices" of smaller dimensionality, where slices correspond to the concatenation of feature descriptor vectors from each column of image patches. Next, discriminant analysis is performed separately on each slice by performing the following three steps: PCA, within-class whitening, and between-class discriminant analysis. Finally, PCA is

applied to the new feature vector to remove redundant information among the feature slices to extract the final feature vector.

To train the LFDA, we use a training set consisting of pairs of a corresponding sketch and photo of n subjects (which are the n classes). This results in a total of $2n$ training images with two supports for each subject i: the image feature representation of the sketch $\Phi_s^i = F(I_s^i)$ and the photo $\Phi_p^i = F(I_p^i)$. We combine these feature vectors as a column vector in training matrices and refer to them as $X^s = [\Phi_s^1\ \Phi_s^2 \dots \Phi_s^n]$ for the sketch, $X^p = [\Phi_p^1\ \Phi_p^2 \dots \Phi_p^n]$ for the photo, and $X = [\Phi_s^1 \dots \Phi_s^n\ \Phi_p^1 \dots \Phi_p^n]$ for the photo and sketch combined.

The first step in LFDA is to separate the image feature vector into multiple sub-vectors or slices. Given the MxN array of patches consisting of SIFT or MLBP descriptors, we create one slice for each of the N patch columns. With a d-dimensional feature descriptor, each of the N slices is of dimensionality $(M \cdot d)$. We call this a "slice" because it is similar to slicing an image into N pieces. After separating the feature vectors into slices, the training matrices now become $X_k^s \in \mathbb{R}^{M \cdot d, n}$, $X_k^p \in \mathbb{R}^{M \cdot d, n}$, and $X_k \in \mathbb{R}^{M \cdot d, 2n}$ $(k = 1 \dots N)$, which are all mean centered.

We next reduce the dimensionality of each training slice matrix X_k using the PCA matrix $W_k \in \mathbb{R}^{M \cdot d, r}$ with r eigenvectors. The purpose is to remove the noisy features which are usually associated with the trailing eigenvectors with the smallest eigenvalues. In our experiments we use the 100 eigenvectors with the largest eigenvalues (which preserves about 90% of the variance). The discriminant extraction proceeds by generating the mean projected class vectors

$$Y_k = W_k^{\mathrm{T}}(X_k^s + X_k^p)/2 \tag{2.1}$$

which are used to center the sketch and photo training instances of each class by

Table 2.1: Rank-1 recognition rates for matching viewed sketches using the CUHK public dataset. The standard deviation across the five random splits for each method in the middle and right columns is less than 1%.

Baseline	
Method	*Rank-1 Accuracy (%)*
FaceVACS [1]	90.37
BP Synthesis [149]	96.30
SIFT Descriptor-based [53]	97.87
Without LFDA	
Method	*Rank-1 Accuracy (%)*
SIFT	97.00
MLBP	96.27
SIFT + MLBP	97.33
LFDA	
Method	*Rank-1 Accuracy (%)*
SIFT	99.27
MLBP	98.60
SIFT + MLBP	**99.47**

$$\begin{aligned} \tilde{X}_k^s &= W_k^{\mathrm{T}} X_k^s - Y_k \\ \tilde{X}_k^p &= W_k^{\mathrm{T}} X_k^p - Y_k \end{aligned} \tag{2.2}$$

To reduce the intra-personal variation between the sketch and photo, a whitening transform is performed. Whitening the within-class scatter matrix reduces the dimensionality by discarding features that represent the principal intra-personal variations, which in this case corresponds to intra-personal differences between sketches and photos. To do so, we recombine the training instances into $\tilde{X}_k = [\tilde{X}_k^s \; \tilde{X}_k^p]$. PCA analysis is performed on $\tilde{X}_k$, such that the computed PCA projection matrix $\tilde{V}_k \in \mathbb{R}^{100\times100}$ retains all of the data variance from $\tilde{X}_k$. Let $\Lambda_k \in \mathbb{R}^{100\times100}$ be a diagonal matrix whose entries are the eigenvalues of the corresponding PCA eigenvectors $\tilde{V}_k$. The

whitening transform matrix is $V_k = \left(\Lambda_k^{-\frac{1}{2}} V_k^{\mathrm{T}}\right)^{\mathrm{T}}$.

The final step is to compute a projection matrix that maximizes the intra-person scatter by performing PCA on $V^{\mathrm{T}} Y_k$ (which is the whitening transform of the mean class vectors). Using all but one of the eigenvectors in the PCA projection matrix, the resultant projection matrix is denoted as $U_k \in \mathbb{R}^{100\times 99}$. This results in the final projection matrix for slice k

$$\Psi_k = W_k V_k U_k \tag{2.3}$$

With each local feature-based discriminant trained, we match sketches to photos using the nearest neighbor matching on the concatenated slice vectors. We first separate the feature representation of an image into individual slices

$$\Phi = [\Phi(1)^{\mathrm{T}}\ \Phi(2)^{\mathrm{T}} \ldots \Phi(N)^{\mathrm{T}}\ ]^{\mathrm{T}} \tag{2.4}$$

where $\Phi(i) \in \mathbb{R}^{M\cdot d}$ is the i-th slice feature vector. We then project each slice using the LFDA projection matrix Ψ_k yielding the new vector representation $\varphi \in \mathbb{R}^{M\cdot 99}$

$$\varphi = \left[(\Psi_k^{\mathrm{T}}\, \Phi(1))^{\mathrm{T}}\ (\Psi_k^{\mathrm{T}}\, \Phi(2))^{\mathrm{T}} \ldots (\Psi_k^{\mathrm{T}}\, \Phi(N))^{\mathrm{T}}\ \right]^{\mathrm{T}} \tag{2.5}$$

With the LFDA representation of a sketch φ_s and photo φ_p, the normed distance $||\varphi_s - \varphi_p||$ is used to select the gallery photo with the minimum distance to the probe sketch.

The proposed LFDA algorithm is a simple yet effective method. From the results in Section 2.4, we can clearly see that LFDA is able to significantly improve the recognition performance over the basic feature-based sketch matching framework. Similar to other variants of LDA that are designed to handle the small sample size problem [45] [85] [148], LFDA has several advantages over traditional linear discrimi-

nant analysis (LDA). First, LFDA is more effective in handling large feature vectors. The idea of segregating the feature vectors into slices allows us to work on more manageable sized data with respect to the number of training images. Second, because the subspace dimensionality is fixed by the number of training subjects, when dealing with the smaller sized slices the LFDA algorithm is able to extract a larger number of meaningful features. This is because the dimensionality of each slice subspace is bounded by the same number of subjects as a subspace for the entire feature representation.

2.4 Viewed Sketch Matching Results

In order to compare our proposed LFDA framework to published methods on sketch matching, we evaluated our method using viewed sketches from the CUHK dataset[2] [149]. This dataset consists of 606 corresponding sketch/photo pairs that was drawn from three face datasets: (1) 123 pairs from the AR face database [89], (2) 295 pairs from the XM2VTS database [92], and (3) 188 pairs from the CUHK student database [137]. Each of these sketch images were drawn by an artist while looking at the corresponding photograph of the subject. Two examples of these viewed sketches are shown in Figure 2.1(a). For the methods presented in this chapter, all results shown are the recognition rates averaged over five separate random splits of 306 training subjects and 300 test subjects.

The results of viewed sketch matching experiment are summarized in Table 2.1. The first column of the table shows the baseline methods, which includes the top two performing methods in the literature [53] [149] (each used 306 training subjects and 300 test subjects) and Cognitec's FaceVACS commercial face recognition engine [1]. FaceVACS has been shown [53] to perform at the same level as earlier solutions

[2]The CUHK Face Sketch Database is available for download at:
`http://mmlab.ie.cuhk.edu.hk/facesketch.html`

specifically trained for viewed sketch recognition [138]. In the second column the matching accuracies achieved by directly comparing SIFT and MLBP feature vectors Φ are listed. The method 'SIFT + MLBP' indicates a sum of score fusion [123] of the match scores from SIFT matching and MLBP matching. While both the SIFT and MLBP methods offer similar levels of performance, using LFDA (third column) the accuracy increases to the point where (on average) fewer than two sketches are incorrectly identified out of the 300 sketches in the probe set.

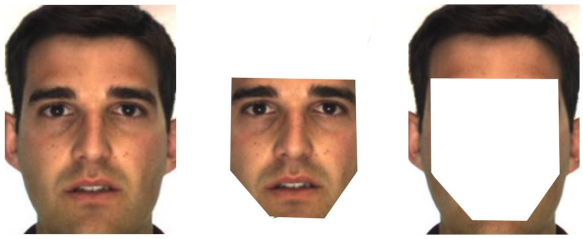
While LFDA was able to reduce the error in half, the use of LDA actually induced higher error. In the same experiment shown in Table 2.1, we applied LDA on the entire feature vector Φ instead of breaking it into slices and performing LDA on each slice vector as is done in LFDA. The accuracy of LDA+SIFT was 95.47%, LDA+MLBP was 91.53%, and (SIFT+MLBP)+LDA was 97.07%. In each case LDA actually lowered the accuracy from the LFDA case. The decrease in accuracy observed when applying the standard LDA is due to the small sample size problem and the resulting curse of dimensionality [115]. Given our large feature representation (for a 32 pixel patch size, the SIFT representation contains 19,712 components and the MLBP representation contains 36,344 components), the subspace projections are over fit to the training data. Because LFDA is an ensemble method, it is better suited to overcome this overfitting problem. Other LDA variants have been shown to handle the small sample size problem as well, such as RSLDA [148] and regularized LDA (R-LDA) [85].

2.5 Matching Forensic Sketches

The available methods for matching forensic sketches to photos is limited. Uhl and Lobo [144] proposed a now antiquated method of matching sketches drawn by forensic artists using photometric standardization and facial features. Yuen and Man [156]

Table 2.2: Demographics of the 159 forensic sketch images and the 10,159 mugshot gallery images.

	Forensic Sketches	Mugshot Gallery
Caucasian	58.49%	46.43%
African American	31.45%	46.93%
Other	10.06%	6.64 %
Male	91.19%	84.33%
Female	8.81%	15.52%
Unknown	0.00%	0.03%



(a) (b) (c)

Figure 2.4: An example of the internal (b) and external (c) features of the face image in (a). Humans tend to use the internal facial features for recognizing faces they are familiar with, and the external features for recognizing faces they are unfamiliar with [155]. Witnesses of a crime are generally unfamiliar with the culprit, therefore the external facial features should be more salient in matching forensic sketches.

matched lab generated forensic composites to photographs based on point distribution models.

2.5.1 Forensic Sketch Database

In our study we used a dataset consisting of 159 forensic sketches, each with a corresponding photograph of the subject that was later identified by the law enforcement agency to belong to the suspect. All of these sketches were drawn by forensic sketch artists working with witnesses who provided verbal descriptions after crimes were committed by an unknown culprit. The corresponding photographs (mugshots) are the result of the subject later being identified, possibly due to citizens coming forward to provide clues. The forensic sketch data set used here comes from four

different sources: (1) 73 images from forensic sketch artist Lois Gibson [31], (2) 43 images from forensic sketch artist Karen Taylor [139], (3) 39 forensic sketches provided by the Michigan State Police Department, and (4) 4 forensic sketches provided by the Pinellas County Sheriff's Office. In addition to these 159 corresponding forensic sketch and photo pairs, we also made use of a dataset of 10,159 mugshot images provided by the Michigan State Police to enlarge the gallery size. Thus, the matching experiments attempt to replicate real world scenarios where a law enforcement agency would query a large gallery of mugshot images with a forensic sketch. Examples of the forensic sketches used in our experiments are shown in Figures 2.1, 2.9, 2.10, and 2.10.

In some cases a witness's memory and hence the description of a suspect is inaccurate. This causes forensic sketches drawn from such witness's descriptions to be of poor quality in terms of not accurately capturing all the facial features of the suspect. For most of these sketches, it is unlikely that they can be successfully matched automatically to the corresponding photos because they barely resemble the subject. For this reason, we separated our forensic sketches into two categories: *good* quality and *poor* quality. This separation was performed subjectively by looking at the corresponding pairs (sketch and photo) and labeling them as "good" if the sketch possessed a reasonable resemblance to the subject in the photo, and labeling them as "poor" if the sketch was grossly inaccurate. We believe this leads to a more accurate portrayal of the performance of proposed automatic sketch to photo matching. Figure 2.1 shows the difference between *good* quality and *poor* quality sketches.

2.5.2 Human Memory and Forensic Sketches

A distinct difference between a viewed sketch and a forensic sketch is that the forensic sketch may have many inaccuracies due to the witness's inability to correctly remember the suspect's face. A body of psychological research exists that focuses on a

person's ability to successfully recall the appearance of an individual she is unfamiliar with and whom she viewed only momentarily [17, 27, 155]. A consistent finding of these studies is that the facial features used to recognize someone depends on the level of familiarity. In this respect, facial features are separated into internal and external features (see Figure 2.4).

When we are familiar with the person we are attempting to recognize (e.g. a co-worker, family member, or celebrity), we predominantly make use of the internal facial features for identification [17, 155]. These features include the nose, eyes, eye brows, and mouth. Most research in automatic face recognition has observed that these internal features are also the most discriminative areas of the face. [67]. When we are attempting to recognize someone who is unfamiliar to us, the external features of the face are predominantly used to establish identity [17, 155]. External features consist of the outer region of the face, including the chin, hairstyle, and general shape of the face.

Frowd et al. [27] studied whether humans are best able to match forensic sketches using the internal or external features of the face. In their experiments, test subjects were shown the photograph of a celebrity they were unfamiliar with and given approximately one minute to remember the appearance. Two days later the subjects worked with a forensic sketch artist to draw a sketch of the person they viewed earlier in the photograph. Using these composites, a separate set of subjects that had familiarity with the same celebrities were asked to identify two different versions of the sketches: (i) sketches with only the interior regions of the face shown, and (ii) sketches with only the exterior regions of the face shown. The experiments concluded that higher identification rates were achieved using the exterior regions of the face [27].

Frowd et al.'s results are based on a controlled experiment, so we must tread lightly in using them for automated face recognition. One of the most important properties for a biometric trait is permanence [40], which the external regions of the

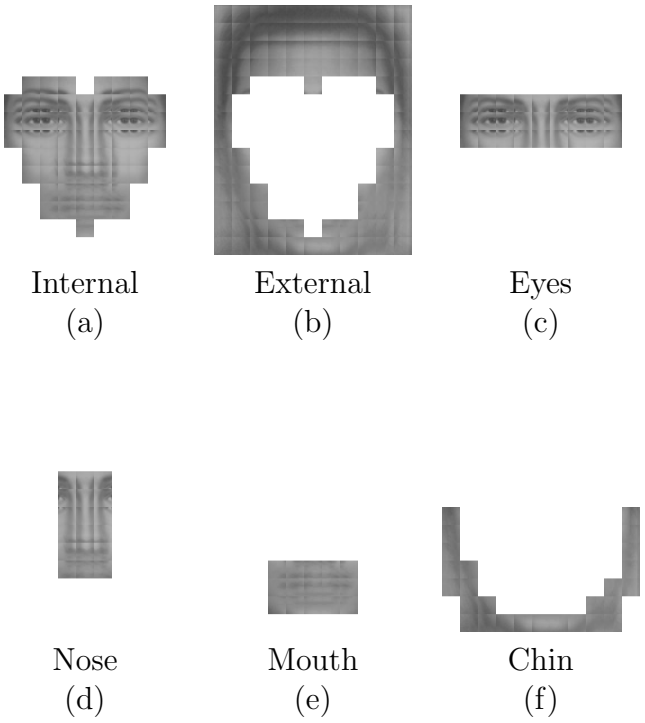


Figure 2.5: Masks used for region based forensic sketch matching. Shown above are the mean photo patches of each patch used for a particular region. The mosaic effect is due to the fact that face patches are extracted in an overlapping manner.

face do not satisfy well. By growing or removing a beard, changing hairstyles, or donning headgear, a person can drastically change the appearance of their external facial features. Therefore, assigning a higher prior probability to the decisions made from external forensic sketch regions over internal regions may not be a wise choice.

2.5.3 Forensic Sketch Region Saliency

Due to the observation in the human cognition studies that different regions of the face have different saliency, we measure the performance of automatic sketch matching using only certain regions of the face. For our feature-based framework, it is quite easy to implement this by only selecting the patches in the face that correspond to a given region. We considered six separate face regions for localized identification: (1) internal, (2) external, (3) eyes, (4) nose, (5) mouth, and (6) chin. Figure 2.5 shows the patches used for each of these face regions (with patch size $s = 32$) and the

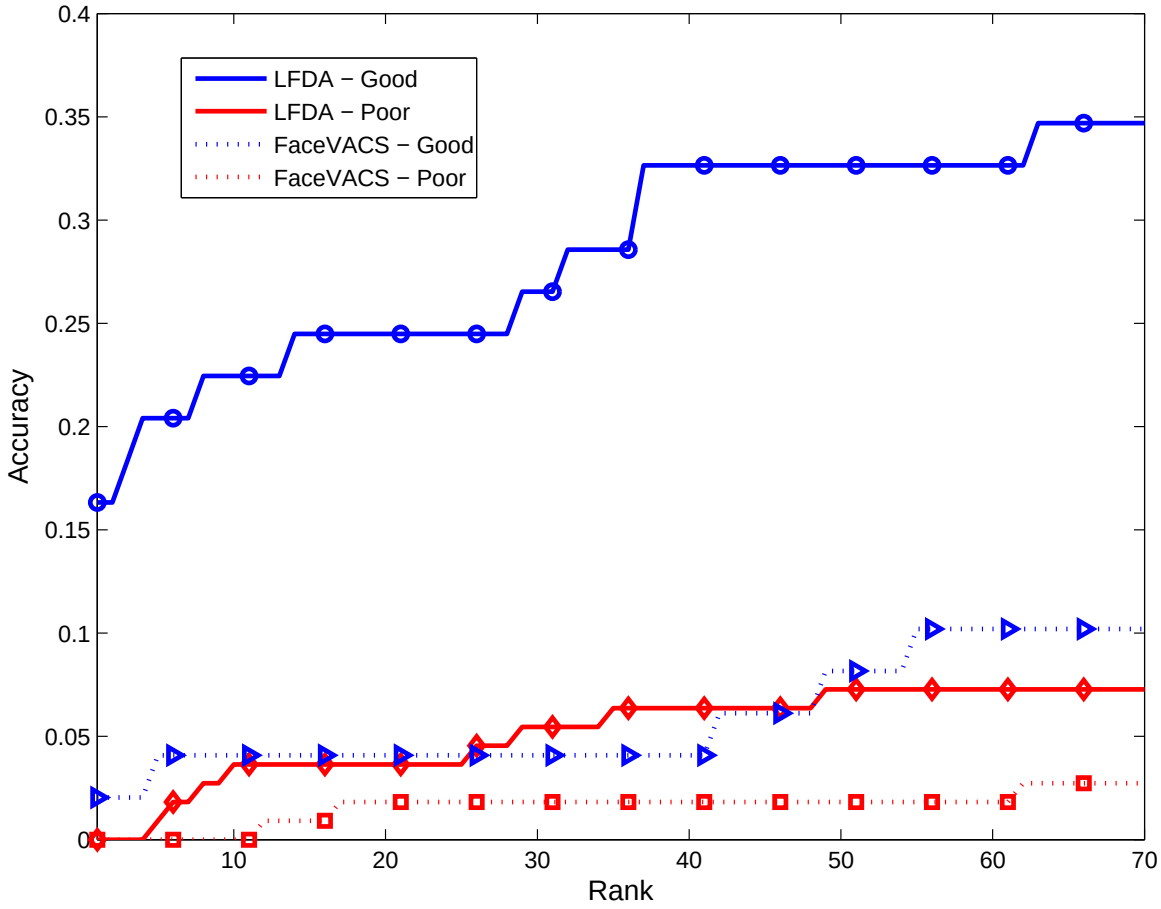


Figure 2.6: Performance of matching forensic sketches that were labeled as *good* (49 sketches) and *poor* (110 sketches) against a gallery of 10,159 mugshot images without using race/gender filtering.

average intensity for each patch (the mean patch). Thus, when matching using one of the masks, we performed distance matching using only the patches shown in each mask.

In Section 2.6 we will show the results of forensic sketch matching using only these face regions.

2.5.4 Large-Scale Forensic Sketch Matching

Matching forensic sketches to large mugshot galleries is different in several respects from traditional face identification scenarios. When presenting face recognition results in normal recognition scenarios, we are generally concerned with exactly identifying

the subject in question in a fully automated manner. For example, when preventing multiple passports from being issued to the same person, human interaction should be limited to only ambiguous cases. This is due to the large volume of requests such a system must process. The same is true for matching arrested criminals against existing mugshot databases to confirm their identity. However, when matching forensic sketches it is not critical for the top retrieval result to be the correct subject, as long as it is in the top R retrieved results, say $R = 50$. This is because the culprit being depicted in a forensic sketch typically has committed a heinous crime (e.g murder, rape, armed robbery) that will receive a large amount of attention from investigators. Instead of accepting or dismissing only the top retrieved photo, law enforcement officers will consider the top R retrieval results as potential suspects. Generally, many of the returned subjects can be immediately eliminated as suspects for various reasons, such as if they are currently incarcerated or deceased. The remaining candidates can each then be investigated for their culpability of committing the crime. This scenario is also true of crimes in which a photograph of a suspect is available. Investigators will consider the top R retrieval results instead of only the highest match. Based on the practice followed in forensics, we would like R to be around 50; that is, we are mainly concerned with whether or not the true subject is within the top 50 retrieved images.

In order improve the accuracy of matching forensic sketches, we utilize ancillary or demographic information provided by the witness, to be used as a soft biometric [42]. For example, suppose the witness reports that the race of the culprit is Caucasian, then we can eliminate all non-Caucasian members of the gallery to not only speed up the matching but also to improve the matching performance. The same is true for gender: if the suspect is reported to be a female then we disregard any male subjects in the gallery. To use this approach, we manually labeled all of the 10,159 mugshot images and all the forensic sketch/photo pairs in our database with race and gender.

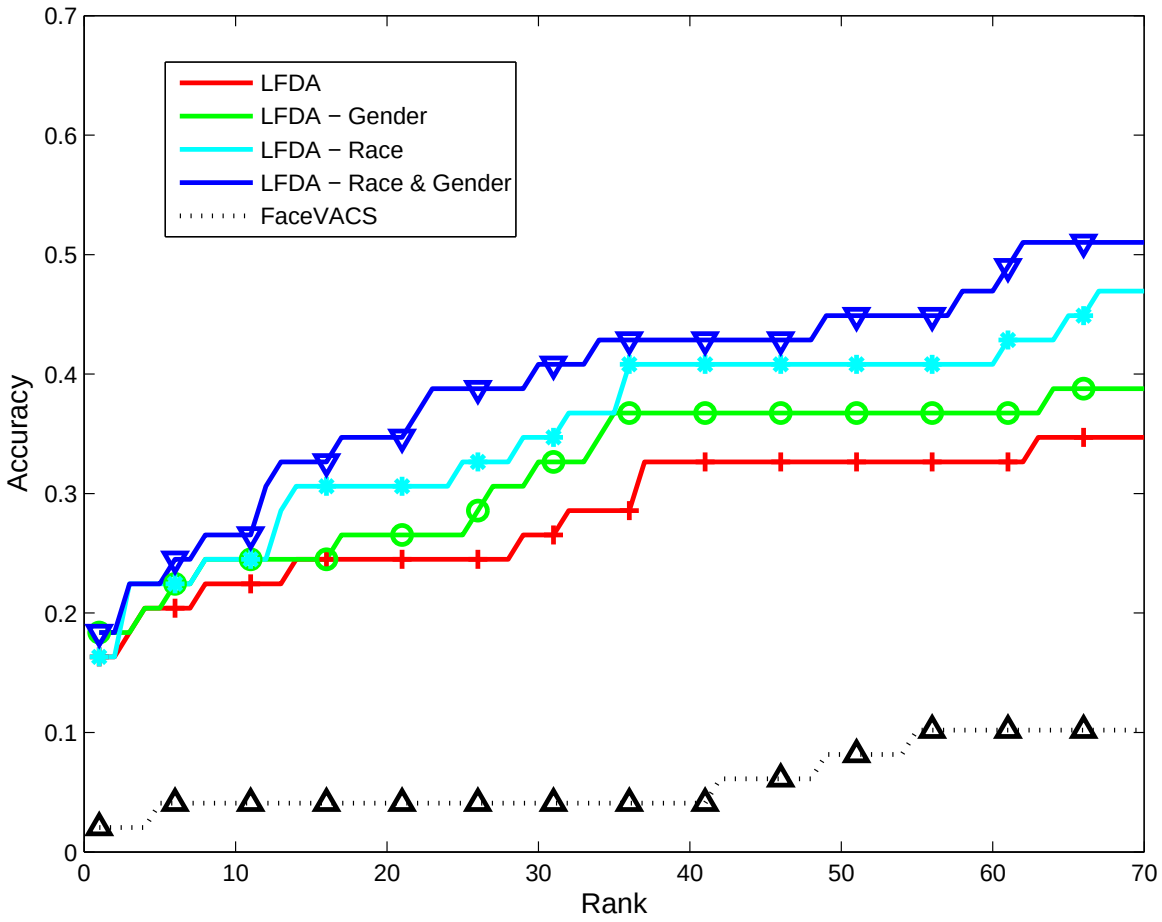


Figure 2.7: Performance of matching *good* sketches with and without using ancillary demographic information (race and gender) to filter the results.

For gender, we considered one of three possible categories: male, female, and (in rare cases) unknown. For race we considered one of three categories: Caucasian, African-American, and "other". The "other" includes individuals who are of Hispanic, Asian, or multiple races. Table 2.2 lists the percentage of members from each race and gender category in the forensic sketches and the mugshot gallery used in our experiments.

We lack additional ancillary information (e.g., age, height, scars, marks and tattoos) that could potentially be used to further improve the matching accuracy.

2.6 Forensic Sketch Matching Results

Forensic sketch recognition performance using the 159 forensic sketch images (probe set) and 10,159 mugshot images (gallery) will now be presented. In these matching experiments we use the *local feature-based discriminant analysis (LFDA)* framework presented in Section 2.3. Our matching uses sum-score fusion of MLBP and SIFT LFDA, as this provided the highest recognition performance for matching viewed sketches (Table 2.1).

The performance of matching sketches classified as *good* and *poor* can be found in Figure 2.6. There is a substantial difference in the matching performance of *good* sketches and *poor* sketches. Despite the fact that *poor* sketches are extremely difficult to match, the CMC plots in Figure 2.6 shows that the proposed method performs roughly the same on the *poor* sketches than a state of the art commercial matcher (FaceVACS) performs on the *good* sketches.

Figure 2.7 shows the recognition performance when race and gender information is used to filter the gallery. By utilizing this ancillary information, we can significantly increase the performance of forensic sketch recognition. We noticed a larger performance gain by using race information than the gender information. This is likely due to the more uniform distribution of race membership than gender membership in our gallery. The use of other demographic information such as age and height should offer further improvements.

Discriminatory information contained in individual face regions (eyes, nose, mouth, etc.) is shown in Figure 2.8. Again, this is achieved by first applying the masks in Figure 2.5 to the face features patches. These results mostly agree with cognitive science research (Section 2.5.2) that indicates that external regions of the face provide more discriminating information in matching forensic sketches. Between the eyes, nose, mouth, and chin, we found the chin to be the most informative region of the face. In fact, only using the chin region for region recognition we were able

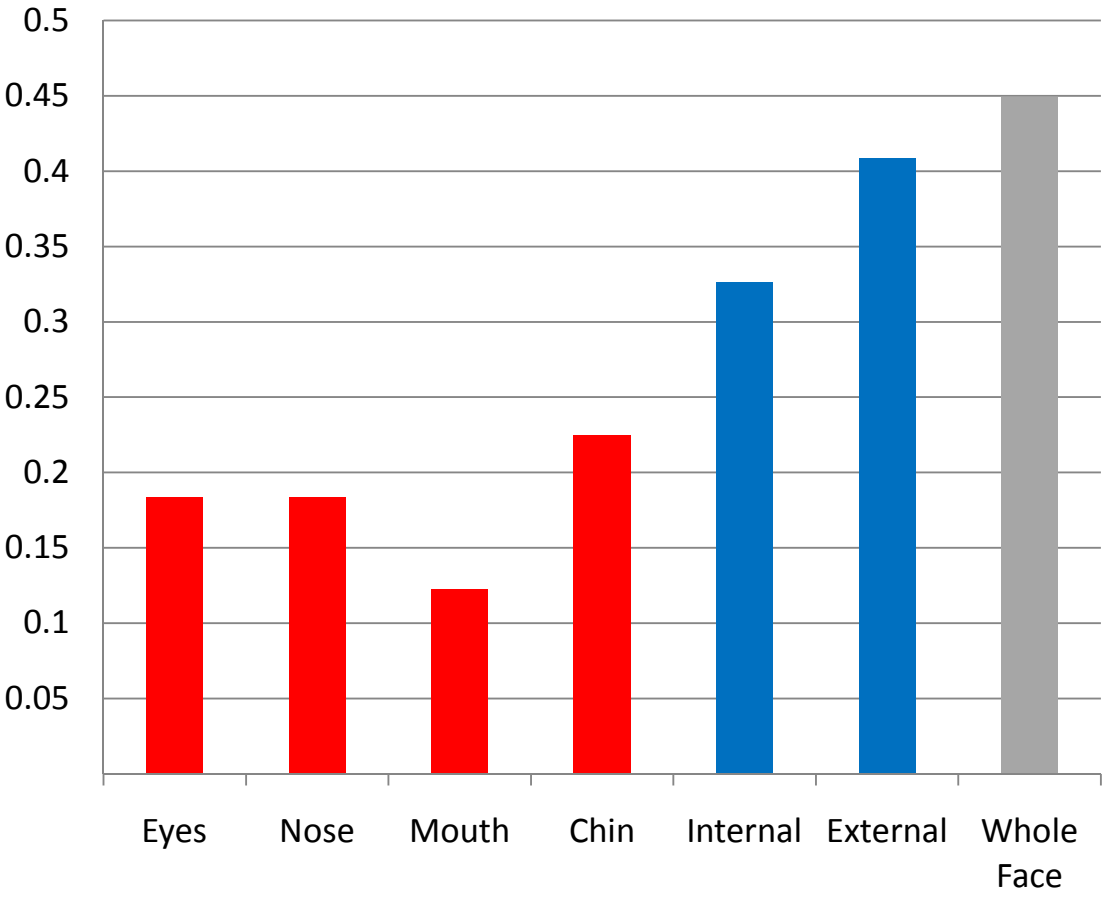


Figure 2.8: Matching performance on the *good* sketches using race/gender filtering with SIFT and MLBP feature-based matching on only specific face regions.

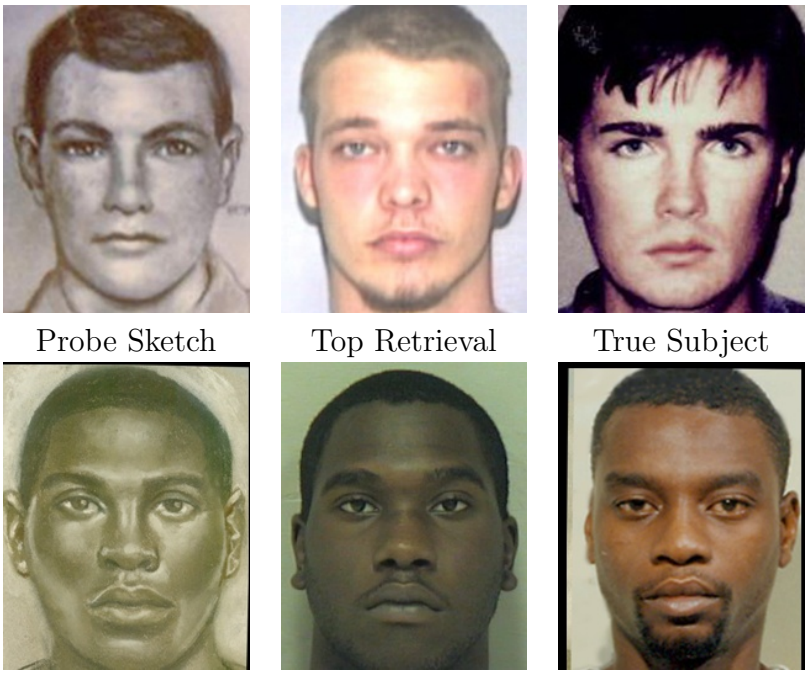


Figure 2.9: Two examples of typical cases in which the true subject photo (third column) was not retrieved at rank 1, but the impostor subject (second column) retrieved at rank 1 visually looks more similar to the sketch (first column) than the true subject.

to achieve Rank-50 accuracy of 22.45% with a gallery size of 10,159 images, which is interesting, given the fact that the chin is not generally regarded as an overly valuable feature in face recognition research.

Examples of failed retrievals are shown in Figure 2.9. While the top retrieved mugshot is not correct in these two examples, the probe sketch appears to be more similar to the top matched photo than the true photograph. This was nearly always the case: *the top retrieved images appeared highly similar to the probe sketch in the incorrect matchings*. This can be explained by the subjective and often incorrect verbal description of the suspect provided by the witness.

Figure 2.10 shows three of the best matches and Figure 2.11 shows three of the worst matches amongst all the *good* sketches using the proposed LFDA recognition method. For each image, we have listed the match rank returned by LFDA and FaceVACS.

One limitation of our study is the small number of forensic sketches in our dataset, but obtaining a large collection of forensics sketches and the mated photographs from law enforcement agencies has not been easy. Not only does a small database limit the evaluation of our method, but it also affects the performance of our local feature-based discriminant analysis. The LFDA needs a reasonably large number of training examples to learn the most discriminative projections. In the case of viewed sketch recognition we used 306 pairs of sketches and photos for training. For the forensic sketches, even if we performed leave-one-out cross validation there would still be only a small number of good quality training samples. For this reason, we trained the discriminant on the viewed sketches when matching forensic sketches. However, we believe that with a larger number of forensic sketches we could more properly train our discriminant and further improve the matching performance. The bottleneck in finding additional forensic sketches for our experiments is in obtaining the photograph mates for the sketches of the suspects who have not yet been identified (cold cases). While forensic sketches exist from numerous crimes, even if there is an eventual identification of the subject, the mated sketch and photo are often not stored together in a central database. We are currently working with various law enforcement agencies to increase our dataset of forensic sketch pairs.

2.7 Summary

We have presented methods and experiments in matching forensic sketches to photographs. Matching forensic sketches is a very difficult problem in heterogeneous face recognition for two main reasons. (1) Forensic sketches are often an incomplete portrayal of the subject's face. (2) We must match across image modalities since the gallery images are photographs and the probe images are sketches.

One of the key contributions of this chapter is using SIFT and MLBP feature

Method	Rank
LFDA	1
FaceVACS	320
Eyes	210
Nose	152
Mouth	1
Chin	33
Internal	3
External	3

Method	Rank
LFDA	1
FaceVACS	299
Eyes	198
Nose	159
Mouth	31
Chin	304
Internal	2
External	31

Method	Rank
LFDA	1
FaceVACS	2131
Eyes	5
Nose	2
Mouth	823
Chin	24
Internal	1
External	6

Figure 2.10: Examples of the three of the best matches using LFDA. Below each example are the rank scores obtained by using the proposed LFDA method, FaceVACS, and component-based matching.

Method	Rank
LFDA	775
FaceVACS	2255
Eyes	166
Nose	298
Mouth	1776
Chin	3508
Internal	101
External	2231

Method	Rank
LFDA	1599
FaceVACS	215
Eyes	3237
Nose	2974
Mouth	1018
Chin	3742
Internal	3012
External	2402

Method	Rank
LFDA	1617
FaceVACS	429
Eyes	1992
Nose	3634
Mouth	3725
Chin	52
Internal	2246
External	692

Figure 2.11: Examples of the three of the worst matches using LFDA. Below each example are the rank scores obtained by using the proposed LFDA method, FaceVACS, and component-based matching.

descriptors to represent both sketches and photos. We have improved the accuracy of this representation by applying an ensemble of discriminant classifiers, and termed this framework *local feature-discriminant analysis (LFDA)*. The LFDA feature-based representation of sketches and photos was clearly shown to perform better on a public domain viewed sketch data set than previously published approaches.

Another major contribution of the chapter is the large-scale experiment on matching forensic sketches. While previous research efforts have focused on viewed sketches, most real world problems only involve matching forensic sketches. Using a collection of 159 forensic sketches, we matched the sketches against a gallery populated with 10,159 mugshot images. Further improvements to the LFDA method were achieved by utilizing ancillary information such as race and gender to filter the 10,159 member gallery. For a fair evaluation of our methods, we used a state-of-the-art face recognition system, FaceVACS [1].

Together, these improvements in forensic sketch recognition advance the state of the art and demonstrate the utility of heterogeneous face recognition which is the focus of this dissertation. In developing a sketch recognition algorithm with substantially improved recognition accuracy, we offer a tool that is critical for assisting law enforcement agencies in apprehending suspects.

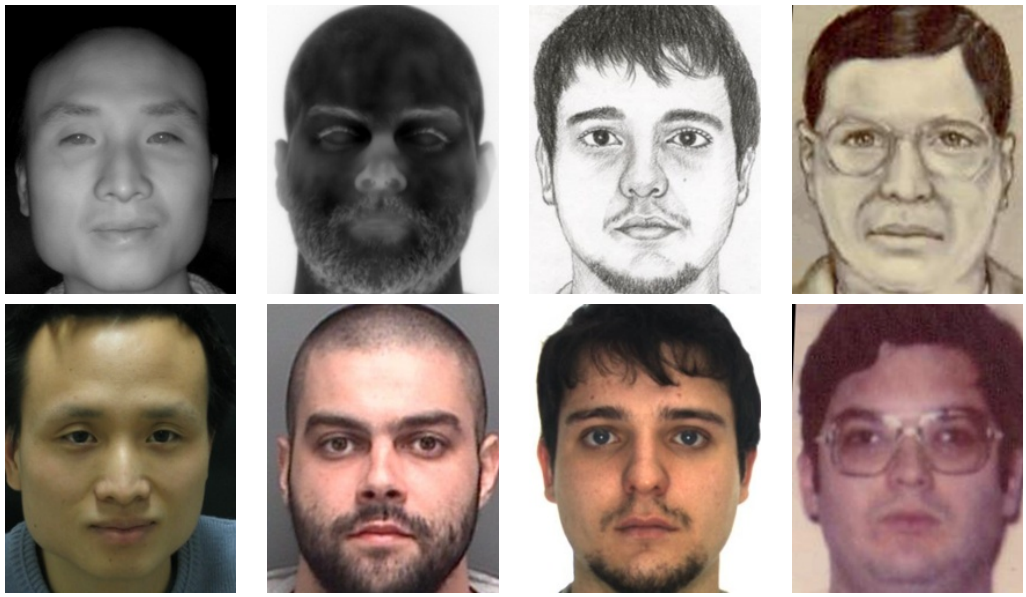
Chapter 3

Heterogenous Face Recognition using Kernel Prototype Similarities

3.1 Introduction

In the previous chapter we discussed a solution to forensic sketch recognition. The solution provided was specific to forensic sketch recognition, and will generally not extend to other face recognition scenarios. The heterogeneous face recognition algorithm presented in this chapter is not built for any specific HFR scenario. Instead, it is designed to generalize to any HFR scenario. Further, this framework can be used for homogeneous face recognition (e.g. visible to visible face recognition) as well. Such a framework offers a strong contribution to the proposed thesis of this dissertation by providing an improvement to the problem of heterogeneous face recognition as a whole.

Again, the motivation behind heterogeneous face recognition is that circumstances exist in which face image to be identified is available only in a particular modality. For example, when a subject's face can only be acquired in nighttime environments, the use of infrared imaging may be the only modality for acquiring a useful face



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

Figure 3.1: Examples images from each of the four heterogenous face recognition scenarios tested in our study, as also shown in Chapter 1. The top row contains probe images from (a) near-infrared, (b) thermal infrared, (c) viewed sketch, and (d) forensic sketch modalities. The bottom row contains the corresponding gallery photograph (visible band face image, called VIS) of the same subject.

image of the subject. Another example is situations in which no imaging system was available to capture the face image of a suspect, as addressed in Chapter 2. In this case a forensic sketch, drawn by a police artist based on a verbal description provided by a witness or the victim, is likely to be the only available source of a face image. Despite continued progress in the accuracy of face recognition systems [110], most commercial off the shelf (COTS) face recognition systems (FRS) are not designed to handle HFR scenarios. The need for face recognition systems specifically designed for the task of matching heterogeneous face images is of substantial interest.

This chapter proposes a unified approach to heterogeneous face recognition that (i) achieves high accuracy on multiple HFR scenarios, (ii) does not necessitate feature descriptors that are invariant to changes in image modality, (iii) facilitates recognition using different feature descriptors in the probe and gallery modalities, and (iv) naturally extends to additional HFR scenarios due to properties (ii) and (iii) above.

3.2 Related Work

3.2.1 Heterogeneous Face Recognition

A flurry of research has emerged providing solutions to various heterogeneous face recognition problems. This began with sketch recognition using viewed sketches, and has continued into other modalities such as near-infrared (NIR) and forensic sketches. In this section we will highlight a representative selection of studies in heterogeneous face recognition as well as studies that use kernel based approaches for classification.

Tang et al. spearheaded the work in heterogeneous face recognition with several approaches to synthesize a sketch from a photograph (or vice-versa) [82, 138, 149]. Tang and Wang initially proposed an eigen-tranformation method [138]. Later, Liu et al. performed the transformation using local linear embedding to estimate the corresponding photo patch from a sketch patch [82]. Wang and Tang proposed a Markov random field model for converting a sketch into a photograph [149]. Other synthesis methods have been proposed as well [29, 157]. A key advantage of synthesis methods is that once a sketch has been converted to a photograph, matching can be performed using existing face recognition algorithms. The proposed prototype framework is similar in spirit to these methods in that no direct comparison between face images in the probe and gallery modalities is needed. The generative transformation-based approaches have generally been surpassed by discriminative feature-based approaches.

A number of discriminative feature-based approaches to HFR have been proposed [12, 52, 59, 77], which have shown good matching accuracies in both the sketch and NIR domains. These approaches first represent face images using local feature descriptors, such as variants of local binary patterns (LBP) [99] and SIFT descriptors [84]. Liao et al. first used this approach on NIR to VIS face recognition by processing face images with a difference of Gaussian filter, and encoding them using multi-block local binary patterns (MB-LBP). Gentle AdaBoost feature selection was

used in conjunction with R-LDA to improve the recognition accuracy. Klare and Jain followed this work on NIR to VIS face recognition by also incorporating SIFT feature descriptors and an RS-LDA scheme [52]. Bhatt et al. introduced an extended uniform circular local binary pattern to the viewed sketch recognition scenario [12]. Klare et al. encoded both viewed sketches and forensic sketches using SIFT and MLBP feature descriptors, and performed local feature-based discriminant analysis (LFDA) to improve the recognition accuracy [59]. Yi et al. [154] offered a local patch-based method to perform HFR on partial NIR face images.

The synthesis method by Li et al. is the only known method to perform recognition between thermal IR and visible face images [71]. The only method to perform recognition between forensic sketches and visible face images is Klare et al. [59], which is also one of two methods, to our knowledge, that has been tested on two different HFR scenarios (viewed sketch and forensic sketch). The other method is Lin and Tang's [78] common discriminant recognition framework which was applied to viewed sketches and near-infrared images. In this work the proposed prototype random subspace framework is tested on four different HFR scenarios.

3.2.2 Kernel Prototype Representation

The core of the proposed approach involves using a relational feature representation for face images (illustrated in Figure 3.2). By using kernel similarites between a novel face pattern and a set of prototypes, we are able to exploit the kernel trick [9], which allows us to generate a high dimensional, non-linear representation of a face image using compact feature vectors.

The benefit of a prototype-based approach is provided by Balcan et al. [9]. Given access to the data distribution and a kernel similarity function, a prototype representation is shown to approximately maintain the desired properties of the high dimensional kernel space in a more efficient representation by using the kernel trick.

While it is not common to refer to kernel methods as prototype representations, in this work we emphasize the fact that kernel methods use a training set of images (which serve as prototypes) to implicitly estimate the distribution of the patterns in a non-linear feature space. One key to our framework is that each prototype has one pattern for each image modality.

The proposed kernel prototype approach is similar to the object recognition method of Quattoni et al. [112]. Kernel PCA [48] and Kernel LDA [46,81] approaches to face recognition have used a similar approach, where a face is represented as the kernel similarity to a collection of prototype images in a high dimensional space. These differ from the proposed method because only a single prototype is used per training subject, and our approach is designed for heterogeneous face recognition. Our earlier work [53] utilized a similar approach but did not exploit the benefit of non-linear kernels, but, like the proposed method, it used a separate pattern from each image modality (sketch and photo) for each prototype.

3.2.3 Proposed Method

The proposed method presents a new approach to heterogeneous face recognition, and extends existing methods in face recognition. The use of a kernel similarity representation is well suited for the HFR problem because a set of training subjects with an image from each modality can be used as the prototypes, and, depending on the modality of a new image (probe or gallery), the image from each prototype subject can be selected from the corresponding modality. Unlike previous feature-based methods, where an image descriptor invariant to changes between the two HFR modalities was needed, the proposed framework only needs descriptors that are effective within each domain. Further, the proposed method is effective even when different feature descriptors are used in the probe and gallery domains. The proposed prototype framework is described in detail in Section 3.4.

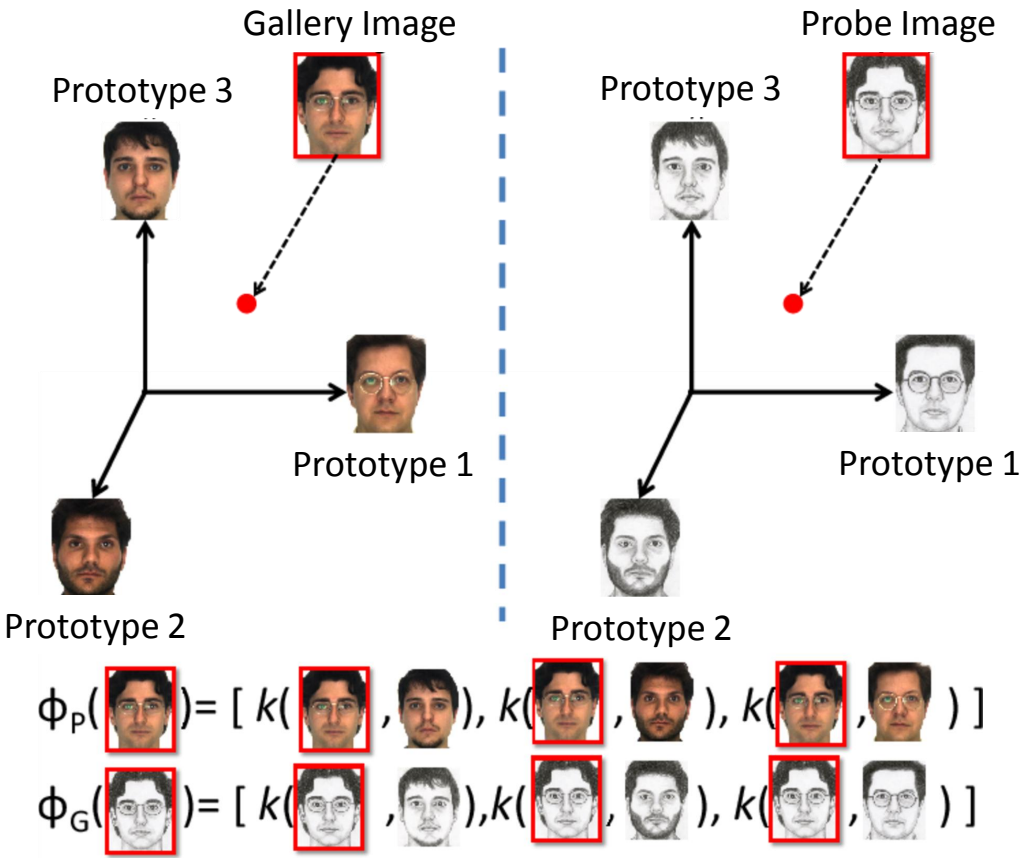


Figure 3.2: The proposed face recognition method describes a face as a vector of kernel similarities to a set of prototypes. Each prototype has one image in the probe and gallery modalities.

The accuracy of the HFR system is improved using a random subspace framework in conjunction with linear discriminant analysis, as described in Section 3.5. The previous method of feature-based random subspaces [52] is revisited in Section 3.6. Experimental results on four different heterogeneous face recognition scenarios (thermal, near-infrared, viewed sketch, and forensic sketch) are provided in Section 3.7, and all the results are benchmarked with a commercial face matcher.

We demonstrate the strength of the propsed framework on many different HFR scenarios, however the parameters controlling the framework are the same across all tested scenarios. This is due to the fact that the contribution of this work is a generic framework for improving solutions to the general HFR problem. Future use of the proposed framework will benefit from tuning the parameters to a specific scenario.

3.3 Preprocessing and Representation

All face images are initially represented using a feature-based representation. The use of local feature descriptors has been argued to closely resemble the postulated representation of the human visual processing system [122], and they have been shown to be well suited for face recognition [56].

3.3.1 Geometric Normalization

The first step in representing face images using feature descriptors is to geometrically normalize the face images with respect to the location of the eyes. This step reduces the effect of scale, rotation, and translation variations. The eye locations for the face images from all modalities are automatically estimated using Cognitec's FaceVACS SDK [1]. The only exceptions are the thermal face images where the eyes are manually located for both the proposed method and the FaceVACS baseline. For thermal images, the available eye detectors do not work.

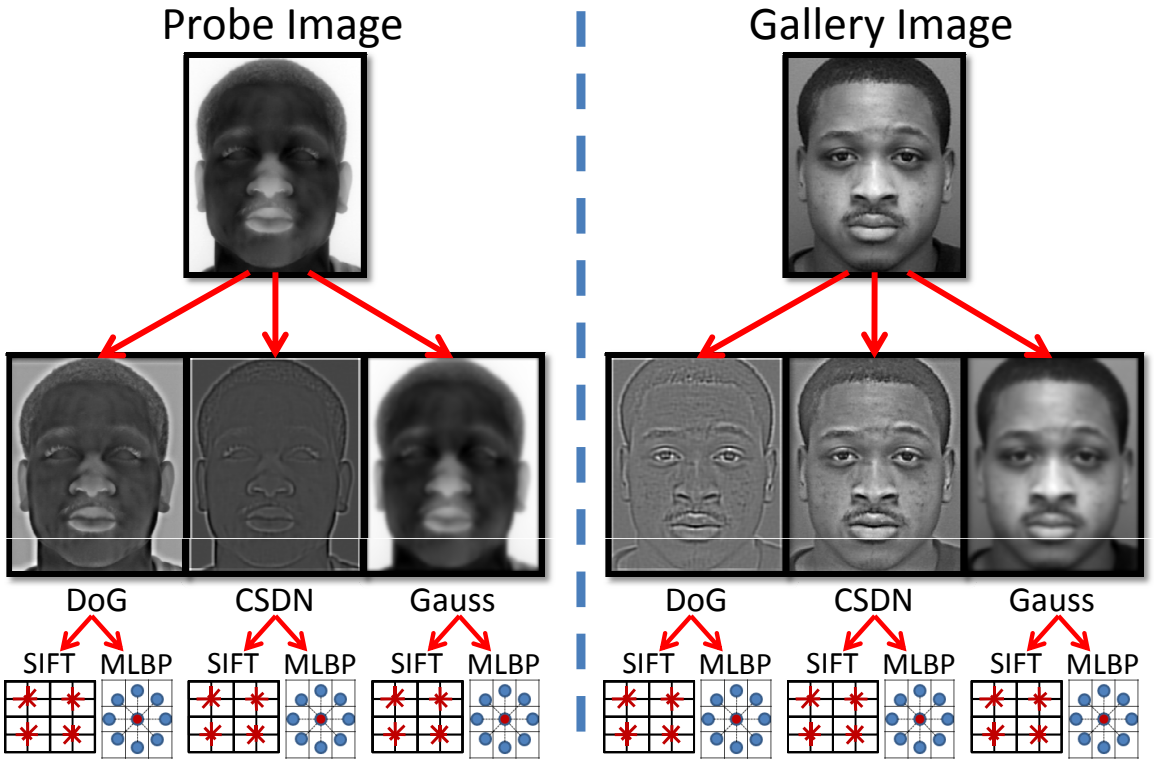


Figure 3.3: Example of thermal probe and visible gallery images after being filtered by a difference of Gaussian, center surround divisive normalization, and Gaussian image filters. The SIFT and MLBP feature descriptors are extracted from the filtered images, and kernel similarities are computed within this image descriptor representation.

Face images are geometrically normalized by (i) performing planar rotation to set the angle between the eyes to 0 degrees, (ii) scaling the images so that the distance between the two pupils is 75 pixels, and (iii) cropping the images to a height of 250 pixels and a width of 200 pixels, with the eyes horizontally centered and vertically placed at row 115.

3.3.2 Image Filtering

Face images are filtered with three different image filters. These filters are intended to help compensate for both intensity variations within an image domain (such as non-uniform illumination changes), as well appearance variations between image domains. The second aspect is of particular importance for the direct random subspace

framework (Section 3.6). An example of the effects of each image filter can be seen in Figure 3.3.

The three image filters used are:

1. *Difference of Gaussian* A difference of Gaussian (DoG) image filter has been shown by Tan and Triggs to improve face recognition performance in the presence of varying illumination [136], as well as in an NIR to VIS matching scenario by Liao et al. [77]. A difference of Gaussian image is generated by convolving an image with a filter obtained by subtracting a Gaussian filter of width σ_1 from a Gaussian filter of width σ_2 ($\sigma_2 > \sigma_1$). In this chapter, $\sigma_1 = 2$ and $\sigma_2 = 4$.

2. *Center-Surround Divisive Normalization* Meyers and Wolf [93] introduced the center-surround divisive normalization (CSDN) filter in conjunction with their biologically inspired face recognition framework. The CSDN filter divides the value of each pixel by the mean pixel value in the sxs neighborhood surrounding the pixel. The non-linear nature of the CSDN filter is seen as a compliment to the DoG filter. In our implementation $s = 16$.

3. *Gaussian* The Gaussian smoothing filter has long been used in image processing applications to remove noise contained in high spatial frequencies while retaining the remainder of the signal. The width of the filter used in our implementation is $\sigma = 2$.

3.3.3 Local Descriptor Representation

Once an image is geometrically normalized and filtered using one of the three filters, local feature descriptors are extracted from uniformly distributed patches across the face. In this work we use two different feature descriptors to represent the face image: the SIFT descriptor [84] and Local Binary Patterns (LBP) [99]. The SIFT feature descriptor has been used effectively in face recognition [56], sketch to VIS

matching [59], and NIR to VIS matching [52]. LBP features have a longer history of successful use in face recognition. Ahonen et al. originally proposed their use for face recognition [3], Li et al. demonstrated their use in NIR to NIR face matching [72], and they have also been successfully applied to several HFR scenarios [12, 52, 59, 77].

The SIFT and LBP feature representations are effective in describing face images due to their ability to encode the structure of the face and their stability in the presence of minor external variations [56]. Each feature descriptor describes an image patch as a d-dimensional vector that is normalized to sum to one. The face image is divided into a set of N overlapping patches of size 32x32. Each patch overlaps its vertical and horizontal neighbors by 16 pixels. With a face image of size 200x250, this results in a total of 154 total patches.

Multi-scale local binary patterns (MLBP) [59], a variant of the LBP descriptor, is used in place of LBP in this work. MLBP is the concatenation of LBP feature descriptors with radii $r = \{1, 3, 5, 7\}$.

Let I be a (normalized and filtered) face image. Let $f_{F,D}(I, a)$ denote the local feature descriptor extracted from image I at patch a, $1 \leq a \leq N$ using image filter F and feature descriptor D. The DoG, CSDN, and Gaussian image filters are, respectively, referred to as F_d, F_c, and F_g. The MLBP and SIFT descriptors are, respectively, referred to as D_m and D_s. The SIFT descriptor yields a 128-dimensional feature descriptor, $f_{F,D_s}(I, a) \in \mathbb{R}^{128}$. The LBP descriptor yields a 59-dimensional feature descriptor, resulting in a 236-dimensional MLBP feature descriptor ($f_{F,D_m}(I, a) \in \mathbb{R}^{236}$). Finally, we have

$$f_{F,D}(I) = \left[f_{F,D}(I, 1)^{\mathrm{T}}, \ \ldots, \ f_{F,D}(I, N)^{\mathrm{T}} \right]^{\mathrm{T}} \tag{3.1}$$

which is the concatenation of all N feature descriptors. Thus, $f_{F,D_s}(I) \in \mathbb{R}^{128 \cdot N}$ and $f_{F,D_m}(I) \in \mathbb{R}^{236 \cdot N}$.

Using the three filters and two descriptors, we have six different representations

available for face image I, namely $f_{F_d,D_m}(I)$, $f_{F_c,D_m}(I)$, $f_{F_g,D_m}(I)$, $f_{F_d,D_s}(I)$, $f_{F_c,D_s}(I)$, and $f_{F_g,D_s}(I)$.

3.4 Heterogeneous Prototype Framework

The heterogeneous prototype framework begins with images from the probe and gallery modalities represented by (possibly different) feature descriptors for each of the N image patches, as described in the previous section. For compactness, let $f(I)$ represent $f_{F,D}(I)$. The similarity between two images is measured using a kernel function $k : f(I) \times f(I) \rightarrow \mathbb{R}$.

Let $\mathcal{T}_1$ be a set of training images consisting of n_{t_1} subjects. The training set contains a probe image P_i and gallery image G_i for each of the n_{t_1} subjects. That is

$$\mathcal{T}_1 = \{P_1, G_1, \ldots, P_{n_{t_1}}, G_{n_{t_1}}\} \tag{3.2}$$

For both the probe and gallery modalities, two positive semi-definite kernel matrices K^P and K^G are computed between the training subjects. The probe kernel matrix is $K^P \in \mathbb{R}^{n_{t_1}, n_{t_1}}$, and the gallery kernel matrix is $K^G \in \mathbb{R}^{n_{t_1}, n_{t_1}}$. The entry in the i-th row and j-th column of K^P and K^G are

$$K^P(i,j) = k(\ f(P_i)\ ,\ f(P_j)\ ) \tag{3.3}$$

$$K^G(i,j) = k(\ f(G_i)\ ,\ f(G_j)\ ) \tag{3.4}$$

where $k(\ \cdot\ ,\ \cdot\ )$ is the kernel similarity function. All the experiments in this chapter use the third degree polynomial kernel $k(f(P_i), f(G_i)) = (f(P_i)^{\mathrm{T}} \cdot f(G_i))^3$, which was empirically choosen over a radial basis function kernel and a second degree polynomial kernel. Again, a generic framework is being presented, and parameter choices such

as the kernel function should be optimized when this framework is engineered into a solution for a specific problem.

Let P and G, respectively, be new probe and gallery face images, i.e. $(P, G \notin \mathcal{T}_1)$. The function $\phi'(P)$ returns a vector containing the kernel similarity of image P to each image P_i in $\mathcal{T}_1$. For gallery image G, $\phi'(G)$ returns a vector of kernel similarities to the gallery prototypes G_i. Thus, face images are represented as the relational vector $\phi'(P) \in \mathbb{R}^{n_{t1}}$ for a probe image, and $\phi'(G) \in \mathbb{R}^{n_{t1}}$ for a gallery image. More precisely, we have

$$\phi'(P) = \Big[\ k(f(P), f(P_1)),\ \ldots,\ k(f(P),\ f(P_{n_{t_1}}))\Big]^{\mathrm{T}} \tag{3.5}$$

$$\phi'(G) = \Big[\ k(f(G), f(G_1)),\ \ldots\ , k(f(G), f(G_{n_{t_1}}))\Big]^{\mathrm{T}} \tag{3.6}$$

Because the feature vectors $\phi'(P)$ and $\phi'(G)$ are a measure of the similarity between the image and the prototype training images, the feature spaces for similarity computation do not have to be the same for the probe and gallery modalities. For example, the probe images could be represented using $F_{F,D_s}(P)$ and the gallery images could be represented using $F_{F,D_m}(G)$. Despite the fact that the SIFT and MLBP feature descriptors are heterogeneous features, the relational representation allows them to be represented in a common feature space. This is based on the assumption that

$$k(f(P), f(P_i)) \ \approx\ k(f(G), f(G_i)) \tag{3.7}$$

In practice we find that Eq. (3.7) does not precisely hold. To compensate for this, we introduce a method called the "R" transform to better align the probe and gallery modalities. The "R" transform uses a matrix R to align the probe prototype feature space with the gallery prototype feature space by:

$$R = K^G((K^P)^{\mathrm{T}}K^P)^{-1}(K^P)^{\mathrm{T}} \tag{3.8}$$

We prove in Appendix A that the R transform is, in fact, a special case of Tang and Wang's eigen-transformation method [138]. Thus, while this transformation was originally applied to synthesize the appearance of a sketch in the photo domain [138], we improve this linear transformation method by incorporating (i) a non-linear feature space (i.e., the kernel prototype similarities), and (ii) a feature descriptor based representation (i.e., the LBP or SIFT representation used to measure the kernel similarities). We do not call the R transform an eigen-transformation because our special case allows for a simpler solution that does not make use of an eigen-decomposition. The importance of this R transform step is experimentally demonstrated in Section 3.7. It is important to note that the scale (or distribution) of K^P and K^G will already be similar because the σ parameter in the RBF kernel is tuned for each modality. Any extreme input values to the system (e.g. a non-face image), will cause the kernel similarity to degenerate to 0, and thus allowing the system to remain stable with respect to scale.

The strength of the R transformation lies in its ability to leverage the constraint that the prototype representation will be n_{t_1}-dimensional and the number of training samples at this phase being n_{t_1}. This allows the R transformation to exactly align the probe prototype feature space to the gallery feature space (with respect to the training set). While this would cause concern that the solution is too tightly fit to the training data, the extension of random sampling provided below alleviates concerns of this being the case.

The benefit of the R transformation is demonstrated quantitatively in the experimental results. Qualitatively, the R transformation is seen as a method to handle additional heterogeneous properties remaining in the new prototype similarity vectors. Despite the fact that $\phi'(\cdot)$ offers a common representation for both modalities,

issues such as the similarities in each modality having different scales (e.g. from the use of differed descriptors in the probe and gallery modalities) are addressed by the R transformation. Using R, we now introduce the final prototype based representation $\phi(\cdot)$ as

$$\phi(P) = R \cdot \phi'(P) \tag{3.9}$$

$$\phi(G) = \phi'(G) \tag{3.10}$$

We alter the tersely presented notation to $\phi_{F,D}(I)$ to specify which feature descriptor and image filter are initially being used to represent the image I. For example, $\phi_{F_C,D_S}(I)$ denotes the prototype similarity of image I when represented using the CSDN image filter and SIFT descriptors.

3.4.1 Discriminant Analysis

A second training set is used to enhance the discriminative capabilities of the prototype representation. This independent training set $\mathcal{T}_2$ consists of probe and gallery images of n_{t_2} subjects such that $\forall\{P'_i, G'_i\} \in \mathcal{T}_2,\ \{P'_i, G'_i\} \notin \mathcal{T}_1$.

A linear subspace of the prototype representation $\phi(\cdot)$ is learned using linear discriminant analysis (LDA) [10] on the images in $\mathcal{T}_2$. LDA (and its variants) has consistently demonstrated its ability to improve the recognition accuracy of various algorithms. The benefits of LDA in the context of face recognition have been demonstrated on image pixel representations [10,147], global Gabor features [75], and image descriptors [59,77].

We learn the linear projection matrix W by following the conventional approach for high dimensional data, namely by first applying PCA, followed by LDA [10]. In all experiments the PCA step was used to retain 99.0% of the data variance. Let X

be a matrix whose columns contain the prototype representation of each image in $\mathcal{T}_2$,

$$X = \left[\phi(P'_1),\ \phi(G'_1),\ \ldots,\ \phi(P'_{n_{t_2}}),\ \phi(G'_{n_{t_2}})\right] \tag{3.11}$$

Let X' denote the mean-centered version of X. The initial step involves learning the subspace projection matrix W'_1 by performing principal component analysis (PCA) on X' to reduce the dimensionality of the feature space. Next, the within-class and between-class scatter matrices of ${W'_1}^{\mathrm{T}} \cdot X'$, respectively, S_W and S_B, are computed. The dimension of the subspace W'_1 is such that S_W will be of full rank. The scatter matrices are built using each subject as a class, where one image each from the probe and gallery modality represents each class. Lastly, the matrix W'_2 is learned by solving the generalized eigenvalue problem

$$S_B \cdot W'_2 = \Lambda \cdot S_W \cdot W'_2 \tag{3.12}$$

This yields the LDA projection matrix W, where

$$W = \left({W'_2}^{\mathrm{T}} \cdot {W'_1}^{\mathrm{T}}\right)^{\mathrm{T}} \tag{3.13}$$

Letting μ denote the mean of X, the final representation for an unseen probe or gallery image I using the prototype framework is $W^{\mathrm{T}} \cdot (\phi(I) - \mu)$. Subsequent uses of W in this chapter will assume the appropriate removal of the mean μ from $\phi(I)$ for terseness.

3.5 Random Subspaces

3.5.1 Motivation

The proposed heterogeneous prototype framework uses training data to (i) define the prototypes, (ii) learn the prototype transformation matrix R, and (iii) learn the linear subspace projection matrix W.

The reliance on training data raises two (somewhat exclusive) issues in the prototype representation framework. The first issue is that the number of subjects in $\mathcal{T}_1$ (i.e. the number of prototypes) is generally too small for an expressive prototype representation. Balcan et al. demonstrated that the number of prototypes does not need to be large (with respect to the margin) to approximately replicate the data distribution [9]. However, their applications primarily dealt with binary classification and a small number of features. When applying a prototype representation to face recognition, a large number of classes (or subjects) and features are present. The small sample size problem implies that the number of prototypes needed to approximate the underlying data distribution should be large [115].

The second issue is also related to the small sample size problem [115]. This common problem in face recognition arises from too few training subjects to learn model parameters that are not susceptible to generalization errors. In the heterogeneous prototype framework this involves learning the R and W matrices that generalize well.

A number of solutions exist to the small sample size problem in face recognition. Most are designed to handles deficiencies in the subspace W, such as dual-space LDA [147], and direct LDA [45]. Regularization methods such as R-LDA [85] also address degenerative properties of W, and could potentially be extended to the learned matrix R as well. However, these methods do not address the issue of too few prototypes for an expressive representation.

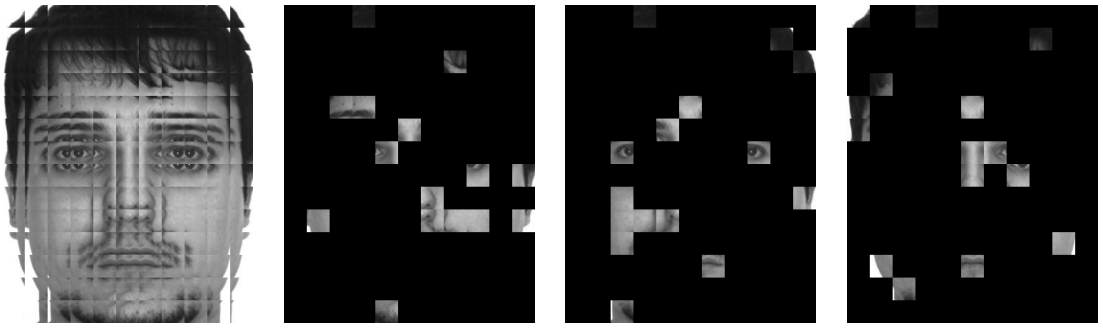
Another approach to handle deficiencies in learning parameters is the use of random subspaces [37]. The random subspace method samples a subset of features and performs training in this reduced feature space. Multiple sets (or bags) of randomly sampled features are generated, and for each bag the parameters are learned. This approach is similar to the classical bagging classification scheme [15], where the training instances are randomly sampled into bags multiple times and training occurs on each bag separately. Ensemble methods such as Ho's random subspaces [37] and Breiman's bagging classifiers have been demonstrated to increase the generalization ability of an arbitrary classifier [125].

Wang and Tang demonstrated the effectiveness of random sampling LDA (RS-LDA) for face recognition. Their approach combined random subspaces and bagging by sampling both features and training instances. For each random sample space, a linear subspace was learned. Klare and Jain utilized this approach in the HFR scenario of NIR/VIS by using multiple subset samples of face patches described by local feature descriptors [52].

We consider random sampling ideal for the prototype recognition framework because it is able to satisfactorily address the two limitations: (i) the number of prototypes is multiplied by the number of bags, which improves the expressiveness of the prototype representation, and (ii) the use of an ensemble method improves deficiencies in the W and R matrices. Further unification of these two separate problems into a single solution offers a simpler framework.

3.5.2 Prototype Random Subspaces

The prototype random subspace (P-RS) framework uses B different bags (or samples) of the N face patches. Each sample consists of $\alpha \cdot N$ patches, $0 \leq \alpha \leq 1$. For bag $b, b = 1 \ldots B$, we have the integer vector $\kappa_b \in \mathbb{Z}^{\alpha \cdot N}$, where each component of κ_b is a unique randomly sampled value from $1 \ldots N$. It is assumed that α is selected such



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

Figure 3.4: The process of randomly sampling image patches is illustrated. (a) All image patches. (b), (c), (d) Bags of randomly sampled patches. The kernel similarity between SIFT and MLBP descriptors at each patch of an input image and the prototypes of corresponding modality are computed for each bag. Images are from [89]

that $\alpha \cdot N$ is an integer. An example of randomly sampled face patches is shown in Figure 3.4.

Let $f(I, \kappa_b)$ denote the concatenation of the $\alpha \cdot N$ descriptors from the randomly selected patch indices in κ_b. That is,

$$f(I, \kappa_b) = \left[f(I, \kappa_b(1))^{\mathrm{T}}, \ldots, f(I, \kappa_b(\alpha \cdot N))^{\mathrm{T}} \right]^{\mathrm{T}} \tag{3.14}$$

Letting K_b^P and K_b^G denote the probe and gallery kernel similarity matrices for bag b, we modify Eqs. (3.3) and (3.4) to

$$K_b^P(i,j) = k(f(P_i, \kappa_b)\ ,\ f(P_j, \kappa_b)) \tag{3.15}$$

$$K_b^G(i,j) = k(f(G_i, \kappa_b)\ ,\ f(G_j, \kappa_b)) \tag{3.16}$$

The preliminary prototype representation $\phi'(\cdot)$ is now modified to $\phi'(\cdot, \cdot)$ as

$$\phi'(P,\kappa_b) = [\ k(f(P,\kappa_b), f(P_1,\kappa_b))\ ,\ \ldots\ , \\ k(f(P,\kappa_b), f(P_{n_{t_1}},\kappa_b))\ ]^{\mathrm{T}} \tag{3.17}$$

$$\phi'(G,\kappa_b) = [\ k(f(G,\kappa_b), f(G_1,\kappa_b))\ ,\ \ldots\ , \\ k(f(G,\kappa_b), f(G_{n_{t_1}},\kappa_b))\ ]^{\mathrm{T}} \tag{3.18}$$

A separate transformation matrix R_b is now learned for each bag as

$$R_b = K_b^G \cdot (K_b^P)^{-1} \tag{3.19}$$

resulting in the final prototype representation (modification of Eqs. (3.9) and (3.10)) as

$$\phi(P,\kappa_b) = R_b \cdot \phi'(P,\kappa_b) \tag{3.20}$$

$$\phi(G,\kappa_b) = \phi'(G,\kappa_b) \tag{3.21}$$

Linear discriminant analysis is performed separately for each bag. Using training set $\mathcal{T}_2$, we learn B subspace projection matrices $W_b,\ b = 1 \ldots B$.

A new face image I is represented in the random subspace prototype framework as $\Phi(I)$, where $\Phi(I)$ is the concatenation of each linearly projected prototype representation from each of the B random subspace bags. That is,

$$\Phi(I) = \left[\left(W_1^{\mathrm{T}} \cdot \phi(I,\kappa_1) \right)^{\mathrm{T}},\ \ldots,\ \left(W_B^{\mathrm{T}} \cdot \phi(I,\kappa_B) \right)^{\mathrm{T}} \right]^{\mathrm{T}} \tag{3.22}$$

For terseness we have omitted the subscripts F and D in the above equations. For example, in Eq. (3.22), $\Phi_{F,D}(I)$ is abbreviated to $\Phi(I)$ by omitting image filter F

Global Parameters
# of bags B, random sample vectors κ_b,
image filter F, feature descriptor D

Training
Input: Training sets $\mathcal{T}_1 = \{P_1, G_1, \ldots, P_{n_{t_1}}, G_{n_{t_1}}\}$,
$\mathcal{T}_2 = \{P'_1, G'_1, \ldots, P'_{n_{t_2}}, G'_{n_{t_2}}\}$
Output: $R_1, \ldots R_B, W_1, \ldots, W_B$
-*FOR* $b = 1 \ldots B$:
- Compute kernel matrices K_b^P, K_b^G
using prototypes in $\mathcal{T}_1$ *Eqs. (3.15), (3.16)*
- Solve R_b using K_b^P and K_b^G *Eq. (3.19)*
- FOREACH image I in $\mathcal{T}_2$:
  - Compute $\phi_{F,D}(I, \kappa_b))$ *Eqs. (3.20), (3.21)*
- Using all $I \in \mathcal{T}_2$, learn LDA subspace W_b
using representation $\phi_{F,D}(I, \kappa_b)$

Face Enrollment
Input: Image $I', \mathcal{T}_1(prototypes), R_1, \ldots, R_B$,
$W_1, \ldots, W_B$
Output: Φ
- *FOR* $b = 1 \ldots B$:
- *IF* I' is probe:
  - $\phi'_{F,D}(I') = [k\left(f_{F,D}(I', \kappa_b), f_{F,D}(P_1, \kappa_b)\right), \ldots,$
$k(f_{F,D}(I', \kappa_b), f_{F,D}(P_{n_{t_1}}, \kappa_b))]$ *Eq. (3.17)*
  - $\phi_{F,D}(I') = R_b \cdot \phi'(I', \kappa_b)$ *Eq. (3.20)*
- *ELSE* I' is gallery:
  - $\phi_{F,D}(I') = [k\left(f_{F,D}(I', \kappa_b), f_{F,D}(G_1, \kappa_b)\right), \ldots,$
$k(f_{F,D}(I', \kappa_b), f_{F,D}(G_{n_{t_1}}, \kappa_b))]$ *Eq. (3.18)*
- $\Phi_b = W_b^{\mathrm{T}} \cdot \phi_{F,D}(I')$
- Concatenate vectors $\Phi = [\Phi_1; \ldots; \Phi_B]$ *Eq. (3.22)*

Figure 3.5: Proposed Prototype Random Subspace framework algorithm. Following the offline training phase, a face image I' is enrolled and the vector Φ is returned for matching.

and descriptor D to represent I.

A summary of the training and image enrollment steps can be found in Figure 3.5.

3.5.3 Recognition

Given a probe face image P and a gallery face image G, we define their similarity $S(P, G)$ using the cosine similarity measure

$$S(P, G) = \frac{\langle \Phi(P), \Phi(G) \rangle}{||\Phi(P)|| \cdot ||\Phi(G)||} \tag{3.23}$$

Further, we let $S^{F_1,D_1}_{F_2,D_2}(P, G)$ denote the similarity between the probe P represented using filter F_1 and descriptor D_1, and gallery image G represented in terms of filter F_2 and descriptor D_2. That is

$$S^{F_1,D_1}_{F_2,D_2}(P, G) = \frac{\langle \Phi_{F_1,D_1}(P),\ \Phi_{F_2,D_2}(G) \rangle}{||\Phi_{F_1,D_1}(P)|| \cdot ||\Phi_{F_2,D_2}(G)||} \tag{3.24}$$

This similarity measure facilitates recognition using a threshold for a verification scenario (claimed identity for the probe is true or false), or a nearest neighbor matcher for an identification scenario (which one of N identities (classes) should be assigned to the probe).

3.5.4 Score Level Fusion

The proposed framework naturally lends to fusion of the different feature representations. For example, given one image filter F and two feature descriptors D_1 and D_2, we can utilize the following sum of similarity scores between probe image P and gallery image G: $\{S^{F_1,D_1}_{F_1,D_1}(P, G) + S^{F_2,D_2}_{F_2,D_2}(P, G) + S^{F_1,D_1}_{F_2,D_2}(P, G) + S^{F_2,D_2}_{F_1,D_1}(P, G)\}$. Min-max score normalization is performed prior to fusion.

3.6 Baselines

3.6.1 Commercial Matcher

The accuracy of the proposed prototype random subspace framework is compared against Cognitec's FaceVACS [1] COTS FRS. Comparing the accuracy of our system against this leading COTS FRS offers an unbiased baseline for each HFR scenario. FaceVACS was chosen because in our internal tests it excels at HFR scenarios (with respect to other commercial matchers). For example, the accuracy of FaceVACS on NIR to VIS [52] and Viewed Sketch to VIS [59] performed at par with some previously published HFR methods.

3.6.2 Direct Random Subspaces

In addition to a commercial face recognition system, the proposed prototype recognition system is compared against a recognition system that directly measures the difference between probe and gallery images using a common feature descriptor representation. As discussed previously, most recent approaches to heterogeneous face recognition involve directly measuring the similarity between two face images from alternate modalities using feature descriptors [12, 52, 59, 77].

The random subspace framework from [52] is used as the baseline because it is most similar to the proposed prototype framework, thus helping to isolate the difference between using kernel prototype similarities versus directly measuring the similarity. Further, because most of the datasets tested in Section 3.7 are in the public domain, the proposed framework may also be compared against any other published method on these data sets.

To briefly summarize the direct random subspace (D-RS) approach using our notation, for each bag b the D-RS framework represents an image as $f_{F,D}(I, \kappa_b)$. LDA is performed on each bag to learn the projection matrix $\tilde{W}_b$. Because only one

Table 3.1: Rank-1 accuracies for the proposed Prototype Random Subspace (P-RS) method across five recognition scenarios using an additional 10,000 subjects in the gallery.

	Rank-1 Accuracy (%)			
Method	NIR	Thermal	Sketch	Forensic *
P-RS	88.4 ± 4.99	55.3 ± 2.62	99.4 ± 0.54	19.6 ± 6.06
D-RS	90.1 ± 2.71	20.1 ± 2.23	97.2 ± 1.03	28.7 ± 4.09
(P-RS) + (D-RS)	91.9 ± 2.91	57.4 ± 2.25	99.6 ± 0.41	26.8 ± 9.66
FaceVACS	79.7 ± 3.75	20.7 ± 1.54	82.4 ± 2.39	4.2 ± 3.38

** Results for forensic sketch are the Rank-50 accuracy.*

training set is needed, LDA is learned from the images in $\mathcal{T}_1$ and $\mathcal{T}_2$ combined. The final representation $\Psi(\cdot)$ is the concatenation of the projected vector on the subspace for each bag

$$\Psi_{F,D}(I) = \Big[\Big(\tilde{W}_1^{\mathrm{T}} \cdot f_{F,D}(I, \kappa_1)\Big)^{\mathrm{T}}, \ \ldots, \Big(\tilde{W}_B^{\mathrm{T}} \cdot f_{F,D}(I, \kappa_B)\Big)^{\mathrm{T}}\Big]^{\mathrm{T}} \tag{3.25}$$

The dissimilarity $\tilde{S}$ between probe image P and gallery image G (each represented with filter F and descriptor D) is

$$\tilde{S}_{F,D}(P, G) = ||\Psi_{F,D}(P) - \Psi_{F,D}(G)||_2 \tag{3.26}$$

Unlike P-RS, D-RS must use the same D for the probe and gallery images. This is obvious as $f_{f,D_1}(I)$ and $f_{f,D_2}(I)$ will be of different dimensionality, and also have a different interpretation.

D-RS will be used in conjunction with the six filter/descriptor representations presented in Section 3.3 (SIFT + DoG, MLBP + CSDN, etc.). Results will be presented from the sum-score fusion of the min-max normalized scores from these six representations.

Table 3.2: Rank-1 accuracies for the proposed Prototype Random Subspace (P-RS) method on a standard photograph to photograph matching scenario using an additional 10,000 subjects in the gallery.

Method	Rank-1 Accuracy (%) Standard
P-RS	95.0 ± 1.58
D-RS	94.0 ± 1.30
(P-RS) + (D-RS)	95.3 ± 1.42
FaceVACS	99.5 ± 0.31

3.7 Experiments

The results reported below use the parameter values $\alpha = 0.1$ and $B = 200$. A third degree polynomial kernel was used to compute the prototype similarity and 99.0% of the variance was retained in the PCA step of LDA.

3.7.1 Databases

Five different matching scenarios are tested in this chapter: four heterogeneous face recognition scenarios, and one standard (homogeneous) face recognition scenario. Example images from each of HFR dataset can be found in Figure 3.1. Results shown on each dataset are the average recognition accuracy and the standard deviation over five random splits of training and testing subjects. No subject that was used in training was used for testing.

Dataset 1 - Near-Infrared to Visible (*Fig. 3.1(a)*) The first dataset consists of 200 subjects with probe images captured in the near-infrared spectrum (780-1,100 nm) and gallery images captured in the visible spectrum. Portions of this dataset are publicly available for download[1]. This dataset was originally used by Li et al. [72, 77]. Our experiments used only one NIR and one VIS image per subject,

[1] `http://www.cbsr.ia.ac.cn/english/Databases.asp`

making the scenario more difficult than previous experiments which benefited from multiple images per subject in training and gallery enrollment [52,77]. The data was split as follows: $n_{t_1} = 67$ subjects were used for training set $\mathcal{T}_1$, $n_{t_1} = 66$ subjects were used for training set $\mathcal{T}_2$, and the remaining 67 subjects were used for testing.

Dataset 2 - Thermal to Visible (*Fig. 3.1(b)***)** The second dataset is a private dataset collected by the Pinellas County Sheriff's Office, and consists of 1,000 subjects with thermal infrared probe images and visible (mug shot) gallery images. The thermal infrared images were collected using a FLIR Recon III ObservIR camera, which has sensitivity in the range of 3-5 μm and 8-12 μm. The data was split as follows: $n_{t_1} = 333$ subjects were used for training set $\mathcal{T}_1$, $n_{t_1} = 334$ subjects were used for training set $\mathcal{T}_2$, and the remaining 333 subjects were used for testing.

Dataset 3 - Viewed Sketch to Visible (*Fig. 3.1(c)***)** The third dataset is the CUHK sketch dataset[2], which was used by Tang and Wang [138, 149]. The CUHK dataset consists of 606 subjects with a viewed sketch image for probe and a visible photograph for gallery. A viewed sketch is a hand drawn sketch of a face which is drawn while looking at a photograph of the subject. The photographs in the CUHK dataset are from the AR [89], XM2VTS [92], and CUHK student [138,149] datasets. The 606 subjects were equally divided to form the training sets $\mathcal{T}_1$, $\mathcal{T}_2$, and the test set.

Dataset 4 - Forensic Sketch to Visible (*Fig. 3.1(d)***)** The fourth and final heterogeneous face dataset consists of real-world forensic sketches and mug shot photos of 159 subjects. This dataset is described in [59]. Forensic sketches are drawn by an artist based only on an eye witness description of the subject. The forensic sketch dataset is a collection of images from Gibson [31], Taylor [139], the Michigan

[2]The CUHK dataset is publicly available for download at http://mmlab.ie.cuhk.edu.hk/facesketch.html

State Police, and the Pinellas County Sheriff's Office. Each sketch contains a suspect involved in a real crime, and the mug shot photo was only available after the subject had later been identified by means other than face recognition. Forensic sketches contain incomplete information regarding the subject, and are one of the most difficult HFR scenarios because the sketches often do not closely resemble the photograph of the true suspect. Here 53 different subjects each are used in $\mathcal{T}_1$, $\mathcal{T}_2$, and the test set.

Dataset 5: Standard Face Recognition A fifth non-heterogeneous (i.e. homogeneous) dataset is used to demonstrate the ability of the proposed approach to operate in standard face recognition scenarios. The dataset consists of one probe and one gallery photograph of 876 subjects, where 117 subjects were from the AR dataset [89], 294 subjects were from the XM2VTS dataset [92], 193 subjects from the FERET dataset [109], and 272 subjects were from a private data set collected at the University of Notre Dame. This is the same dataset used in [56].

Enlarged Gallery A collection of 10,000 mug shot images were used in certain experiments to increase the size of the gallery. These mug shot images were provided by the Michigan State Police, and were also used in [59]. Any experiment using these additional images will have a gallery with the number of testing subjects plus these additional 10,000 mug shot images. Experiments with a large gallery are meant to present results that more closely resemble real-world face matching scenarios.

3.7.2 Results

Tables 3.1 and 3.2 lists the results of P-RS, D-RS, and FaceVACS for each dataset using the additional 10,000 gallery images for each experiment. The results for P-RS are the fusion of the match scores from $\{S^{F_d,D_s}_{F_d,D_s} + S^{Fc,Ds}_{Fc,Ds} + S^{Fg,Ds}_{Fg,Ds} + S^{F_d,Dm}_{F_d,Dm} + S^{Fc,Dm}_{Fc,Dm} + S^{Fg,Dm}_{Fg,Dm}\}$, i.e. the same features are used in the probe and gallery images. Similarly, D-RS is the fusion of the match scores from $\{\tilde{S}_{F_d,D_s} + \tilde{S}_{Fc,Ds} + \tilde{S}_{Fg,Ds}$

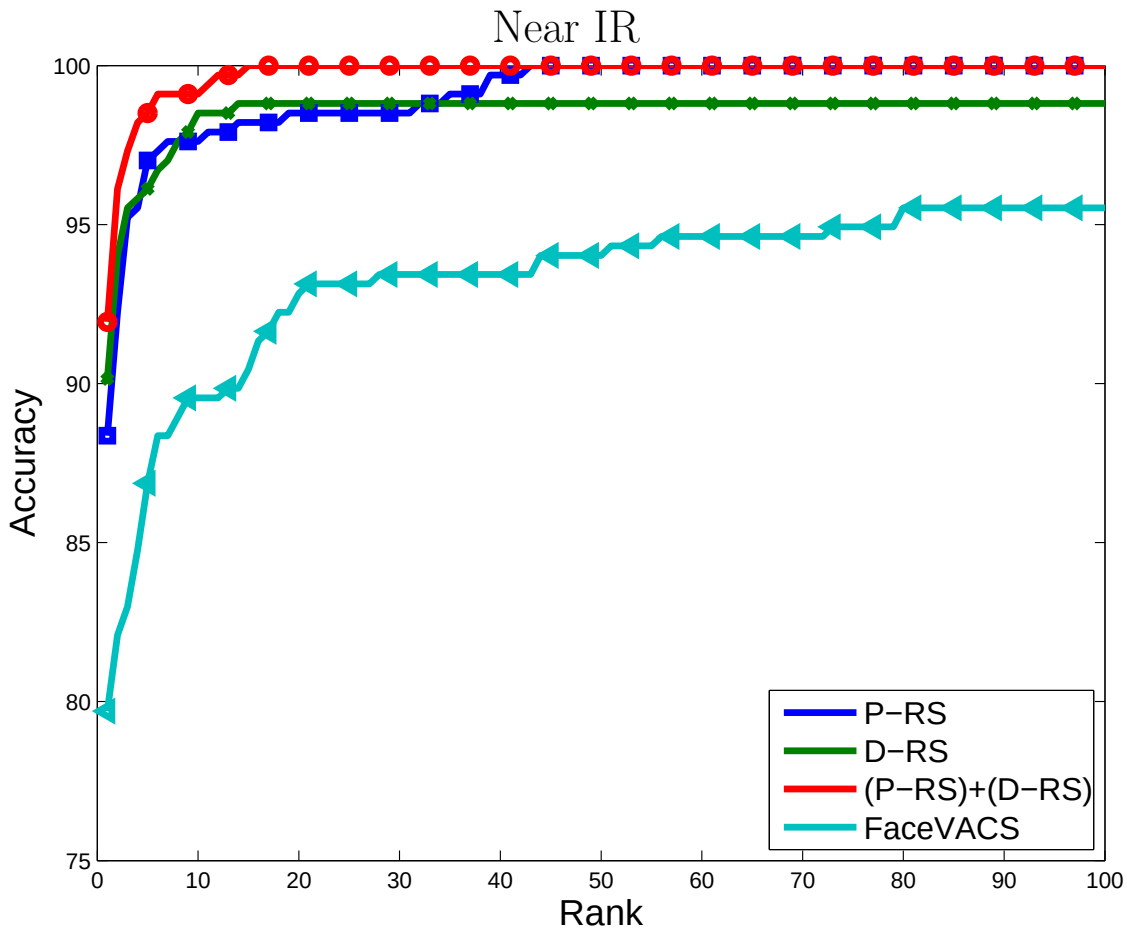


Figure 3.6: CMC plot for the NIR HFR scenario. Results use an additional 10,000 gallery images to better replicate real world matching scenarios. Listed are the accuracies for the proposed Prototype Random Subspace (P-RS) method, the Direct Random Subspace (D-RS) method [52], the sum-score fusion of P-RS and D-RS, and Congitec's FaceVACS system [1].

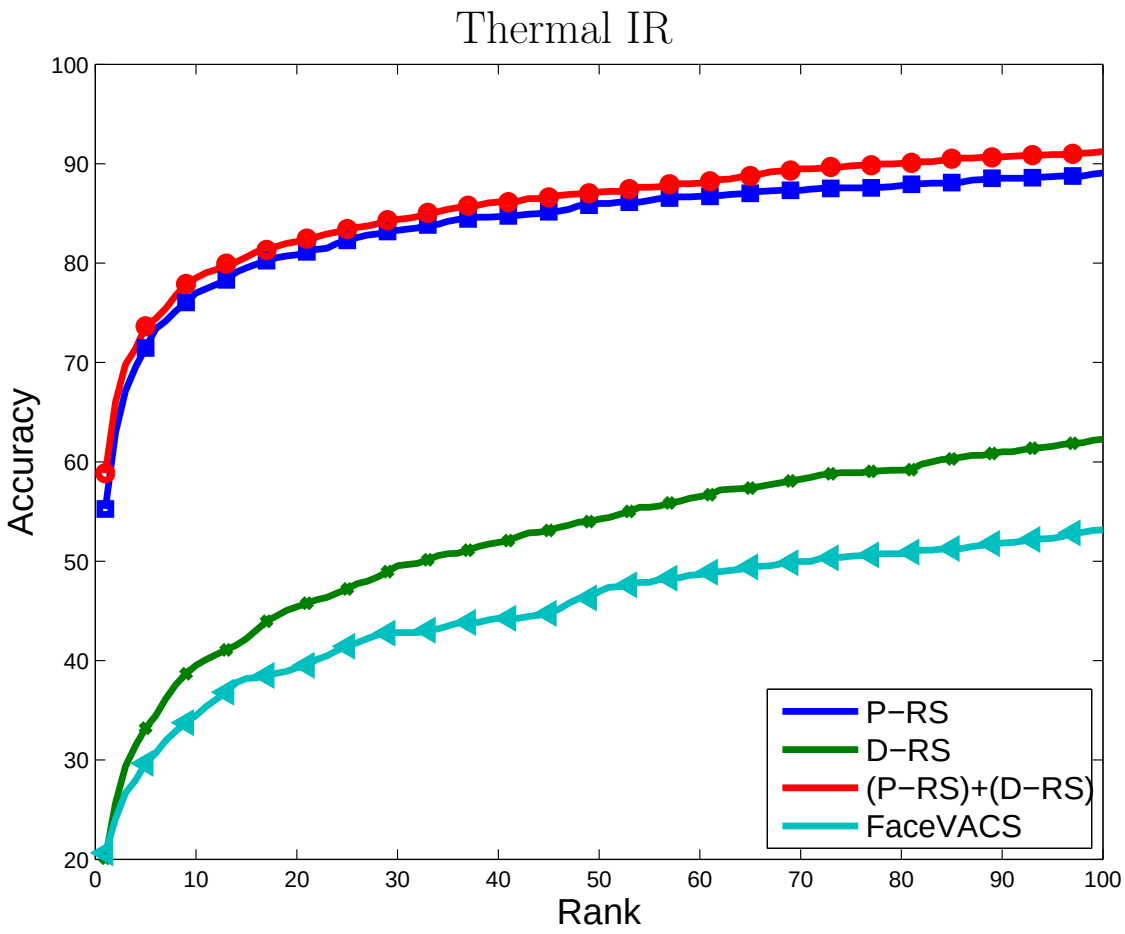


Figure 3.7: CMC plot for the thermal HFR scenario. Results use an additional 10,000 gallery images.

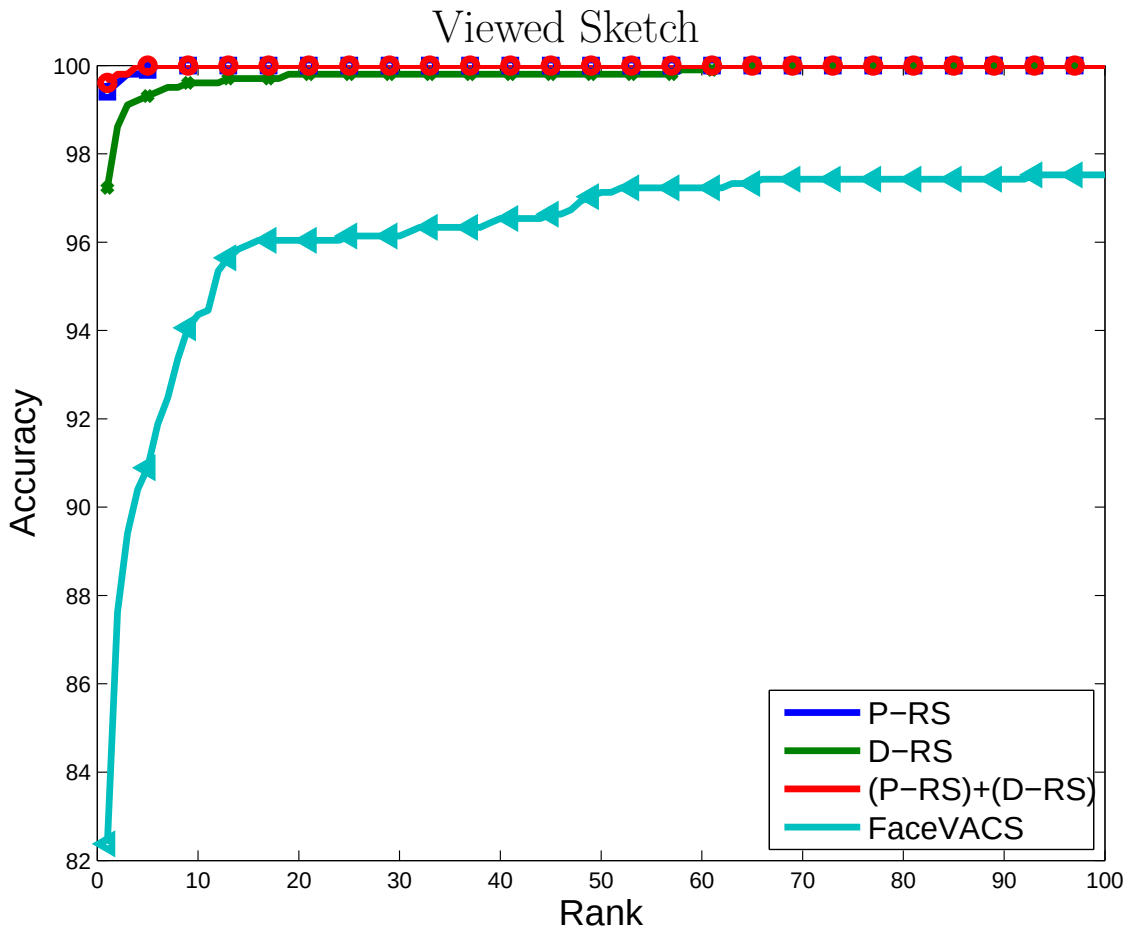


Figure 3.8: CMC plot for the viewed sketch HFR scenario. Results use an additional 10,000 gallery images.

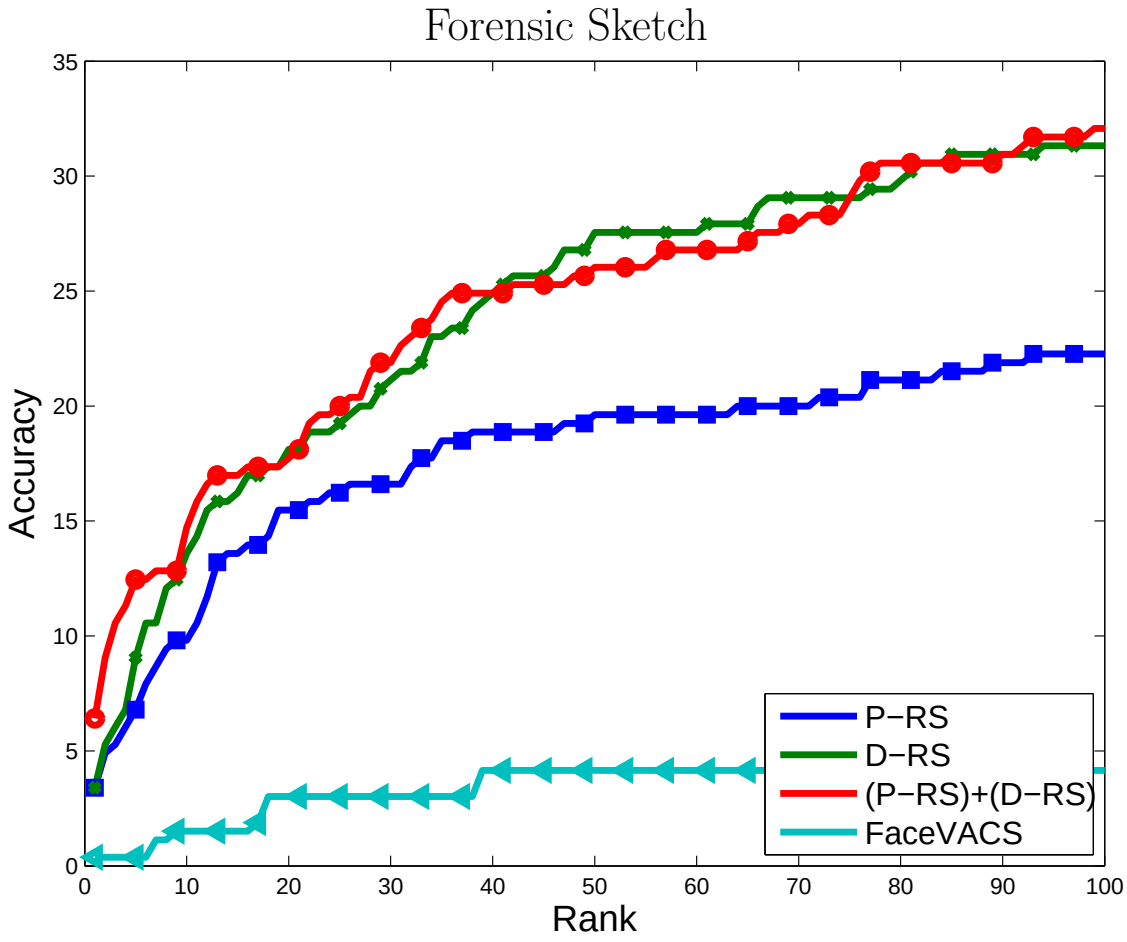


Figure 3.9: CMC plot for the forensic sketch HFR scenario. Results use an additional 10,000 gallery images.

$+\tilde{S}_{F_d,D_m} + \tilde{S}_{Fc,D_m} + \tilde{S}_{Fg,D_m}\}$. Results from these same matchers are also displayed in CMC (cumulative match characteristic) plots in Figures 3.6, 3.7, 3.8, and 3.9. Again, the P-RS method represents face images using their similarity to a set of prototype subjects, while the D-RS method directly measures the similarity between two face images using SIFT and LBP descriptors.

The CMC results of matching NIR face images to standard face images are shown in Figure 3.6. The Rank-1 accuracy of 88.4% from Table 3.1 and Figure 3.6 demonstrates that the proposed P-RS matcher is able to perform at a similar level as D-RS and FaceVACS. FaceVACS was earlier benchmarked as performing at the same level as the top methods [52]. Thus, the proposed P-RS method is on par with leading methods in NIR to VIS matching.

The CMC results of matching thermal face images to standard face images are shown in Figure 3.7. P-RS is able to achieve an average Rank-1 accuracy of 55.3%. By comparison, it is observed that the D-RS method achieves a Rank-1 accuracy of only 20.1% and FaceVACS has a Rank-1 accuracy of 20.7%. This drastic improvement demonstrates the benefit of P-RS's notable property of not requiring a feature descriptor that is invariant to changes in the probe and gallery modalities. A Rank-1 accuracy of 55.3% still falls short of the accuracy desired in lights out systems, however, the examples in Figures 3.1 and 3.12 show that even humans would have difficulty in this recognition task. The only previous method on thermal to visible matching achieved a Rank-1 accuracy of 50.06% but it was evaluated on only 47 subjects in the gallery [71]. By contrast, the Rank-1 accuracy of 55.3% of the proposed P-RS method used a gallery consisting of 10,333 subjects.

The CMC results of matching viewed sketch face images to standard face images are shown in Figure 3.8. P-RS achieved near perfect accuracy with an average Rank-1 accuracy of 99.4%. Other methods have also achieved nearly 99% Rank-1 accuracy [59, 157], though the results in Figure 3.8 are based on a gallery with over 10,000

subjects compared to a gallery size of less than $1,000$ in previous studies.

The CMC results of matching forensic sketch face images to standard face images are shown in Figure 3.9. For forensic sketches the Rank-50 accuracy is most relevant because the Rank-1 accuracy is too low to be useful in practice: forensic investigators generally examine roughly the top 50 retrieved matches from a query. It is observed that this is the one scenario in which P-RS (Rank-50 accuracy of 19.6%) was outperformed by D-RS (Rank-50 accuracy of 28.7%). The only previous method to publish results on forensic sketch matching also used the same extended gallery and achieved a Rank-50 accuracy of 13.4% [59] (this number is the weighted average of a 32.65% Rank-50 accuracy on 49 *good* sketches and a 8.16% accuracy on 110 *poor* sketches). It is important to note that the matcher in [59] was trained on viewed sketches, and not forensic sketches like P-RS and D-RS.

The decreased accuracy of P-RS compared to D-RS on the forensic sketch dataset is attributed to two factors. The primary factor is the small size of the data set. While both methods utilize learning, D-RS is able to leverage the a priori knowledge that SIFT and MLBP perform well for direct similarity measurement. Further, D-RS is able to use both training sets to learn the LDA subspaces. By contrast, P-RS must use the first training set to develop the prototypes. An additional reason for P-RS's lower accuracy on forensic sketch matching is that these sketches are often not completely accurate due to the inability of a witness to adequately describe the face of a suspect, which impacts the assumption in Eq. (3.7). Despite these limitations, P-RS still achieved approximately four times accuracy improvement over a leading COTS FRS.

Examples cases where (i) P-RS succeeds but FaceVACS fails and (ii) P-RS fails but FaceVACS succeeds are shown for the two most difficult HFR scenarios (thermal and forensic sketch) in Figure 3.12.

Figures 3.10 and 3.11 demonstrate the ability of the P-RS framework to perform

NIR

	Gallery Features					
Probe Features	DoG SIFT	DoG MLBP	CSDN SIFT	CSDN MLBP	Gauss SIFT	Gauss MLBP
DoG SIFT	77.9	**72.8**	**77.9**	**76.7**	**66.6**	**58.8**
DoG MLBP	**60.6**	85.4	**59.1**	**78.8**	**48.7**	**57.0**
CSDN SIFT	**75.8**	**63.9**	81.8	**76.7**	**74.0**	**68.1**
CSDN MLBP	**66.6**	**77.9**	**69.6**	84.8	**69.0**	**75.5**
Gauss SIFT	**62.4**	**49.0**	**72.5**	**72.5**	72.2	**66.9**
Gauss MLBP	**52.8**	**56.1**	**59.1**	**70.4**	**63.0**	67.5

(a)

Thermal

	Gallery Features					
Probe Features	DoG SIFT	DoG MLBP	CSDN SIFT	CSDN MLBP	Gauss SIFT	Gauss MLBP
DoG SIFT	50.8	**46.1**	**49.1**	**47.7**	**36.0**	**34.7**
DoG MLBP	**42.0**	57.8	**35.9**	**49.7**	**23.6**	**30.0**
CSDN SIFT	**46.7**	**35.7**	50.2	**47.3**	**40.1**	**36.4**
CSDN MLBP	**41.9**	**49.0**	**44.7**	55.1	**33.6**	**37.1**
Gauss SIFT	**29.0**	**20.5**	**37.4**	**33.2**	36.0	**32.1**
Gauss MLBP	**24.4**	**22.6**	**31.6**	**36.3**	**28.8**	32.6

(b)

Figure 3.10: Rank-1 accuracies (%) on the NIR and thermal modalities using the proposed P-RS framework. The rows list the features used to represent the probe images, and the columns list the features for the gallery images. The non-diagonal entries in each table (in bold) use different feature descriptor representations for the probe images than the gallery images. These results demonstrate another "heterogeneous" aspect of the proposed framework: recognition using heterogeneous features between the probe and gallery images.

Viewed Sketch

	Gallery Features					
Probe Features	DoG SIFT	DoG MLBP	CSDN SIFT	CSDN MLBP	Gauss SIFT	Gauss MLBP
DoG SIFT	98.6	**95.6**	**98.3**	**95.1**	**96.8**	**89.4**
DoG MLBP	**87.5**	96.3	**82.3**	**92.5**	**60.8**	**66.8**
CSDN SIFT	**98.4**	**91.0**	98.7	**95.1**	**97.6**	**93.2**
CSDN MLBP	**85.8**	**92.3**	**87.6**	96.2	**75.1**	**81.2**
Gauss SIFT	**96.6**	**74.2**	**98.3**	**91.6**	97.8	**94.6**
Gauss MLBP	**90.5**	**83.7**	**94.8**	**95.1**	**95.6**	98.0

(a)

Forensic Sketch

	Gallery Features					
Probe Features	DoG SIFT	DoG MLBP	CSDN SIFT	CSDN MLBP	Gauss SIFT	Gauss MLBP
DoG SIFT	6.4	**7.2**	**8.7**	**6.4**	**7.5**	**9.1**
DoG MLBP	**6.0**	10.6	**6.0**	**9.4**	**5.7**	**8.7**
CSDN SIFT	**7.5**	**4.5**	5.7	**6.4**	**5.7**	**6.0**
CSDN MLBP	**6.8**	**8.3**	**8.7**	11.3	**5.7**	**4.9**
Gauss SIFT	**4.9**	**3.8**	**6.4**	**5.3**	7.5	**9.1**
Gauss MLBP	**6.4**	**5.3**	**7.9**	**5.3**	**9.1**	10.9

(b)

Figure 3.11: Rank-1 accuracies (%) on the viewed sketch and forensic sketch modalities using the proposed P-RS framework. The rows list the features used to represent the probe images, and the columns list the features for the gallery images. The non-diagonal entries in each table (in bold) use different feature descriptor representations for the probe images than the gallery images. These results demonstrate another "heterogeneous" aspect of the proposed framework: recognition using heterogeneous features between the probe and gallery images.

P-RS: *Rank 1*
FaceVACS: *Rank 7622*
(a)

P-RS: *Rank 891*
FaceVACS: *Rank 1*
(b)

P-RS: *Rank 1*
FaceVACS: *Rank 7622*
(c)

P-RS: *Rank 891*
FaceVACS: *Rank 1*
(d)

Figure 3.12: Examples of thermal recognition not successfully matched by (a) FaceVACS, and (b) the proposed P-RS method. Examples of forensic sketch recognition not successfully matched by (c) FaceVACS, and (d) P-RS. In each image pair the left and right images are the probe and gallery, respectively.

recognition using different feature descriptors for the probe and gallery images. Figure 3.10 lists the Rank-1 accuracy for the NIR and thermal HFR scenarios, and Figure 3.11 lists the same for the viewed sketch and forensic sketch scenarios. These scores are averaged over five random training/testing splits but do not use the additional 10,000 gallery images. The columns indicate each of the six different image filter and feature descriptor combinations used to represent the gallery, and the rows indicate the representations used for the probe images. Thus, the non-diagonal entries for each scenario are when the probe and gallery images are represented with different features. The accuracy is generally higher when the same features are used for faces in the probe and gallery (i.e. the diagonal entries). Various levels of accuracy are achieved when using different image features, ranging from poor to high.

The ability to perform face recognition with the probe and gallery images using different representations is a property that previous feature-based methods did not possess. This property is important to mention because it demonstrates the proposed method's ability to generalize to other unknown HFR scenarios. For example, in the case of thermal to visible recognition, if a local feature descriptor is developed that performs at a very high level in matching thermal to thermal, it can be incorporated into this framework even if it does not work well in the visible domain. As other HFR scenarios are attempted (such as matching 3D depth map to 2D visible photograph), this property could prove extremely useful in overcoming the hurdle of finding a feature descriptor that is invariant to changes between the two domains, which feature-based methods rely on.

Table 3.3 lists the Rank-1 accuracy (without the additional gallery) for each scenario with and without various components of the prototype random subspace framework (namely LDA, the transformation matrix R, and random subspaces, RS). The improvement in recognition accuracy when using the R tranformation quantitatively demonstrates the importance of this step in our algorithm.

The proposed P-RS framework also generalizes to standard face recognition scenarios. Using the standard dataset, Figure 3.14(a) compares the accuracy of P-RS, D-RS, and FaceVACS. FaceVACS clearly outperforms P-RS and D-RS as it is consistently one of the top performers in NIST face recognition benchmarks. However, using four different face datasets we see that P-RS and D-RS both achieve Rank-1 accuracies around 95% with 10,876 subjects in the gallery, compared to 99.5% accuracy for FaceVACS. In Figure 3.14(b) the results of matching using different feature descriptors in the probe and gallery domain are shown. The ability to match probe and gallery images using different feature representations is novel and could benefit situations in which only the face templates, instead of the face image are available.

The proposed P-RS method is computationally scalable to meet the demands of real world face recognition systems. Running in Matlab and using a single core from a 2.8GHz Intel Xeon processor, the breakdown of compute time is needed to enroll a single face image is as follows. Image filtering requires roughly 0.008 sec for DoG, 1.1 sec for CSDN, and 0.004 sec for Gauss. The image MLBP and SIFT features descriptors each take roughly 0.35 sec to compute. Because each image filtering is performed once, and each feature descriptor is computed three times (once for each filter), computing all six filter/descriptor combinations takes around 3.2 sec. The

Table 3.3: Effect of each component in the P-RS framework on recognition accuracy. Components tested are LDA, the transformation matrix R, and random subspaces (RS). Listed are the average Rank-1 accuracies for each scenario without the additional 10,000 gallery images.

	LDA				No LDA			
	R		No R		R		No R	
	RS	No RS	RS	No RS	RS	No RS	RS	No RS
NIR	0.904	0.901	0.725	0.699	0.722	0.472	0.600	0.319
Thermal	0.643	0.637	0.508	0.474	0.217	0.174	0.178	0.120
Viewed Sketch	0.994	0.992	0.981	0.970	0.970	0.867	0.939	0.618
Forensic Sketch	0.136	0.147	0.143	0.140	0.102	0.064	0.094	0.068

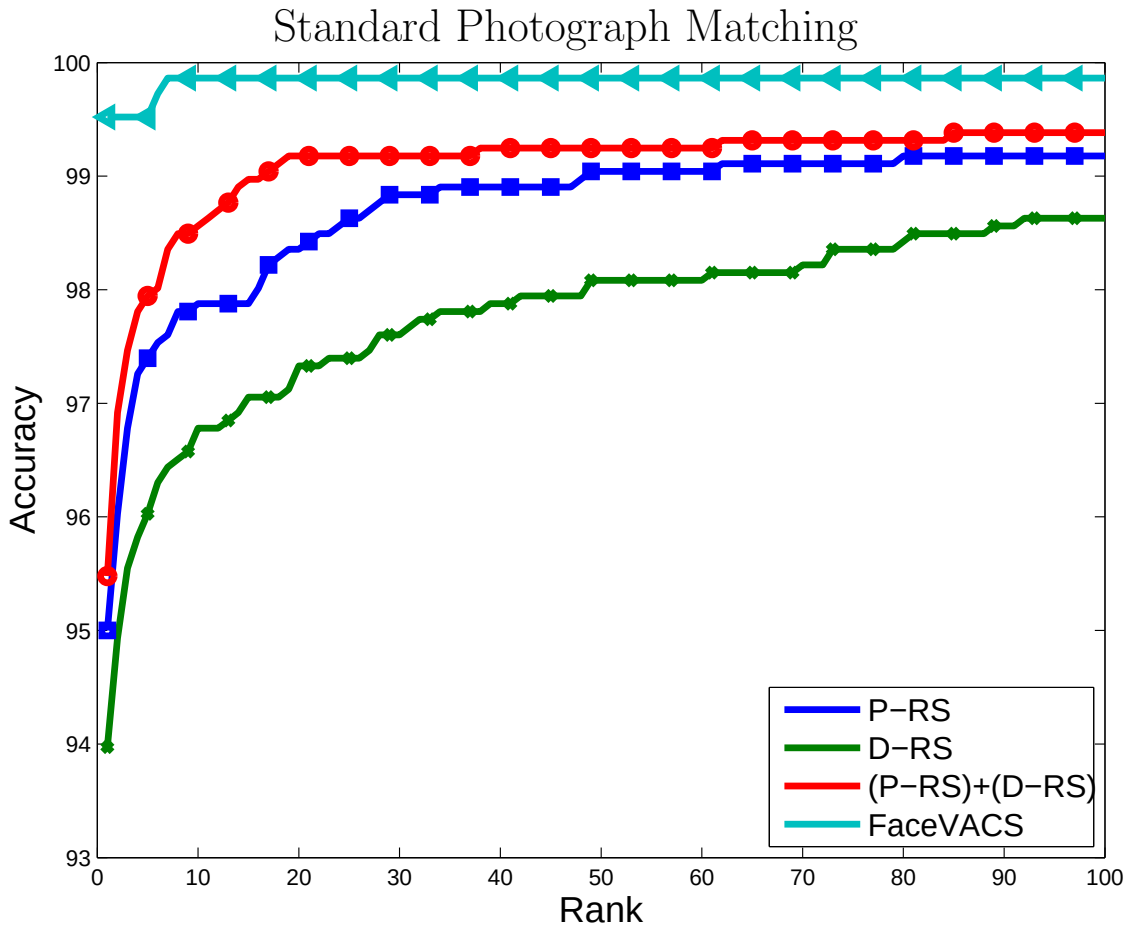


Figure 3.13: CMC plot of matcher accuracies with an additional 10,000 gallery images when photos are used for both the probe and gallery (i.e. non-heterogeneous face recognition).

	Gallery Features					
Probe Features	DoG SIFT	DoG MLBP	CSDN SIFT	CSDN MLBP	Gauss SIFT	Gauss MLBP
DoG SIFT	93.0	**85.6**	**90.8**	**84.4**	**81.8**	**72.9**
DoG MLBP	**78.4**	92.9	**73.0**	**89.2**	**53.5**	**61.6**
CSDN SIFT	**90.7**	**79.8**	93.6	**88.4**	**90.4**	**85.7**
CSDN MLBP	**73.8**	**87.3**	**79.7**	94.1	**73.4**	**81.6**
Gauss SIFT	**80.5**	**53.3**	**89.8**	**81.0**	93.2	**90.6**
Gauss MLBP	**71.0**	**65.7**	**84.9**	**88.6**	**89.9**	95.1

Figure 3.14: Face recognition results (%) when photos are used for both the probe and gallery (i.e. non-heterogeneous face recognition). The layout is the same as in Figure 3.10 (i.e. results shown are when different features are used to represent the probe and gallery images).

prototype random subspace representation with 200 bags takes roughly 0.3 sec to compute for a single filter/descriptor combination. Thus, all six filter/descriptor combinations take roughly 1.8 sec. In total, a face image needs around 5.0 sec. to enroll in Matlab. With a gallery of n_g subjects and the final feature vector Φ of size d', identification of a subject takes $O(d' \cdot n_g)$ time. Depending on the number of bags, the number of prototypes for each scenario, and the variance retained in the PCA step, d' is of the order $1,000$.

3.8 Summary

A method for heterogeneous face recognition, called Prototype Random Subspaces (P-RS), is proposed. Probe and gallery images are initially filtered with three different image filters, and two different local feature descriptors are then extracted. A training set of prototypes is selected, in which each prototype subject has an image in both the gallery and probe modalities. The non-linear kernel similarity between an image and the prototypes is measured in the corresponding modality. A random subspace framework is employed in conjunction with LDA subspace analysis to further improve the recognition accuracy.

The proposed method leads to excellent matching accuracies across four different HFR scenarios (near infrared, thermal infrared, viewed sketch, and forensic sketch). Results were compared against a leading commercial face recognition engine. In most of our experiments the gallery size was increased with an additional 10,000 subjects to better replicate real matching scenarios. In addition to excellent matching accuracies, one key benefit of the proposed P-RS method is that different feature descriptors can be used to represent the probe and gallery images. Finally, the P-RS method performed comparable to a leading commercial face recognition engine on a visible to visible matching scenario (i.e. non-heterogeneous face recognition).

Future work will focus on (i) improving the accuracy of each of the tested HFR scenarios separately, (ii) improving the runtime complexity of the prototype representation, and (iii) incorporating additional HFR scenarios. Tailoring the P-RS parameters and learning weighted fusion schemes for each HFR scenario separately should offer further accuracy improvements. Another potential technique to improve the recognition accuracies is to allow the P-RS method to leverage multiple training samples per subject. That is, in the conducted experiments each training subject has only one image per modality. However, in many operational scenarios training data will be available with multiple images per modality for each subject. This additional information will improve the ability to estimate the within-class scatter in our discriminant analysis. We will also continue to improve aspects of the D-RS method (which achieved high accuracy on several of the HFR scenarios), such as the similarity metrics and image filters. Improvements to the runtime complexity is of the P-RS method should be explored through an examination of the maximum number of prototypes needed to achieve the highest recognition accuracy on a given scenario, and kernel approximation methods such as the Nystrom method [150]. One additional HFR scenario that should be considered is 3D to 2D face matching. P-RS should be particularly impact fool in this scenario because heterogeneous features will be required to represent faces in the 3D and 2D modalities.

While a vast majority of previous algorithms for heterogeneous face recognition have been designed for a specific application (e.g. sketch recognition [12, 59, 82, 138, 149] and near-infrared recognition [52, 77]), the algorithm presented in this chapter generalizes to any HFR scenario. By providing a generalized approach to heterogeneous face recognition this chapter offers a strong contribution the field of heterogeneous face recognition. As new heterogeneous face recognition scenarios are attempted (such as much depth images acquired from LIDAR), the P-RS algorithm proposed in this chapter should allow for success in these future endeavors.

Chapter 4

Face Recognition Across Time Lapse

4.1 Introduction

In addition to changes in image modality, another variate that is known to greatly impact face recognition performance is the alteration in facial appearance that occurs through the aging process [120].

This chapter will look at the performance of aging-invariant face recognition systems in both aging and non-aging scenarios. By showing that aging invariant face recognition systems do not generalize to non-aging scenarios, we are able to pose aging-invariant face recognition as a heterogeneous face recognition problem. That is, the identifiable facial features for faces that have not undergone aging are largely heterogeneous from the identifiable features of faces that have aged. This is evidenced by the different discriminative subspaces that are learned from aged face images than non-aged face images.

Unlike the pose, expression, and illumination, aging factors cannot be constrained in order to improve face recognition performance. For example, many years may pass

Figure 4.1: Multiple images of the same subject are shown, along with the match score (obtained by a leading face recognition system) between the initial gallery seed and the image acquired after a time lapse. As the time lapse increases, the recognition score decreases. This phenomenon is a common problem in face recognition systems. The work presented in this chapter (i) demonstrates this phenomenon on the largest aging dataset to date, and (ii) demonstrates that solutions to improve face recognition performance across large time lapse impact face recognition performance in scenarios without time lapse.

before a released prisoner is recidivate, resulting in a large time lapse between the mug shot image in the gallery and the current booking image (probe). Similarly, a U.S. passport is valid for ten years, and most state driver's licenses only need to be renewed every five to ten years. Thus, in many critical applications the success of face recognition technology may be impacted by the large time lapse between a probe image and its true mate in the gallery.

Over the past five years there has been a growing interest in understanding the impact of aging on face recognition performance and proposing solutions to mitigate any negative impact from aging. A major contributor to these advances has been the availability of the MORPH database by Ricanek et al. [116, 121]. The MORPH database consists of two albums, which, in total, contains roughly 100,000 images of about 25,000 subjects. The MORPH dataset has facilitated studies on synthetic aging [104,135], age invariant face recognition [76,104], age estimation [35], and aging analysis [107]. A broader examination of facial aging methods in the literature can be found in the summary provided by Ramanathan et al. [114].

Various approaches for improving face recognition performance in the presence of aging can be dichotomized into two groups. The first contains generative synthesis methods which seek to learn an aging model that can estimate the appearance of an aged face from an input image. While these methods have shown some success in mimicking the aging process [104, 129, 135], generative methods are challenging due to the large number of parameters that must be estimated. Synthesis methods also rely on the appearance of the face in order to simulate the aging process, which can suffer from the minor pose and illumination variations that are encountered in large datasets. Further, synthesis methods do not handle the problem of face recognition, and need a separate face engine to perform matching. Of course, this also speaks to one of the advantages of synthetic aging methods: they can be easily integrated with existing face recognition engines.

An alternative solution to improving face recognition performance across time lapse is through discriminative learning methods [76, 79, 86, 113]. Such methods seek to find the weighted combination of features that are most stable across a particular time lapse. Discriminative approaches are able to leverage both the wide range of facial feature representations [56], as well as the family of learning methods in face recognition. Beginning with Belhuemer's FisherFaces approach [10], discriminative learning approaches have been critical to the advancement of face recognition over the past two decades.

Li et al. used a discriminative random subspace method that outperforms a leading commercial face recognition engine on the MOPRH dataset [76]. This work helped demonstrate that a face recognition system could be trained to improve performance in the presence of aging [76]. While these contributions helped advance the state of the art in face recognition in the presence of time lapse, they also raise another question regarding the design of face recognition systems: *does the learned subspace for face recognition across time lapse impact the face recognition performance in a*

non-aging scenario? In other words, while we can improve face recognition performance in the presence of a large time lapse between the probe and gallery images, we also want to maintain the performance on two images with minimal time lapse. This question, to our knowledge, has not yet been addressed before.

The contributions of the research presented in this chapter are motivated by the need to answer the question posed above. This question is answered by providing the largest study to date on the impact of aging on face recognition performance. Leveraging a dataset of 200,000 mug shot images from 64,000 subjects, we demonstrate (i) a degradation in face recognition performance from two leading commercial-of-the-shelf (COTS) face recognition systems (FRS) on match sets partitioned by the amount of time lapse occurring between the probe and gallery images, and (ii) training to improve performance on a particular time lapse range impacts performance on other time lapse ranges. These findings suggest that face recognition systems should update face templates after a certain time interval has passed from the original acquisition date in order to maximize the benefit of age invariant subspaces without impacting face recognition in non-aging scenarios.

The remainder of this chapter is outlined as follows. In Section 4.2 we discuss the face dataset used in this research. In Section 4.3 we revisit the random subspace framework and discuss how it was adopted for this work. Section 4.4 presents experiments on the impact of training for age invariant face recognition, as well as the computational demands that were born from undertaking such a large scale study.

4.2 Dataset

This study leverages a set of 200,000 mug shot images from roughly 64,000 subjects collected in the state of Florida, U.S.A. Each image contains a subject id and an image acquisition date, which enables the time lapse between any two images to be

determined. The 200,000 images are a subset of a larger 1.5 million image dataset available to us; these 200,000 images were selected so that different time lapse ranges were equally represented in this study.

The time lapse ranges (between a probe and gallery image) analyzed in this study were (i) 0 to 1 year, (ii) 1 to 5 years, (iii) 5 to 10 years, and (iv) more than 10 years. Training sets for each of these time lapse ranges are generated so that each range has 8,000 subjects. The only exception is the 10+ year time lapse range, where only around 2,000 subjects were available in the database in the database for training. Similarly, test sets were generated to represent each of the time lapse ranges listed above. For each time lapse range, 12,000 subjects were used for testing. Again, the 10+ year time lapse test set contained only 2,000 subjects, similar to the training set. For each subject in the study, their oldest face image was used as the gallery seed image. Multiple probe images that fell within the time lapse range for a subject were often available as well. For example, the 1 to 5 year test set contained 12,000 gallery images and 33,443 probe images, where each probe image was taken between one to five years after the corresponding gallery image.

All parameter tuning in this work was performed using the training set. This was performed by using the first half of the training set to train on different parameter values and the second half of the training set to determine the optimal parameter combination (with respect to face recognition performance). Thus, the second half of the training set also served as a validation set.

The analysis performed on this dataset is the largest such study reported to date. Further, because the images are pulled from a larger pool of an operational database of mug shot images, the study is unique in that it controls the time lapse variate so that the same number of subjects are available to analyze 5 to 10 years aging as 0 to 1 year aging (for example). As such, measuring the performance of COTS FRS on this dataset will provide a convincing demonstration of how commercially

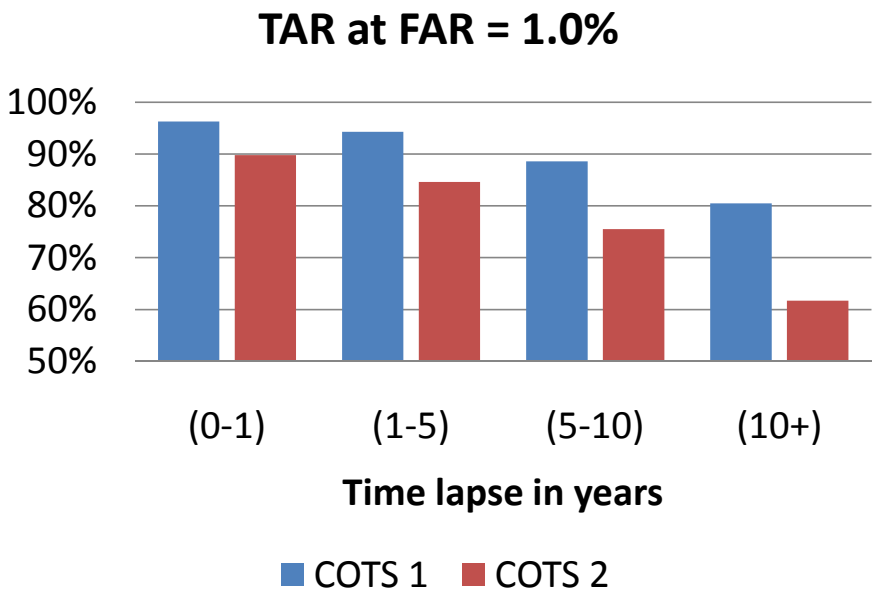


Figure 4.2: The performance of two commercial face recognition systems as a function of time lapse between probe and gallery images.

available face recognition technology performs in the presence of aging. Because both Li et al. [76] and Ling et al. [79] have been able to surpass the performance of COTS FRS by performing discriminative learning on face images with time lapse, it is generally accepted that face recognition performance degrades monotonically as the time between image acquisition increases.

We analyzed to performance of two commercial-of-the-shelf face recognition systems: Cognitec's FaceVACS SDK [1], and PittPatt's Face Recognition SDK [2]. Both matchers were competitive participants in the 2010 NIST Multi-Biometrics Evaluation (MBE) [34]. Recognition results reported here list the two matchers as "COTS 1" and "COTS 2" in order to make anonymous each matcher's performance relative to the other.

Figure 4.2 shows the matching accuracies of the two COTS matchers as a function of the time lapse between the probe and gallery image. The decrease in performance as the time lapse increases clearly shows the difficulty face recognition systems have with age variation.

4.3 Random Subspace Face Recognition

In this work we adopt a random subspace linear discriminant analysis (RS-LDA) algorithm, based on Wang and Tang's original face subspace method [148]. More recently, Li et al. [76] have used a variant of this approach to improve face recognition in the presence in aging. Klare and Jain have also demonstrated the benefit of RS-LDA on a heterogeneous face recognition scenario [52, 54].

RS-LDA is based on the FisherFace linear discriminant analysis algorithm [10], where a linear subspace Ψ is learned from the original feature space by solving the generalized eigenvalue problem $S_b \cdot \Psi = \Lambda \cdot S_w \cdot \Psi$ with the between-class and within-class matrices S_b and S_w built from a set of training images. In RS-LDA, multiple subspaces $\Psi_b,\ b = 1 \ldots B$, are learned using both randomly sampled subsets of the original feature space as well as randomly sampled subjects from the set of training instances. The motivation for using RS-LDA over standard LDA is due to degenerative properties that often manifest in S_w (which must be of full rank to solve $S_w^{-1} \cdot S_b$). While Level 2 facial feature representations [56] (such the local binary patterns [99] used in this work) offer improved recognition accuracies, they also increase the dimensionality of the facial feature vectors. This in turn increases the likelihood that S_w is degenerate, and further necessitates the need for a method such as RS-LDA. Other LDA variants offer solutions to this small sample size problem [45, 85], however RS-LDA is preferred due to the ease of implementation and wider range of successful applications in face recognition [52, 54, 76, 148].

The approach used in this work is mostly based on the method by Li et al. [76], however we had to modify their method in order to reduce the computational requirements because the number of images handled in this experiment is an order of magnitude larger than their work. Again, the intent of this work is not to provide a method that can improve on commercially available face recognition technology (this capability has already been demonstrated [76, 79]). Instead, we wish to understand

how training a face recognition system to improve recognition accuracies on a particular time lapse scenario performs on scenarios with a larger or smaller amount of time lapse than training time lapse.

4.3.1 Face Representation

We represent face images in this experiment with multi-scale local binary patterns (MLBP), which is the concatenation of local binary patterns [99] of radius 1, 3, 5, and 7. Ahonen et al. first demonstrated the effectiveness of representing face images with LBP descriptors [3].

In order to represent a face with MLBP feature descriptors, the face is first geometrically normalized using the eye locations to (i) perform planar rotation so the angle between the eyes is 0 degrees, (ii) scale the face so the inter-pulilary distance between eyes is 75 pixels, and (iii) crop the face 250x200 pixels. Once geometrically normalized, MLBP feature descriptors are densely sampled from patches of size 24x24 across the face, with an overlap of 12 pixels. In total, this yields 285 MLBP descriptors representing the face. The size of the patch (24x24) was selected by using the training set to perform parameter validation.

To reduce the total feature vector size, principal component analysis (PCA) was performed on one half of the training set to learn a subspace for each of the 285 MLBP feature sampling locations. The second half of the training set was used to determine the minimum energy variation that needed to be retained without impacting face recognition performance. It was determined that 98% of the variance could be retained without impacting the recognition performance. The original MLBP descriptor is 236 dimensional ($4 \cdot 59$). After PCA dimensionality reduction, the descriptor size, on average, was reduced to 99 dimensions at each of the 285 sampling locations. After the dimensionality of the MLBP descriptor for each face patch was reduced, all descriptors are concatenated together, resulting in a feature vector of dimensionality

$d = 28,187$. Without this PCA step, the feature dimensionality would have been $67,260$.

4.3.2 Random Subspaces

A total of B random LDA subspaces Ψ_b are learned from B random samples of the d-dimensional feature space. The eigenvalues corresponding to each feature dimension extracted from the PCA step were used to weight the random sampling so that features with higher variation energy will have a larger likelihood of being selected. The benefit of this approach was confirmed by evaluation on the validation set. The number of features sampled with the weighted random sampling was controlled by the parameter ρ $(0 < \rho < 1)$ in order to select $d' = \rho \cdot d$ features at each stage $b = 1, \ldots, B$. Additionally, from the N training subjects available, a subset of size $N' < N$ was randomly sampled to build the between-class scatter matrix $S^b_{\text{Btwn}} \in \mathbb{R}^{d',d'}$ and the within-class scatter matrix $S^w_{\text{Wthn}} \in \mathbb{R}^{d',d'}$ at each stage b. Finally, we learn the subspace Ψ_b as

$$\Psi_b = \underset{\Psi'}{\text{argmax}} \frac{||\Psi'^{\text{T}} \cdot S^b_{\text{Btwn}} \cdot \Psi'||}{||\Psi'^{\text{T}} \cdot S^b_{\text{Wthn}} \cdot \Psi'||} \tag{4.1}$$

After learning the set of B subspaces $\Psi_b,\ b = 1 \ldots B$, a new face image is represented as the concatenation of the each of the B subspace projections. The dissimilarity between two faces is then measured by their L-2 norm.

Despite reducing the feature dimensionality and only using a ρ percent of the $(d' = \rho \cdot d)$ features, we still have a feature vector that is too large to accurately solve Eq. 4.1. To resolve this, a second PCA step was applied at each stage b to perform feature reduction on the d'-dimensional feature vector. This second PCA step was performed by retaining $0 < p < 1$ percent of the variance in the training instances at stage b.

The parameters in the RS-LDA framework are the number of training subjects to use at each stage (N'), the percentage of features to sample at each stage (ρ), the number of random sample stages (B), and the percentage of variance in the PCA step for each stage (p). Using the training set for validation to find the highest recognition accuracies, the following parameters values were selected: $N' = 300$, $\rho = 0.45$, $B = 20$, and $p = 0.95$.

4.4 Experiments

Figure 4.2 shows the negative correlation between face recognition accuracy and the amount of time lapse between probe and gallery image capture. A strong case has been made to handle this problem by training discriminative face recognition systems [76, 79]. Here we will use the random subspace framework developed in Section 4.3 to understand if training a face recognition system to improve performance on aging impacts the standard face recognition scenarios.

Using the training set splits discussed in Section 4.2, we trained five different versions of RS-LDA matcher using the algorithm presented in Section 4.3.

- The first RS-LDA matcher was trained on the 8,000 training subjects with 0 to 1 year time lapse between probe and gallery image.
- The second matcher was trained on 8,000 subjects with 1 to 5 year time lapse.
- The third matcher was trained on 8,000 subjects with 5 to 10 year time lapse.
- A fourth matcher was trained on 2,000 subjects with over 10 years times lapse (again, only 2,000 subjects were available with such a large time lapse).
- A final matcher was trained using 8,000 subjects whose time lapse was equally distributed amongst the four time lapse splits considered above. Thus, this

Test set: 0 to 1 year time lapse

RS-LDA trained on (time lapse in years):					Baselines:		
(0-1)	(1-5)	(5-10)	(10+)	(All)	MLBP Only	COTS1	COTS2
94.5%	94.1%	93.1%	91.8%	94.1%	71.2%	96.3%	89.8%

# of Match Comparions: *19,996*
# of Non-Match Comparions: *239,572,034*

(a)

Test set: 1 to 5 year time lapse

RS-LDA trained on (time lapse in years):					Baselines:		
(0-1)	(1-5)	(5-10)	(10+)	(All)	MLBP Only	COTS1	COTS2
90.3%	**90.5%**	89.1%	87.7%	90.2%	62.9%	94.3%	84.6%

# of Match Comparions: *33,443*
# of Non-Match Comparions: *401,282,557*

(b)

Test set: 5 to 10 year time lapse

RS-LDA trained on (time lapse in years):					Baselines:		
(0-1)	(1-5)	(5-10)	(10+)	(All)	MLBP Only	COTS1	COTS2
75.2%	81.2%	**82.0%**	80.4%	81.3%	46.7%	88.6%	75.5%

# of Match Comparions: *24,036*
# of Non-Match Comparions: *215,795,208*

(c)

Test set: 10+ year time lapse

RS-LDA trained on (time lapse in years):					Baselines:		
(0-1)	(1-5)	(5-10)	(10+)	(All)	MLBP Only	COTS1	COTS2
65.6%	72.2%	**72.4%**	71.0%	71.2%	39.2%	80.5%	61.7%

# of Match Comparions: *6,221*
# of Non-Match Comparions: *12,995,669*

(d)

Figure 4.3: The true accept rates (TAR) at a fixed false accept rate (FAR) of 1.0% across datasets with different amounts of time lapse between the probe and gallery images. Four different RS-LDA subspaces were trained on a separate set of subjects with the different time lapse ranges tested above. The results suggests the need for multiple recognition subspaces depending on the time lapse.

matcher trained on subjects with 0 year time lapse up to 17 years (17 years is the maximum time lapse in the 10+ aging set).

Figure 6.1 shows the accuracy on each of the four test sets using the five trained systems. The first test set (Fig. 6.1(a)) has 0 to 1 year time lapse between the probe and gallery images for 12,000 subjects. The results show that the best performance from the five trained systems is the system trained on 0 to 1 year time lapse. As the time lapse between the training set and the test set increases, the face recognition accuracy decreases. These results help provide the following answer to the question originally posed: *training a face recognition system to improve on face aging does seem to reduce its performance when facial aging has not occurred.*

The recognition performance on face images that have 1 to 5 years time lapse (Fig. 6.1(b)) shows the best performance from the five RS-LDA systems is the system trained on 1 to 5 year lapse. However, the performance from 0 to 1 year time lapse training is not much lower. In fact, the difference between training and testing on 0 to 1 year and 1 to 5 years is rather minimal. This is likely due the fact that only minor aging changes have occurred in these time spans.

The recognition performance on face images with 5 to 10 years time lapse (Fig. 6.1(c)) shows how learning can help improve recognition accuracies in the presence of a large amount of aging. The true accept rate improves by nearly 7.0% when trained on the 5 to 10 years set than with the 0 to 1 year training set. Thus, the feature subspaces learned on data with minimal aging did not generalize well to data with larger amounts of aging.

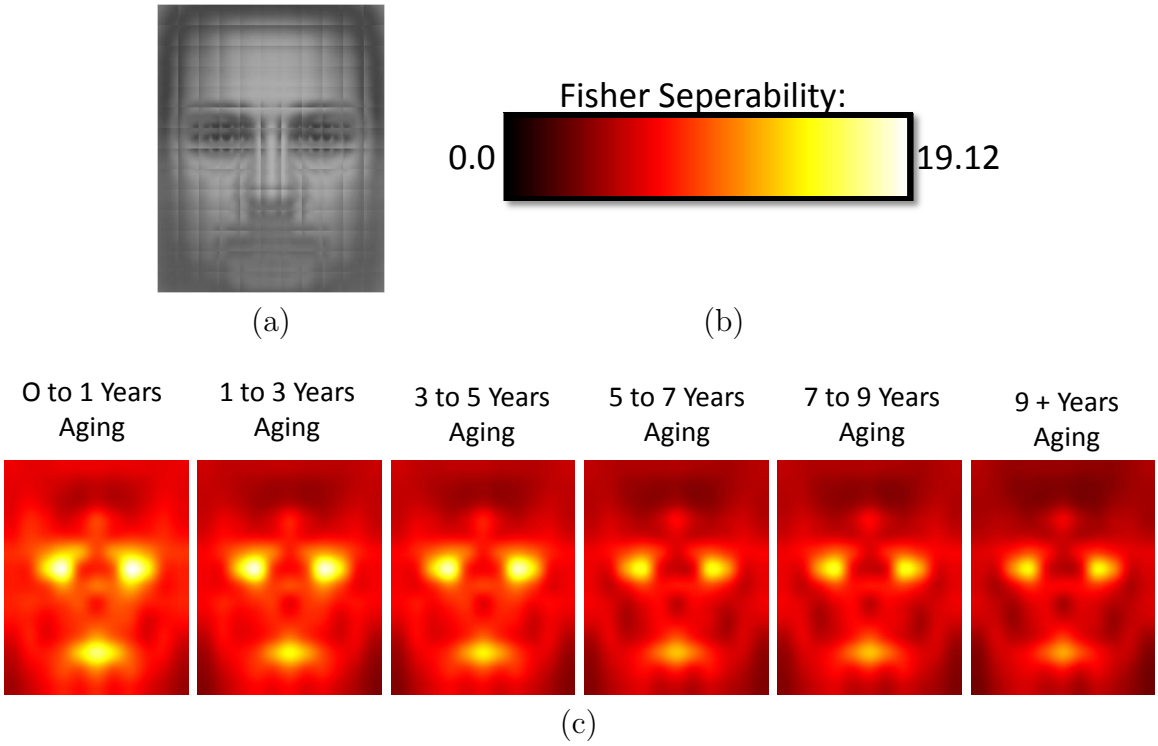
The recognition results on aging over 10 years (Fig. 6.1(c)) is the only scenario in which the subspace trained on the same time lapse as tested on did not offer the highest results. However, the 10+ year subspace only had 2,000 subjects to train on while the other subspaces had 8,000 subjects available for training. This could also be explained by the complex nature of face aging that manifests itself in different

ways for different individuals especially when the time lapse is large.

In each testing scenario the subspace labeled (All) is the one trained on 8,000 subjects exhibiting all the time lapse ranges considered above. While this subspace never had the top accuracy with respect to the other RS-LDA subspaces, it consistently performed well on all time lapses. This indicates that choosing training with equally distributed amount of time lapse is a viable solution when learning multiple subspace models is not reasonable.

The performance of COTS1 exceeded the RS-LDA system in each testing scenario. However, the RS-LDA system was purposely designed to be relatively simple to help facilitate the scope of this study. Incoporating additional features such as the SIFT descriptors and multiple patch sizes that Li et al. used in their aging-invariant recognition system [76] would result in improved performance. Despite this, the role of the training set in training RS-LDA is clearly established when examining the performance of the RS-LDA subspaces over the baseline MLBP only performance. MLBP only makes use of the initial MLBP feature representation to measure the (dis)similarity between the faces, but does not perform training. Through the use of RS-LDA the recognition accuracy is improved substantially.

The large time lapse dataset with a large number of subjects presented in this study also enabled us to examine which regions of the face remained the most persistent or retained the most discriminative power over time. To examine this stability, we measured the Fisher separability at each patch where the MLBP feature descriptors were computed. For a given face patch, we measured the Fisher separability as the ratio of the sum of eigenvalues from the between-class scatter to the sum of eigenvalues from the within-class scatter. This indicates the inherent separability provided by the Level 2 MLBP features at different regions of the face. These Fisher separability values at different time lapses are shown in Figure 4.4. The results show that while, as expected, the inherent separability decreases for each facial region as



(a) (b)

(c)

Figure 4.4: Inherent separability of different facial regions with aging. (a) The mean pixel values at each patch where MLBP feature descriptors are computed. (b) The scale of the Fisher separability criterion used. (c) The heat map showing Fisher separability values at each image patch across different time lapses. As time lapse increases, the eyes and mouth regions seem to be the most stable sources of identifiable information.

time increases, the mouth region has more discriminative information than the nose region, especially with the progression of time or aging. This also confirms the discriminative information contained in the region of the face around the eyes. Such information could be useful in explicitly assigning different weights to different face regions.

4.4.1 Computational Demands

Future work will attempt to leverage the additional face images contained in the 1.5 million mug shot image dataset available to us. However, one of the major difficulties we anticipate in this analysis is the computational demands for processing such a wide corpus of data. In this section we briefly highlight some of the challenges of processing a large scale face database.

In this study, each of the roughly 120,000 test images used were enrolled by the Cognitec's and PittPatt's FRS. After enrollment, 869 million match comparisons were performed by each matcher to measure the performance on each time lapse data set.

The analysis of RS-LDA on the MLBP feature representation used all the 200,000 images. This, in turn, required all images to be geometrically normalized using the eye locations automatically detected by the FaceVACS system. Once the images were aligned, the MLBP feature descriptors was extracted. With a 236-dimensional MLBP descriptor extracted at 285 patches across each face, roughly 48Gb of space was needed for storing these features.

For analyzing RS-LDA performance on each of the five time lapse training sets, a total of 869 million test set comparisons needed to be performed five times, resulting in a total of 4.34 billion comparisons. Other computational demands arose from the training of the RS-LDA subspaces on various sets of 8,000 subjects, perform-

ing parameter validation on four different parameter combinations[1] in the RS-LDA framework, and generating the ROC curves for each score matrix.

Machines with large amounts of RAM were also required to efficiently process the data. For example,the covariate analysis necessary for RS-LDA needed the MLBP features from all 8,000 subjects to be loaded into the main memory. For testing, a major bottleneck occurred loading the MLBP feature descriptors from each of the 12,000 subjects from disk. This made it necessary to keep the MLBP features in memory as each of the 20 random subspaces were being processed (as opposed to releasing the memory as each image was projected into one of the subspaces).

Efficient code design helped overcome some of these computational challenges. However, this study was primarily made possible by MSU's High Performance Computing Center (HPCC), which provides a cloud computing service where at times over 40 different compute nodes, each with over 10gb of RAM, were used at the same time to meet the computational demands of this study.

4.5 Conclusions

This chapter presents studies on the largest face aging dataset reported to date. These results demonstrate that (i) face recognition systems degrade monotonically as the time lapse between face images to be matched increases, (ii) training to improve face recognition performance in the presence of aging can lower its performance in non-aging scenarios, and (iii) the best performance on a particular amount of time lapse is achieved by training a system on that particular time lapse. Indeed, we see that face recognition across time lapse is similar to more traditional heterogeneous face recognition problems in that a different sets of feature subspaces are necessary to maximize the recognition accuracies. Similar to heterogeneous face recognition

[1]Recognition accuracies based on training on roughly 10,000 subjects and testing on 10,000 subjects was explored on over two hundred parameter combinations.

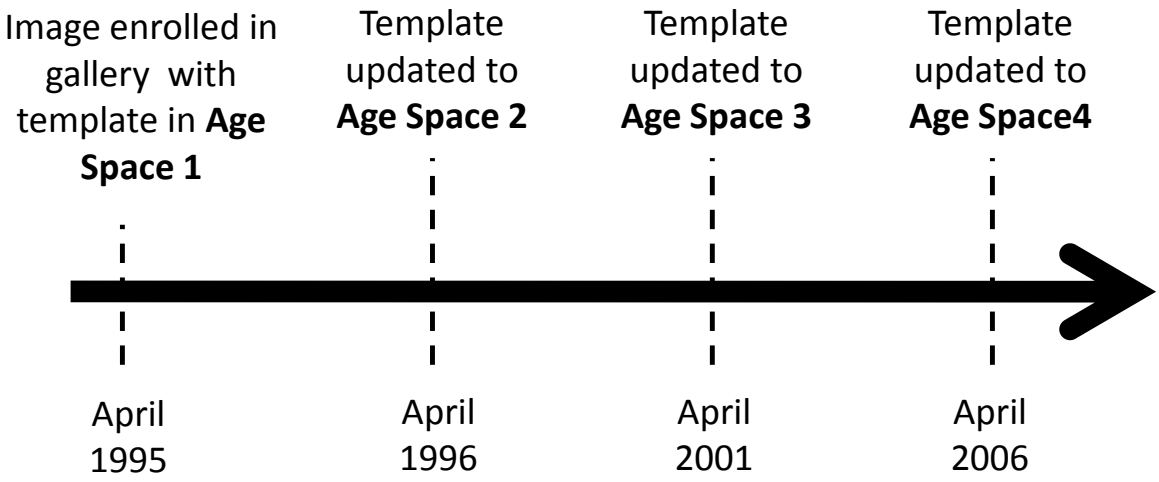


Figure 4.5: The ability to improve face recognition performance by training on the same time lapse being tested on suggests face recognition systems should update templates over time. For example, at fixed intervals from the original acquisition date the template is updated to reside in a subspaces trained for the time lapse that has occurred since acquisition. Probe images would be projected into each subspace and matched in the subspace corresponding to each gallery image.

between different image modalities, these feature subspaces do not generalize well to the more constrained case (i.e. minimal time lapse).

The findings presented in this chapter suggest a periodic update of face templates (see Figure 4.5). With a significant time lapse, updating the face template to reside in a subspace designed to capture the most discriminative features is likely to help improve the recognition performance in the presence of aging without compromising performance in cases where only a minimal amount of aging has occurred. Thus, much like heterogeneous face recognition between images from differing modalities, expanding face recognition algorithms to also handle different time lapses between face images requires multiple system configurations that are designed for specific recognition scenarios (e.g. matching faces with large amounts of aging, matching sketches, infrared images to photos).

Chapter 5

Face Recognition Performance: Role of Demographic Information

5.1 Introduction

In the previous chapter we examined heterogeneity with respect to differences in age of the same person. In this chapter we will examine heterogeneity with the respect to differences in the race/ethnicity, gender, and age of different persons. That is, previously we examined the impact of within-class demographic variations (namely, age). The chapter presents the complement study: an examination of the impact of between-class demographic variations.

As discussed in the first chapter, sources of errors in automated face recognition algorithms are generally attributed to the well studied variations in pose, illumination, and expression [108], collectively known as PIE. Other factors such as image quality (e.g., resolution, compression, blur), time lapse (facial aging), and occlusion also contribute to face recognition errors [43]. Previous studies have also shown within a specific demographic group (e.g., race/ethnicity, gender, age) that certain cohorts are more susceptible to errors in the face matching process [34, 111]. However, there has

yet to be a comprehensive study that investigates whether or not we can train face recognition algorithms to exploit knowledge regarding the demographic cohort of a probe subject.

This study presents a large scale analysis of face recognition performance on three different demographics (see Figure 5.1): (i) race/ethnicity, (ii) gender, and (iii) age. For each of these demographics, we study the performance of six face recognition algorithms belonging to three different types of systems: (i) three commercial off the shelf (COTS) face recognition systems (FRS), (ii) face recognition algorithms that do not utilize training data, and (iii) a trainable face recognition algorithm. While the COTS FRS algorithms leverage training data, we are not able to re-train these algorithms; instead they are black box systems that output a measure of similarity between a pair of face images. The non-trainable algorithms use common feature representations to characterize face images, and similarities are measured within these feature spaces. The trainable face recognition algorithm used in this study also outputs a measure of similarity between a pair of face images. However, different versions of this algorithm can be generated by training it with different sets of face images, where the sets have been separated based on demographics. Both the trainable algorithms, and (presumably) the COTS FRS, initially use some variant of the non-trainable representations.

The study of COTS FRS performance on each of the demographics considered is intended to augment previous experiments [34, 111] on whether these algorithms, as used in government and other applications, exhibit biases. Such biases would cause the performance of commercial algorithms to vary across demographic cohorts. In evaluating three different COTS FRS, we confirmed that not only do these algorithms perform worse on certain demographic cohorts, they consistently perform worse on the same cohorts (females, Blacks, and younger subjects).

Even though biases of COTS FRS on various cohorts were observed in this study,

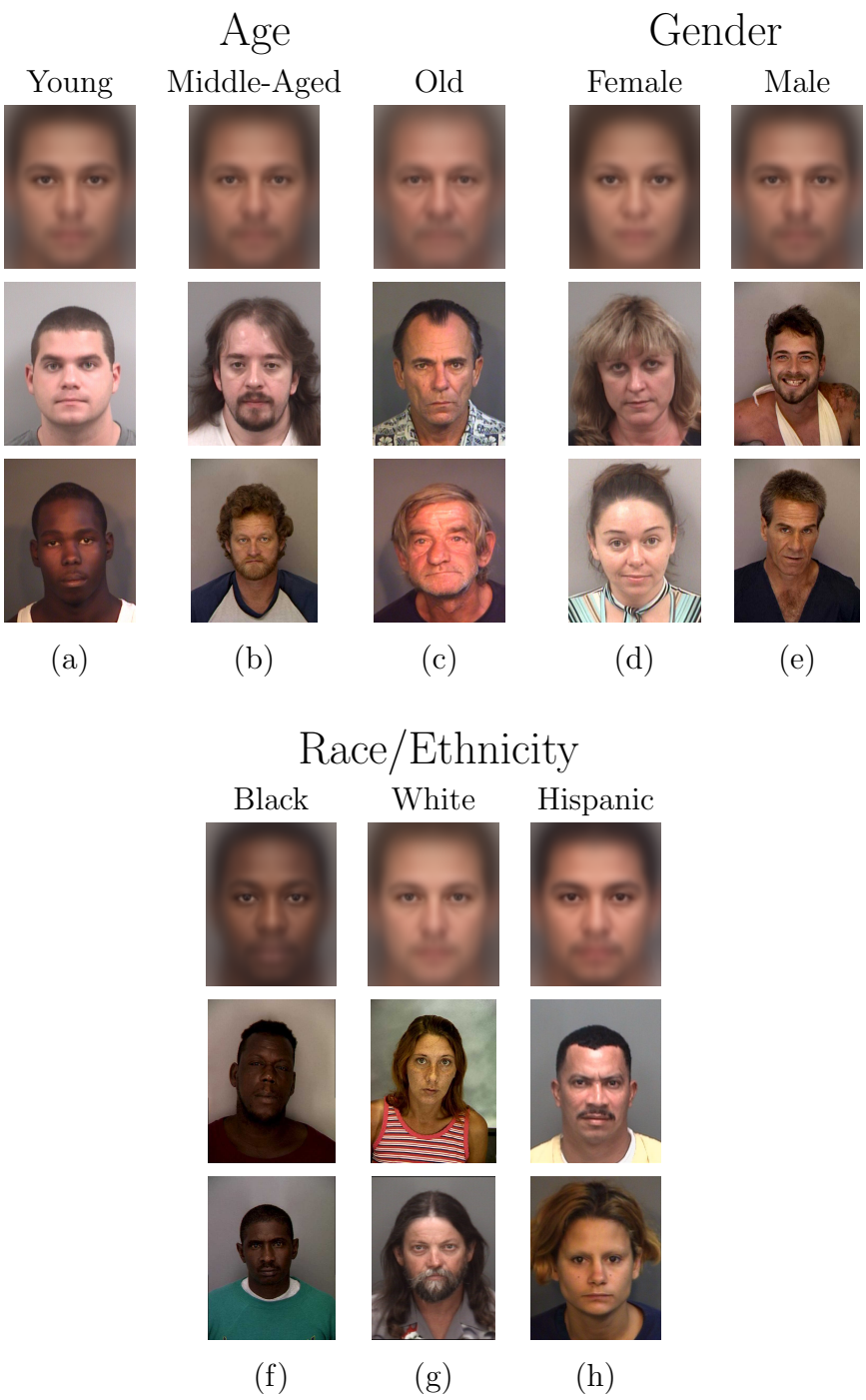


Figure 5.1: Examples of the different demographics studied. (a-c) Age demographic. (d-e) Gender demographic. (f-h) Race/ethnicity demographic. Within each demographic, the following cohorts were isolated: (a) ages 18 to 30, (b) ages 30 to 50, (c) ages 50 to 70, (d) female gender, (e) male gender, (f) Black race, (g) White race, and (h) Hispanic ethnicity. The first row shows the "mean face" for each cohort. A "mean face" is the average pixel value computed from all the aligned face images in a cohort. The second and third rows show different sample images within the cohorts.

these algorithms are black boxes that offer little insight into to why such errors manifest on specific demographic cohorts. To understand this, we also study the performance of non-commercial trainable and non-trainable face recognition algorithms, and whether statistical learning methods can leverage this phenomenon.

By studying non-trainable face recognition algorithms, we gain an understanding of whether or not the errors are inherent to the specific demographics. This is because non-trainable algorithms operate by measuring the (dis)similarity of face images based on a specific feature representation that, ideally, encodes the structure and shape of the face. This similarity is measured independent of any knowledge of how face images vary for the same subject and between different subjects. Thus, cases in which the non-trainable algorithms have the same relative performance within a demographic group as the COTS FRS indicates that the errors are likely due to one of the cohorts being inherently more difficult to recognize.

Relative differences in performance between the non-trainable algorithms and the COTS FRS indicate that the lower performance of COTS FRS on a particular cohort may be due to imbalanced training of the COTS algorithm. We explore this hypothesis by training the Spectrally Sampled Structural Subspace Features (4SF) face recognition algorithm [50] (i.e., the trainable face recognition algorithm used in this study) on image sets that consist exclusively of a particular cohort (e.g., White only). The learned subspaces in 4SF are applied to test sets from different cohorts to understand how unbalanced training with respect to a particular demographic impacts face recognition accuracy.

The 4SF trained subspaces also help answer the following question: to what extent can statistical learning improve accuracy on a demographic cohort? For example, it will be shown that females are more difficult to recognize than males. We will investigate how much training on only females, for example, can improve face recognition accuracy when matching females. Such improvements suggest the use of multiple dis-

criminative subspaces (or face recognition algorithms), with each trained exclusively on different cohorts. The results of these experiments indicate we can improve face recognition performance on the race/ethnicity cohort by using an algorithm trained exclusively on different demographic cohorts. This finding leads to the notion of *dynamic face matcher selection*, where demographic information may be submitted in conjunction with a probe image in order to select the face matcher trained on the same cohort. This framework, illustrated in Figure 5.2, should lead to improved face recognition accuracies.

The remainder of this chapter is organized as follows. In Section 5.2 we discuss previous studies on demographic introduced biases in face recognition algorithms and the design of face recognition algorithms. Section 5.3 discusses the data corpus that was utilized in this study. Section 5.4 identifies the different face recognition algorithms that were used in this study (commercial systems, trainable and non-trainable algorithms). Section 6.7 describes the matching experiments conducted on each demographic. Section 5.6 provides analysis of the results in each experiment and summarizes the contributions of this chapter.

5.2 Prior Studies and Related Work

Over the last twenty years the National Institute of Standards and Technology (NIST) has run a series of evaluations to quantify the performance of automated face recognition algorithms. Under certain imaging constraints these tests have measured a relative improvement of over two orders of magnitude in performance over the last two decades [34]. Despite these improvements, there are still many factors known to degrade face recognition performance (e.g., PIE, image quality, aging). In order to maximize the potential benefit of face recognition in forensics and law enforcement applications, we need to improve the ability of face recognition to sort through facial

images more accurately and in a manner that will allow us to perform more specialized or targeted searches. Facial searches leveraging demographics represents one such avenue for performance improvement.

While there is no standard approach to automated face recognition, most face recognition algorithms follow a similar pipeline [73]: face detection, alignment, appearance normalization, feature representation (e.g., local binary patterns [99], Gabor features [151]), feature extraction [10, 148]), and matching [96]. Feature extraction generally relies on an offline training stage that utilizes exemplar data to learn improved feature combinations (such as feature subspaces). For example, variants of the linear discriminant analysis (LDA) algorithm [10, 148] use training data to compute between-class and within-class scatter matrices. Subspace projections are then computed to maximize the separability of subjects based on these scatter matrices.

This study examines the impact of training on face recognition performance. Without leveraging training data, face recognition algorithms are not able to discern between noisy facial features and facial features which offer consistent cues to a subject's identity. As such, automated face recognition algorithms are ultimately based on statistical models of the variance between individual faces. These algorithms seek to minimize the measured distance between facial images of the same subject, while maximizing the distance between the subject's images and those of the rest of the population. However, the feature combinations discovered are functions of the data used to train the recognition system. If the training set is not representative of the data a face recognition algorithm will be operating on, then the performance of the resulting system may deteriorate. For example, the most distinguishing features for Black subjects may differ from White subjects. As such, if a system was predominantly trained on White faces, and later operated on Black faces, the learned representation may discard information useful for discerning Black faces.

The observation that the performance of face recognition algorithms could suffer

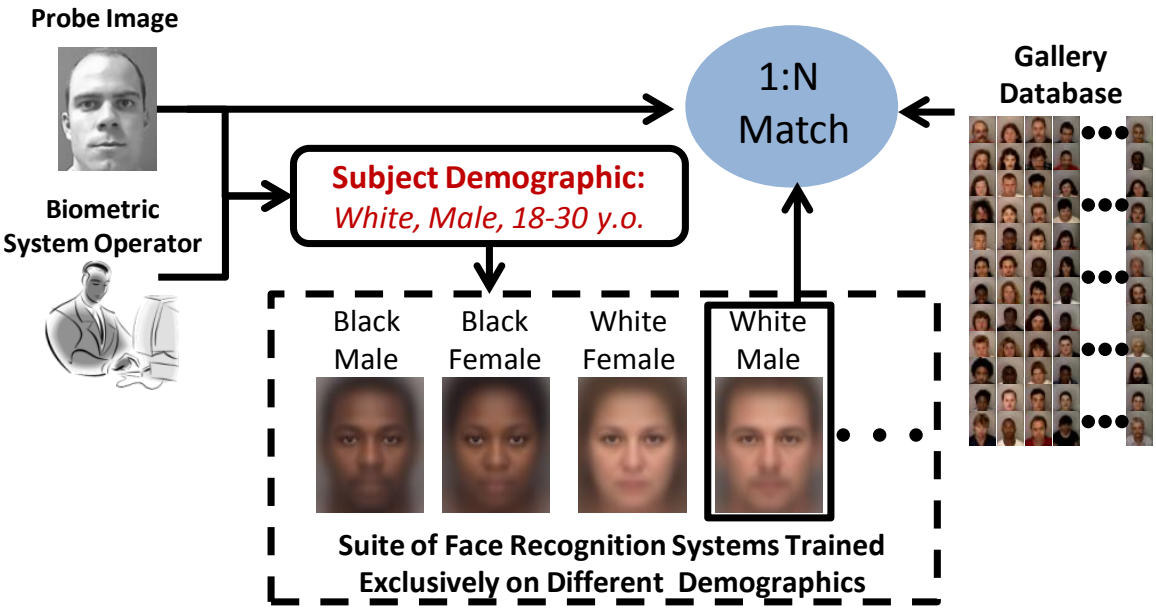


Figure 5.2: Dynamic face matcher selection. The findings in this study suggest that many face recognition scenarios may benefit from multiple face recognition systems that are trained exclusively on different demographic cohorts. Demographic information extracted from a probe image may be used to select the appropriate matcher, and improve face recognition accuracy.

if the training data is not representative of the test data is not new. One of the earliest studies reporting this phenomenon is not in the automated face recognition literature, but instead in the context of human face recognition. Coined the "other-race effect", humans have consistently demonstrated a decreased ability to recognize subjects from races different from their own [14, 127]. While there is no generally agreed upon explanation for this phenomenon, many researchers believe the decreased performance on other races is explained by the "contact" hypothesis, which postulates that the lower performance on other races is due to a decreased exposure [20]. While the validity of the contact hypothesis has been disputed [97], the presence of the "other-race effect" has not.

From the perspective of automated face recognition, Phillips et als findings in the 2002 government sponsored NIST Face Recognition Vendor Test (FRVT) is believed to be the first finding that face recognition algorithms have different recognition

accuracies depending on a subject's demographic cohort [111]. Among other findings, this study demonstrated for commercial face recognition algorithms on a dataset containing roughly 120,000 images that (i) female subjects were more difficult to recognize than male subjects, and (ii) younger subjects were generally more difficult to recognize than older subjects.

More recently, Grother et als measured the performance of seven commercial face recognition algorithms and three university face recognition algorithms in the 2010 NIST Multi-Biometric Evaluation [34]. The experiments conducted also concluded that females were more difficult to recognize than males. This study also measured the recognition accuracy of different races and ages.

Previous studies have investigated what impact the distribution of a training set has on recognition accuracy. Furl et als [28] and O'Toole et als [100] conducted studies to investigate the impact of cross training and matching on White and Asian races. Similar training biases were investigated by Klare and Jain [57], who showed that aging-invariant face recognition algorithms suffer from decreased performance in non-aging scenarios.

The study in [100] was motivated by a rather surprising result in the 2006 NIST Face Recognition Vendor Test (FRVT) [110]. In this test, the various commercial and academic face recognition algorithms tested exhibited a common characteristic: algorithms which originated in East Asia performed better on Asian subjects than did algorithms from the West. The reverse was true for White subjects: algorithms developed in the western hemisphere performed better. O'Toole et als suggested that this discrepancy was due to the different racial distribution in the training sets for the Western and Asian algorithms.

The impact of these training sets on face recognition algorithms cannot be overemphasized; face recognition algorithms do not generally rely upon explicit physiological models of the human face for determining match or non-match. Instead, the measure

of similarity between face images is based on statistical learning, generally in the feature extraction stage [10, 80] or during the matching stage [96].

In this work, we expand on previous studies to better demonstrate and understand the impact of a training set on the performance of face recognition algorithms. While previous studies [28, 100] only isolated the race variate, and only considered two races (i.e., Asian and White), this study explores both the inherent biases and training biases across gender, race (three different races/ethnicities) and age. To our knowledge, no studies have investigated the impact of gender or subject age for training face recognition algorithms.

5.3 Face Database

This study was enabled by a collection of over one million mug shot face images from the Pinellas County Sheriff's Office[1] (examples of these images can be found in Figure 5.1). Accompanying these images are complete subject demographics. The demographics provide the race/ethnicity, gender, and age of the subject in each image, as well as a subject ID number.

Given this large corpus of face images, we were able to use the metadata provided to control the three demographics studied: race/ethnicity, gender, and age. For gender, we partitioned image sets into cohorts of (i) male only, and (ii) female only. For age, we partitioned the sets into three cohorts: (i) young (18 to 30 years old), (ii) middle-age (30 to 50 years old), and (iii) old (50 to 70 years old). There were very few individuals in this database with age less than 18 and older than 70. For race/ethnicity[2], we partitioned the sets into cohorts of (i) White, (ii) Black, and (iii)

[1]The mug shot data used in this study was acquired in the public domain through Florida's "Sunshine" laws. Subjects shown in this manuscript may or may not have been convicted of a criminal charge, and thus should be presumed innocent of any wrongdoing.

[2]Racial identifiers (i.e. White, Black, and Hispanic) follow the FBI's National Crime Information Center code manual.

Table 5.1: Number of subjects used for training and testing for each demographic category. Two images per subject were used. Training and test sets were disjoint. A total of 102,942 face images were used in this study.

Demographic	Cohort	# Training	# Testing
Gender	Female	7995	7996
	Male	7996	7998
Race	Black	7993	7992
	White	7997	8000
	Hispanic	1384	1425
Age	18 to 30	7998	7999
	30 to 50	7995	7997
	50 to 70	2801	2853

Hispanic[3]. A summary of these cohorts and the number of subjects available for each cohort can be found in Table 5.1. Asian, Indian, and Unknown race/ethnicities were not considered because an insufficient number of samples were available.

For each of the eight cohorts (i.e., male, female, young, middle-aged, old, White, Black, and Hispanic), we created independent training and test sets of face images. Each set contains a maximum of 8,000 subjects, with two images (one probe and one gallery) for each subject. Table 5.1 lists the number of subjects included for each set. Cohorts far less than 8,000 subjects (i.e., Hispanic and older) reflect a lack of data available to us. Cases with cohorts containing only slightly fewer than 8,000 subjects are the result of removing a few images that could not be successfully enrolled in the COTS FRS.

The dataset of mug shot images did not contain a large enough number of Asian subjects to measure that particular race/ethnicity cohort. However, studies by Furl et al. [28] and O'Toole et al. [100] investigated the impact of the Whites and East Asians. As previously discussed, these studies concluded that algorithms developed

[3]Hispanic is not technically a race, but instead an ethnic category.

in the Western Hemisphere did better on White subjects and Asian algorithms did better on Asian subjects.

5.4 Face Recognition Algorithms

In this section we will discuss each of the six face recognition algorithms used in this study. We have organized these algorithms into commercial algorithms (Sec. 5.4.1), non-trainable algorithms (Sec. 5.4.2), and trainable algorithms (Sec. 5.4.3).

5.4.1 Commercial Face Recognition Algorithms

Three commercial face recognition algorithms were evaluated in this study: (i) Cognitec's FaceVACS v8.2, (ii) PittPatt v5.2.2, and (iii) Neurotechnology's MegaMatcher v3.1. The results in this study obfuscate the names of the three commercial matchers.

These commercial algorithms are three of the ten algorithms evaluated in the NIST sponsored Multi-Biometrics Evaluation (MBE) [34]. As such, these algorithms are representative of the state of the art performance in face recognition technology.

5.4.2 Non-Trainable Face Recognition Algorithms

Two non-trainable face recognition algorithms were used in this study: (i) local binary patterns (LBP), and (ii) Gabor features. Both of these methods operate by representing the face with Level 2 facial features (LBP and Gabor), where Level 2 facial features are features that encode the structure and shape of the face, and are critical to face recognition algorithms [56].

These non-trainable algorithms perform an initial geometric normalization step (also referred to as alignment) by using the automatically detected eye coordinates (eyes were detected using FaceVACS SDK) to scale, rotate, and crop a face image. After this step, the face image has a height and width of 128 pixels. Both algorithms

are custom implementations by the authors.

Local Binary Patterns

A seminal method in face recognition is the use of local binary patterns [99] (LBP) to represent the face [3]. Local Binary Patterns are Level 2 features that represent small patches across the face with histograms of binary patterns that encode the structure and texture of the face.

Local binary patterns describe each pixel using a p-bit binary number. Each bit is determined by sampling p pixel values at uniformly spaced locations along a circle of radius r, centered at the pixel being described. For each sampling location, the corresponding bit receives the value 1 if it is greater than or equal to the center pixel, and 0 otherwise.

A special case of LBP, called the uniform LBP [99], is generally used in face recognition. Uniform LBP assigns any non-uniform binary number to the same value, where uniformity is defined by whether more than u transitions between the values 0 and 1 occur in the binary number. In the case of $p = 8$ and $u = 2$, the uniform LBP has 58 uniform binary numbers, and the 59th value is reserved for the remaining $256 - 58 = 198$ non-uniform binary numbers. Thus, each pixel will take on a value ranging from 1 to 59. Two different radii are used ($r = 1$ and $r = 2$), resulting in two different local binary pattern representations that are subsequently concatenated together (called Multi-scale Local Binary Patterns, or MLBP).

In the context of face recognition, LBP values are first computed at each pixel in the (normalized) face image as previously described. The image is tessellated into patches with a height and width of 12 pixels. For each patch i, a histogram of the LBP values $S'_i \in \mathbb{Z}^{d_s}$ is computed (where $d_s = 59$). This feature vector is then normalized to the feature vector $S_i \in \mathbb{R}^{d_s}$ by $S_i = \frac{S'_i}{\sum_i^{d_s} S'_i}$. Finally, we concatenate the N vectors into a single vector x of dimensionality $d_s \cdot N$.

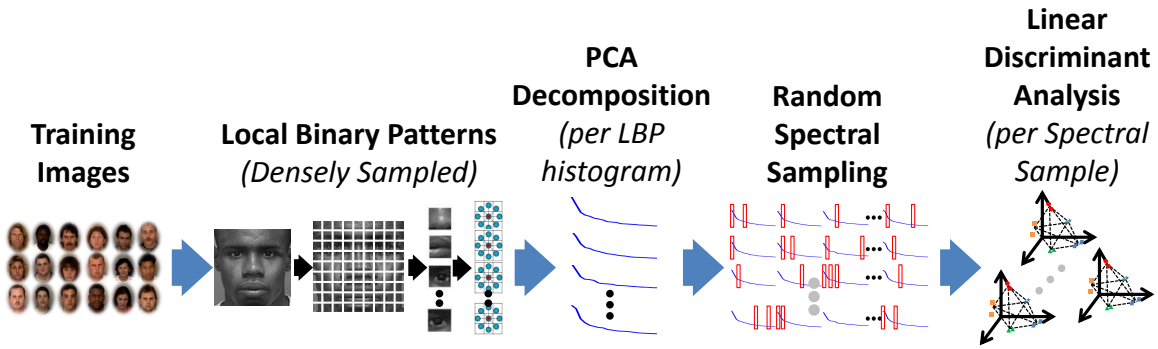


Figure 5.3: Overview of the Spectrally Sampled Structural Subspace Features (4SF) algorithm. This custom algorithm is representative of state of the art methods in face recognition. By changing the demographic distribution of the training sets input into the 4SF algorithm, we are able to analyze the impact the training distribution has on various demographic cohorts.

In our implementation, the illumination filter proposed by Tan and Triggs [136] is used prior to computing the LBP codes in order to suppress non-uniform illumination variations. This filter resulted in improved recognition performance.

Gabor Features

Gabor features are one of the first Level 2 facial features [56] to have been used with wide success in representing facial images [80, 128, 151]. One reason Gabor features are popular for representing both facial and natural images is their similarity with human neurological receptor fields [94, 122].

A Gabor image representation is computed by convolving a set of Gabor filters with an image (in this case, a face image). The Gabor filters are defined as

$$G(x, y, \theta, \eta, \gamma, f) = \frac{f^2}{\pi\gamma\eta} e^{-\left(\frac{f^2}{\gamma^2}x'^2 + \frac{f^2}{\gamma^2}y'^2\right)} e^{\left(j2\pi f x'\right)} \tag{5.1}$$

$$x' = x\cos\theta + y sin\theta \tag{5.2}$$

$$y' = -x sin\theta + y\cos\theta \tag{5.3}$$

where f sets the filter scale (or frequency), θ is the filter orientation along the major axis, γ controls the filter sharpness along the major axis, and η controls the sharpness along the minor axis. Typically, combinations across the following values for the scale f and orientation θ are used: $f = \{0, 1, \ldots, 4\}$ and $\theta = \{\pi/8, \pi/4, 3\pi/8, \ldots, \pi\}$. This creates a set (or bank) of filters with different scales and orientations. Given the bank of Gabor filters, the input image is convolved with each filter, which results in a Gabor image for each filter. The combination of these scale and orientation values results in 40 different Gabor filters, which in turn results in 40 Gabor images (for example).

In this chapter, the recognition experiments using a Gabor image representation operate by: (i) performing illumination correction using the method proposed by Tan and Triggs [136], (ii) computing the phase response of the Gabor images with $f = \{1, 2\}$, and $\theta = 0, \pi/4, \pi/2, 3\pi/4$, (iii) tessellating the Gabor image(s) into patches of size 12x12, (iv) quantizing the phase response (which ranges from 0 to 2π) into 24 values and computing the histogram within each patch, and (v) concatenating the histogram vectors into a single feature vector. Given two (aligned) face images, the distance between their corresponding Gabor feature vectors is used to measure the dissimilarity between the two face images.

5.4.3 Trainable Face Recognition Algorithm

The trainable algorithm used in this study is the Spectrally Sampled Structural Subspace Features algorithm [50], which is abbreviated as 4SF@. This algorithm uses multiple discriminative subspaces to perform face recognition. After geometric normalization of a face image using the automatically detected eye coordinates (eyes were detected using FaceVACS SDK), illumination correction is performed using the illumination correction filter presented by Tan and Triggs [136]. Face images are then represented using histograms of local binary patterns at densely sampled face

patches [3] (to this point, 4SF is the same as the non-trainable LBP algorithm described in Sec. 5.4.2). For each face patch, principal component analysis (PCA) is performed so that 98.0% of the variance is retained. Given a training set of subjects, multiple stages of weighted random sampling is performed, where the spectral densities (i.e., the eigenvalues) from each face patch are used for weighting. The randomly sampled subspaces are based on Ho's original method [37], however the proposed approach is unique in that the sampling is weighted based on the spectral densities. For each stage of random sampling, LDA [10] is performed on the randomly sampled components. The LDA subspaces are learned using subjects randomly sampled from the training set (i.e., bagging [15]). Finally, distance-based recognition is performed by projecting the LBP representation of face images into the per-patch PCA subspaces, and then into each of the learned LDA subspaces. The sum of the Euclidean distance in each subspace is the dissimilarity between two face images. The 4SF algorithm is summarized in Figure 5.3.

As shown in the experiments conducted in this study, the 4SF algorithm performs on par with several commercial face recognition algorithms. Because 4SF is initially the same approach as the non-trainable LBP matcher, the improvement in recognition accuracies (in this study) between the non-trainable LBP matcher and the 4SF algorithm clearly demonstrates the ability of 4SF to leverage training data. Thus, a high matching accuracy and the ability to leverage training data make 4SF an ideal face recognition algorithm to study the effects of training data on face recognition performance. The 4SF algorithm was developed in house.

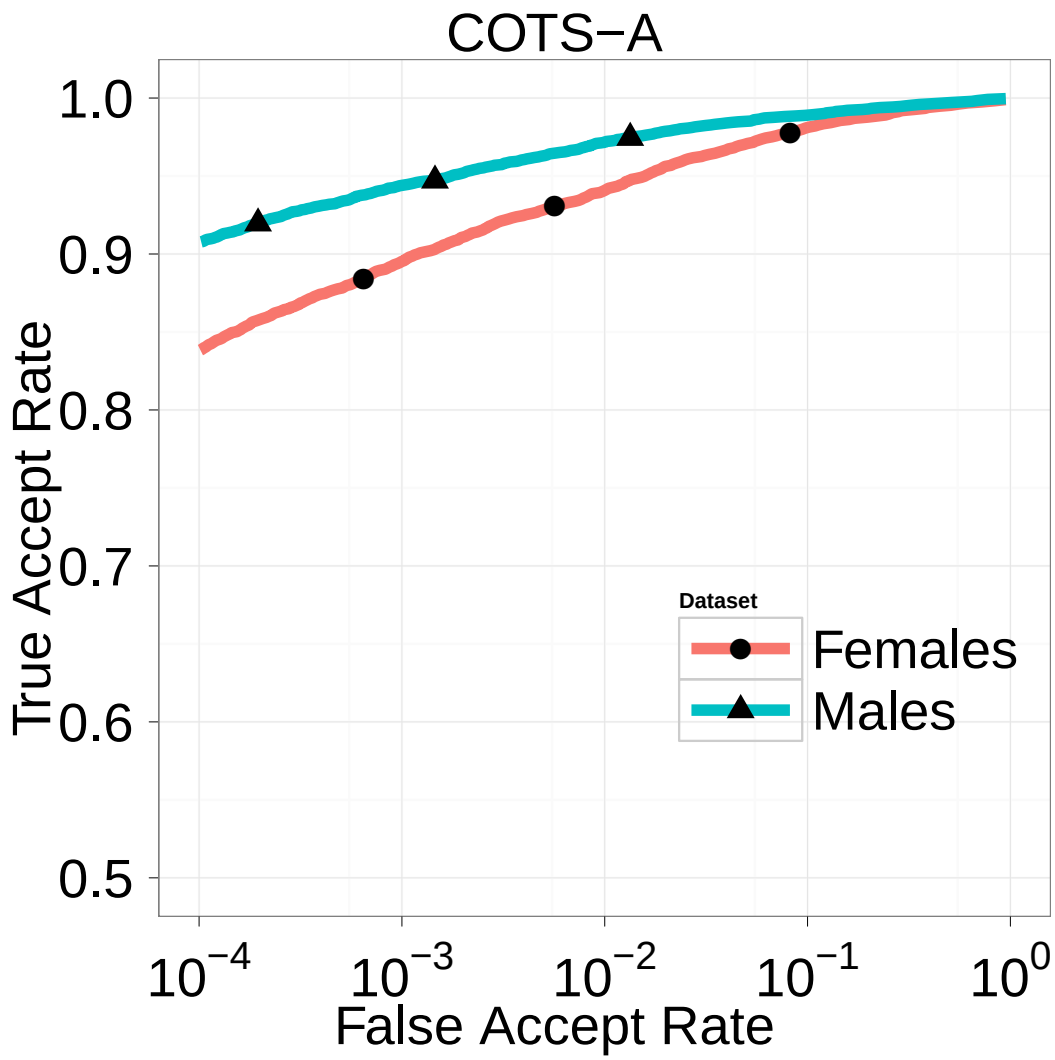


Figure 5.4: Performance of the COTS-A commerical face recognition system on datasets seperated by cohorts within the gender demographic.

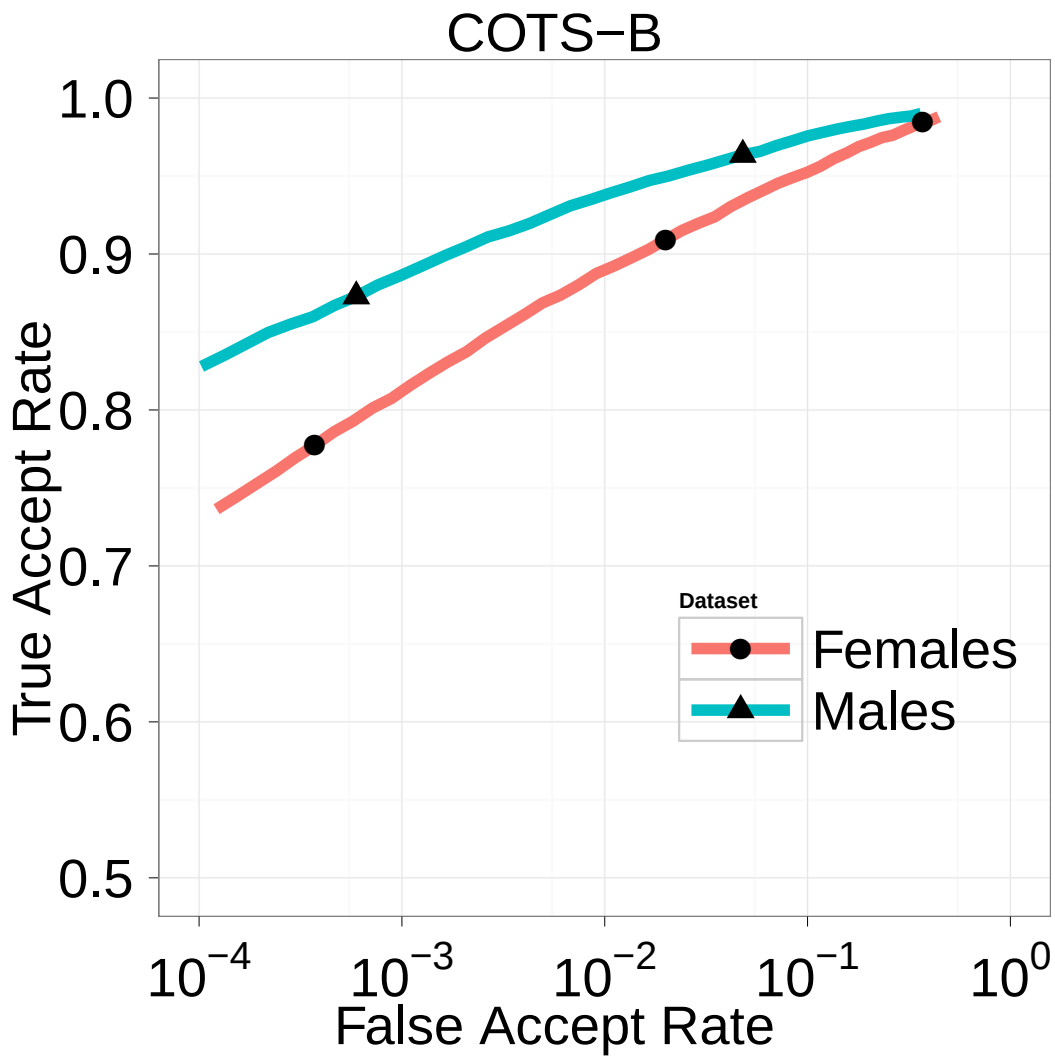


Figure 5.5: Performance of the COTS-B commerical face recognition system on datasets seperated by cohorts within the gender demographic.

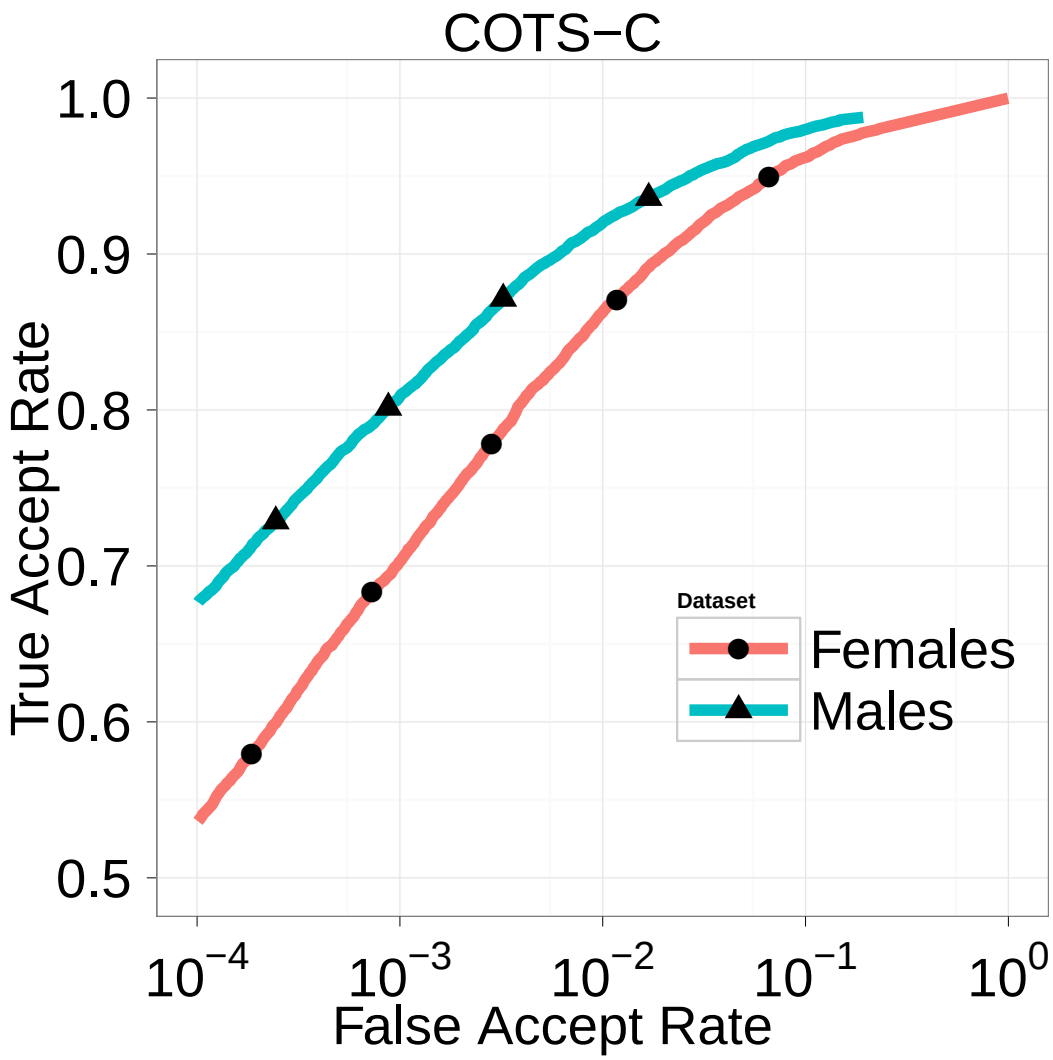


Figure 5.6: Performance of the COTS-C commerical face recognition system on datasets seperated by cohorts within the gender demographic.

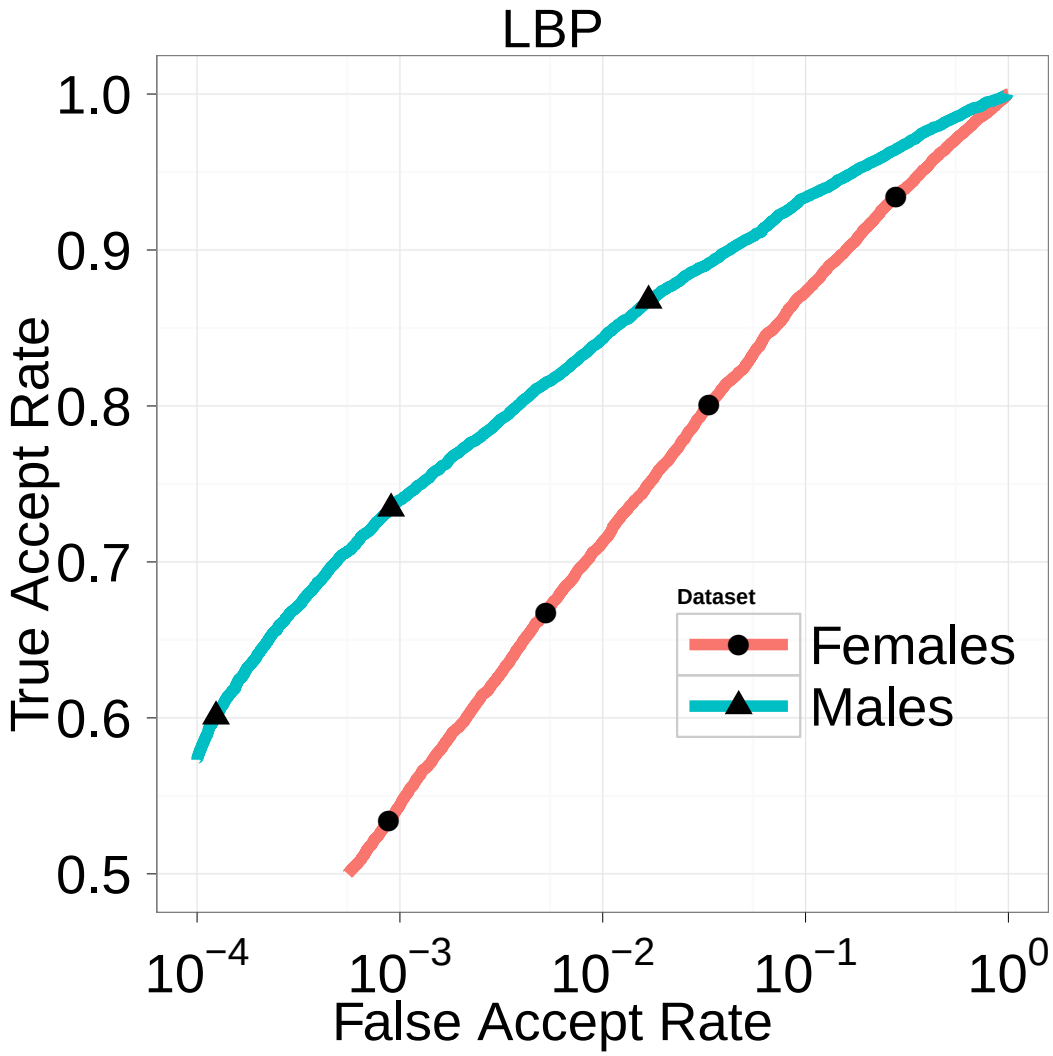


Figure 5.7: Performance of the local binary pattern-based non-trainable face recognition system on datasets seperated by cohorts within the gender demographic.

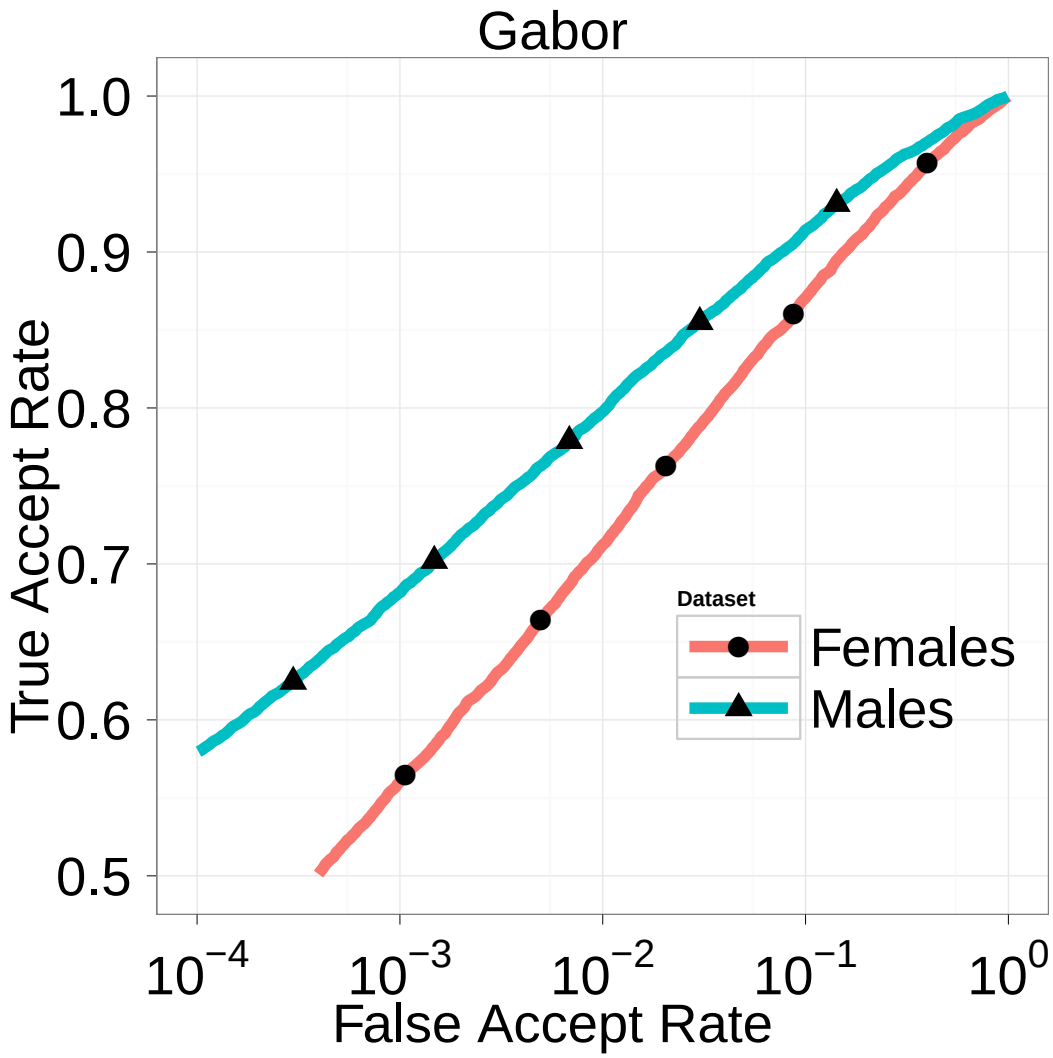


Figure 5.8: Performance of the Gabor-based non-trainable face recognition system on datasets seperated by cohorts within the gender demographic.

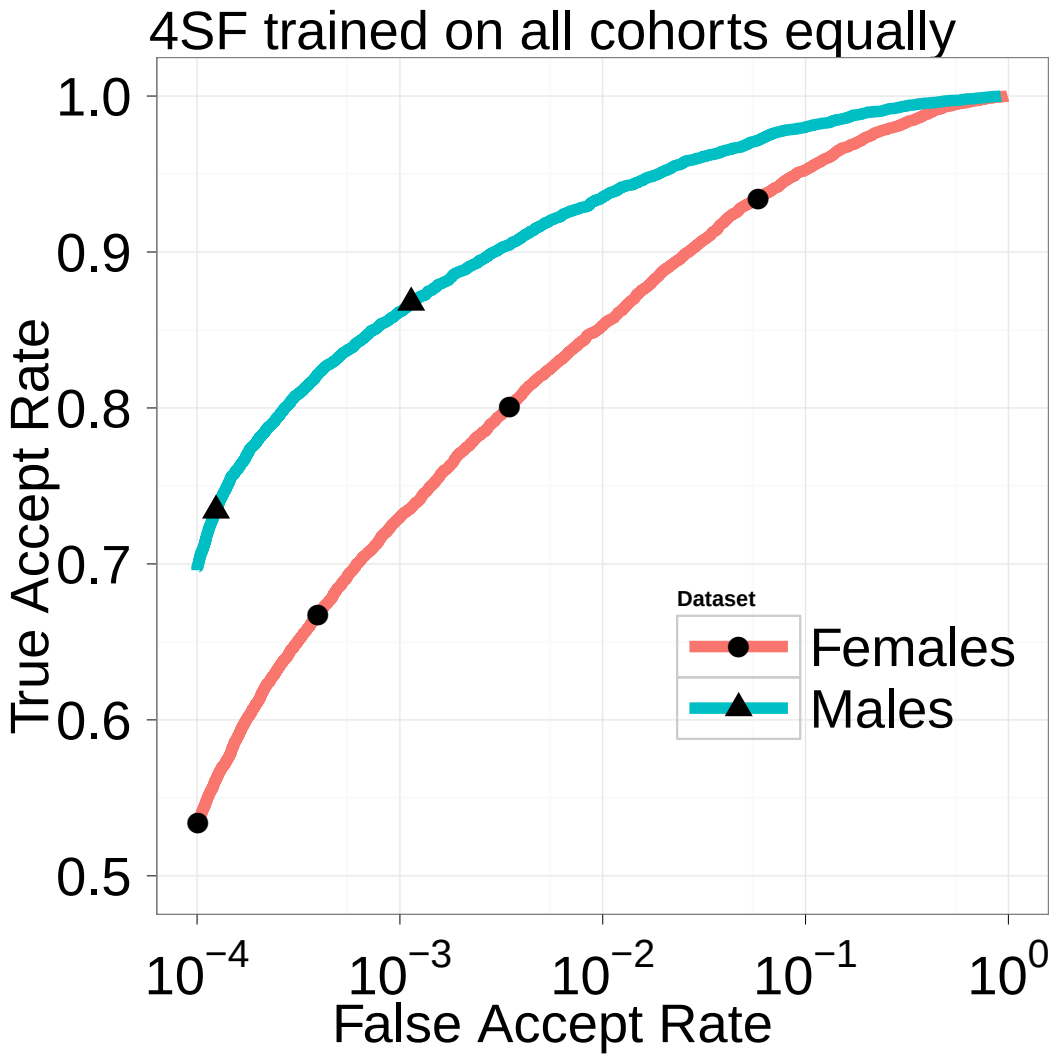


Figure 5.9: Performance of the 4SF algorithm trained on an equal number of samples from each gender on datasets seperated by cohorts within the gender demographic.

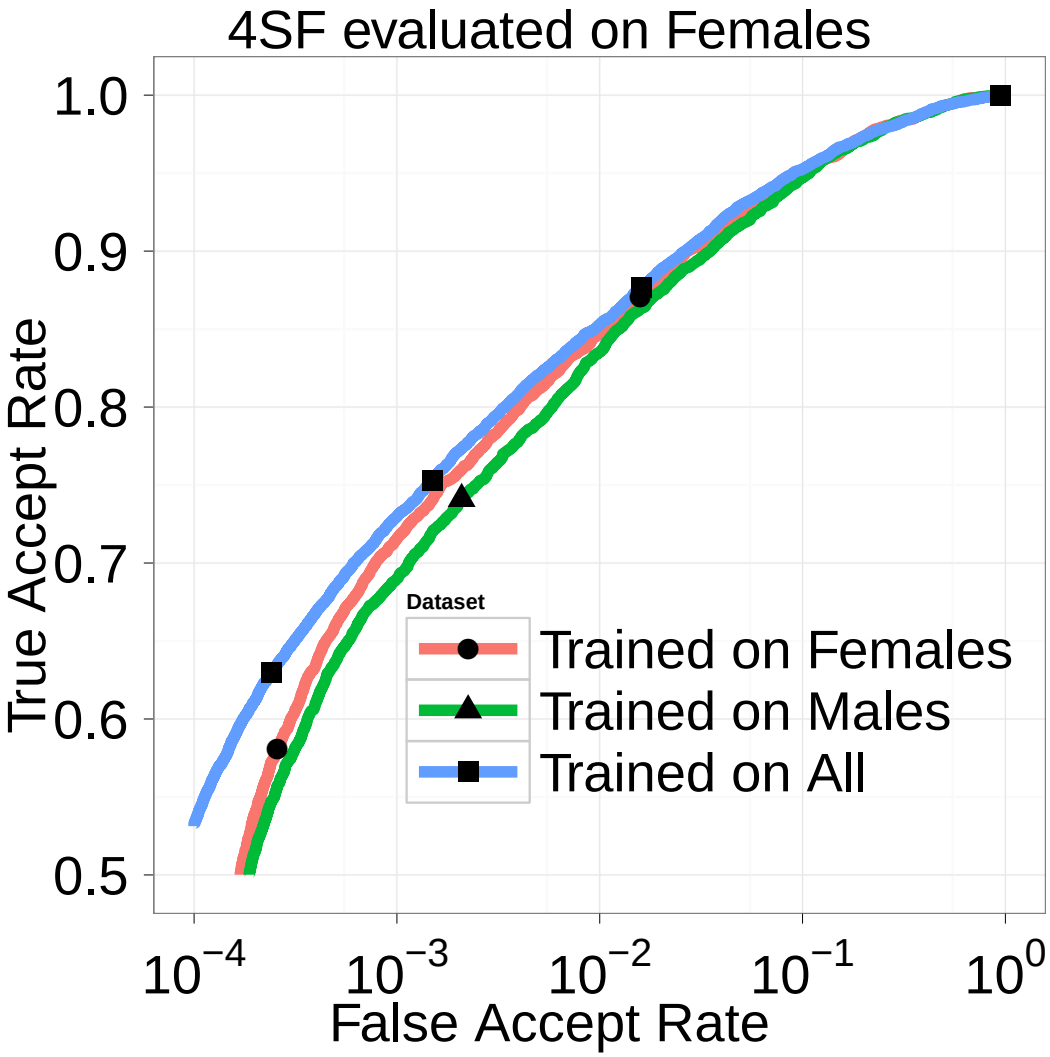


Figure 5.10: Performance of the different trained versions of the 4SF algorirthm on the Females cohort.

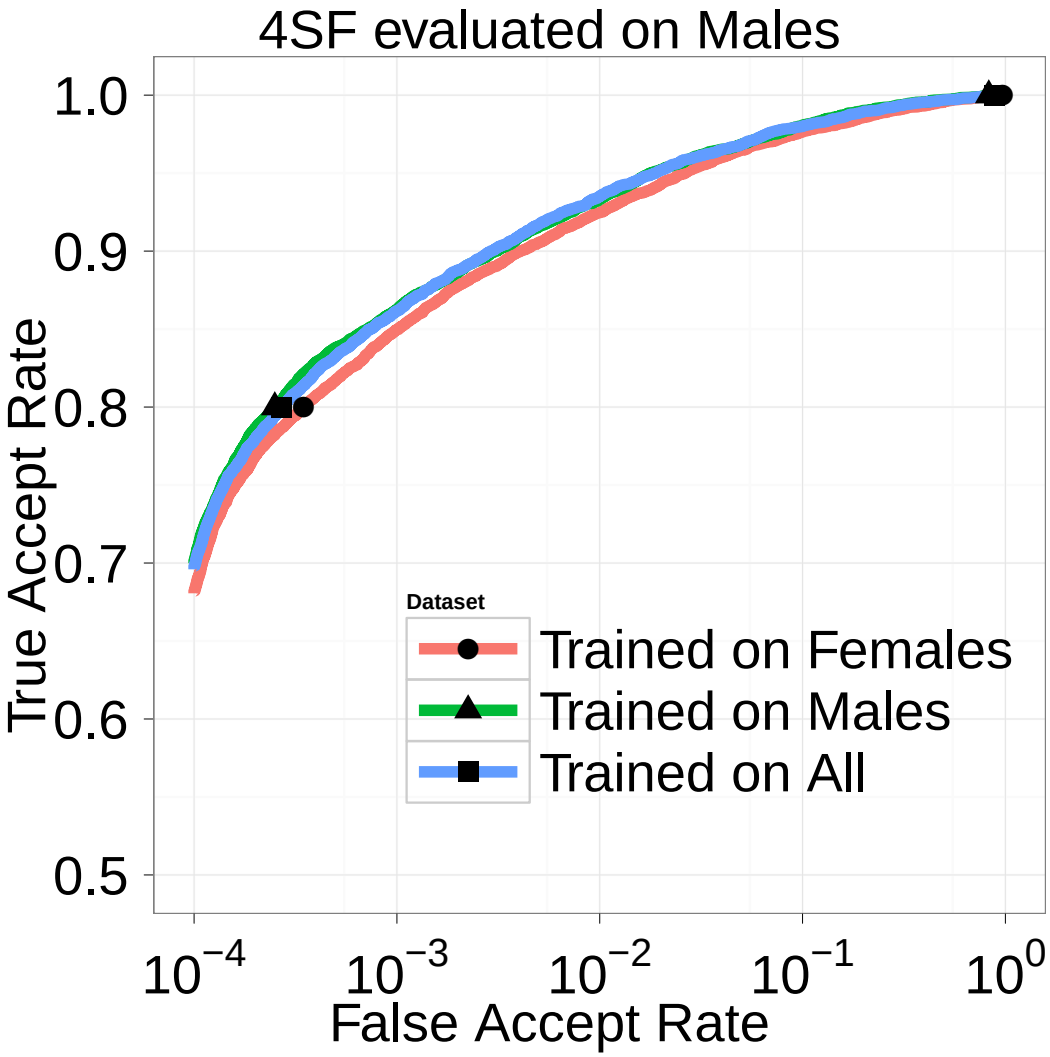


Figure 5.11: Performance of the different trained versions of the 4SF algorithm on the Male cohort.

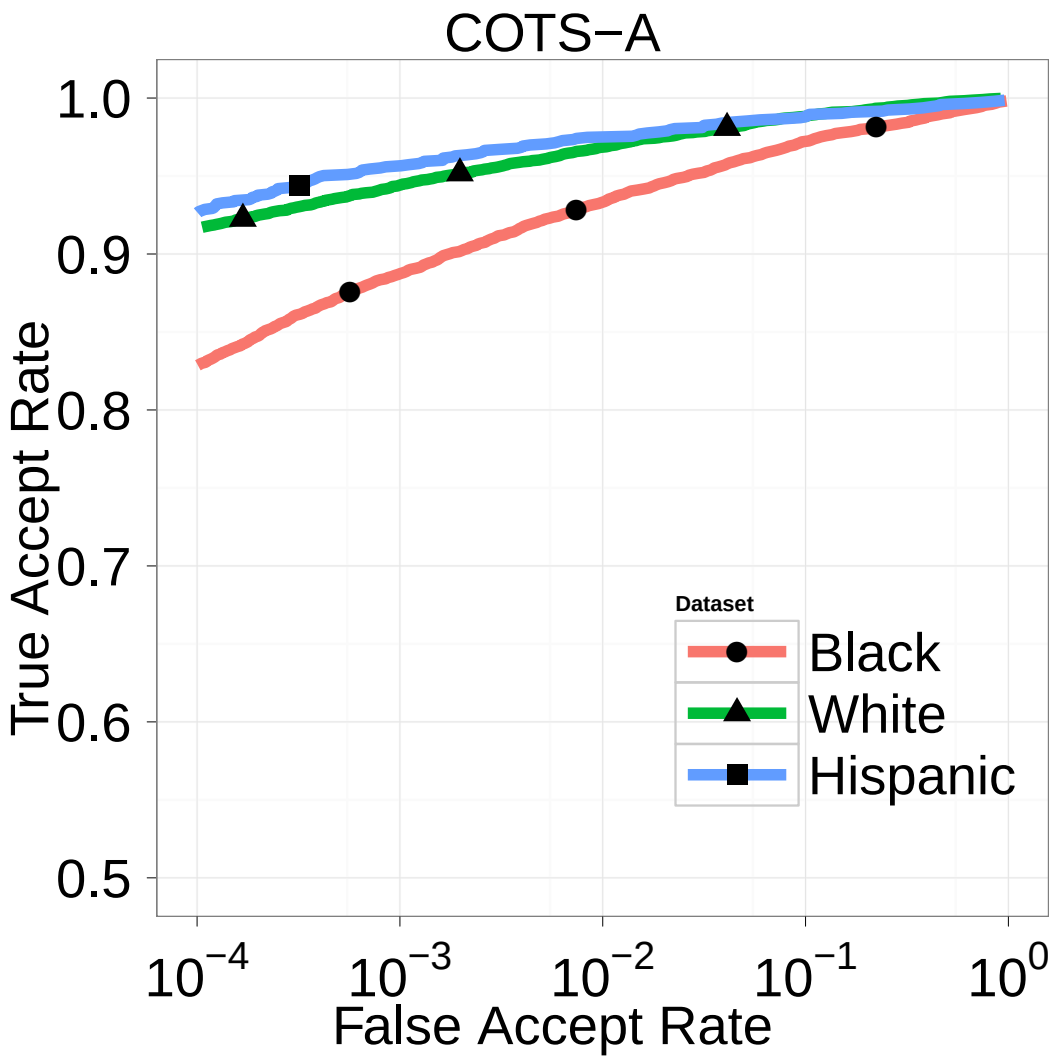


Figure 5.12: Performance of the COTS-A commercial face recognition system on datasets seperated by cohorts within the race demographic.

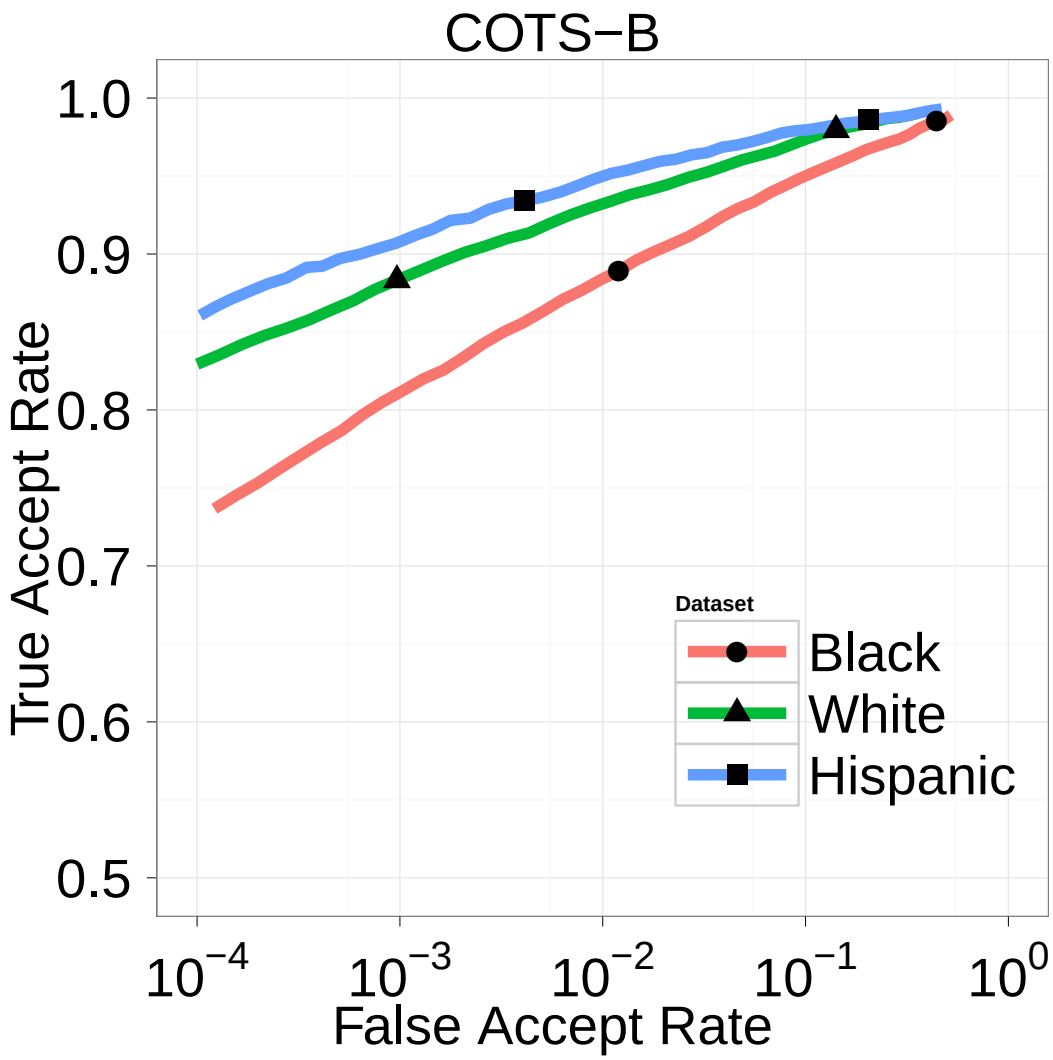


Figure 5.13: Performance of the COTS-B commercial face recognition system on datasets seperated by cohorts within the race demographic.

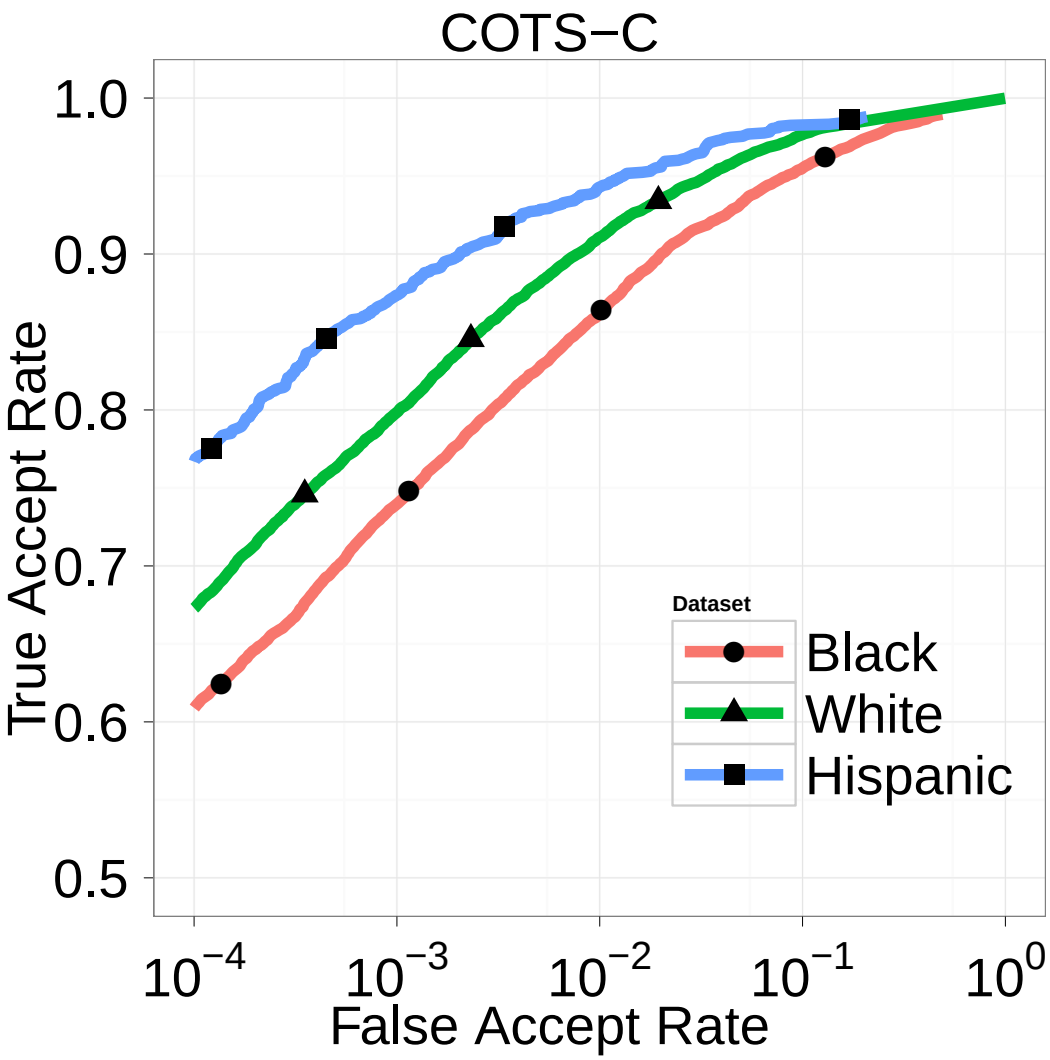


Figure 5.14: Performance of the COTS-C commercial face recognition system on datasets seperated by cohorts within the race demographic.

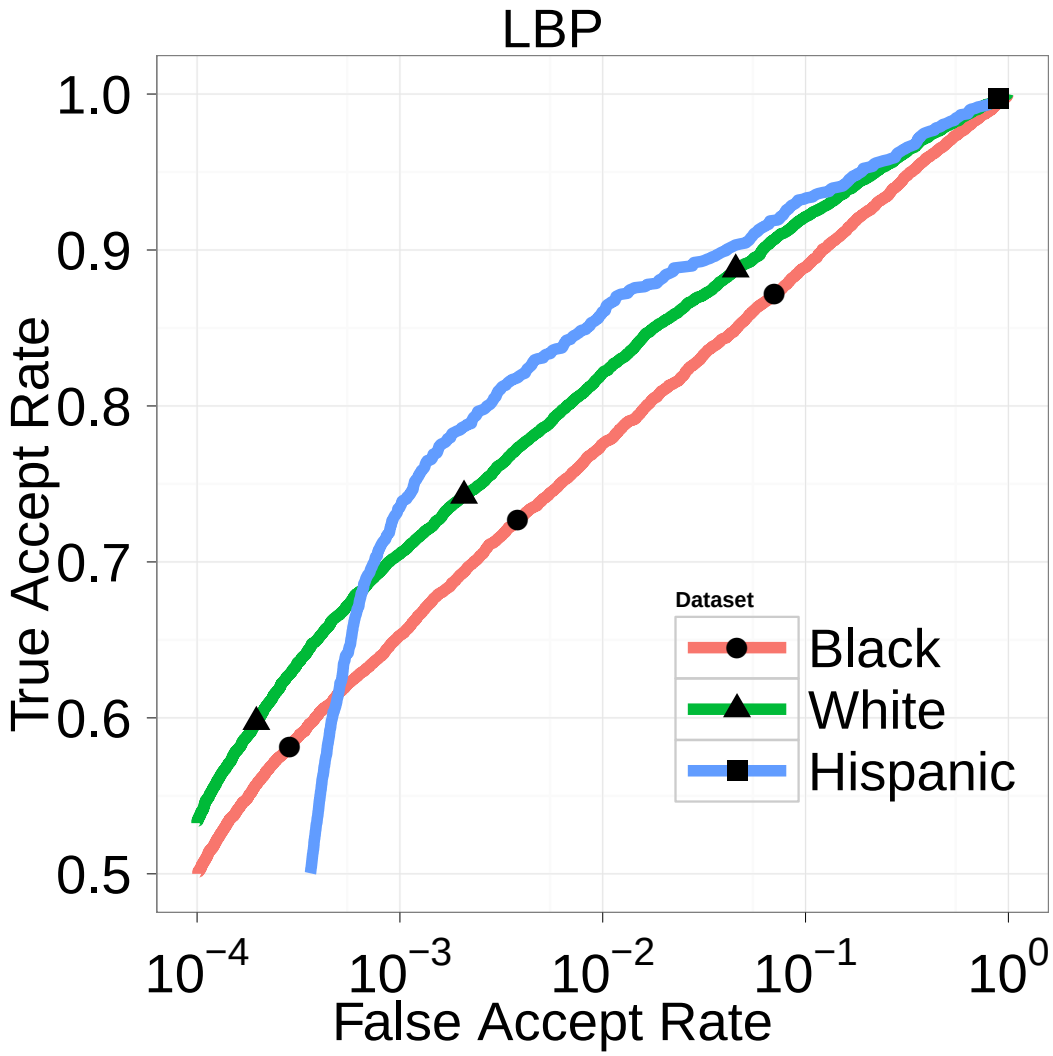


Figure 5.15: Performance of the local binary pattern-based non-trainable recognition system on datasets seperated by cohorts within the race demographic.

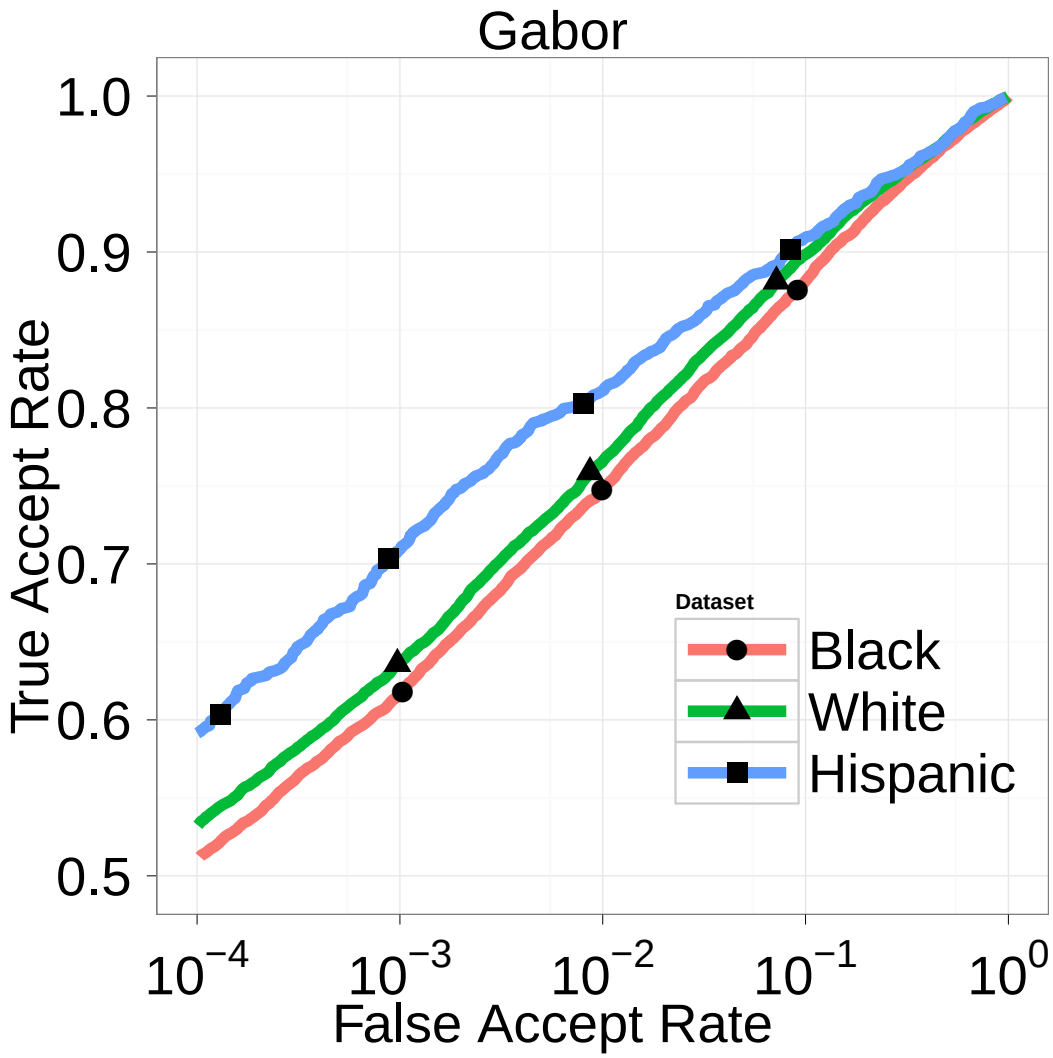


Figure 5.16: Performance of the Gabor-based non-trainable recognition system on datasets seperated by cohorts within the race demographic.

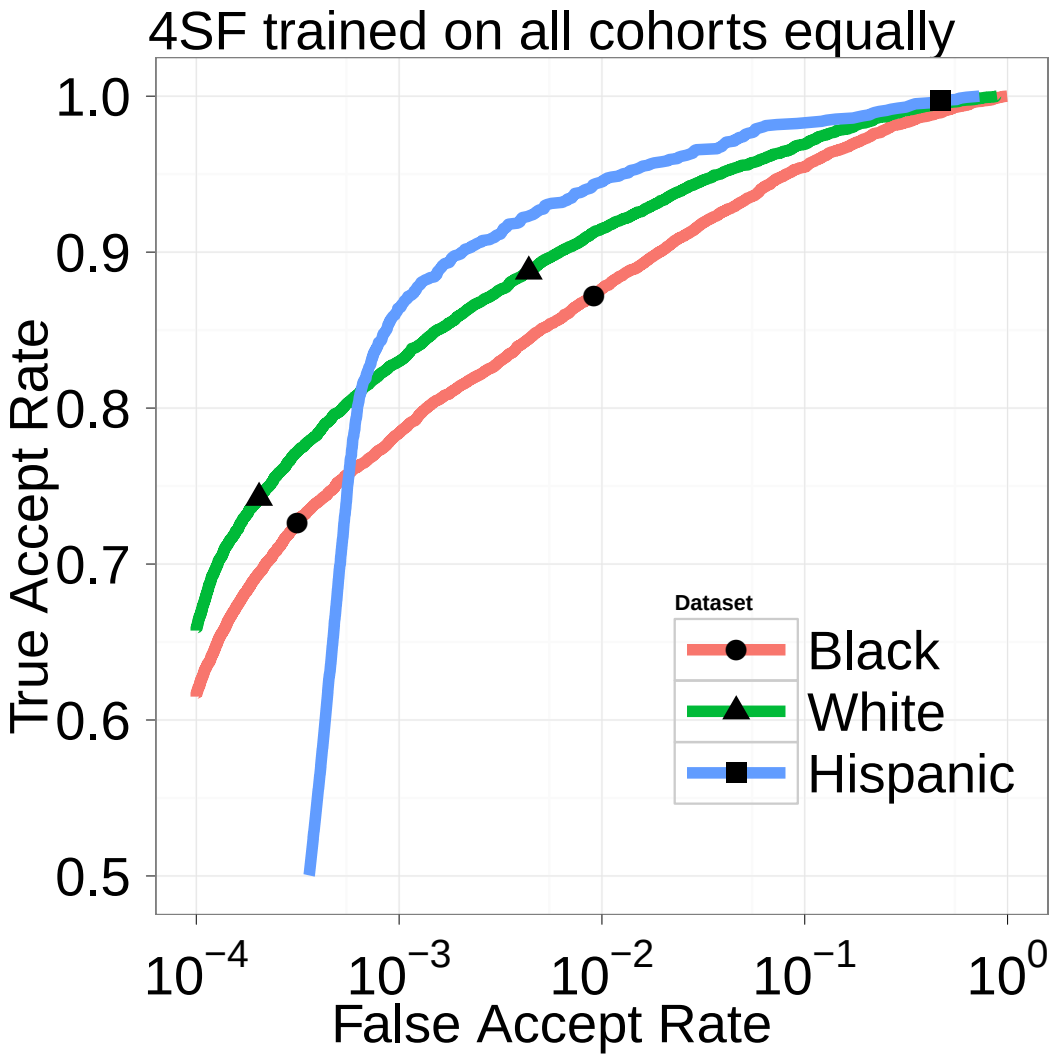


Figure 5.17: Performance of the 4SF algorithm trained on an equal number of samples from each race on datasets seperated by cohorts within the race demographic.

5.5 Experimental Results

For each demographic (gender, race/ethnicity, and age), three separate matching experiments are conducted. The results of these experiments are presented per demographic. Figures 5.4 to 5.11 delineate the results for all the experiments on the gender demographic. Figures 5.12 to 5.20 delineate the results for all experiments on the race/ethnicity demographic. Finally, Figures 5.21 to 5.29 delineate the results for all experiments on the age demographic. The true accept rate at a fixed false accept

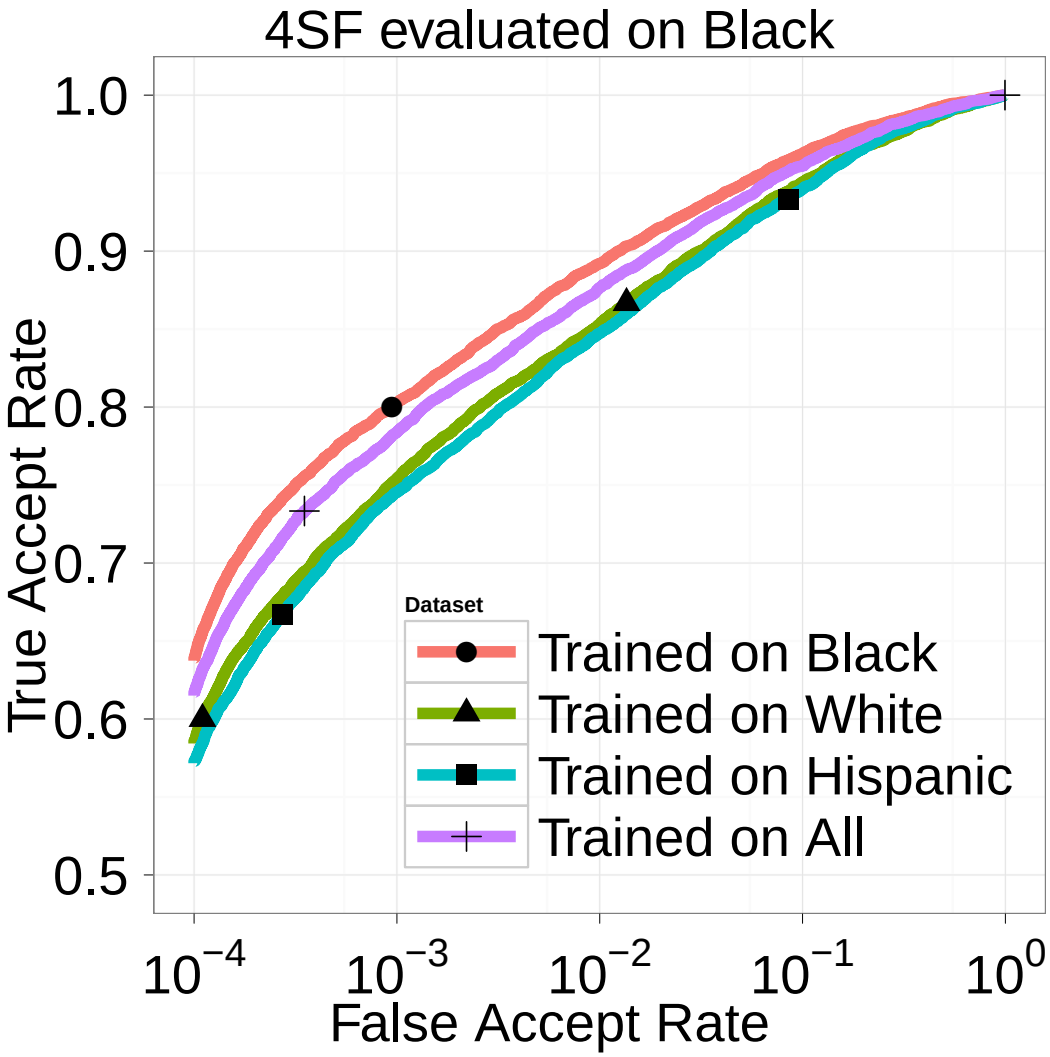


Figure 5.18: Performance of the different trained versions of the 4SF algorithm on the Black cohort.

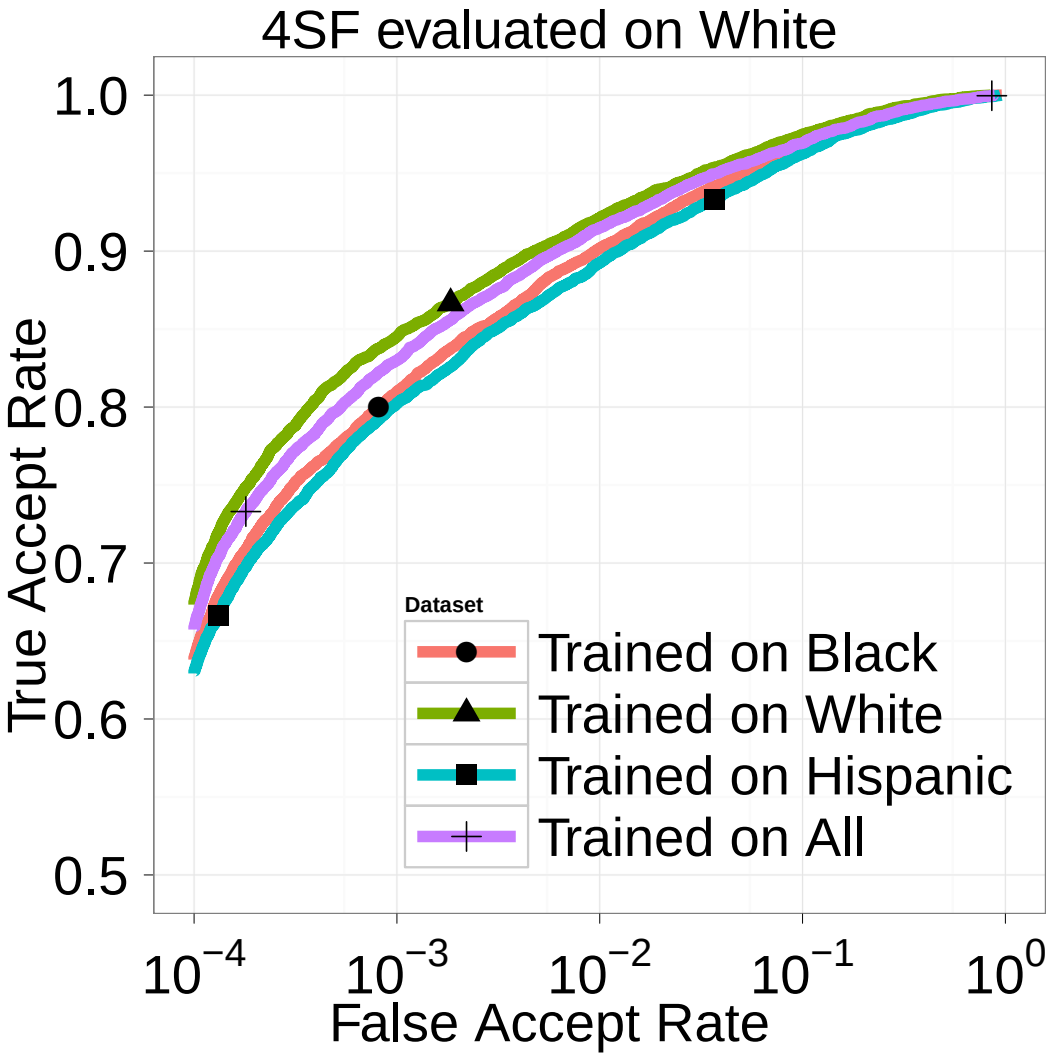


Figure 5.19: Performance of the different trained versions of the 4SF algorithm on the White cohort.

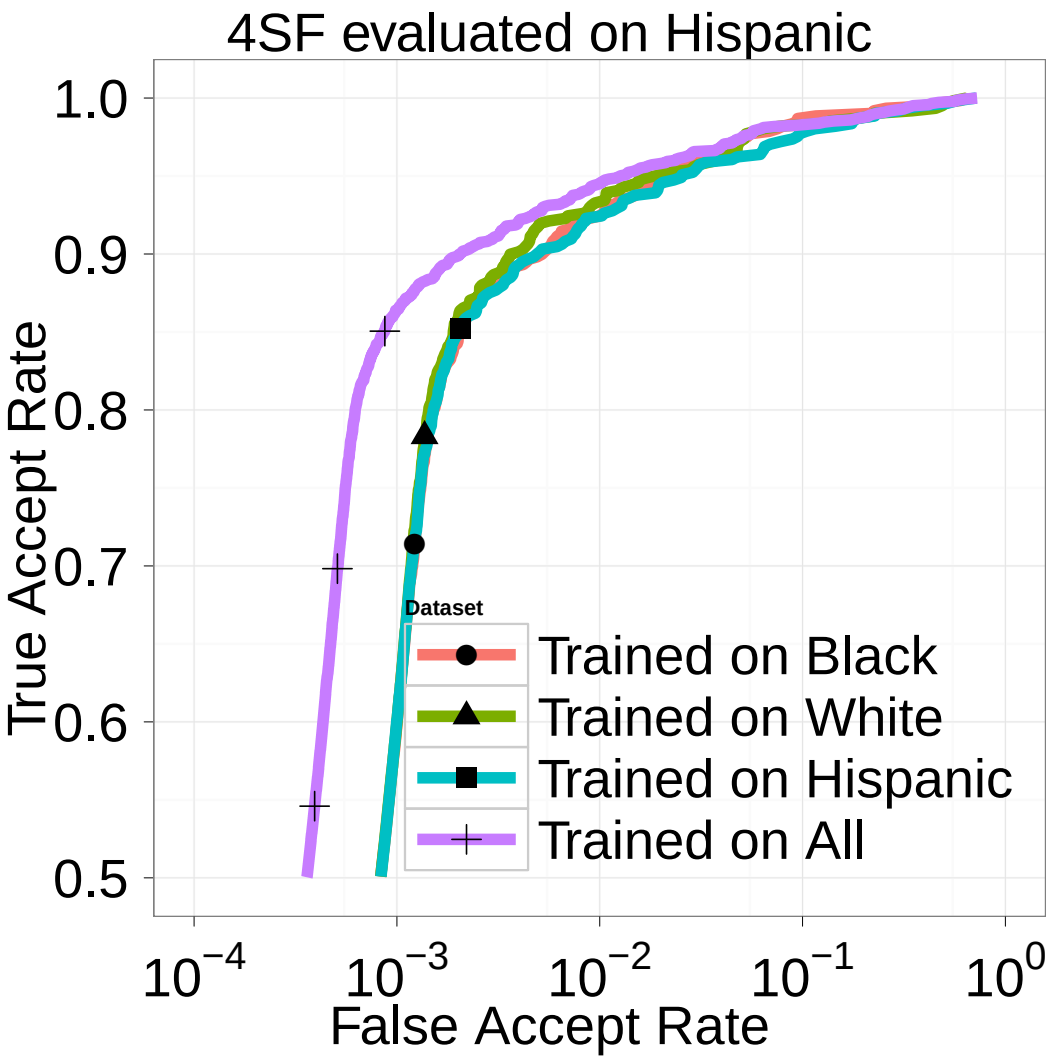


Figure 5.20: Performance of the different trained versions of the 4SF algorithm on the Hispanic cohort.

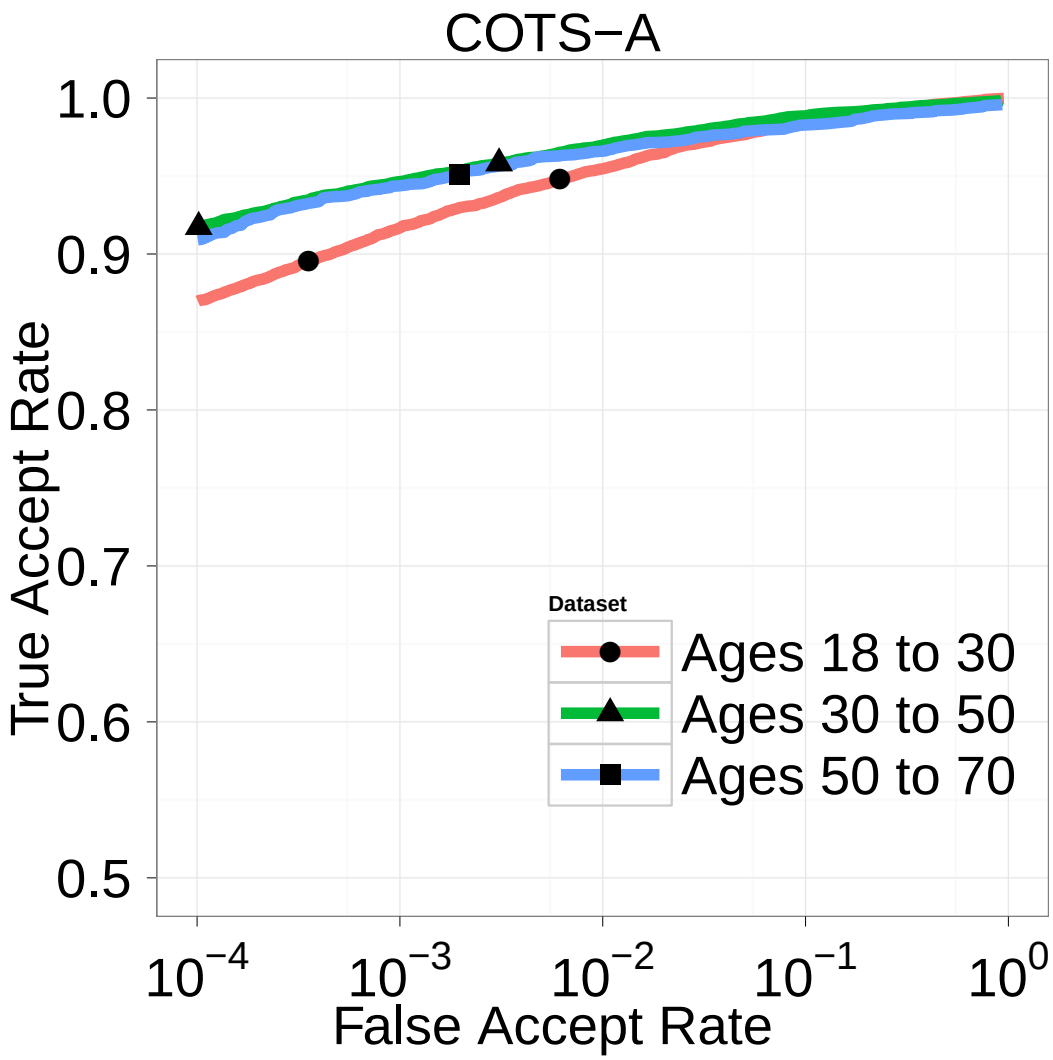


Figure 5.21: Performance of the COTS-A commercial face recognition system on datasets seperated by cohorts within the age demographic.

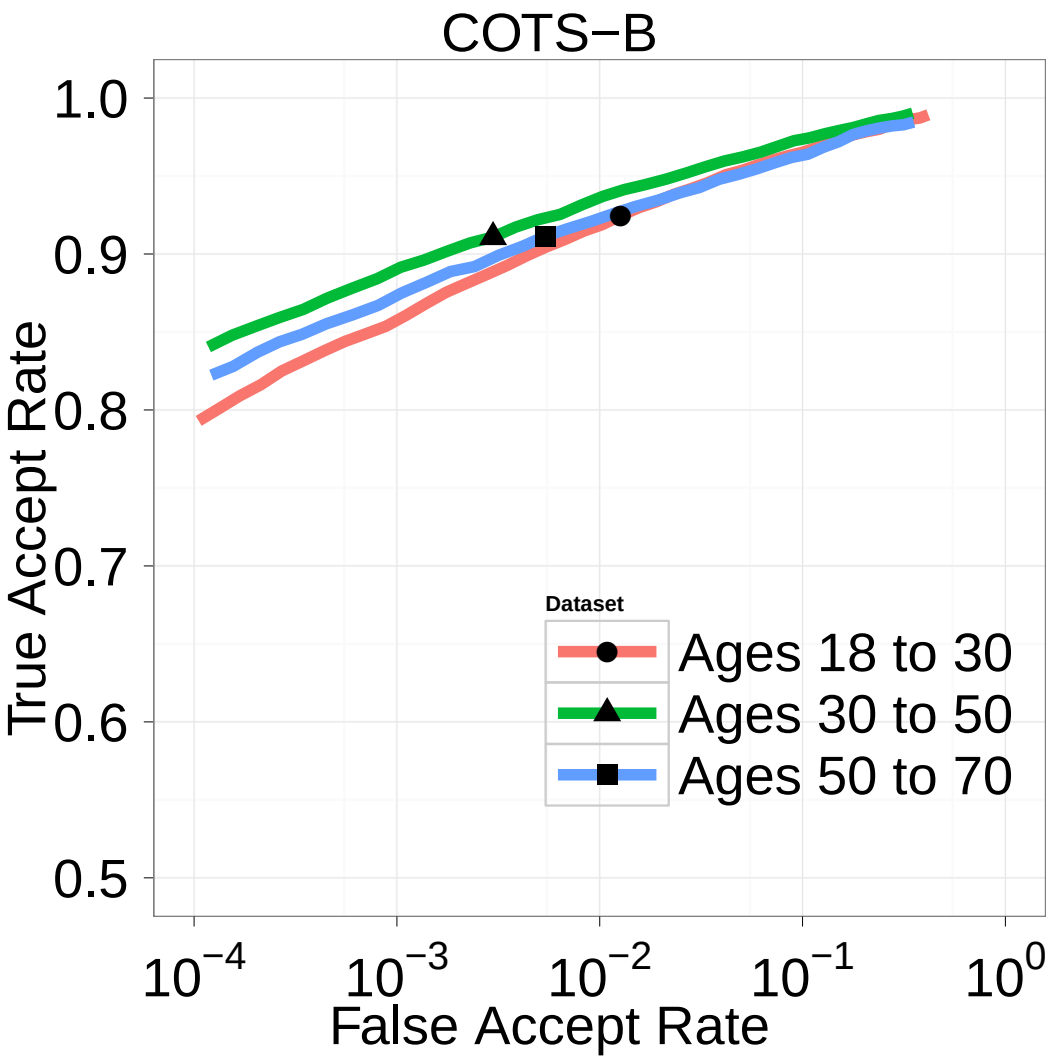


Figure 5.22: Performance of the COTS-B commercial face recognition system on datasets seperated by cohorts within the age demographic.

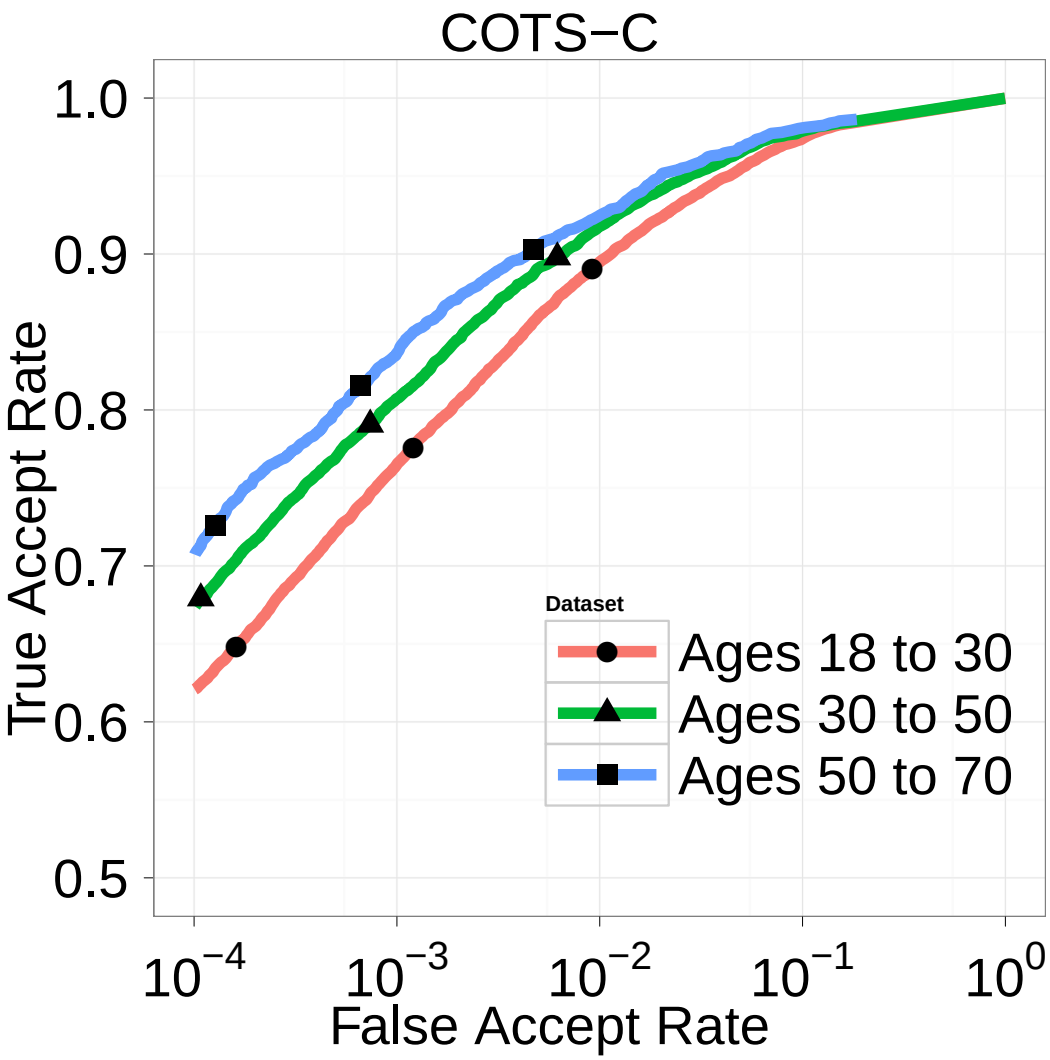


Figure 5.23: Performance of the COTS-C commerical face recognition system on datasets seperated by cohorts within the age demographic.

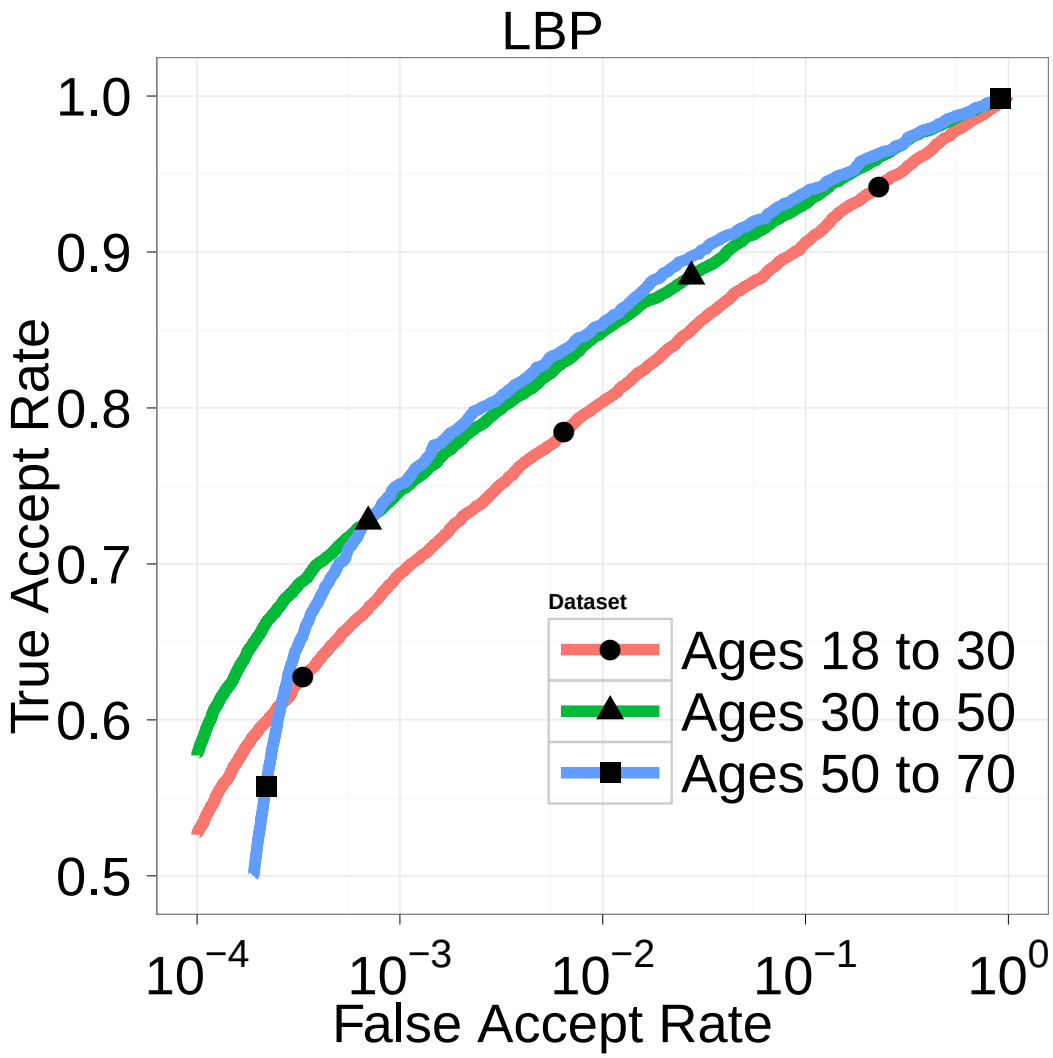


Figure 5.24: Performance of the local binary pattern-based non-trainable face recognition system on datasets seperated by cohorts within the age demographic.

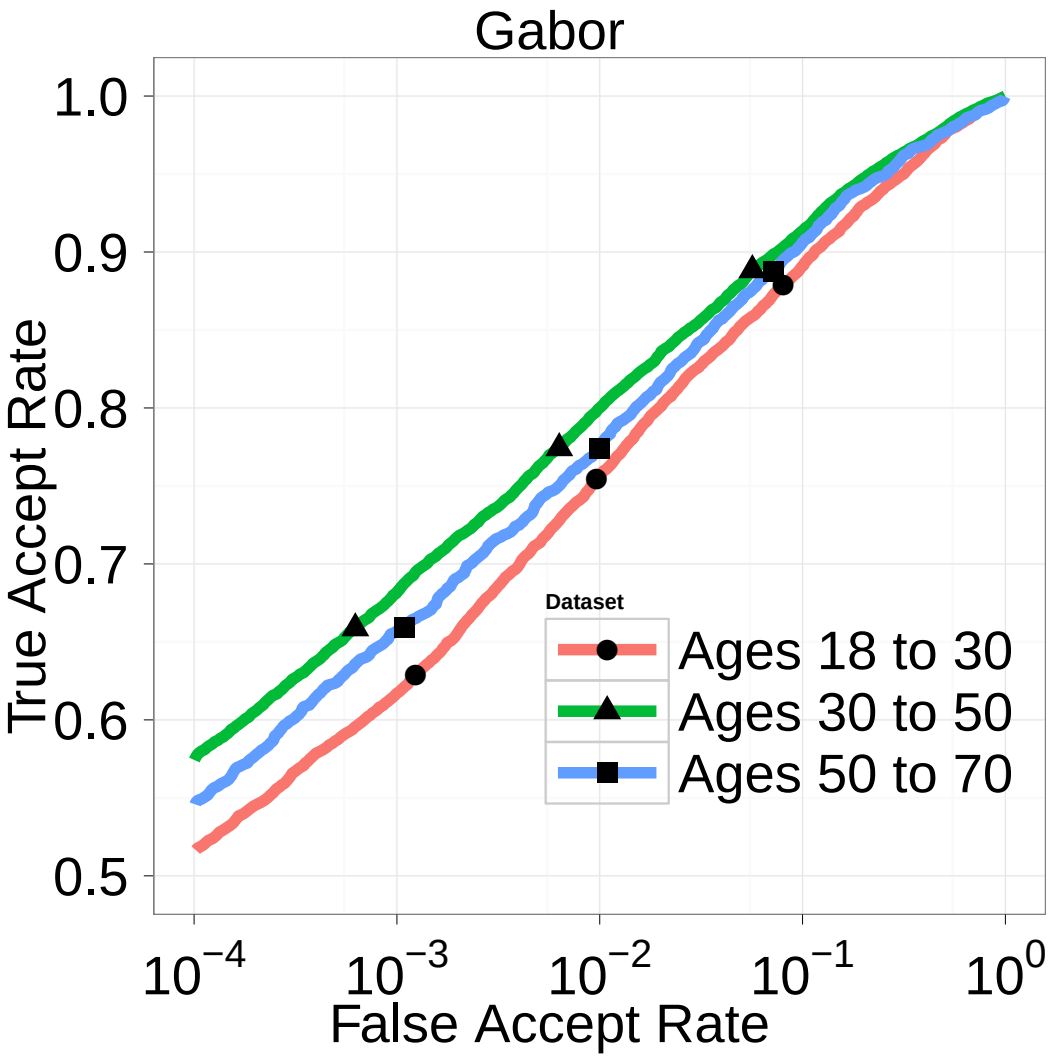


Figure 5.25: Performance of the Gabor-based non-trainable face recognition system on datasets seperated by cohorts within the age demographic.

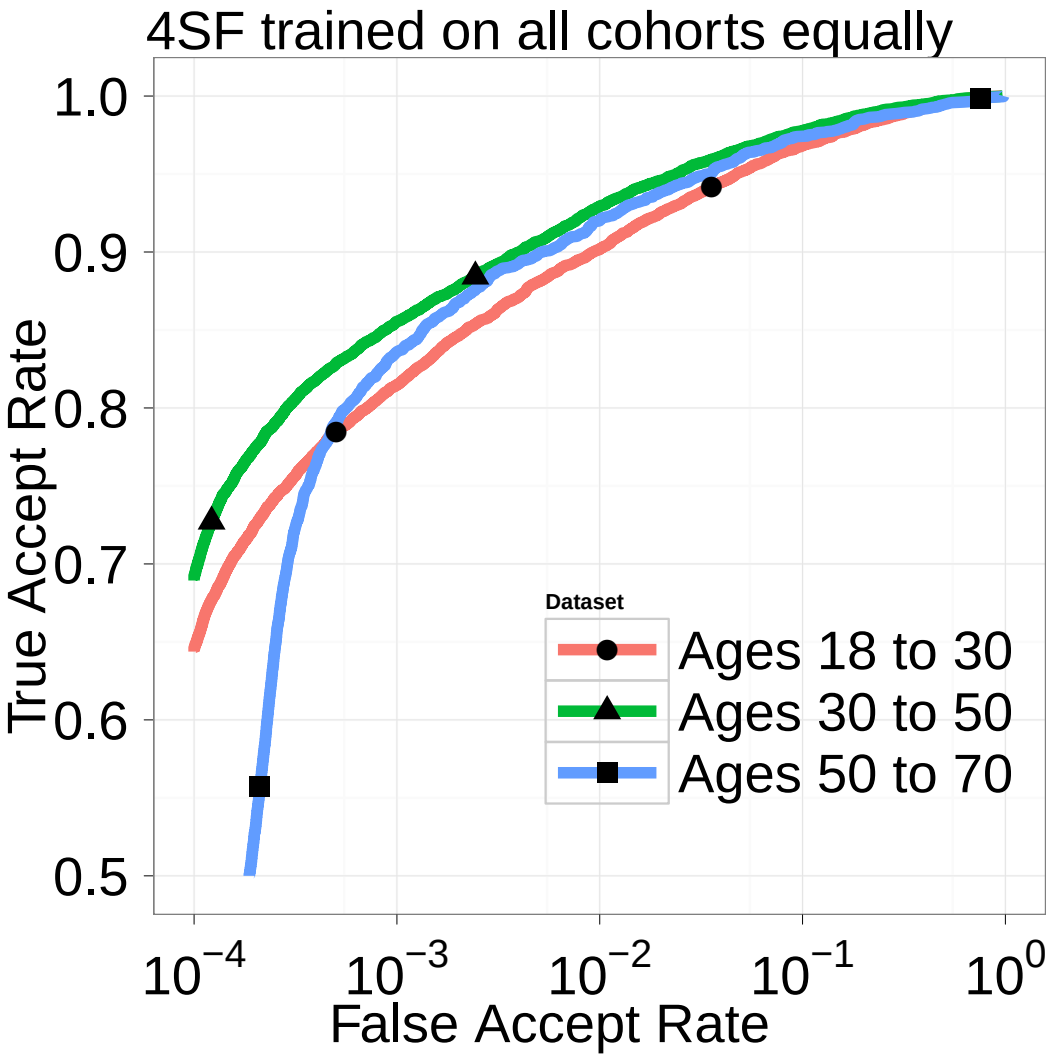


Figure 5.26: Performance of the 4SF algorithm trained on an equal distribution of samples acoress age on datasets seperated by cohorts within the age demographic.

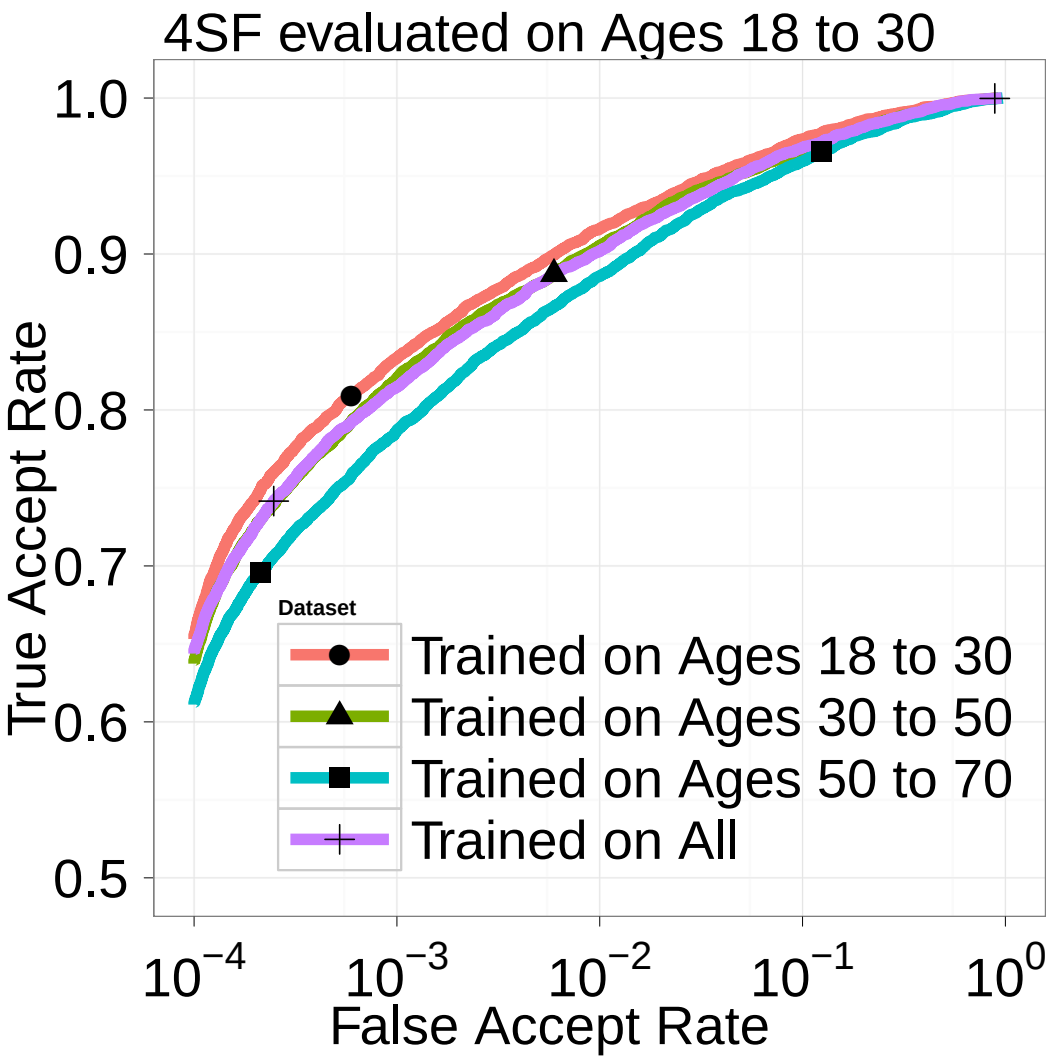


Figure 5.27: Performance of the different trained versions of the 4SF algorithm on the Ages 18 to 30 cohort.

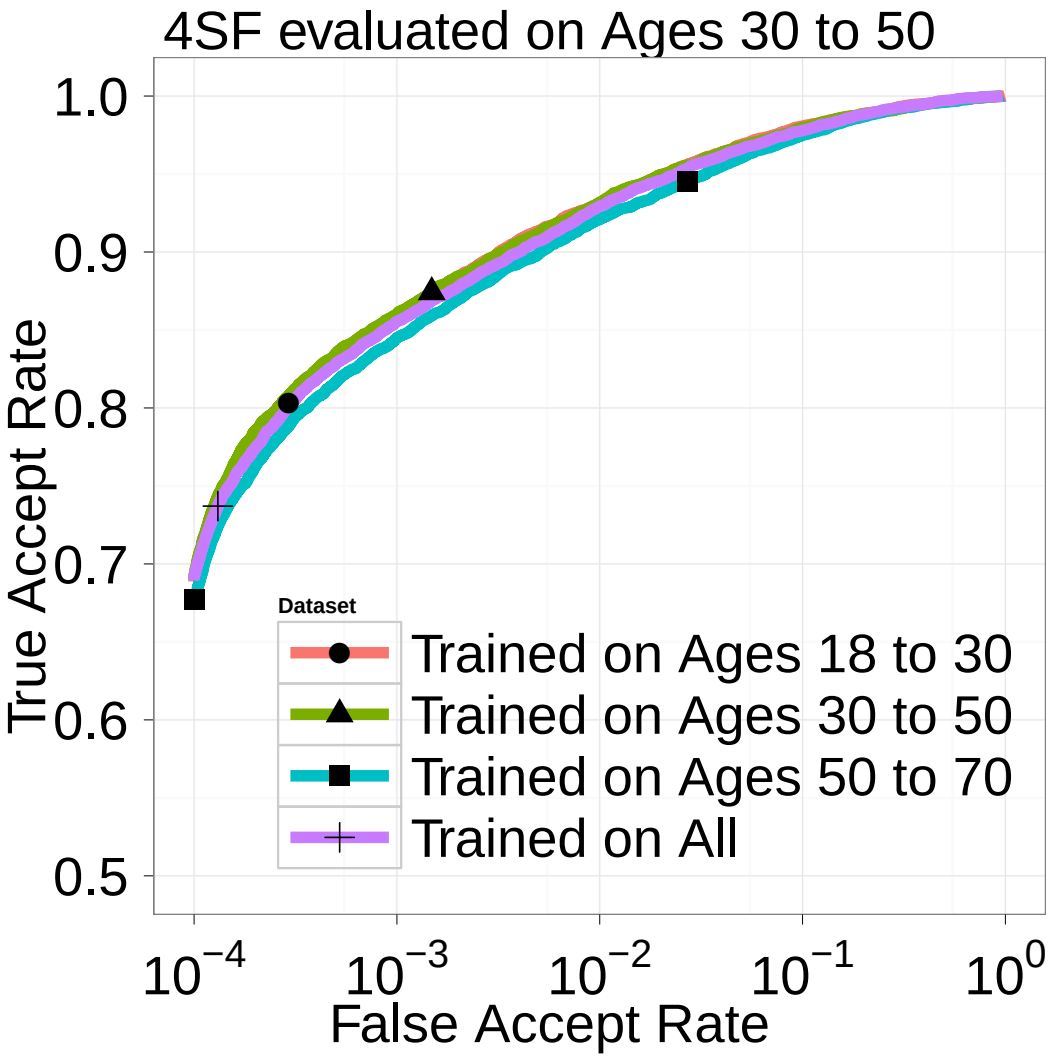


Figure 5.28: Performance of the different trained versions of the 4SF algorithm on the Ages 30 to 50 cohort.

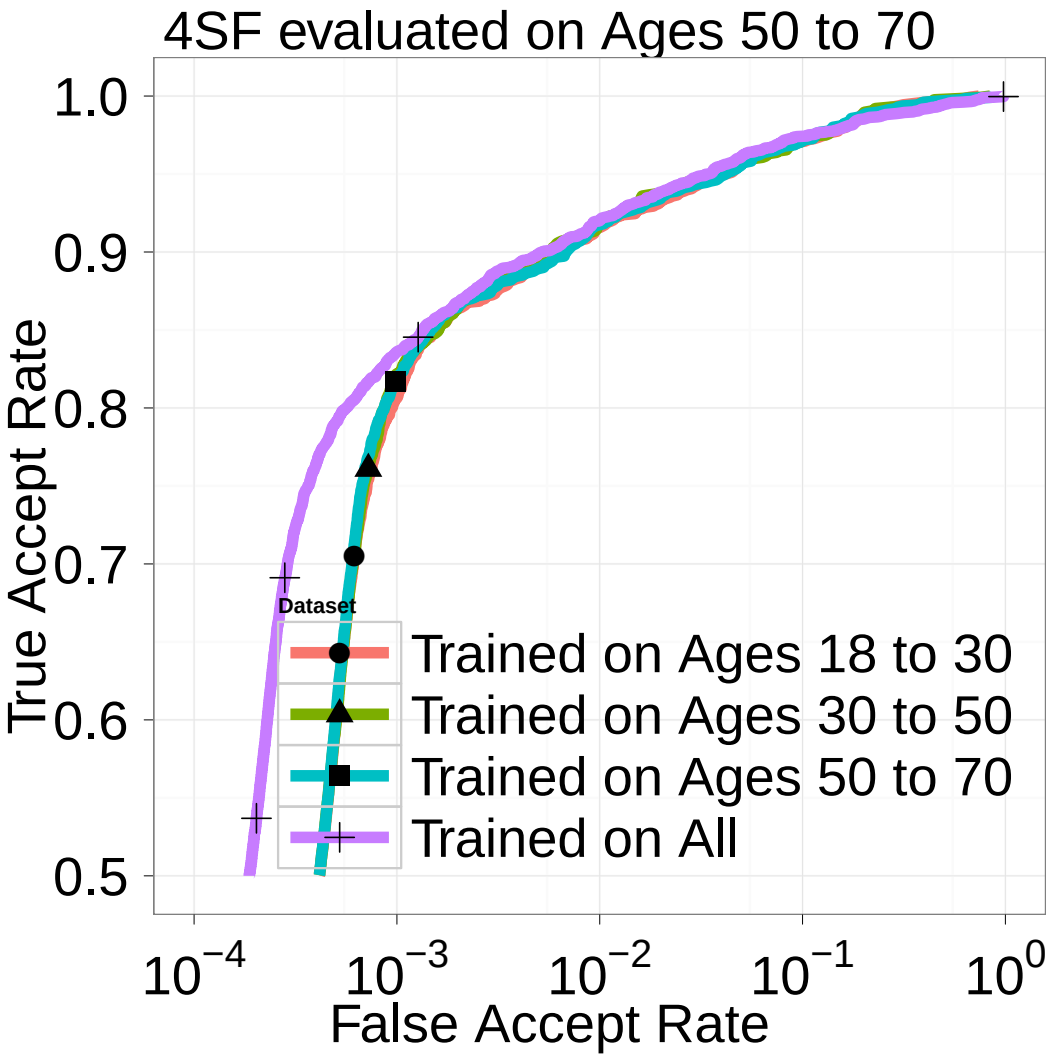


Figure 5.29: Performance of the different trained versions of the 4SF algorithm on the Ages 50 to 70 cohort.

rate of 0.1% for all the aforementioned plots are summarized in Tables 5.2, 5.3, and 5.4.

The first experiment conducted on each demographic measures the relative performance within the demographic cohort for each COTS FRS@. That is, for a particular commercial matcher (e.g., COTS-A), we compare it's matching accuracy on each cohort within that demographic. For example, on the gender demographic, this experiment will measure the difference in recognition accuracy for commercial matchers on males versus females. The results from this set of experiments can be found in Figures 5.4, 5.5, and 5.6 for the gender demographic, Figures 5.12, 5.13, 5.14 for the race/ethnicity demographic, and Figures 5.21, 5.22, and 5.23 for the age demographic. The second experiment conducted on each demographic cohort measures the relative performance within the cohort for non-trainable face recognition algorithms. Because the non-trainable algorithms do not leverage statistical variability in faces, they are not susceptible to training biases. Instead, they reflect the inherent (or a priori) difficulty in recognizing cohorts of subjects within a specific demographic group. The results from this set of experiments can be found in Figures 5.7 and 5.8 for the gender demographic, Figures 5.15 and 5.16 for the race/ethnicity demographic, and Figures 5.24 and 5.25 for the age demographic.

The final experiment investigates the influence of the training set on recognition performance. Within each demographic cohort, we train several versions of the 4SF algorithm (one for each cohort). These differently trained versions of the 4SF algorithm are then applied to separate testing sets from each cohort within the particular demographic. This enables us to understand within the gender demographic (for example), how much training exclusively on females (i) improves performance on females, and (ii) decreases performance on males. In addition to training 4SF exclusively on each cohort, we also use a version of 4SF trained on an equal representation of specific demographic cohorts (referred to as "Trained on All"). For example, in the

gender demographic, this would mean that for "All", 4SF was trained on 4,000 male subjects and 4,000 female subjects. The results from this set of experiments can be found in Figures 5.9 to 5.11 for the gender demographic, Figures 5.17 to 5.20 for the race/ethnicity demographic, and Figures 5.26 to 5.29 for the age demographic.

5.6 Analysis

In this section we provide an analysis of the findings of the experiments described in Section 6.7. A strength of this study is the large face dataset leveraged; accuracies measured on each cohort (except Hispanic and Old cohorts) are from roughly 8,000 subjects.

5.6.1 Gender

Each of the three commercial face recognition algorithms performed significantly worse on the female cohort than the male cohort (see Figures 5.4, 5.5, and 5.6). Additionally, both non-trainable algorithms (LBP and Gabor) performed significantly worse on females (see Figures 5.7 and 5.8).

The agreement in relative accuracies of the COTS FRS and the non-trainable LBP method on the gender demographic suggests that the female cohort is more difficult to recognize using frontal face images than the male cohort. That is, if the results in the COTS algorithms were due to imbalanced training sets (i.e., training on more males than females), then the LBP matcher should have yielded similar matching accuracies on males and females. Instead, the non-trained LBP and Gabor matchers performed worse on the female cohort. When training on males and females equally (Figure 5.9), the 4SF algorithm also did significantly worse on the female cohort. Together, *these results strongly suggest that the female cohort is inherently more difficult to recognize.*

The results of the 4SF algorithm on the female cohort (Figure 5.10) offer additional

Table 5.2: Listed are the true accept rates at a fixed false accept rate of 0.1% for each matcher on the gender demographic.

	Females	**Males**
COTS-A	89.5	94.4
COTS-B	81.6	89.3
COTS-C	70.3	80.9
LBP	54.4	74.0
Gabor	56.0	68.2
4SF trained on All	73.0	86.2
4SF trained on Females	71.5	85.0
4SF trained on Males	69.0	86.3

evidence about the nature of the discrepancy. The performance of training on only females is not higher than the performance of training on a mix of males and females (labeled "All"). Further, the difference in performance when training on only males versus training on only females is much lower than the difference in performance between males and females on the non-trainable algorithm. In other words, the difficulty in recognizing females seems to be due to a higher ratio of inter-class variance to intra-class variance in the initial face image representations.

Different factors may explain why females appear more difficult to recognize than males. One explanation may be that because females often use cosmetics (i.e., makeup), and males generally do not, there is a higher within-class variance in females. This hypothesis is supported by the match score distributions for males and females (see Figures 5.30 and 5.31). A greater difference in the true match distributions is noticed when compared to the false match distributions. The increased dissimilarities between images of the same female subjects demonstrate intra-class variability. Again, a cause of this may be due to the application of cosmetics.

Table 5.3: Listed are the true accept rates at a fixed false accept rate of 0.1% for each matcher on the race dataset.

	Black	**White**	**Hispanic**
COTS-A	88.7	94.4	95.7
COTS-B	81.3	89.0	90.7
COTS-C	74.0	79.8	87.3
LBP	65.3	70.5	73.5
Gabor	61.6	63.7	70.9
4SF trained on All	78.4	83.0	86.3
4SF trained on Black	80.2	81.0	59.8
4SF trained on White	75.4	84.5	59.9
4SF trained on Hispanic	74.5	80.2	60.1

Table 5.4: Listed are the true accept rates at a fixed false accept rate of 0.1% for each matcher on the age dataset.

	18 to 30 y.o.	**30 to 50 y.o.**	**50 to 70 y.o.**
COTS-A	91.7	94.6	94.4
COTS-B	86.1	89.1	87.5
COTS-C	76.5	80.7	83.6
LBP	69.4	74.7	75.1
Gabor	61.7	68.2	65.7
4SF trained on All	81.5	85.6	83.6
4SF trained on 18 to 30 y.o.	83.3	85.9	80.7
4SF trained on 30 to 50 y.o.	82.1	86.0	82.2
4SF trained on 50 to 70 y.o.	78.7	84.5	82.0

5.6.2 Race

When examining the race/ethnicity cohort, all three commercial face recognition algorithms achieved the lowest matching accuracy on the Black cohort (see Figures 5.12 to 5.14). The two non-trained algorithms had similar results (Figures 5.15 and 5.16).

When matching against only Black subjects (Figure 5.18), 4SF has higher accuracy when trained exclusively on Black subjects (about a 5% improvement over the system trained on Whites and Hispanics only). Similarly, when evaluating 4SF on only White subjects (Figure 5.19), the system trained on only the White cohort had the highest accuracy. However, when comparing the 4SF algorithm trained equally on all race/ethnicity cohorts (Figure 5.17), we see that the performance on the Black cohort is still lower than on the White cohort. Thus, even with balanced training, the Black cohort still is more difficult to recognize.

The key finding in the training results shown in Figures 5.17 to 5.20 is the ability to improve recognition accuracy by training exclusively on subjects of the same race/ethnicity. Compared to balanced training (i.e., training on "All"), the performance of 4SF when trained on the same race/ethnicity it is recognizing is higher. Thus, by merely changing the distribution of the training set, we can improve the recognition rate by nearly 2% on the Black cohort and 1.5% on the White cohort (see Table 5.3).

The inability to effectively train on the Hispanic cohort is likely due to the insufficient number of training samples available for this cohort. However, the biogeographic ancestry of the Hispanic ethnicity is generally attributed to a three-way admixture of Native American, European, and West Black populations [88]. Even with an increased number of training samples, we believe this mixture of races would limit the ability to improve recognition accuracy through race/ethnicity specific training.

5.6.3 Age Demographic

All three commercial algorithms had the lowest matching accuracy on subjects grouped in the ages 18 to 30 (see Figures 5.21 to 5.23). The COTS-A matcher performed nearly the same on the 30 to 50 year old cohort as the 50 to 70 year old cohort. However, COTS-B had slightly higher accuracy on 30 to 50 age group than 50 to 70 age group, while COTS-C performed slightly better on 50 to 70 than 30 to 50 age groups.

The non-trainable algorithms (Figures 5.24 and 5.25) also performed the worst on the 18 to 30 age cohort.

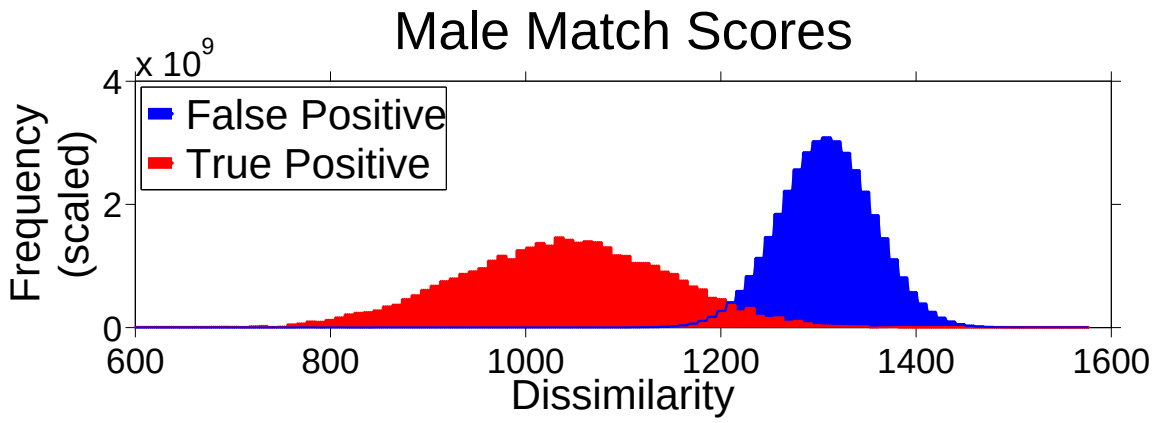
When evaluating 4SF on only the 18 to 30 year old cohort (Figure 5.27) and the 30 to 50 year old cohort (Figure 5.28), the highest performance was achieved when training on the same cohort. Table 5.4 helps elaborate on the exact accuracies. Similar to race, we were able to improve recognition accuracy by merely changing the distribution of the training set.

When comparing the 4SF system that is trained with equal number of subjects from all age cohorts, the performance on the 18 to 30 year old cohort is the lowest. This is consistent with the accuracies of the commercial face recognition algorithms.

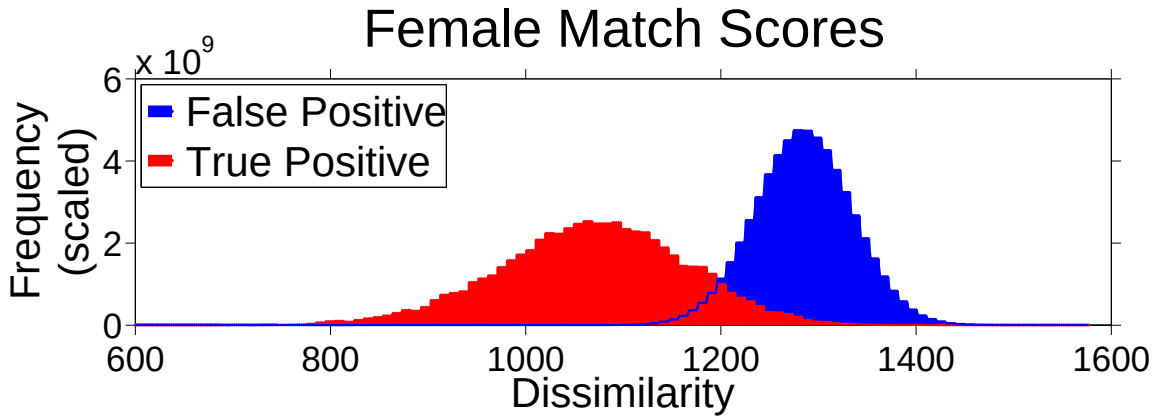
The less effective results from training on the 50 to 70 year old cohort is likely due to an small number of training subjects. This is consistent with the training results on the Hispanic cohort, which also had a small number of training subjects.

5.6.4 Impact of Training

The demographic distribution of the training set generally had a clear impact on the performance of different demographic groups. Particularly in the case of race/ethnicity, we see that training on a set of subjects from the same demographic cohort as being matched offers an increase in the True Accept Rate (TAR). This finding is particularly important because in most operational scenarios, particularly those

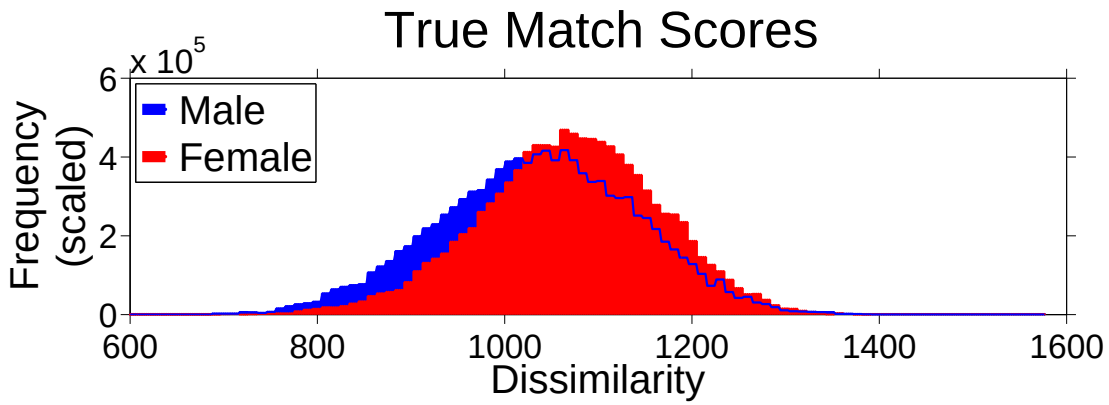


(a)

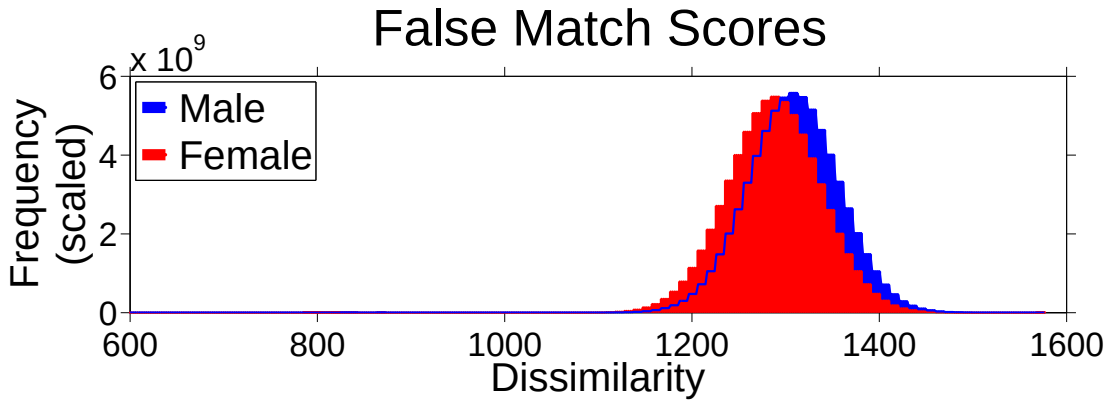


(b)

Figure 5.30: Match score distributions for the (a) male and (b) female genders using the 4SF system trained with an equal number of male and female subjects. All histograms are aligned on the same horizontal axis.



(a)



(b)

Figure 5.31: Geniune and impostor score distributions for the male and female genders using the 4SF system trained with an equal number of male and female subjects. The increased distances (dissimilarities) for the true match comparisons in the female cohort suggest increased within-class variance in the female cohort. All histograms are aligned on the same horizontal axis.

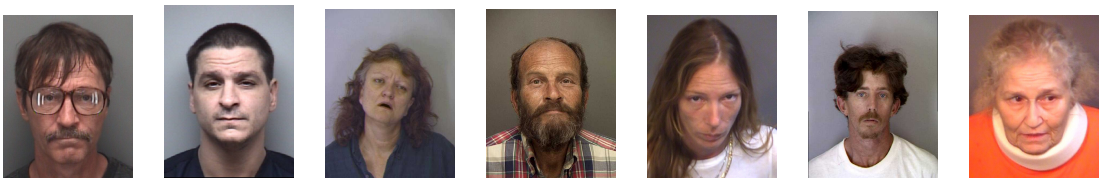
dealing with forensics and law enforcement, the use of face recognition is not being done in a fully automated, "lights out" mode. Instead, an operator is usually interacting with a face recognition system, performing a one-to-one verification task, or exploring the gallery to group together candidates in clusters for further exploitation. Each of these scenarios can benefit from the use of demographic-enhanced matching algorithms, as described below.

Scenario 1 - 1:N Search In many large face recognition database searches, the objective is to have the true match candidates ranked high enough to be found by the analyst performing the candidate adjudication. While it will not always be the case, under many conditions, the analyst will be able to categorize the demographics of the probe image based on age, gender, and/or race/ethnicity. In such a situation, if the analyst has the option to select a different matching algorithm that has been trained for that specific demographic group, then improved matching results should be expected. An schematic of this is shown in Figure 5.2. This individual could be searched using an algorithm trained on male, Whites, and aged 18 to 30. If a true match is not found using that algorithm, then a more generic algorithm might be used as a follow up to further search the gallery. Note that this scenario does not require that the gallery images be pre-classified based on specific demographic information. Instead, the algorithm should simply generate higher match scores for subjects that share the characteristics of that demographic cohort. We call this method of face search *dynamic face matcher selection*. In cases where the demographic is unclear (e.g., a mixed race/ethnicity subject), the matcher trained on all cohorts equally can be used. Examples of improved retrieval instances through applying this technique can be found in Figure 5.32.

Scenario 2 - 1:1 Verification It is often the case that investigators will identify a possible match to a known subject and will request an analyst to perform a 1:1

Probe Images:

Gallery Mates:



Retrieval Rank for 4SF Trained on all cohorts equally:

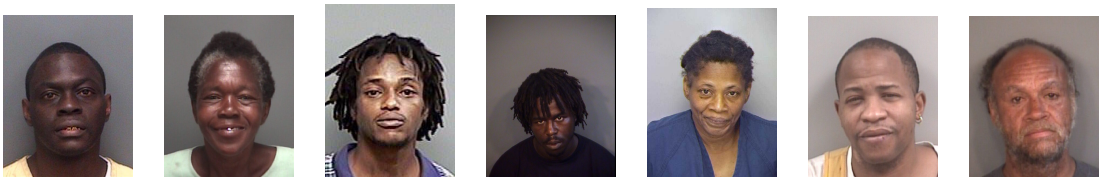
873	866	763	679	628	5	3

Retrieval Rank for 4SF Trained on White cohort exclusivly:

10	42	48	20	10	48	42

(a)

Probe Images:



Gallery Mates:

Retrieval Rank for 4SF Trained on all cohorts equally:

820	.112	730	640	574	7	6

Retrieval Rank for 4SF Trained on Black cohort exclusively:

20	34	43	9	43	42	41

(b)

Figure 5.32: Shown are examples where dynamic face matcher selection improved the retrieval accuracy. The final two columns show the less frequent case where such a technique reduced the retrieval accuracy. Retrieval ranks are out of roughly 8,000 gallery subjects for each cohort. Leveraging demographic information (such as race/ethnicity in this example) allows a face recognition system to perform the matching using statistical models that are tuned to the differences within the specific cohort.

verification of the match. This also happens as a result of a 1:N search, once a potential match to a probe is identified. In either case, the analyst must reach a determination of match or no-match. In fully automated systems, this decision is based on a numerical similarity threshold. In some environments, the analyst is prevented from seeing the similarity score out of concern that his judgment will be biased. But in others, the analyst is permitted to incorporate this into his analysis. In either case, it is anticipated that an algorithm trained on a specific demographic group will return higher match scores for true matches than one that was more generic. As a result, the analyst is more likely to get a hit and the 1:1 matching results process will be improved.

Scenario 3 - Verification at Border Crossings The results presented here provide support for further testing of additional demographic groups, potentially including specific country or geographic-region of origin. Assuming such demographics proved effective at improving match scores, then use of dynamic face matcher selection could be extended to immigration or border checks on entering subjects to verify that their passport or other documents accurately reflects their country of origin.

Scenario 4 - Face Clustering Another analyst-driven application involves the exploitation of large sets of uncontrolled face imagery. Images encountered in intelligence or investigative applications often include large sets of videos or arbitrary photographs taken with no intention of enrolling them in a face recognition environment. Such image sets offer a great potential for development of intelligence leads by locating multiple pictures of specific individuals and giving analysts an opportunity to link subjects who may be found within the same photographs. Clustering methods are now being used on these datasets to group faces that appear to represent the same subject. Implementations of such clustering methods today usually rely upon a single algorithm to perform the grouping and an analyst must perform the quality

control step to determine if a particular cluster contains only a single individual. By combining multiple demographic-based algorithms into a sequential analysis, it may be possible to improve the clustering of large sets of face images and thereby reduce the time required for the analyst to perform the adjudication of individual clusters.

5.7 Conclusions

In this chapter we examined face recognition performance on different demographic cohorts on a large operational database of 102,942 face images. Three demographics were analyzed: gender (male and female), race/ethnicity (White, Black, and Hispanic), and age (18 to 30 years old, 30 to 50 years old, and 50 to 70 years old).

For each demographic cohort, the performances of three commercial face recognition algorithms were measured. The performances of all three commercial algorithms were consistent in that they all exhibited lower recognition accuracies on the following cohorts: females, Blacks, and younger subjects (18 to 30 years old).

Additional experiments were conducted to measure the performance of non-trainable face recognition algorithms (local binary pattern-based and Gabor-based), and a trainable subspace method (the Spectrally Sampled Structural Subspace Features (4SF) algorithm). These experiments offered additional evidence to form hypotheses about the observed discrepancies between certain demographic cohorts.

Some of the keys findings in this study are:

- The female, Black, and younger cohorts are more difficult to recognize for all matchers used in this study (commercial, non-trainable, and trainable).
- Face recognition performance on race/ethnicity and age cohorts generally improve when training exclusively on that same cohort.
- The above finding suggests the use of *dynamic face matcher selection*, where multiple face recognition systems, trained on different demographic cohorts, are

available as a suite of systems for operators to select based on the demographic information of a given query image (see Figure 5.2).

- In scenarios where dynamic matcher selection is not possible, training face recognition systems on datasets that are well distributed across all demographics is critical to reduce face matcher vulnerabilities on specific demographic cohorts.

Finally, as with any empirical finding, additional ways to exploit the findings of this research are likely to be found. Of particular interest is the observation that women appear to be more difficult to identify through facial recognition than men. If we can determine the cause of this difference, it may be possible to use that information to improve the overall matching performance.

The experiments conducted in this chapter should have a significant impact on the design of face recognition algorithms. Similar to the large body of research on algorithms that improve face recognition performance in the presence of other variates known to compromise recognition accuracy (e.g., pose, illumination, and aging), the results in this study should motivate the design of algorithms that specifically target different demographic cohorts within the race/ethnicity, gender and age demographics. By focusing on improving the recognition accuracy on such confounding cohorts (i.e., females, Blacks, and younger subjects), researchers should be able to further reduce the error rates of state of the art face recognition algorithms and reduce the vulnerabilities of such systems used in operational environments.

Chapter 6

Towards Automated Caricature Recognition

6.1 Introduction

Among the remarkable capabilities possessed by the human visual system, perhaps none is more compelling than our ability to recognize a person from a caricature. A caricature is a face image in which certain facial attributes and features have been exaggerated to a degree that is often beyond realism, and yet the face is still recognizable (see Fig. 6.1). As Leopold et al. discussed [69], the caricature generation process can be conceptualized by considering each face to lie in a face space. In this space, a caricature face beyond the line connecting the mean face[1] and a subject's face. In other words, a caricature is an extrapolated version of the original face.

Despite the (often extreme) exaggeration of facial features, the identity of a subject in a caricature is generally obvious, provided the original face is known to the viewer. In fact, studies have suggested that people may be better at recognizing a familiar person through a caricature portrait than from a veridical portrait[2] [90, 118].

[1] A mean face is the average appearance of all faces.
[2] A verdical portrait is a highly accurate facial sketch of a subject.

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

Figure 6.1: Examples of caricatures (top row) and photographs (bottom row) of four different personalities. Shown above are: (a) Angelina Jolie (drawn by Rok Dovecar), (b) Adam Sandler (drawn by Dan Johnson), (c) Bruce Willis (drawn by Jon Moss), and (d) Taylor Swift (drawn by Pat McMichael).

So why is it that an exaggerated, or extrapolated, version of a face can be so easy to recognize? Studies in human cognition have suggested this phenomenon is correlated to how humans represent and encode facial identity [90]. Empirical studies suggest that this representation involves the use of prototype faces, where a face image is encoded in terms of its similarity to a set of prototype face images [69, 130, 145]. Under this assumption, the effectiveness of a caricature would be due to its ability to emphasize deviations from prototypical faces. This would also explain why faces that are "average" looking, or typical, are more difficult to recognize [145].

Automated face recognition, despite its significant progress over the past decade [34], still has many limitations. State of the art face recognition algorithms are not able to meet the performance requirements in uncontrolled and non-cooperative face matching scenarios, such as surveillance. We believe clues on how we can better compute the similarity between faces may be found through investigating the caricature

matching process [51].

In this chapter we further expand our contributions of heterogeneous face recognition by studying the process of automatically matching a caricature to a facial photograph. To accomplish this, we define a set of qualitative facial attributes (e.g. "nose to mouth distance") that are used to encode the appearance of a face (caricature or photograph). These features, called "qualitative features", are generally on a nominal scale (and occasionally on an ordinal scale) that characterize when a particular facial attribute is either typical or atypical (deviates from the mean face). Statistical learning is performed to learn feature weighting and the optimal subset of these features.

While several methods exist for automating the caricature generation process, to our knowledge, this is the first attempt to automate the caricature recognition process. In addition to posting impressive accuracies on this difficult heterogeneous face recognition task, we are also releasing a caricature recognition dataset, experimental protocol, and qualitative features to the research community. Through the design and performance evaluation of caricature recognition algorithms, it is our belief that we will help advance the state of automatic face recognition through the discovery of additional facial representations and feature weightings [5].

6.2 Related Work

Caricature recognition belongs to a face recognition paradigm known as heterogeneous face recognition (HFR) [54], which has been well discussed in this dissertation. In brief, heterogeneous face recognition is the task of matching two faces from alternate modalities.

Solutions to heterogeneous face recognition problems generally follow one of two approaches. The first approach, popularized by Wang and Tang [149], seeks to syn-

thesize an image from one of the modalities (e.g. sketch) in the second modality (e.g. photograph). Once this synthesis has occurred, standard matching algorithms can be applied in the now common modality.

The second approach to HFR is to densely sample feature descriptors (such as local binary patterns (LBP) [99]) from the images in each modality. The feature descriptor is selected such that it varies little when moving between the imaging modalities, while still capturing key discriminative information. A benefit of this feature-based approach is that it facilitates statistical subspace learning (such as linear discriminant analysis (LDA) [10] and its variants) to further improve the class separability. This approach has been successfully used by Liao et al. [77], Klare and Jain [52, 54, 59], and Bhatt et al. [12].

In the context of caricature recognition, an image feature descriptor-based approach is challenged because the descriptors from the caricature and photograph may not be highly correlated due to misalignment caused by feature exaggerations (e.g. the nose in the caricature may extend to where the mouth or chin is in the photograph). However, the application of LDA, in a manner similar to other HFR studies [52, 59], somewhat compensates for these misalignments. Further, LDA offers a solution to the intra-artist variability through the modeling of the within-class scatter. For these reasons, the study in this chapter also makes use of the image feature descriptor-based approach in addition to the qualitative feature based approach (see Section 6.6).

A major contribution of this of this chapter is the definition of a set of categorical, or nominal, facial attributes. This approach is similar to the attribute and simile features proposed by Kumar et al. [65], who demonstrated the benefit of this nominal feature representation for recognizing face images. While we present a similar representation, the features proposed here have been carefully defined by a professional artist with experience in drawing caricatures.

A number of methods in graphics have been developed for automatically generat-

ing caricature images [6,7,16,64,70]. However, to our knowledge, no previous research on matching caricatures to photographs has been conducted. The method proposed by Hsu and Jain [38] was the closest attempt, where facial photographs were matched by first synthesizing them into caricature drawings. Klare and Jain considered the task of matching facial carvings [62] and avatar face images [153], both of which exhibited some facial disproportions that are similar to caricatures.

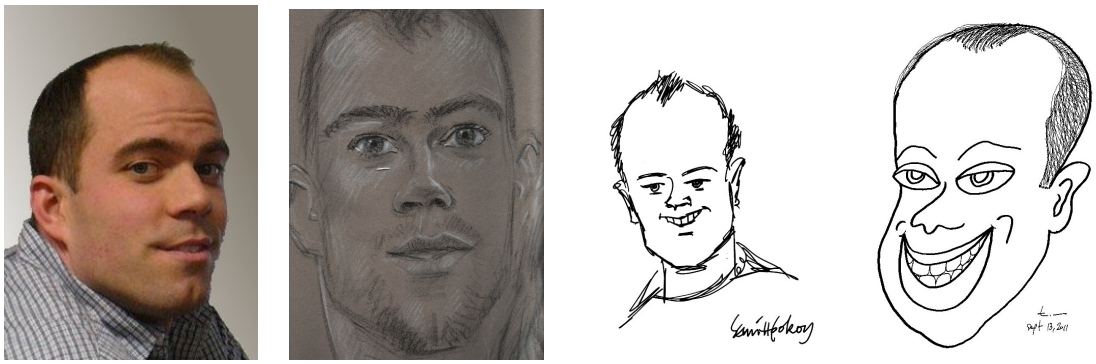
6.3 Caricature Dataset

In this section we describe the dataset that was used in this study. Future studies comparing accuracies on this dataset should follow the protocol detailed in Section 6.7.

The dataset consists of pairs of a caricature sketch and a corresponding facial photograph from 196 subjects (see Fig. 1 for examples). Two sources were used to collect these images. The first was through contacts with various artists who drew the caricatures. For these images, permission was granted to freely distribute the caricatures. In total 89 caricatures were collected from this source.

The second source of caricature images was from Google Image searches. For these caricatures, the url of the image was recorded, and is included in the dataset release (along with the actual image). There were 107 pairs from this source.

The corresponding face image for each subject was provided by the caricature artist for caricatures from the first source, and by Google Image search for the second source. When selecting face photographs, care was taken to find images that had minimal variations in pose, illumination, and expression. However, such "ideal" images do not always exist. Thus, many of the PIE (pose, illumination and expression) factors still persist.



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

Figure 6.2: Different forms of facial sketches (b-d). (a) Photograph of a subject. (b) Portrait sketch. (c) Forensic sketch drawn by Semih Poroy from a verbal description. (d) Caricature sketch.

6.4 Qualitative Feature Representation

We define a set of categorical facial features for representing caricature images and face photographs. These features were developed by one of the authors who is a cartoonist (in addition to being a professor of electrical engineering [4]).

While face images are typically encoded by high-dimensional numerical features (such as local binary patterns [99]), the tendency of a caricature to exaggerate distinctive facial features [117] makes such numerical encodings not appropriate for representing caricatures images. Instead, the proposed qualitative features describe facial features that a caricature artist may portray as to whether or not it is present. Thus, if "large distance between the nose and mouth" is a feature the artist chooses to emphasize, the proposed representation is able to capture this without being impacted by exactly how much the artist extrapolates this distance from the norm [5].

A caricaturist can be likened to a "filter" that only retains useful information in a face for identification. As a filter, the artist uses his talent to analyze a face, eliminate insignificant facial features, and capture the identity though exaggeration of the prominent features. Most of the caricaturists start with the description of the general shape of the head. They assemble the eyes, nose, eyebrows, lips, chin and

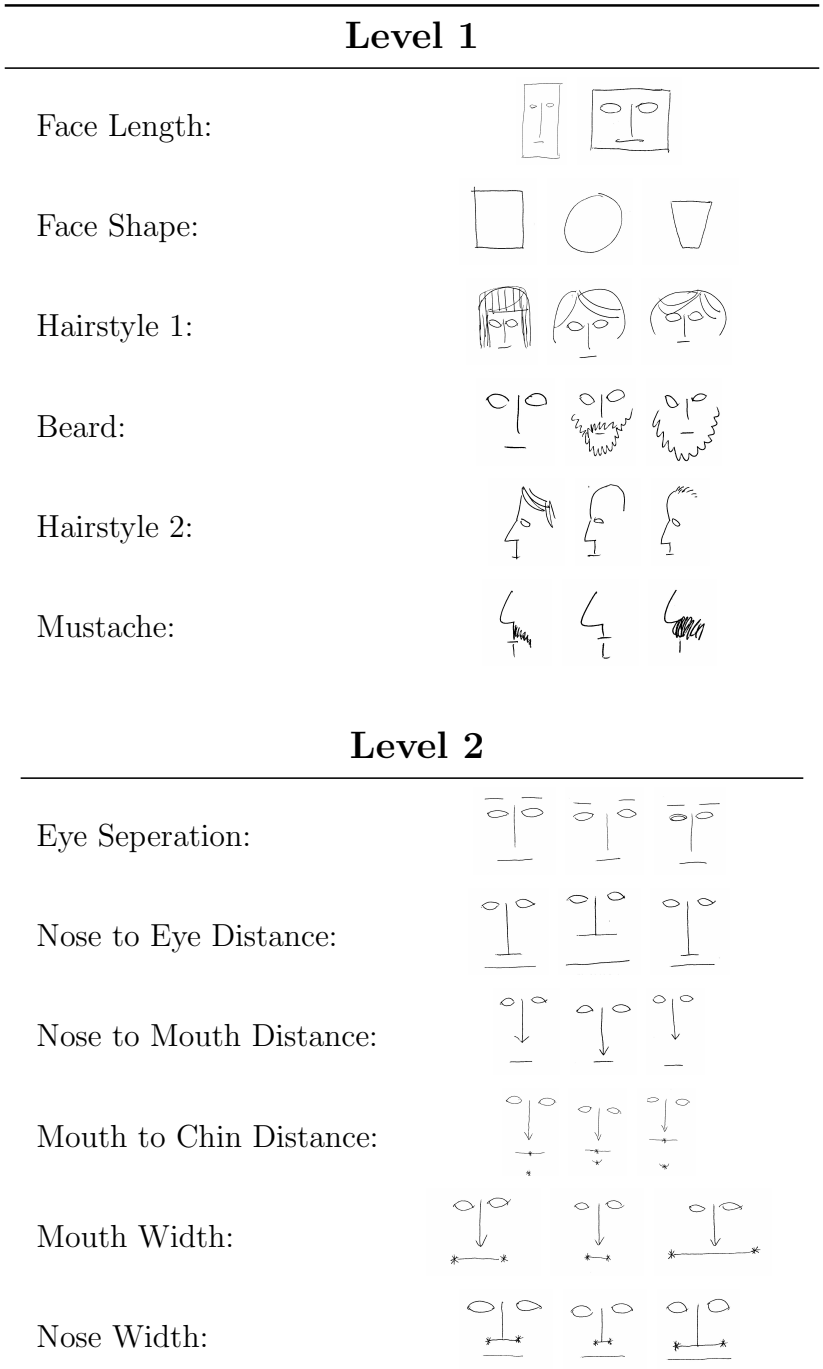


Figure 6.3: Illustration of features numbered one through twelve in the set of twenty five qualitative features used to represent both caricatures and photographs. The similarity between sketches and photographs were measured within this representation.

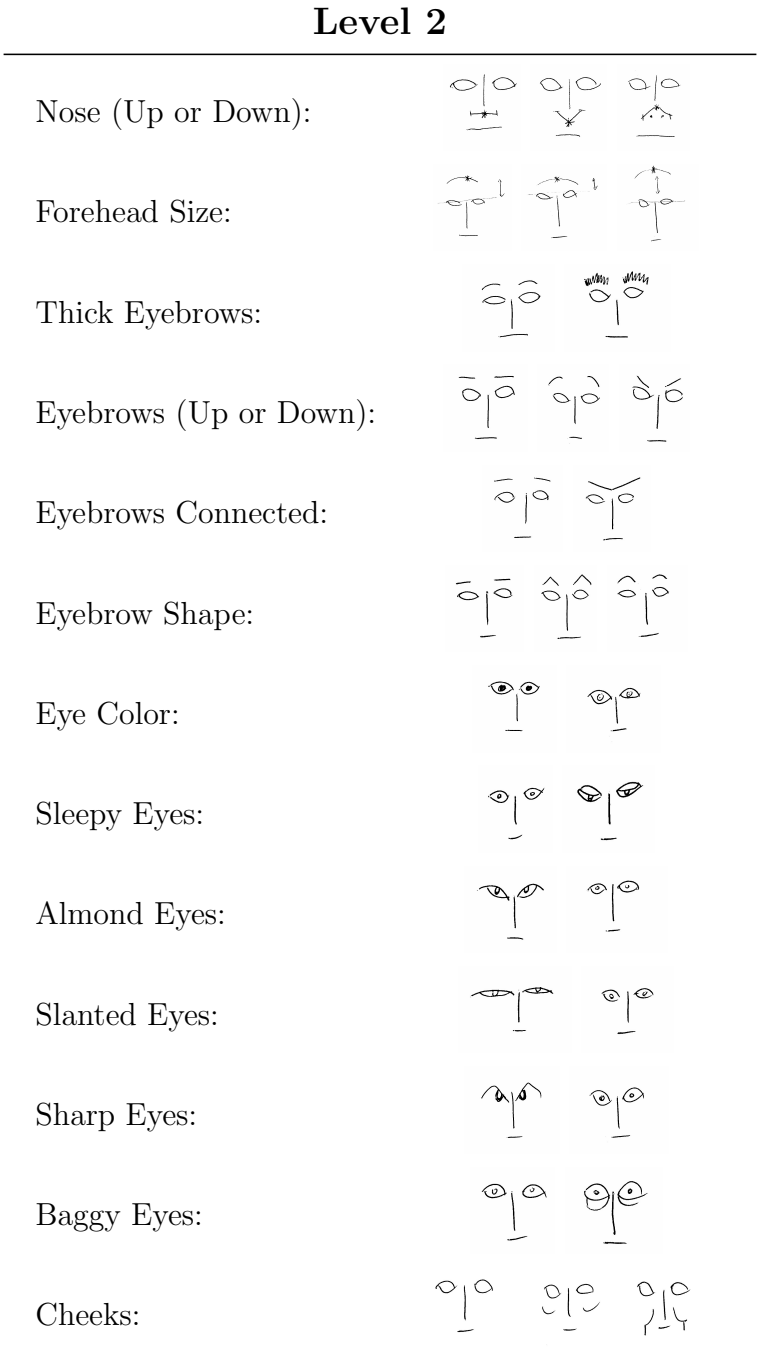


Figure 6.4: Illustration of the features numbered thirteen through twentyfour in the set of twenty five qualitative features used to represent both caricatures and photographs. The similarity between sketches and photographs were measured within this representation.

ears with some exaggerations in geometrically correct locations (always maintaining the appropriate ratios amongst them); finally, they include the hair, moustache and beard (depending on the gender or their presence in the face).

In this study, following the caricaturists methodology [117], we define a set of 25 qualitative facial features that are classified into two levels (see Figure 6.4). The first level (Level 1) is defined for the general shapes and sizes of the facial components and the second level (Level 2) is defined for the size and appearance of facial components, as well as ratios amongst the locations of different components (e.g. distance of the mouth from the nose).

6.4.1 Level 1 Qualitative Features

Level 1 features describe the general appearance of the face. These features can be more quickly discerned than Level 2 features. In standard face recognition tasks, Level 1 features are less informative than Level 2 features [56] due to their lack of persistence and uniqueness. However, in caricature recognition experiments these features are shown to be the most informative (see Section 6.7).

The length of the face is captured by the Face Length feature (narrow or elongated). The shape of the face is described by the Face Shape feature (boxy, round, or triangular). Two different features are used to capture the hair style, with values including: short bangs, parted left, parted right, parted middle, bald, nearly bold, thin middle, and curly. Facial hair is represented with the Beard feature (none, normal, Abraham Lincoln, thin, thick, and goatee) and Mustache feature (normal, none, thin, and thick) features. See Figure 6.4 for visual examples of these features.

6.4.2 Level 2 Features

Specific facial details are captured by the Level 2 facial features. Level 2 facial features will offer more precise descriptions of specific facial components (such as the eyes,

nose, etc.) compared to Level 1 features.

Several binary features are used to represent the appearance of the eyes. These include whether or not the eyes are dark, sleepy, "almond" shaped, slanted, sharp, or baggy. Similarly, the eyebrows are represented by their thickness, connectedness, direction (up or down), and general shape (normal, rounded, or pointed). The nose is represented by its width (normal, thin or wide) and direction (normal, points up, or points down). The mouth is characterized by its width (normal, thin, or wide). The cheeks are described as being either normal, thin, fat or baggy.

Several features are used to capture the geometric relationships among the facial components. They describe the distance between the nose and the eyes, the nose and the mouth, and the mouth and the chin. Two additional features describe the distance between the eyes, and the length of the forehead.

6.4.3 Feature Labeling

Each image (caricature and photo) was labeled with qualitative features by annotators provided through Amazon's Mechanical Turk[3]. Several annotators combined to label the entire set of image pairs with each of the 25 qualitative features. Each annotator was asked to label a single image with a single feature value at a time. Thus, the annotator was shown an image of either a caricature or a photograph, and each of the possible feature values (along with their verbal description) for the current feature being labeled.

To compensate for differences in annotator opinions on less obvious image/feature combinations, each image was labeled three times by three different annotators. Thus, given 25 qualitative features and three labelers per feature, a total of 75 feature labels were available per image. In all, 29,400 labeling tasks were performed through this crowdsourcing method (costing roughly $300 USD).

[3] `https://www.mturk.com/`

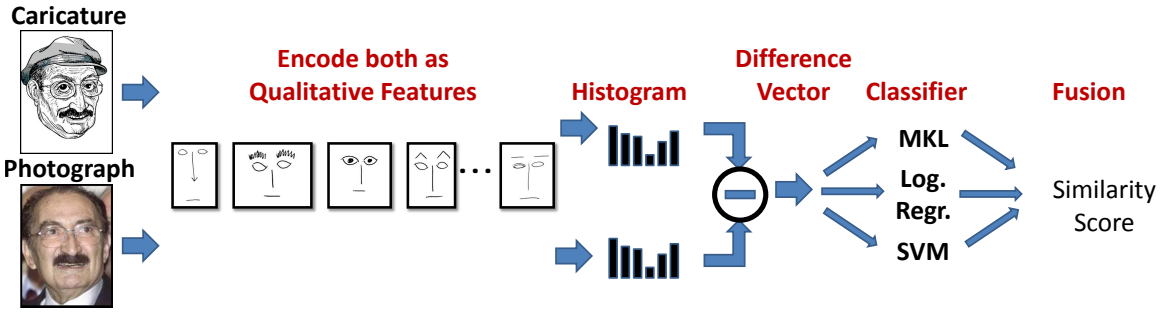


Figure 6.5: Overview of the caricature recognition algorithm.

6.5 Matching Qualitative Features

With each image labeled with 25 qualitative attributes u times ($u = 3$, see Sec. 6.4.3), each image (photo or caricature) can be represented by a $u \times 25$ matrix $\mathbf{C} \in Z_+^{u \times 25}$. Note that the matrix elements are nonnegative integers since each feature is of categorical type.

In order to improve the matching performance, we adopt machine learning techniques for feature subset selection and weighting. To facilitate this, we encode the categorical attributes into binary features by using r_i bits for each attribute, where r_i is the number of possible choices for the i^{th} attribute. For example, $r_i = 2$ for "*Thick Eyebrows*" and $r_i = 3$ for "*Nose Width*" (see Figure 6.4).

Ideally, the binary valued feature vector should lead to a vector with only one non-zero element per feature. However, the annotators may give contradicting annotations (e.g. one subject can be labeled as having a "Wide Nose" and "Normal Nose" by two different annotators). Hence, we accumulate the binary valued feature vectors into histogram feature vectors. Thus, a single feature will no longer be represented by an r_i-bit binary number, but instead by an r_i-dimensional feature vector. Each component will have a minimum value of 0 and a maximum value of u. Finally, for each image, we concatenate the 25 individual attribute histograms to get a 77-dimensional feature vector ($\mathbf{x} \in \mathbb{Z}_+^{77}, ||\mathbf{x}||_1 = 25u$). Given this representation, the simplest method for matching is to perform nearest neighbor search with Euclidean

distance (referred to as NN_{L_2}).

Next, we convert the caricature-photo matching problem into a binary classification task by calculating the absolute difference vector for every possible caricature-photo pair in the training set. In the binary classification setting, the difference vector for the caricature and photo pair of the same subject (i.e. a true match) is labeled as '1' whereas the difference vector for caricature-photo pair of two different subjects (i.e. a false match) is labeled as '-1'. This gives us n positive samples (genuine matches) and $n^2 - n$ negative samples (imposter matches), where n is the number of subjects in the training set.

With the caricature recognition problem reformulated as a binary classification task, we leverage a fusion of several binary classifiers. Let $\{(\mathbf{x_i}, y_i), \mathbf{x_i} \in \mathbb{R}^d, y_i \in \{-1, 1\}, i = 1, 2, ..., m\}$ be the m pairs of difference vectors, where $d = 77$. Again, if $\mathbf{x}_i$ is a difference vector between a caricature and photograph of the same subject then $y_i = 1$, otherwise $y_i = -1$.

6.5.1 Logistic Regression

Logistic regression seeks to find a function that maps the difference vectors to their numerical label ($+1$ or -1). The output of this regression can be interpreted as a similarity score, which facilitates fusion and receiver operator characteristic (ROC) analysis.

The objective function of the logistic regression is as follows

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{m} \{-y_i \mathbf{x_i}'\boldsymbol{\beta} + \log(1 + \exp(\mathbf{x_i}'\boldsymbol{\beta}))\} + \lambda R(\boldsymbol{\beta}), \tag{6.1}$$

where $\boldsymbol{\beta}$ is the vector of the feature weights to be learned, $R(\boldsymbol{\beta})$ is a regularization term (to avoid overfitting and impose structural constraints) and λ is a coefficient to control the contribution of the regularizer to the cost. Two different regularizers are

Method	FAR=10.0%	FAR=1.0%
Qualitative Features (no learning):		
NN_{L_2}	39.2 ± 5.4	
Qualitative Features (learning):		
Logistic Regression	50.3 ± 2.4	11.3 ± 2.9
MKL	39.5 ± 3.2	7.4 ± 3.9
NN_{MKL}	46.6 ± 3.9	10.3 ± 3.6
SVM	52.6 ± 5.0	12.1 ± 2.8
Logistic Regression+NN_{MKL}+SVM	56.9 ± 3.0	15.5 ± 4.6
Image Descriptors (learning):		
LBP with LDA	33.4 ± 3.9	11.5 ± 2.5
Qualitative Features + Image Descriptors:		
Logistic Regression+NN_{MKL}+ SVM+LBP with LDA	61.9 ± 4.5	22.7 ± 3.5

Table 6.1: Average verification accuracies of the proposed qualitative, image feature-based, and baseline methods. Shown are the true accept rates (TAR) at fixed false accept rates (FAR) of 1.0% and 10.0%. Average accuracies and standard deviations were measured over 10 random splits of 134 training subjects and 62 testing subjects (subjects in training and test sets are different).

commonly used: (i) the L_1-norm regularizer, $R(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$ (also know as Lasso [141]), which imposes sparseness on the solutions by making most of the coefficients to be equal to zero for large values of λ, and (ii) the L_2-norm regularizer, $R(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2$, which leads to non-sparse solutions.

Our experimental results with the implementation of [49] favored the L_2-norm regularizer, which we refer to in Section 6.7 as *Logistic Regression*. Having solved for $\boldsymbol{\beta}$ using a gradient descent method, we compute the similarity value of the difference vector $\mathbf{x}$ between a caricature and photograph as: $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}\boldsymbol{\beta}))$.

Method	Rank-1	Rank-10
Qualitative Features (no learning):		
NN_{L_2}	12.1 ± 5.2	52.1 ± 7.1
Qualitative Features (learning):		
Logistic Regression	17.7 ± 4.2	62.1 ± 3.8
MKL	11.0 ± 3.9	50.5 ± 4.0
NN_{MKL}	14.4 ± 2.9	59.5 ± 3.9
SVM	20.8 ± 5.6	65.0 ± 3.8
Logistic Regression+NN_{MKL}+SVM	23.7 ± 3.5	70.5 ± 4.4
Image Descriptors (learning):		
LBP with LDA	15.5 ± 4.6	42.6 ± 4.6
Qualitative Features + Image Descriptors:		
Logistic Regression+NN_{MKL}+ SVM+LBP with LDA	32.3 ± 5.1	74.8 ± 3.4

Table 6.2: Average identification accuracies of the proposed qualitative, image feature-based, and baseline methods. Average accuracies and standard deviations were measured over 10 random splits of 134 training subjects and 62 testing subjects (subjects in training and test sets are different).

6.5.2 Multiple Kernel Learning and SVM

One limitation of the logistic regression method is that it is restricted to finding linear dependencies between the features. In order to learn non-linear dependencies we use support vector machines (SVM) and multiple kernel learning (MKL) [8].

Given m training images, we let $\{\mathbf{K}_j \in \mathbb{R}^{m \times m}, j = 1, \ldots, 25\}$ represent the set of base kernels. $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_s)^\top \in \mathbb{R}^s_+$ denotes the coefficients used to combine these base kernels, and $\mathbf{K}(\mathbf{p}) = \sum_{j=1}^{s} \mathbf{p}_j \mathbf{K}_j$ is the combined kernel matrix. We learn the coefficient vector $\mathbf{p}$ by solving the convex-concave optimization of the MKL dual formulation [66]:

$$\min_{\mathbf{p} \in \Delta} \max_{\alpha \in \mathcal{Q}} \widehat{\mathcal{L}}(\boldsymbol{\alpha}, \mathbf{p}) = \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2} (\boldsymbol{\alpha} \circ \mathbf{y})^\top \mathbf{K}(\mathbf{p}) (\boldsymbol{\alpha} \circ \mathbf{y}), \tag{6.2}$$

where $\circ$ denotes the Hadamard (element-wise) product, $\mathbf{1}$ is a vector of all ones, and $\mathcal{Q} = \{\boldsymbol{\alpha} \in [0, C]^m\}$ is the domain for dual variables $\boldsymbol{\alpha}$. Note that this formulation can be considered as the dual formulation of SVM for the combined kernel.

One popular choice for domain Δ is $\Delta_2 = \left\{\mathbf{p} \in \mathbb{R}^s_+ : \|\mathbf{p}\|_2 \leq 1\right\}$. Often the L_1 norm is used to generate a sparse solution, however, in our application, the small sample size impacted the accuracy of this approach.

For MKL, each individual attribute is considered as a separate feature by constructing one kernel for each attribute (resulting in 25 base kernels). Our MKL classifier was trained using an off-the-shelf MKL tool [131].

Once this training is complete, we are able to measure the similarity of a caricature and photograph (represented as the difference vector $\mathbf{x}$) by: $f(\mathbf{x}) = \sum_{i=1}^{n_s} \boldsymbol{\alpha}_i y_i \mathbf{K}_p(\mathbf{x}_i, \mathbf{x})$, where n_s is the number of support vectors, and $\mathbf{K}_p(\cdot)$ is the combined matrix. In Section 6.7, we refer to this method as *MKL*.

In addition to the MKL algorithm, we also use the standard SVM algorithm [19] by replacing the multiple kernel matrix $\mathbf{K}_p(\cdot)$ with a single kernel $\mathbf{K}(\cdot)$ that utilizes

all feature components together. In Section 6.7, we refer to this approach as *SVM*. Both the MKL and SVM algorithms used RBF kernels (the kernel bandwidth was determined empirically).

Finally, we introduce a method known as the nearest neighbor MKL (NN_{MKL}). Because the vector $\mathbf{p}$ in Eq. 6.2 assigns weight to each of the 25 qualitative features, we can explicitly use these weights to perform weighted nearest neighbor matching. Thus, the dissimilarity between a caricature and photograph is measured as the sum of weighted differences between each of the qualitative feature vectors.

6.6 Image Descriptor-based Recognition

As discussed, encoding facial images with low level feature descriptors such as local binary patterns [99] is challenged with respect to matching caricatures to photograph due to the misalignments caused from the feature exaggeration in caricatures. However, since this approach has seen success in matching facial sketches to photographs [12, 54, 59], we also employ a similar technique for the caricature matching task.

The first step in the image descriptor-based algorithm is to align the face images using the two eye locations. These locations are marked manually due to the wide variations of pose in both the photographs and (especially) the caricatures. Using the center of the eyes, we performed planar rotation to fix the face upright, scaled the image to 75 pixels between the eyes, and cropped the image to a height of 250 pixels and a width of 200 pixels.

For both caricatures and photographs, we densely sampled local binary pattern histograms from image patches of 32 by 32 pixels. Next, all of the LBP histograms computed from a single image are concatenated into a single feature vector. Finally, we performed feature-based random subspace analysis [52] by randomly sampling the

feature space b times. For each of the b subspaces, linear discriminant analysis (LDA) is performed to extract discriminative feature subspaces [10]. In Section 6.7 we will refer to this method as *LBP with LDA*.

6.7 Experimental Results

The 196 pairs of caricatures and photographs (see Section 6.3), were randomly split such that 134 pairs (roughly 2/3rd) were made available for training, and 62 pairs (roughly 1/3rd) was available for testing. These sets were non-overlapping (i.e. no subject used in training was used for testing). We partitioned the data into training and testing sets 10 different times, resulting in 10 different matching experiments. The results shown in this section are the mean and standard deviation of the matching accuracies from those 10 random partitions. The precise splits used for these experiments are included with the release of the caricature image dataset.

The performance of each matching algorithm was measured using both the cumulative match characteristic (CMC) and the receiver operating characteristic (ROC) curves. For the CMC scores, we list the Rank-1 and Rank-10 accuracies. With 62 subjects available for testing, the gallery size was 62 images (photographs), and the scores listed are the average rank retrieval when querying the gallery with the 62 corresponding caricatures. The ROC analysis is listed as the true accept rate (TAR) at fixed false accept rates (FAR) of 1.0% and 10.0%.

Table 6.1 and Table 6.2 lists the identification and retrieval accuracies (respectively) for each of the recognition algorithms discussed in this work. Even without learning the qualitative features (NN_{L_2}) still had a higher accuracy than the image descriptor-based method (LBP with LDA). Thus, while image descriptor-based methods work well in matching verdical sketches to photographs [59], the misalignments caused by the exaggerations in the caricatures challenge this method. At a false ac-

cept rate of 10.0%, several of the proposed learning methods (*Logistic Regression,* NN_{MKL}*, and SVM*) are able to improve the accuracy of the qualitative features by around 10%. Despite the inability of the *MKL* method to improve the matching accuracy, using the weights from MKL with the nearest neighbor matcher (NN_{MKL}) improves the matching accuracy.

Because the classification algorithms used in this study output numerical values that indicate the similarity of a caricature image and a photograph, we are able to leverage fusion techniques to further improve the accuracy. Fusion of algorithms in Table 6.1 and Table 6.2 are denoted by the a '+' symbol between algorithms names. This indicates the use of sum of score fusion with min-max score normalization [123].

Using only qualitative features, the matching accuracy (at FAR=10.0%) was improved to nearly 57% (using *Logistic Regression+*NN_{MKL}*+SVM*). While the image descriptor-based method performed poorly with respect to the qualitative features, it proved valuable when added to the fusion process: *Logistic Regression+*NN_{MKL}*+SVM+LBP with LDA* had an accuracy of 61.9%.

Using the estimated $\mathbf{p}$ vector in the multiple kernel learning (MKL) algorithm, we are able to interpret the relative importance of each of the qualitative features. Since each component of $\mathbf{p}$ corresponds to the weight assigned to each of the 25 qualitative features, we can loosely interpret this vector to understand which features provided the most discriminative information. Figure 6.6 lists the weights for each of the 25 facial features. Surprisingly, we see that the Level 1 qualitative features are more discriminative than the Level 2 facial features. While this is counter to a standard face recognition task [56], caricatures are different in nature than face images. We believe the relative importance of Level 1 facial features in this setting is akin to the information an artist filters from the face.

6.8 Summary

This chapter introduced a challenging new problem in heterogeneous face recognition: matching facial caricatures to photographs. Unlike the other heterogeneous face recognition scenarios encountered in this paper, the development of a caricature recognition system does not have direct societal benefits. However, the indirect benefits of such research may be substantial. Given the human ability to ascertain identify from these extremely exaggerated sketches, designing common facial representations for both caricatures and photographs is akin to mimic human facial representations.

In order to facilitate research in caricature matching, we released the initial dataset of 196 pairs of caricatures and photographs used in this study in order to allow other researchers to study this problem.

A major contribution of this research is the definition of a set of qualitative facial features for representing both caricatures and photographs. Given these representations, a suite of statistical learning algorithms were adopted to learn the most salient combinations of these features from a training set of caricature and photograph pairs.

Feature Name	Weight	Feature Name	Weight
Hairstyle 1	*2.86*	*Almond Eyes*	*0.21*
Beard	*0.85*	*Nose (Up or Down)*	*0.21*
Mustache	*0.81*	*Face Shape*	*0.20*
Hairstyle 2	*0.70*	*Forehead Size*	*0.19*
Eyebrows (Up or Down)	*0.45*	*Eye Color*	*0.18*
Nose to Mouth Distance	*0.43*	*Sleepy Eyes*	*0.14*
Eye Seperation	*0.43*	*Sharp Eyes*	*0.13*
Nose Width	*0.42*	*Baggy Eyes*	*0.12*
Face Length	*0.27*	*Nose to Eye Distance*	*0.12*
Cheeks	*0.27*	*Thick Eyebrows*	*0.11*
Mouth Width	*0.26*	*Eyebrows Connected*	*0.10*
Mouth to Chin Distance	*0.23*	*Slanted Eyes*	*0.10*
Eyebrow Shape	*0.22*		

Figure 6.6: The multiple kernel learning (MKL) weights ($\mathbf{p}$), scaled by 10, for each of the qualitative features. Higher weights indicate more informative features.

Chapter 7

Summary and Conclusions

This thesis studied the problem of heterogeneous face recognition across both specific and broad applications. The primary contributions were realized by advancing facial feature representations to better handle heterogeneous face images, and adapting feature extraction algorithms (i.e. statistical learning) to better leverage training data exemplar to a particular heterogeneous face matching task.

7.1 Contributions

In Chapter 2 we developed a framework for matching forensic sketches to facial photographs that offered the following contributions:

- Presented the first large-scale experiment on automated identification using operational forensic sketches.
- In encoding sketches and photographs with SIFT and LBP feature descriptors, we proposed the first feature-based approach to automated sketch recognition.
- Developed a recognition framework called local feature-based discriminant analysis, which demonstrated a substantial improvement in matching viewed

sketches to photos over both previously published algorithms and a state of the art commercial face recognition system.

- Applied race and gender filtering to further improve sketch recognition accuracy.

The prototype-based framework in Chapter 3 offered the following contributions:

- Presented a prototype-based representation for heterogeneous face images. This approach represents images from each modality as a vector of their similarities to a common set of prototypes.

- Improved the recognition accuracy of the prototype features by applying linear discriminant analysis on randomly sampled prototype features, resulting in a final framework called prototype-random subspaces (P-RS).

- The P-RS framework provides a method for computing inter-modality similarities by using only intra-modality similarity measures, thus extending it (conceptually) to any heterogeneous face recognition scenario in which intra-modality similarities can be computed.

- Demonstrated the ability of the P-RS framework to perform face recognition using feature templates from alternate facial feature representation (e.g. matching LBP to SIFT).

The studies of facial aging presented in Chapter 4 offer the following contributions:

- Performed the largest facial aging study to date by using a dataset of 200,000 mug shot images from 64,000 subjects with time lapses up to 17 years between images of the same subject.

- Demonstrated a degradation in face recognition performance from two leading commercial-of-the-shelf (COTS) face recognition systems (FRS) as a function of the amount of time lapse that has occurred between two face images.

- Showed that training a face recognition system on a particular amount of time lapse resulted in the highest recognition accuracy on that time lapse.

- The previous finding suggests to the notion of periodically updating face templates to reside in feature subspaces that are trained for the amount of time that has lapsed since the image was acquired.

The study on the impact of the demographic distributions on face recognition performance from over 50,000 subjects presented in Chapter 5 offered the following contributions:

- A demonstration that female, black, and younger cohorts are more difficult to recognize for six different face matchers (commercial, non-trainable, and trainable).

- That training a matcher exclusively on the images of a particular racial cohort will result in improved recognition accuracies on that cohort.

- The notion of dynamic face matcher selection is presented, where multiple face recognition systems that are each trained on different demographic cohorts are available for an operator to use in matching with the goal of improving face retrieval performance.

Finally, the study on caricature recognition in Chapter 6 offered the following contributions:

- The first study on matching caricatures to photographs was presented.

- Developed a set of qualitative facial features for representing both caricatures and photographs. The features resulted in the highest accuracy for matching caricatures to their photographs.

- Adapted classification algorithms to categorize the difference vector between a caricature and a photograph as being either a match or non-match.

- Collected a database of caricature and photograph pairs that are freely available to other researchers.

7.2 Future Work

Though contributions were offered across a range of challenges in heterogeneous face recognition, the studies presented in this thesis also segue into many new research challenges that are both within the scope of heterogeneous face recognition and beyond.

The studies on forensic sketch recognition developed a system that had near perfect accuracy on accurate viewed sketches, while the accuracy on real world forensic sketches was significantly lower. The chief similarity between these two types of data is that they are both hand drawn sketches. Thus, we see that the true heterogeneous aspect of this problem (i.e. matching a sketch to a photo) is largely solved. However, the difference between viewed and forensic sketches is that forensic sketches rely on the human memory. While many researchers are still seeking to improve the trivial problem of sketch recognition on viewed sketches, it is only through a more in depth examination of forensic sketches that we make the gains needed to close the gap between the two forms of sketches. Researchers must make greater use of the previous findings in the cognitive science literature on human witness memory to help shape the next generation of sketch recognition algorithms.

The prototype-based framework for performing heterogeneous face recognition demonstrated that heterogeneous facial features could be used to match heterogeneous face images. While this was notably demonstrated on the difficult scenario of matching thermal images to photographs (a task which humans struggle with), more

ambitious scenarios can now be considered. One such example is the matching of face depth maps acquired from a LIDAR (Light Detection And Ranging) sensor. LIDAR sensors have demonstrated the ability to acquire high resolution images across large distances, making for an ideal heterogeneous face recognition scenario in intelligence and law enforcement applications.

The prototype framework may also be extended to problems outside the scope of face recognition. For example, the prototype-based representation should be explored in the context of cross-lingual retrieval, as well as heterogeneous image retrieval challenges (such as querying a database of large images with much smaller thumbnail images).

Both the study on facial aging and demographics demonstrated the tight coupling between face matcher accuracy and the dataset with which it was trained on. Though quite intuitive, this idea that the demographic distribution of a dataset could be controlled to generate several different versions of a face recognition system had not previously been explored. While the experimental findings in the two chapters on aging and demographics indirectly demonstrated the benefit of these approaches (namely, template update over time and the use of dynamic face matcher selection), more explicit experiments should be conducted. While such follow on studies are perhaps more suited for system-related research (as opposed to pattern recognition research), they are, none the less, important.

Perhaps the study on caricature recognition presented in this thesis offers the most avenues for future research. The release of the dataset will hopefully prompt other researchers to explore orthogonal ideas on how to solve the problem. The union of many different approaches to this problem should in turn yield a wide set of approaches that can then be extended to the standard face recognition problem. Within our approach of encoding caricatures and photographs using qualitative features, several new avenues for research exist. One of which is the application of the qualitative

features to face matching and retrieval.

7.3 Conclusions

A thesis comprised of a set of studies on heterogeneous face recognition has been presented. In each study contributions are made to improve face recognition performance given heterogeneous forms of data. The results of these studies is an improvement to specific problems that are of interest in law enforcement, defense, and intelligence applications.

Appendices

Appendix A

"R" Transform is a Special Case of Eigen-transform

Given the matrix of probe features K_p and gallery features K_g, Tang and Wang's eigen-transformation method [138] performs a transformation from the probe face modality to the gallery face modality by first performing the eigen-decomposition using the dominant eigenvector method

$$(K_p K_p^{\mathrm{T}}) K_p V_p = K_p V_p \Lambda_p \tag{A.1}$$

$$(K_g K_g^{\mathrm{T}}) K_g V_g = K_g V_g \Lambda_g \tag{A.2}$$

Here K_p and K_g are the matrices containing the kernel prototype similarities from the training instances as described in Eqs. (3.3) and (3.4). Given a feature vector ϕ' from the probe modality, the eigen-transformation method transforms (or synthesizes) ϕ' to the vector ϕ in the gallery modality by

$$U_p = K_p V_p \Lambda_p^{-1/2} \tag{A.3}$$

$$b_p = U_p^{\mathrm{T}} \phi' \tag{A.4}$$

$$c_p = V_p \Lambda_p^{-1/2} b_p \tag{A.5}$$

$$\phi = K_g c_p \tag{A.6}$$

We will now prove that Eq. (A.6) is equivalent to the R transform shown in Eq. (3.9), given the special case that K_p and K_g are square matrices (as is always the case in our framework). For terseness, $\phi(P)$ and $\phi'(P)$ (from Eq. (A.6)) are simply represented as ϕ and ϕ' (respectively).

Expanding Eq. (A.6) we have

$$\phi = K_g c_p \tag{A.7}$$

$$= K_g V_p \Lambda_p^{-1/2} b_p \tag{A.8}$$

$$= K_g V_p \Lambda_p^{-1/2} (\Lambda_p^{-1/2})^{\mathrm{T}} V_p^{\mathrm{T}} K_p^{\mathrm{T}} \phi' \tag{A.9}$$

$$= K_g V_p \Lambda_p^{-1} V_p^{\mathrm{T}} K_p^{\mathrm{T}} \phi' \tag{A.10}$$

Now, going back to Eq. (A.6), we see that we transform ϕ' to ϕ as

$$\phi = R\, \phi' \tag{A.11}$$

$$= K_g (K_p^{\mathrm{T}} K_p)^{-1} K_p^{\mathrm{T}} \phi' \tag{A.12}$$

If Eq. (A.6) was equivalent to Eq. (3.9), then, from Eq. (A.10) and Eq. (A.12), we see that it must be true that $V_p \Lambda_p^{-1} V_p^{\mathrm{T}} = (K_p^{\mathrm{T}} K_p)^{-1}$. By definition, we have

$$K_p^{\mathrm{T}} K_p = V_p \Lambda_p V_p^{\mathrm{T}} \tag{A.13}$$

Because V_p is a square, orthonormal matrix (and, thus, $V_p^{\mathrm{T}} = V_p^{-1}$), we see that

$$\begin{aligned} (K_p^{\mathrm{T}} K_p)^{-1} &= (V_p \Lambda_p V_p^{\mathrm{T}})^{-1} && \text{(A.14)} \\ &= V_p \Lambda_p V_p^{\mathrm{T}} && \text{(A.15)} \end{aligned}$$

Thus, given the special case that K_p and K_g are square matrices, the proposed "R" transform in Chapter 3 is in fact equivalent to Tang and Wang's eigen-transformation method [138].

Bibliography

Bibliography

[1] FaceVACS Software Developer Kit, Cognitec Systems GmbH, http://www.cognitec-systems.de.

[2] PittPatt Face Recognition SDK, Pittsburgh Pattern Recognition, http://www.pittpatt.com.

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 28(12):2037–2041, Dec. 2006.

[4] T. Akgul. Introducing the cartoonist, Tayfun Akgul. *IEEE Antennas and Propagation Magazine*, 49(3):162, 2007.

[5] T. Akgul. Can an algorithm recognize montage portraits as human faces? *IEEE Signal Processing Magazine*, 28(1):160 –158, 2011.

[6] E. Akleman. Making caricatures with morphing. In *Proc. ACM SIGGRAPH*, 1997.

[7] E. Akleman. Modeling expressive 3d caricatures. In *Proc. ACM SIGGRAPH*, 2004.

[8] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[9] M.-F. Balcan, A. Blum, and S. Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65:79–94, 2006.

[10] P. Belhumeur, J. Hespanda, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 19(7):711–720, 1997.

[11] A. Bertillon. *The Bertillon System of Identification*. Chicago, IL, 1896.

[12] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. On matching sketches with digital face images. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, pages 1 –7, 2010.

[13] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063 – 1074, sept. 2003.

[14] R. K. Bothwell, J. Brigham, and R. Malpass. Cross-racial identication. *Personality & Social Psychology Bulletin*, 15:1925, 1989.

[15] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[16] S. Brennan. Caricature generator: The dynamic exaggeration of faces by computer. *Leonardo*, 18:170–178, 1985.

[17] V. Bruce, Z. Henderson, K. Greenwood, P. Hancock, A. Burton, and P. Miller. Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, 5(4):339–360, 1999.

[18] V. Bruce and A. Young. Understanding face recognition. *British Journal of Psychology*, 77(3), 1986.

[19] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[20] P. Chiroro and T. Valentine. An investigation of the contact hypothesis of the own-race bias in face recognition. *Quarterly Journal of Experimental Psychology, A, Human Experimental Psychology*, 48A:879894, 1995.

[21] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995.

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[23] N. R. Council. *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press, 2009.

[24] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

[25] G. Edmond, K. Biber, R. Kemp, and G. Porter. Laws looking glass: Expert identification evidence derived from photographic and video images. *Current Issues in Criminal Justice*, 20(3):337–377, 2009.

[26] M. P. Evison and R. W. Vorder Bruegge, editors. *Computer-aided Forensic Facial Comparison*. CRC Press, 2010.

[27] C. Frowd, V. Bruce, A. McIntyr, and P. Hancock. The relative importance of external and internal features of facial composites. *British Journal of Psychology*, 98(1):61–77, 2007.

[28] N. Furl, P. J. Phillips, and A. J. O'Toole. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive Science*, 26(6):797 – 815, 2002.

[29] X. Gao, J. Zhong, J. Li, and C. Tian. Face sketch synthesis algorithm based on e-hmm and selective ensemble. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(4):487–496, April 2008.

[30] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 23(6):643 –660, jun 2001.

[31] L. Gibson. *Forensic Art Essentials*. Elsevier, 2008.

[32] M. A. Goodale and A. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

[33] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 26(4):449 –465, 2004.

[34] P. J. Grother, G. W. Quinn, and P. J. Phillips. MBE 2010: Report on the evaluation of 2d still-image face recognition algorithms. *National Institute of Standards and Technology, NISTIR*, 7709, 2010.

[35] G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, 2010.

[36] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: global versus component-based approach. In *Proc. of Int. Conf. on Computer Vision*, 2001.

[37] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 20(8):832–844, Aug 1998.

[38] R.-L. Hsu and A. Jain. Generating discriminating cartoon faces using interacting snakes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(11):1388 – 1398, 2003.

[39] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Technical Report 07-49, University of Massachusetts, Amherst*, 2007.

[40] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. Circuits and Systems for Video Technology*, 14(1):4–20, Jan. 2004.

[41] A. K. Jain, Y. Chen, and M. Demirkus. Pores and ridges: High-resolution fingerprint matching using level 3 features. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(1):15–27, 2007.

[42] A. K. Jain, S. C. Dass, K. Nandakumar, and K. N. Soft biometric traits for personal recognition systems. In *Proc. of International Conference on Biometric Authentication*, 2004.

[43] A. K. Jain, B. Klare, and U. Park. Face matching and retrieval: Applications in forensics. *IEEE Multimedia*, 19(1):20–28, 2012.

[44] A. K. Jain, B. F. Klare, and U. Park. Face recognition: Some challenges in forensics. In *Proc. Int. Conference on Automatic Face and Gesture Recognition*, 2011.

[45] H. Y. Jie, H. Yu, and J. Yang. A direct LDA algorithm for high-dimensional data – with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.

[46] L. Juwei, K. Plataniotis, and A. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Trans. Neural Networks*, 14(1):117 – 126, 2003.

[47] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.

[48] K. I. Kim, K. Jung, and H. J. Kim. Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters*, 9(2):40 –42, 2002.

[49] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Staistica Sinica*, 16:375–390, 2006.

[50] B. Klare. Spectrally sampled structural subspace features (4SF). In *Michigan State University Technical Report, MSU-CSE-11-16*, 2011.

[51] B. Klare, S. S. Bucak, T. Akgul, , and A. K. Jain. Towards automated caricature recognition. In *Proc. Int. Conference on Biometrics*, 2012.

[52] B. Klare and A. Jain. Heterogeneous face recognition: Matching NIR to visible light images. In *Proc. International Conference on Pattern Recognition*, 2010.

[53] B. Klare and A. Jain. Sketch to photo matching: A feature-based approach. In *Proc. SPIE Conference on Biometric Technology for Human Identification VII*, 2010.

[54] B. Klare and A. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Trans. on Pattern Analysis and Machine Intelligence (under review)*, pages 639–646, 2011.

[55] B. Klare and A. Jain. Matching forensic sketches and mug shots to apprehend criminals. *IEEE Computer*, 44(5):94–96, 2011.

[56] B. Klare and A. K. Jain. On a taxonomy of facial features. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010.

[57] B. Klare and A. K. Jain. Face recognition across time lapse: On learning feature subspaces. In *Int. Joint Conference on Biometrics*, 2011.

[58] B. Klare, Z. Li, and A. Jain. On matching forensic sketches to mugshot photos. In *MSU Technical Report, MSU-CSE-10-3*, 2010.

[59] B. Klare, Z. Li, and A. Jain. Matching forensic sketches to mugshot photos. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011.

[60] B. Klare, A. Paulino, and A. K. Jain. Analysis of facial features in identical twins. In *Int. Joint Conference on Biometrics*, 2011.

[61] B. F. Klare and M. Burge. Assessment of H.264 video compression on automated face recognition performance in surveillance and mobile video scenarios. In *Proc. of SPIE, Biometric Technology for Human Identification VII*, 2010.

[62] B. F. Klare, P. Mallapragada, A. Jain, , and K. Davis. Clustering face carvings: Exploring the devatas of angkor wat. In *Proc. Int. Conference on Pattern Recognition*, 2010.

[63] B. F. Klare and S. Sarkar. Background subtraction in varying illuminations using an ensemble based on an enlarged feature set. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2009.

[64] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami. On kansei facial image processing for computerized facial caricaturing system picasso. In *Proc. IEEE Conference on Systems, Man, and Cybernetics*, 1999.

[65] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. IEEE Int. Conference on Computer Vision*, 2009.

[66] G. Lanckriet, N. Cristianini, P. Bartlett, and L. E. Ghaoui. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[67] P.-H. Lee, G.-S. Hsu, and Y.-P. Hung. Face verification and identification using facial trait code. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1613–1620, June 2009.

[68] Z. Lei and S. Li. Coupled spectral regression for matching heterogeneous faces. In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, pages 1123 –1128, june 2009.

[69] D. A. Leopold, A. J. O'Toole, T. Vetter, and V. Blanz. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, 4:89–94, 2001.

[70] T. Lewiner, T. Vieira, D. Martinez, A. Peixoto, V. Mello, and L. Velho. Interactive 3d caricature from harmonic exaggeration. *Computers and Graphics*, 35(3):586–595, 2011.

[71] J. Li, P. Hao, C. Zhang, and M. Dou. Hallucinating faces from thermal infrared images. In *Proc. Int. Conference on Image Processing*, pages 465 –468, 2008.

[72] S. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 29(4):627 –639, 2007.

[73] S. Z. Li and A. K. Jain, editors. *Handbook of Face Recognition*. Springer, 2nd edition, 2011.

[74] Y. Li, M. Savvides, and V. Bhagavatula. Illumination tolerant face recognition using a novel face from sketch synthesis approach and advanced correlation filters. In *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, 2006.

[75] Z. Li, D. Lin, and X. Tang. Nonparametric discriminant analysis for face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 31(4):755 –761, 2009.

[76] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE Trans. Information Forensics and Security*, 6(3):1028–1037, 2011.

[77] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li. Heterogeneous face recognition from local structures of normalized appearance. In *Proc. Int. Conference on Biometrics*, 2009.

[78] D. Lin and X. Tang. Inter-modality face recognition. In *Proc. of European Conference on Computer Vision*, 2006.

[79] H. Ling, S. Soatto, N. Ramanathan, and D. Jacobs. Face verification across age progression using discriminative methods. *IEEE Trans. on Information Forensics and Security*, 5(1):82 –91, 2010.

[80] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. on Image Processing*, 11(4):467–476, 2002.

[81] Q. Liu, H. Lu, and S. Ma. Improving kernel fisher discriminant analysis for face recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(1):42 – 49, 2004.

[82] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, pages 1005–1010, 2005.

[83] W. Liu, X. Tang, and J. Liu. Bayesian tensor inference for sketch-based facial photo hallucination. In *Proc. of 20th International Joint Conference on Artificial Intelligence*, 2007.

[84] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[85] J. Lu, K. Plataniotis, and A. Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*, 26(2):181 – 191, 2005.

[86] G. Mahalingam and C. Kambhamettu. Age invariant face recognition using graph matching. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010.

[87] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar. Handbook of Fingerprint Recognition. pages 39–40. Springer, 2009.

[88] X. Mao, A. W. Bigham, R. Mei, G. Gutierrez, K. M. Weiss, T. D. Brutsaert, F. Leon-Velarde, L. G. Moore, E. Vargas, P. M. McKeigue, M. D. Shriver, and E. J. Parra. A genomewide admixture mapping panel for hispanic/latino populations. *The American Journal of Human Genetics*, 80(6):1171 – 1178, 2007.

[89] A. Martinez and R. Benavente. The AR face database. In *CVC Technical Report 24*, 1998.

[90] R. Mauro and M. Kubovy. Caricature and face recognition. *Memory & Cognition*, 20(4):433–440, 1992.

[91] G. McCarthy, A. Puce, J. C. Gore, and T. Allison. Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, 9(5):605–610, 1997.

[92] K. Messer, J. Matas, J. Kittler, and K. Jonsson. XM2VTSDB: The extended M2VTS database. In *Proc. of Audio and Video-based Biometric Person Authentication*, 1999.

[93] E. Meyers and L. Wolf. Using biologically inspired features for face processing. *Int. Journal of Computer Vision*, 76(1):93–104, 2008.

[94] E. Meyers and L. Wolf. Using biologically inspired features for face processing. *Int. Journal of Computer Vision*, 76(1):93–104, 2008.

[95] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, Oct. 2005.

[96] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771 – 1782, 2000.

[97] W. Ng and R. C. Lindsay. Cross-race facial recognition: Failure of the contact hypothesis. *Journal of Cross-Cultural Psychology*, 25:217232, 1994.

[98] H. Nizami, J. Adkins-Hill, Y. Zhang, J. Sullins, C. McCullough, S. Canavan, and L. Yin. A biometric database with rotating head videos and hand-drawn face sketches. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2009.

[99] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 24(7):971–987, 2002.

[100] A. O'Toole, P. J. Phillips, A. Narvekar, F. Jiang, and J. Ayyad. Face recognition algorithms and the other-race effect. *Journal of Vision*, 8(6), 2008.

[101] S. Pankanti, S. Prabhakar, and A. K. Jain. On the individuality of fingerprints. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:1010–1025, 2002.

[102] U. Park and A. K. Jain. 3d model-based face recognition in video. In *Proc. International Conference on Biometrics*, 2007.

[103] U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *IEEE Trans. on Information Forensics and Security*, 6(3):1028–1037, 2011.

[104] U. Park, Y. Tong, and A. Jain. Age-invariant face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(5):947–954, 2010.

[105] J. Parris, M. Wilber, B. Helfin, H. Rara, A. E. barkouky Aly Farag, J. Movellan, M. Santana, J. Lorenzo, M. N. Teli, S. Marcel, and C. Atanasoaei. Face and eye detection on hard datasets. In *Int. Joint Conference on Biometrics*, 2011.

[106] D. A. Patterson and J. L. Hennessy. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[107] E. Patterson, A. Sethuram, M. Albert, K. Ricanek, and M. King. Aspects of age variation in facial morphology affecting biometrics. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2007.

[108] P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *Proc. of Automatic Face Gesture Recognition*, 2011.

[109] P. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, 1997.

[110] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 32(5):831 –846, 2010.

[111] P. J. Phillips, P. J. Grother, R. J. Micheals, D. Blackburn, E. Tabassi, and J. M. Bone. Face recognition vendor test 2002: evaluation report. *National Institute of Standards and Technology, NISTIR*, 6965, 2003.

[112] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, 2008.

[113] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Trans. on Image Processing*, 15(11):3349 –3361, 2006.

[114] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131 – 144, 2009.

[115] S. Raudys and A. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 13(3):252 –264, 1991.

[116] A. Rawls and K. Ricanek. Morph: Development and optimization of a longitudinal age progression database. In *Biometric ID Management and Multimodal Communication*, 2009.

[117] L. Redman. *How to draw caricatures*. McGraw-Hill, 1984.

[118] G. Rhodes, S. Brennan, and S. Carey. Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, 19(4):473–497, 1987.

[119] H. T. F. Rhodes. *Alphonse Bertillon, Father of Scientific Detection*. Abelard-Schuman, New York, 1956.

[120] K. Ricanek, A. Sethuram, E. K. Patterson, A. M. Albert, and E. J. Boone. *Craniofacial Aging*. John Wiley & Sons, Inc., 2008.

[121] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *Proc. of Automatic Face and Gesture Recognition*, 2006.

[122] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):10191025, 1999.

[123] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, 2003.

[124] T. Sakai, M. Nagao, and T. Kanade. Computer analysis and classification of photographs of human faces. In *Proc. First USA-JAPAN Computer Conference*, pages 55–62, 1972.

[125] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[126] L. G. Shapiro and G. C. Stockman. *Computer Vision*. Prentice Hall, 2001.

[127] P. N. Shapiro and S. D. Penrod. Meta-analysis of face identication studies. *Psychological Bulletin*, 100:139156, 1986.

[128] L. Shen and L. Bai. A review on gabor wavelets for face recognition. *Pattern Analysis & Applications*, 9:273–292, 2006.

[129] R. Singh, M. Vatsa, A. Noore, and S. Singh. Age transformation for improving face recognition performance. In *Pattern Recognition and Machine Intelligence*, 2007.

[130] R. L. Solso and J. E. McCarthy. Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, 72(4):499–503, 1981.

[131] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

[132] N. Spaun. Facial comparisons by subject matter experts: Their role in biometrics and their training. In *Int. Conference on Advances in Biometrics*, 2009.

[133] N. A. Spaun. Forensic biometrics from images and video at the federal bureau of investigation. In *Int. Conference on Biometrics: Theory, Applications and Systems*, page 13, 2007.

[134] Z. Sun, A. Paulino, J. Feng, Z. Chai, T. Tan, and A. K. Jain. A study of multibiometric traits of identical twins. In *Proc of SPIE, Biometric Technology for Human Identification VII*, 2010.

[135] J. Suo, S.-C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 32(3):385 –401, 2010.

[136] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. on Image Processing*, 19(6):1635 –1650, 2010.

[137] X. Tang and X. Wang. Face sketch synthesis and recognition. In *Proc. of IEEE International Conference on Computer Vision*, pages 687–694, 2003.

[138] X. Tang and X. Wang. Face sketch recognition. *IEEE Trans. Circuits and Systems for Video Technology*, 14(1):50–57, 2004.

[139] K. Taylor. *Forensic Art and Illustration*. CRC Press, 2001.

[140] P. Thompson. Margaret Thatcher: A new illusion. *Perception*, 9(4):483–484, 1980.

[141] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[142] D. Y. Tsao, W. A. Freiwald, R. B. Tootell, and M. S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, Feb 2006.

[143] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[144] R. Uhl and N. Lobo. A framework for recognizing a facial image from a police sketch. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 1996.

[145] T. Valentine and V. Bruce. The effects of distinctiveness in recognising and classifying faces. *Perception*, 15(5):525–535, 1986.

[146] P. Viola and M. J. Jones. Robust real-time face detection. *Int. Journal of Computer Vision*, 57:137–154, 2004.

[147] X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, 2004.

[148] X. Wang and X. Tang. Random sampling for subspace face recognition. *Int. Journal of Computer Vision*, 70(1):91–104, 2006.

[149] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 31(11):1955–1967, Nov. 2009.

[150] C. Williams and M. Seeger. Using the Nystrom method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 15(13), 2001.

[151] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 19(7):775 –779, 1997.

[152] B. Xiao, X. Gao, D. Tao, and X. Li. A new approach for face recognition by sketches in photos. *Signal Processing*, 89(8):1576 – 1588, 2009.

[153] R. Yampolskiy, B. F. Klare, , and A. K. Jain. Face recognition in the virtual world: recognizing avatar faces. In *Proc. of SPIE, Biometric Technology for Human Identification IX*, 2012.

[154] D. Yi, S. Liao, Z. Lei, J. Sang, and S. Li. Partial face matching between near infrared and visual images in mbgc portal challenge. In *Proc. Int. Conference on Biometrics*, pages 733–742. 2009.

[155] A. W. Young, D. Hay, K. H. McWeeny, B. M. Flude, and A. W. Ellis. Matching familiar and unfamiliar faces on internal and external features. *Perception*, 14:737–746, 1985.

[156] P. Yuen and C. Man. Human face image searching system using sketches. *IEEE Trans. Systems, Man and Cybernetics*, 37(4):493–504, July 2007.

[157] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *Proc. European Conference on Computer Vision*. 2010.

[158] J. Zhong, X. Gao, and C. Tian. Face sketch synthesis using e-hmm and selective ensemble. In *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing*, 2007.

[159] G. Zhu, D. L. Duffy, A. Eldridge, M. Grace, C. Mayne, L. O'Gorman, J. F. Aitken, M. C. Neale, N. K. Hayward, A. C. Green, and N. G. Martin. A major quantitative-trait locus for mole density is linked to the familial melanoma gene CDKN2A: A maximum-likelihood combined linkage and association analysis in twins and their sibs. *The American Journal of Human Genetics*, 65(2):483–492, 1999.