Face Recognition: Role of Aging and Quality Covariates

By

Lacey Best-Rowden

A Dissertation

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

Computer Science – Doctor of Philosophy

2016

Abstract

Face Recognition: Role of Aging and Quality Covariates

By

Lacey Best-Rowden

A technology once seen only in television dramas, automatic face recognition systems are now deployed in many important applications. Recognition of individuals from facial images is used for de-duplication of identification cards (*e.g.,* driver's licenses and passports), verification of prisoner identities, and tag suggestions for personal photo collections. Face images acquired in such applications are conducive to the current capabilities of face recognition algorithms; state-of-the-art systems are able to recognize *constrained* face images with close to 99% accuracy. However, the performance of automatic face recognition degrades when processing *unconstrained* face images (*i.e.,* image acquisition is uncontrolled and subjects may be uncooperative). In such scenarios, a face image may simultaneously contain multiple confounding factors, or *covariates*, such as variations in facial pose, illumination, expression, occlusion, resolution, and facial aging.

The first contribution of this dissertation is a framework for matching a collection of unconstrained *face media* (images, videos, 3D model, demographics, facial sketch) when multiple instances of a subject's face are available. This is particularly relevant to forensic investigations where the goal is to identify a "person of interest" based on low quality face images and videos (*e.g.,* captured by surveillance cameras or mobile phones of bystanders) and other information compiled during the investigation (*e.g.,* gender, race, age, facial sketch). While traditional face matching methods generally take a single media (*i.e.,* a still face image, video track, or face sketch) as input, this work considers using the entire gamut of media as a probe to generate a single candidate list for the person of interest. We show that the proposed approach boosts the likelihood of correctly identifying the person of interest through the use of different fusion schemes, 3D face models, and incorporation of quality

measures for fusion and video frame selection.

Secondly, this dissertation proposes an automatic measure of the *quality* of an unconstrained face image, where quality is defined as a measure of the utility of a face image to automatic face recognition. A large database of unconstrained face images is first annotated with target quality labels using two methods: (i) human assessments of face image quality, and (ii) quality values computed from similarity scores. A support vector regression model trained on image features automatically extracted using a deep convolutional neural network is then used to predict the quality of an unseen face image. Results demonstrate that target quality values from human assessments and similarity scores are not highly correlated with each other, but both are useful for applications of face image quality, such as to reject low-quality face images prior to matching and to rank a collection of face images based on quality.

Finally, this dissertation addresses the important problem of facial aging, which is a challenge for both constrained and unconstrained applications. The two underlying premises of automatic face recognition are uniqueness and permanence. We investigate the permanence property by addressing the following: Does face recognition ability of state-of-the-art systems degrade with elapsed time between enrolled and query face images? If so, what is the rate of decline with respect to the elapsed time? While previous studies have reported degradations in accuracy, no formal statistical analysis of large-scale longitudinal data has been conducted. We conduct such an analysis on two mugshot databases, which are the largest facial aging databases studied to date in terms of number of subjects, images per subject, and elapsed times. Longitudinal analysis shows that despite decreasing genuine scores, 99% of subjects can still be recognized at 0.01% FAR up to approximately 6 years elapsed time, and that age, sex, and race only marginally influence these trends. The methodology presented in this dissertation should be periodically repeated to determine age-invariant properties of face recognition as state-of-the-art evolves to better address facial aging.

Acknowledgments

I would like to extend my sincerest gratitude to my advisor, Dr. Anil Jain, to my parents, family, friends, and PRIP lab members. This thesis would not be possible without the overwhelming kindness and support that they have all given me throughout this journey.

Additional thanks to Patrick Grother and Mei Ngan at the National Institute of Standards and Technology (NIST) for collaboration with the longitudinal study of face recognition in this thesis, and to Jane Wankmiller and Sarah Krebs, sketch artists at the Michigan State Police.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Automatic face recognition systems are currently deployed in many important applications. Face recognition plays a key role in identity card de-duplication to prevent a person from obtaining multiple ID cards, such as driver's licenses and passports, under different names. Face recognition is used by the United States Department of Defense (DoD) to assist soldiers in determining friend or foe at security checkpoints and village assessments, and law enforcement officers in the field are able to capture face images with mobile devices, submit them to face recognition system on central servers, and quickly identify people who refuse to give their name, provide false information, or are injured and unresponsive. Face recognition systems are additionally utilized for surveillance purposes and access control to secure locations. Commercial applications of automatic face recognition are also now abundant, including "tag" suggestions on Facebook, organization of personal photo collections, and mobile phone unlock.

Face image acquisition conditions for many of these applications are conducive to the current capabilities of face recognition systems (*i.e.,* relatively controlled environments and/or cooperative subjects). Face images for identification documents require a neutral expression and no facial accessories, a uniform background, and controlled lighting. For example, de-duplication entails frontal-to-frontal face matching of controlled images. In these types

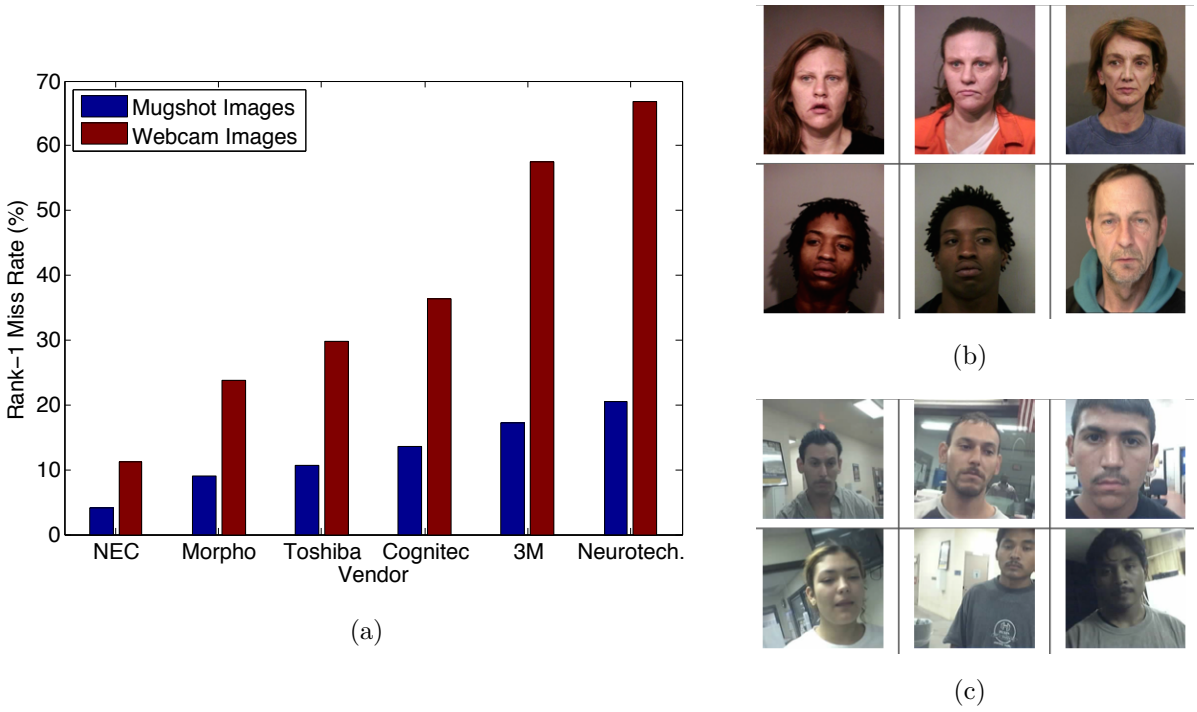


Figure 1.1 (a) Rank-1 miss rates of six vendors for closed-set identification of (b) mugshot and (c) webcam face images against a gallery of mugshot photos 1.6 million individuals, as reported by the NIST FRVT 2013 evaluation [55].

of scenarios, state-of-the-art commercial off-the-shelf (COTS) face recognition systems are highly accurate and have proven to be extremely useful. As of 2013, at least 37 states[1] are using face recognition technology to assist in the detection of fraudulent identification documents; the state of New York alone attributes more than 2,500 arrests in three years to the use of face recognition technology.[2] In terms of accuracy, a large-scale evaluation conducted by the National Institute of Standards and Technology (NIST) [55] demonstrated that error rates of the top performing COTS face recognition systems were lower than 10% for identifying mugshot face images at rank-1 against a gallery database of 1.6 million individuals (see Fig. 1.1).

As the demand for automatic recognition of individuals continues to increase, face offers

[1] www.washingtonpost.com/business/technology/state-photo-id-databases-become-troves-for-police/2013/06/16/6f014bd4-ced5-11e2-8845-d970ccb04497_story.html

[2] http://www.governor.ny.gov/news/governor-cuomo-announces-13000-identity-fraud-cases-investigated-dmv-using-facial-recognition

a number of advantages over other biometric traits (*e.g.,* fingerprint and iris): (i) Recognition by faces is how humans naturally interact with each other, so face images do not contain any information that people do not also disclose to the public on a daily basis. Face recognition tends to be more publicly accepted (compared to fingerprints, for example, which are commonly associated with criminal accusations). (ii) Large legacy face image databases already exist that can be searched against (*e.g.,* passport and driver's license). (iii) The face reveals other attributes (gender, race, age) that can be used as side information. (iv) Face can be captured unobtrusively, at a distance, and in a covert manner, if necessary. (v) No specialized sensors are required; digital cameras are readily available (*i.e.,* in mobile phones) and/or relatively inexpensive.

The above advantages of face biometric lend themselves to new emerging applications of face recognition, which are largely due to the increasing ubiquity of surveillance cameras and mobile imaging devices. According to a 2013 survey, there is one surveillance camera for every 11 people in the UK,[3] and a study conducted in 2014 estimates that the video surveillance market will reach $42 billion by the year 2019.[4] With recent tragic and controversial police-civilian incidents, such as the deaths of Michael Brown in Ferguson, Missouri, and Eric Garner in New York City, many police agencies are now equipping patrol officers with body cameras, and national debates are ensuing about whether police should be required to wear them at all times.[5,6] Personal collections of photos have also skyrocketed, as front-facing cameras on mobile phones sparked the "selfie" boom (*i.e.,* taking a picture of yourself) and an era of constant documentation of personal lives on social media.

Recent tragic events have made use of this increase in available imagery for solving high profile crimes. For example, the 2011 London riots, which resulted in one fatality, had

[3]http://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html

[4]http://www.securitysales.com/article/report_video_surveillance_market_to_reach_42b_by_2019

[5]http://www.npr.org/2015/04/10/398704487/eyewitness-video-a-controversial-tool-for-holding-police-accountable

[6]http://www.msnbc.com/msnbc/missouri-lawmaker-police-body-camera-footage

over 100,000 hours of surveillance footage for law enforcement officials to utilize.[7] The 2013 Boston marathon bombings resulted in four fatalities and more than 250 injured; again, law enforcement acquired a daunting amount of surveillance footage to sift through, as well as images and videos from the mobile phones of bystanders and marathon runners.[8] In both of these cases, large amounts of *manual* resources were immediately devoted to searching for investigative leads from the acquired media, and face images of suspects were released to the public for identification.

Made evident by these recent tragic events, in addition to countless other routine crimes (*e.g.,* robbery, kidnapping, assault), government and law enforcement officials could greatly benefit from *automated* (or semi-automated) face recognition to assist with identification of persons of interest. A face recognition system designed for the 2012 Olympics was available for use in the London riots but did not play a major role in identifying the rioters,[9] and there have been no reports that automatic face recognition was attempted for the Boston bombings. Although, a recent case study demonstrated that a state-of-the-art commercial face recognition system had the potential to identify one of the suspects, Dzhokhar Tsarnaev (the younger brother), at Rank-1 amongst one million mugshot images if he was in the database [73].

The success of face recognition technology in these scenarios is currently limited by the unconstrained nature of the imagery typically available. Accuracies of current COTS systems are highly sensitive to the quality of available face images. The large-scale face recognition evaluation by NIST (FRVT 2013 [55]), also reported that error rates of the top six COTS systems more than doubled when matching lower quality webcam images to the mugshot gallery (see Fig. 1.1). While the feasibility and utility of fully automated face recognition for surveillance purposes are limited, used as an investigative tool, face recognition can still assist law enforcement in searching for a list of suspects for manual examination.

[7] http://www.independent.co.uk/news/uk/crime/more-support-for-cctv-after-riots-2375768.html

[8] http://www.washingtonpost.com/wp-srv/special/national/boston-marathon-bombing-victims/

[9] http://latimesblogs.latimes.com/technology/2011/08/london-riots-facial-recognition-tech-being-used-by-police.html

In unconstrained scenarios where face image acquisition is not well controlled and subjects may be uncooperative (or unknowing), multiple factors which are known to confound the performance of face recognition systems are simultaneously present. Such confounding factors include facial pose, non-uniform illumination, facial expression, as well as occlusion and low image resolution.

- **Pose:** Facial pose can be categorized as in-plane (roll) or out-of-plane rotation (yaw and/or pitch). In-plane rotations can be corrected for with simple 2D transformations. However, when the head is rotated out-of-plane, certain regions of the face become "self-occluded" or no longer visible in the acquired face image (see Fig. 1.2(a)). This results in missing information and makes it difficult to determine correspondences between features of two faces at different poses.

- **Illumination:** For face images acquired in natural settings, ambient lighting can be drastically different depending on the setting (*e.g.,* indoor vs. outdoor) and is affected by daily changes even in a specific environment (*e.g.,* the amount of light coming in from windows on a particular day and time). The angle of the head with respect to the light source also causes changes in how the face is illuminated. Due to the three-dimensional structure of the face, certain angles of illumination can cause severe shadows across the face. Darkening or lightening of facial features causes them to appear very different in a 2D color or grayscale image. Some features may even diminish completely if the illumination is either too strong or too weak (see Fig. 1.2(b)).

- **Expression:** While a neutral or relaxed facial expression is probably the most frequent state of a person's face, face images are often captured mid-conversation, while viewing something surprising, upsetting, etc., or while simply "making a face." Such daily activities cause different expressions involving different facial regions and components (see Fig. 1.2(c)). As facial recognition technology became widely used by Departments of Motor Vehicles (DMVs) across the United States, some DMVs be-

gan enforcing a no smiling rule for new driver's license photos.[10] However, recently DMVs (*e.g.,* Delaware[11]) have started to upgrade their facial recognition technology to systems which are capable of matching face images with high accuracy, regardless of smiling or neutral expression, and have lifted the ban on smiling. Nevertheless, extreme expressions are still challenging for state-of-the-art face recognition systems.

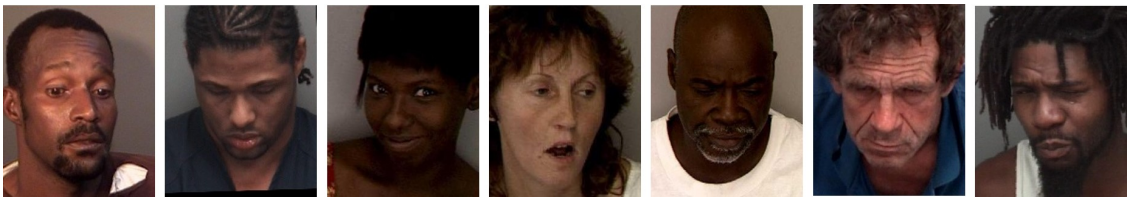
- **Occlusion:** Eyeglasses and sunglasses are a common cause of errors in facial recognition systems because the eye region, which is often highly discriminative, gets occluded. Facial occlusions not only cause missing information, but also extraneous information because it is difficult to detect and mask out occluded facial regions for matching. Even if a person consistently wears eyeglasses, specular reflections that change based on the light source still cause additional intra-person face variation. Other occluding facial accessories, such as baseball caps and hoods, can hide the forehead and eyes and cast shadows on the face. Besides facial accessories, faces can also be occluded by other objects or persons, which is typical of faces in a crowd; to accurately identify such "partial faces" in a crowd is an application of high interest for surveillance purposes.

- **Resolution:** The spatial resolution of a face, irrespective of image resolution, can be measured as the distance (*i.e.,* number of pixels) between the two eyes, also termed interpupillary distance (IPD). Smaller IPDs generally lead to lower face recognition accuracy, but there have also been studies (*e.g.,* [53]) that show that the discrepancy between the IPDs of two face images being compared can cause more errors than the absolute IPD values.

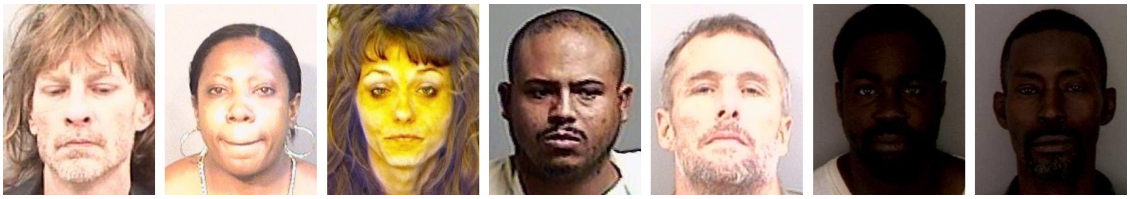
The above factors are typically those assumed present when dealing with "unconstrained face recognition." However, another variation that is known to degrade performance of face recognition systems is facial aging (see Fig. 1.3). Given two face images of the same person

[10] http://usatoday30.usatoday.com/news/nation/2009-05-25-licenses_N.htm

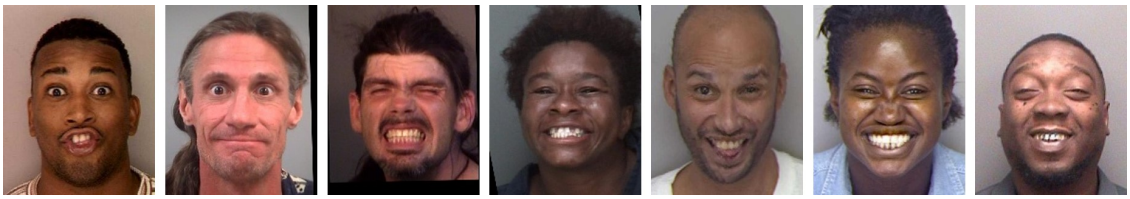
[11] http://www.delawareonline.com/story/news/traffic/burke/2015/01/28/dmv-lifts-ban-smiling-license-photos/22475061/



(a) Facial pose



(b) Illumination



(c) Expression

Figure 1.2 Sources of intra-class variability: (a) pose, (b) illumination, and (c) expression. Although all images shown here are of different people, such variations typically cause two images of the same person to appear very different.

captured multiple years apart, a robust face recognition system should still be able to recognize the two photos as the same person. Unlike the above factors, facial aging cannot be controlled either by the subject or the imaging environment; it is a challenge that can be present in both constrained and unconstrained face recognition scenarios. Facial aging will be discussed later in this chapter. While intra-class variations are a major challenge for face recognition systems, inter-class similarities can also cause errors. For example, it can be difficult (even for humans) to distinguish between persons with kinship relations (particularly twins, see Fig 1.4(a)), and persons that are not related can exhibit strong similarities (see Fig 1.4(b-d)).

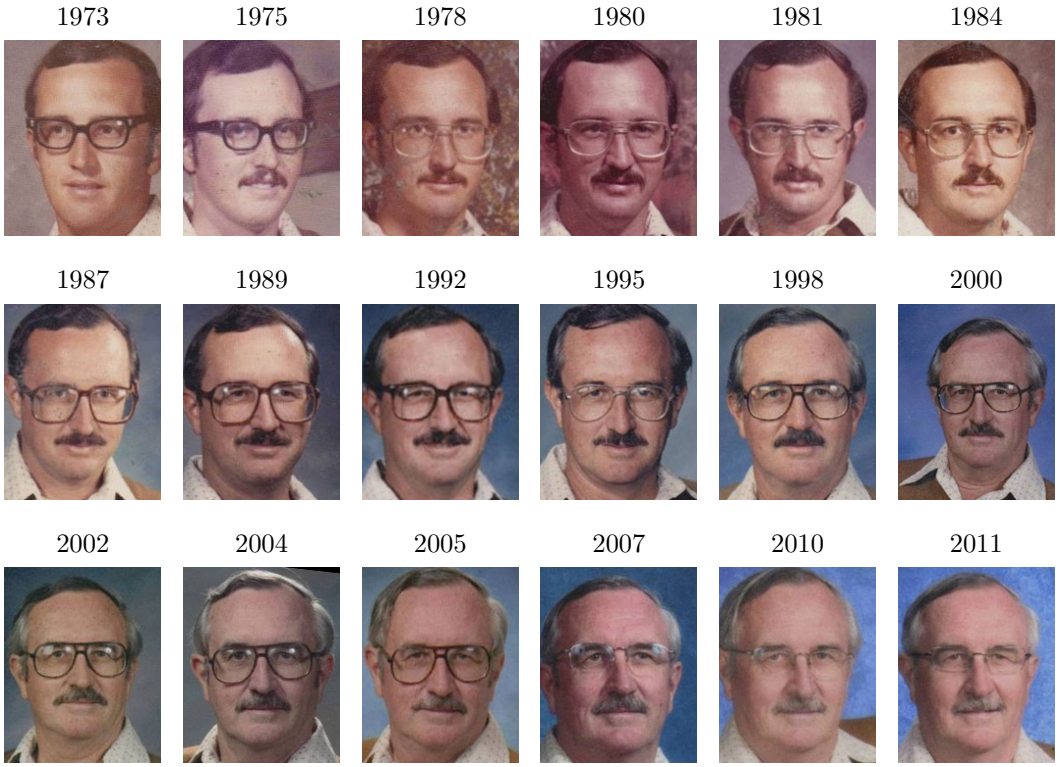
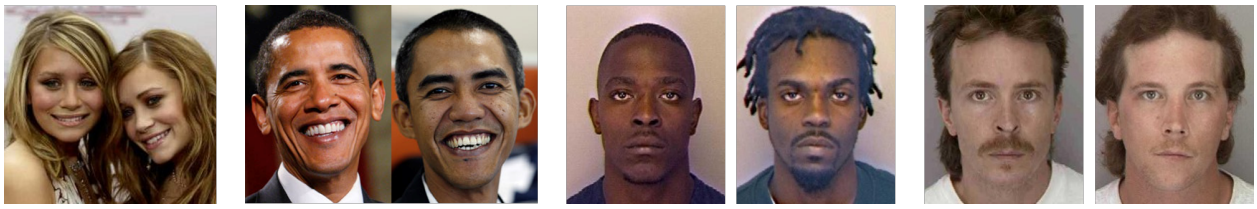


Figure 1.3 A teacher who wears the same outfit for his school picture every year; while the outfit is the same, his face and eyeglasses change over time. The overall quality of the image also changes (*i.e.,* improves over time). Such temporal aspects are additional sources of intra-class variation.
[Images are from: http://fillthewell.com/yearbook-pictures/]



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

Figure 1.4 Sources of inter-class similarity: (a) kinship similarities (in this case, twins) and (b)-(d) different people with no kinship relation who happen to exhibit very similar facial characteristics. This is sometimes referred to as a doppelgänger; (b) shows, as an example, that president Barack Obama (left) has a doppelgänger (right) from Indonesia.
[Images in (b) are from: http://www.theguardian.com/theguardian/2010/dec/05/barack-obama-doppelganger-ilham-anas]

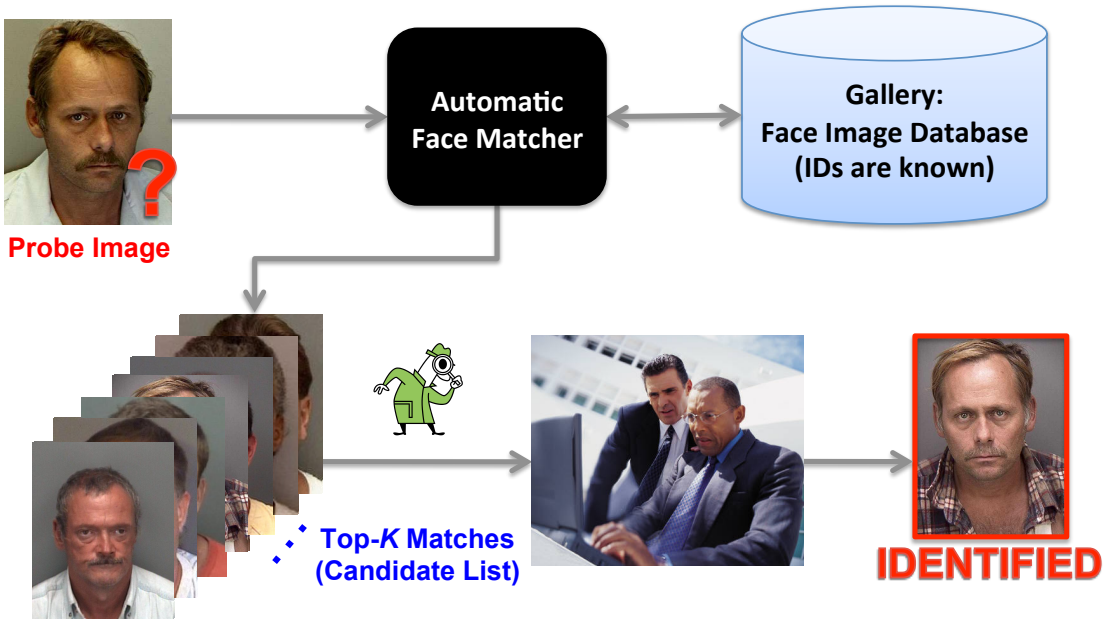


Figure 1.5 A flowchart of automatic face recognition in identification mode. A probe face image (with unknown identity) is matched against all face images enrolled in a gallery database. The top-k most similar identities retrieved from the database are then manually adjudicated by human analysts to determine whether the top-k candidates contain the identity of the probe face image. In verification mode, the probe image would be accompanied by a claimed identity and then only compared to the gallery image with the same identity as that which is claimed by the probe.

1.1 Background

Automatic face recognition operates in different modes depending on the application. Regardless of application, face images labeled with their identities are first enrolled in a database, referred to as the *gallery*. A face recognition system then takes a face image as input (*i.e.*, the *probe* or *query*) and matches it against one or many face images in the database. Face *verification* involves a one-to-one comparison to verify that the probe face image is the identity that it claims to be (*e.g.*, passport and passenger processing at airports, access control for buildings, and mobile phone authentication). Face *identification* involves one-to-many comparisons to retrieve (from the gallery) the identity of a probe face image whose identity is unknown. Because automatic face recognition systems susceptible to errors, in practice, the identity of the probe face image is established by manual adjudication

of the top-k most similar identities (see Fig. 1.5), where k is application dependent (*e.g.,* de-duplication, watch list surveillance, tag suggestions). In some scenarios, the top-k candidate identities are always manually adjudicated (*e.g.,* identification of a suspected criminal in forensics); this can be considered a *closed-set* identification scenario, where we assume that the identity of the probe is present in the gallery. However, *open-set* identification, where the identity of the probe may not be present in the gallery, is more representative of real-world scenarios. For open-set applications, the frequency of false alarms raised for subjects not in the gallery can be reduced by only returning the top candidate matches if they exceed a predetermined threshold (*i.e.,* k is of variable length). This is useful for "lights out" applications where it is impractical for a human analyst to review candidates for every query to the database (*e.g.,* watch list surveillance, especially in high traffic areas).

Whether verification or identification, the primary goal of an automatic face recognition system is to compute a measure of similarity between any two face images. Ideally, faces of the same individual should have higher similarity than faces of different individuals. However, there are multiple components in the face recognition pipeline that have significant impact on the computation of similarity scores and the resulting recognition performance.

1.1.1 Automatic Face Recognition Pipeline

The automatic face recognition pipeline (shown in Fig. 1.6) typically consists of the following sequential components: (i) face detection, (ii) face normalization, (iii) feature extraction and face representation, and (iv) comparison. Each of these components are crucial for achieving accurate and robust face recognition systems, and a significant amount of research has been devoted to each component individually. While state-of-the-art systems perform these steps fully automatically with extremely high accuracies for controlled environment and cooperative subject scenarios (*e.g.,* mugshot face images), open research problems still exist for unconstrained scenarios.

Face Detection: Face detection is the process of automatically determining whether a

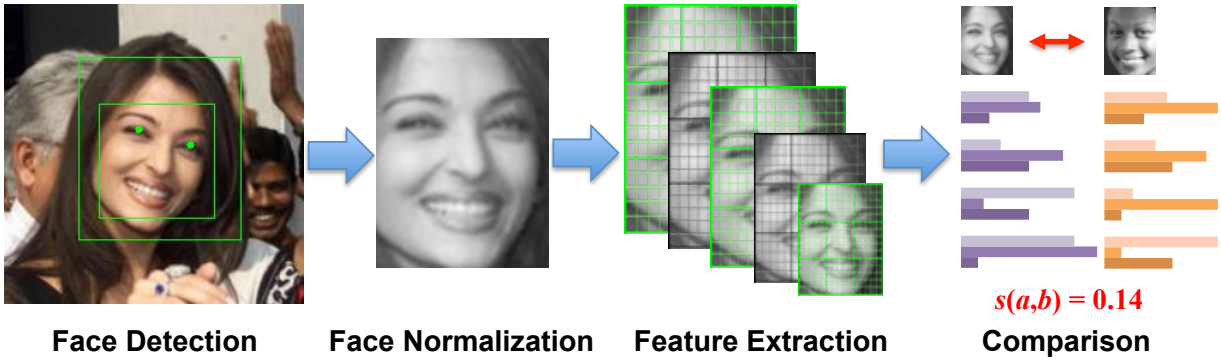


Figure 1.6 The automatic face recognition pipeline typically consists of (i) face detection, (ii) face normalization (to mitigate geometric and photometric variations), (iii) feature extraction, and (iv) comparison of resulting face representations.

face (or multiple faces) exist in an image, and subsequently outputting the locations of all detected faces. While it is a trivial task for humans to locate faces in an image, automatic extraction of face "sub images" from arbitrary images is a challenging task for machines. This is because of large intra-class variability in the appearance of faces (due to location, scale, skin color, etc.), as well as the possible presence of other face-like objects.

Research in face detection has been ongoing for more than two decades, but the seminal work of Viola and Jones [135] is credited with being the first real-time and accurate face detector, enabling many real-world applications. The Viola-Jones algorithm is an appearance-based method that uses simple Haar-like features which are sums of rectangular regions of pixels that respond to contrast differences structures on the face (*e.g.,* the two eyes are typically darker than the bridge of the nose). At the time it was introduced, the novelty of the Viola-Jones face detector was due to three key contributions: (i) fast feature computation using an "integral image", ii) feature subset selection with AdaBoost [46], and iii) fast and accurate rejection of non-faces using an attentional cascade structure [135]. Though the Haar-like features are simple, computation of the over-complete set is expensive (*e.g.,* 160,000 features for a 24×24 window), so the integral image enables the sum of an arbitrary rectangle to be computed with just four lookups [135]. The over-complete set of features is also computationally expensive to be used directly for classification, so multiple

"weak" classifiers are trained sequentially with AdaBoost [46] where each weak classifier is based on a single feature. Because selection of a small number of features sacrifices accuracy for real-time processing, Viola and Jones further use a cascade of weak classifiers which quickly discards non-face regions and allocates more resources to possible faces [135]. Experimental results in [135] demonstrated that a single-stage classifier with 200 features and a cascade of 10 classifiers each with 20 features achieved similar detection rates, but the cascade was 10 times faster.

Since its publication in 2004, the Viola-Jones face detector has greatly influenced research in face detection and is still widely used. However, many other methods have since been proposed that stem from the techniques proposed by Viola and Jones [135], and aim to be more robust to variations in facial pose, illumination, expressions, and occlusions. A survey of face detection approaches is provided in [150], and an in depth evaluation of various detection algorithms on unconstrained faces is given in [36]. Figure 1.7 shows example face detection results from an implementation of the Viola-Jones algorithm and a detector from a COTS face recognition system. The COTS detector performs better than the Viola-Jones algorithm, but errors are still observed for faces with extreme facial pose and occlusions, for example.

Face Normalization: Face normalization seeks to mitigate geometric and photometric variations that can greatly affect the subsequent modules of the face recognition pipeline. To normalize shape, face alignment is often performed to transform all faces to a canonical view. Face alignment aims at determining correspondences between face images based on any number of feature/landmark/fiducial points (*e.g.,* eyes, nose, mouth, contour, etc.). The most common face alignment technique is a simple 2D rigid affine transformation based on the two eye locations to correct for size and in-plane head rotation. However, in unconstrained face recognition, face images may contain out-of-plane rotations, so a simple 2D rotation based on the eyes alone may not be sufficient.

Active Shape Models (ASMs) [38] and Active Appearance Models (AAMs) [37, 93] were

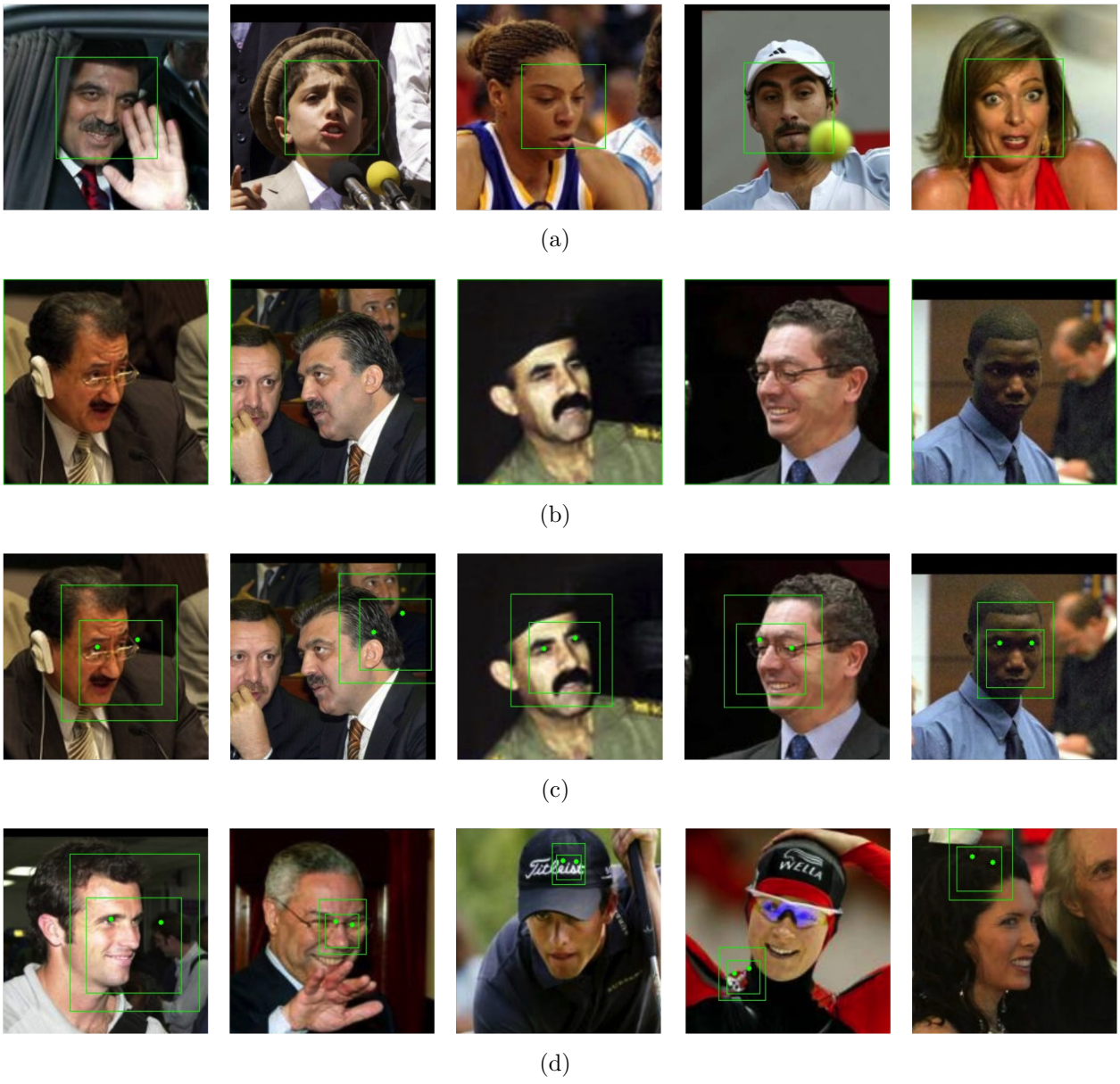


Figure 1.7 Example face detection results. Faces were (a) detected and (b) not detected by an implementation of the Viola-Jones algorithm [135]. Face images in (b) can be better detected by (c) a COTS face recognition system. However, the COTS detector also encounters (d) errors due to occlusion and facial pose, in particular. The small and large rectangles in (c) and (d) show bounding boxes of face and head detections, respectively. The circles are detected eye locations. All face images are from the LFW database [62].

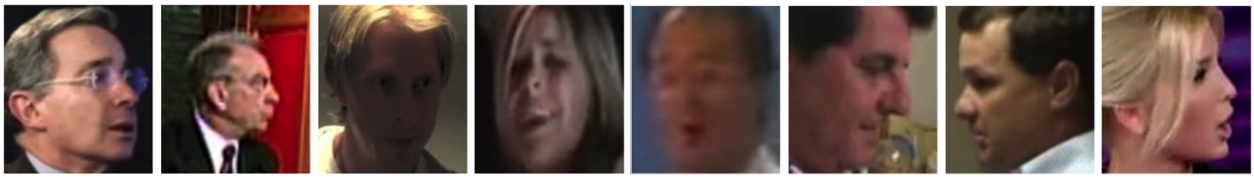


Figure 1.8 Example images from eight face tracks in the YouTube Faces (YTF) database where all images in that track could not be enrolled by one of the COTS matchers. These images display extreme pose and illumination conditions, low resolution, and motion blur.

some of the first statistical models proposed for object (*e.g.,* face) alignment. At their time, ASMs and AAMs were state-of-the-art, mainly due to novelty from learning shape and/or texture variations of a face from labeled training data. While ASMs and AAMs improved specificity of model-based approaches, they did so at the cost of generalization; alignment performance suffers when ASMs or AAMs are trained on a large database and/or fitted to previously unseen instances.

AAM-based methods were predominant for some time, but more robust solutions for landmark localization and alignment have since been proposed. For example, Zhu and Ramanan [153] propose a unified approach for detection, alignment, and landmark localization for faces "in the wild" that discriminatively encodes deformation and 3D structure as mixtures of trees with shared pool of parts [153]. Face alignment can also be done in 3D, with 3D morphable models (3DMMs), for example [26, 27]. Jourabloo and Liu propose a 3DMM-based approach to estimate both 2D and 3D facial landmarks for full pose variations, which additionally allows for estimation of the visibility of 2D landmarks [66]. Additionally, some recent works have shown impressive results for "frontalization" of unconstrained 2D face images with 3D modeling techniques (*e.g.,* [127, 128]), as well as 3D face reconstruction from a collection of unconstrained 2D face images [115].

Face alignment is also associated with "failure to enroll." If landmark points can not be detected, features cannot be extracted which can cause the entire enrollment process to fail. Figure 1.8 shows cropped face images from video frames in the YouTube Faces database where two COTS face matchers failed to enroll the face. Landmark localization and face

alignment are difficult problems, and many face recognition methods are highly dependent on the accuracy of either one or both of these processes; hence, some "alignment-free" methods have been proposed (*e.g.,* [84]).

Feature Extraction and Face Representation: Feature extraction and face representation go hand in hand. The simplest features are the raw pixel values of the face, where the representation is then a rasterized vector of raw pixel values. However, raw pixel values in vector form are not very informative; a significant amount of additional and relevant information exists in a face image that can be used to represent a face and enhance matching results. For example, high-level features, such as the distances between facial components and their relative locations and ratios, in addition to low-level features such as wrinkles and facial marks, can also be encoded to further discriminate between individuals.

Use of additional features seems like an obvious way to improve performance. However, the primary issue with adding more informative features is that the dimensionality of the feature vector becomes increasingly large (and likely redundant). Hence, the representation step typically focuses on compressing features so that they are both compact and highly discriminative. A vast amount of research has been devoted to these tasks (extraction and representation), some of which will be discussed in Section 1.2.

Comparison: Once a compact and discriminative representation of a face image has been obtained, the next step is to compare it to the representations of other face images to compute a measure of similarity. The Euclidean distance between feature vectors can be used; however, a more sophisticated choice of distance metric may significantly improve the recognition rate. Some examples of other distance functions include cosine, Manhattan, Tchebyshev, and correlation, as well as histogram intersection, log-likelihood statistics, chi-square statistics, etc. Distance metric learning has also been applied to face recognition (*e.g.,* [39, 61, 128]), where a distance metric is learned from training data to simultaneously minimize distances between instances of the same class and maximize distances between different classes.

1.2 Research Progression

The concept of identifying individuals based on retained face images dates back to the 19th century when Alphonse Bertillon developed a system for identifying criminals based on anthropometric measurements in 1879 [112]. The Bertillon system, or bertillonage, was introduced in the U.S. in 1887 as the primary method for identifying and tracking criminals.[12] Although it was replaced by fingerprinting in the early 20th century, face images of criminals, now known as mugshots, are still used worldwide.

> ‘‘...according to the method prescribed by Dr. Bertillon, the exact identity of any adult person can be established with so much definiteness that when signalized a second time he can be recognized with infallible certainty by a simple reference to the file in which the former signalment is kept. Even if this file represented the entire population in the country, the process of identifying two correctly-taken signalments by its means could be performed in most cases in a few minutes, without any assistance from a similarity of names.’’ - From publisher’s preface to *Signaletic Instructions Including the Theory and Practice of Anthropometrical Identification* by Alphonse Bertillon, 1896

Partially automated recognition began in the mid 1960s when Woodrow W. Bledsoe came up with a “man-machine” system for identification of individuals based on physiological measurements which were entered by hand (*e.g.,* height, weight, interpupillary distance, etc.), stored in documents, and searched automatically [9]. Bledsoe understood that the results were highly dependent on the angle of the face images, so he learned a transformation from the actual 3D heads of seven individuals and applied this transformation to the measurements of any non-frontal faces; a concept that is still used in current state-of-the-art 3D face models.

[12] http://www.nleomf.org/museum/news/newsletters/online-insider/november-2011/bertillon-system-criminal-identification.html

Since Bledsoe's man-machine system, 50 years of research (see Jain *et al.* for an overview [65]) has been devoted to improving the robustness and efficiency of fully automated face recognition systems (albeit recognition results are often manually adjudicated). Every stage of the pipeline has received substantial research attention and great progress has been made in face detection, alignment and normalization, feature extraction and representation, and comparison. The progression of face recognition from frontal constrained face matching to unconstrained "in the wild" face matching can roughly be delineated by three face representation approaches: (i) holistic, (ii) local, and (iii) learned representations. This section briefly discusses a few methods related to these categories.

1.2.1 Face Databases

First of all, it would not be possible to discuss progress in face recognition research without reference to standardized face image databases and evaluations that have paved the way for such success. While many researchers evaluate proposed methods on in-house databases, research progression in face recognition is primarily facilitated and motivated by the compilation and public release of face image databases. Some of the first standardized databases on which the research community began to evaluate proposed methods are shown in Fig. 1.9.

While databases such as FERET [97], FRGC [97], and AR [92] (example images shown in Fig. 1.9) greatly contributed to advancements in face recognition research, most of them were acquired under relatively controlled conditions and were compiled by research teams for studying specific subproblems of face recognition (*e.g.,* illumination, expression, pose). Such databases allow researchers to directly evaluate performance on face images that exhibit certain variations, but are not very representative of face images encountered in real-world scenarios. As algorithms continued to mature to handle controlled/simulated variations in pose, illumination, expression, and occlusion, more challenging databases were needed.

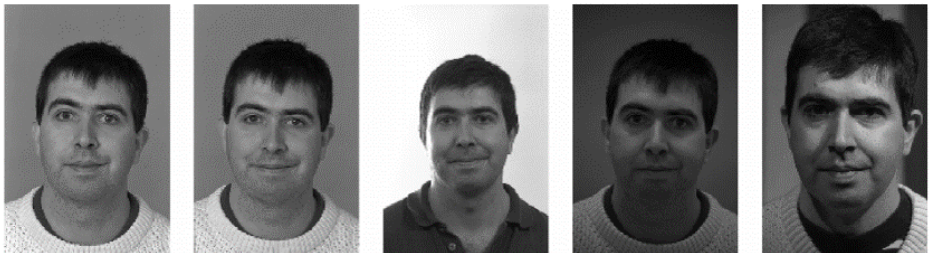
For this reason, Huang *et al.* released the Labeled Faces in the Wild (LFW) database which was compiled by searching the internet for the names of public figures, athletes,

actors/actresses, etc. [62]. The LFW database includes 13,233 face images of 5,749 different people. All face images were automatically detected by an implementation of the Viola-Jones face detector [135], so they are constrained in that respect, but images typically exhibit multiple variations that are challenging for face recognition algorithms. Along with the database, Huang *et al.* released the LFW experimental protocol: 10-fold cross-validation on verification/classification of 300 same and 300 not-same face pairs per split.

1.2.2 Holistic Representation

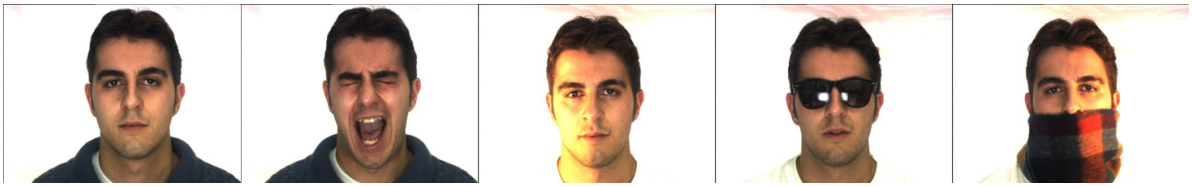
Drawing upon the Sirovich and Kirby [119] discovery that face images could be reconstructed as projections onto a small set of eigenpictures, the Eigenfaces method was one of the first fully automatic face recognition algorithms proposed in 1991 by Turk and Pentland [132]. A low-dimensional "face space" is calculated based on the training set of N face images using principal component analysis (PCA). The face space is the set of $M < N$ eigenvectors corresponding to the largest M eigenvalues of the covariance matrix of the training set. All faces are then represented as the weights associated with their linear projection onto the set of eigenfaces, and dissimilarity is defined as Euclidean distance between two M-dimensional feature vectors. Turk and Pentland also use the distance to face space for automatic face detection; every pixel of an image is projected onto face space to acquire a "face map" where low values (*i.e.,* small distances to face space) indicate the presence of a face. Experiments conducted on 16 subjects, represented by 7 eigenfaces, showed that Eigenface representation was fairly robust to lighting variations (96% identification accuracy) but suffered more errors with changes in head size and pose. Fisherfaces [13] is an extension of Eigenfaces that uses supervised dimensionality reduction to find the subspace that minimizes intra-person and maximizes extra-person variance via linear discriminant analysis (LDA).

These first fully automatic face recognition methods can be categorized as *holistic representations*, as they utilize all the facial pixels together to drive a representation. Holistic methods heavily rely on accurate alignment (typically based on eye locations) which be-



(a) FERET

(b) FRGC



(c) AR

(d) LFW

(e) IJB-A

Figure 1.9 Example face images from different databases: (a) FERET [97], (b) FRGC [97], (c) AR [92], (d) LFW [62], and (e) IJB-A [72]. Databases (a)-(c) contain variations such as illumination, expression, and occlusion to challenge face recognition research, but they are relatively controlled acquisition conditions because such variations are simulated/staged (subjects are typically students and members of research groups). Databases (d) and (e) contain more unconstrained face images (*e.g.,* collected from the internet).

Table 1.1 Face image databases in the public domain

Database	Year	Num. Subj. (Num. Imgs.)	Acquisition Conditions
NIST Mugshot Id [140]	1994	1,573 (3,248)	constrained, operational
FERET [97]	1996	1,199 (14,126)	simulated/staged PIE
Yale [13]	1997	15 (165)	simulated/staged IE
AR [92]	1999	126 (4,000)	simulated/staged IEO
Yale B [48]	2001	10 (5,760)	simulated/staged PIE
CMU PIE [117]	2003	68 (41,368)	simulated/staged PIE
FRGC [97]	2005	>466 (>20,000)	simulated/staged IE
LFW [62]	2007	5,749 (13,233)	unconstrained, web-collected
CMU Multi-PIE [51]	2008	337 (>750,000)	simulated/staged PIE
MEDS [45]	2011	518 (1,219)	constrained, operational
CASIA-WebFace [146]	2014	10,575 (494,414)	unconstrained, web-collected
IJB-A [72]	2015	500 (5,712)	unconstrained, web-collected

comes difficult when faces are encountered that may be non-frontal or contain expression variations, etc. Holistic methods also do not generalize well to new databases and have difficulty with variations not present in the training set (*e.g.,* presence/absence of eyeglasses in training/testing).

1.2.3 Local Representation

Local representations typically perform a dense sampling of features at overlapping patches in the face image and at multiple scales. To incorporate global information, geometric relationships between features are often encoded by concatenating features extracted from either a common set of landmark points or from a grid overlaid on the face. Hence, local representations can also be sensitive to face alignment. Because the resulting set of features is often over-complete with high dimensionality, feature selection (*e.g.,* boosting) or subspace methods (*e.g.,* PCA, LDA) are adopted to achieve a compact face representation.

Liu *et al.* presented a novel augmented Gabor feature vector for face representation and proposed the Gabor-Fisher classifier (GFC) for face recognition [88]. Gabor wavelets had been used for face representation in prior works (*e.g.,* Lades *et al.* [77]), but the novelty of the Liu *et al.* Gabor feature was the concatenation of Gabor filter responses (using five scales and eight orientations) and the subsequent application of PCA to compress the high-dimensional feature vector. They showed that Gabor face representation with PCA performed better than both Eigenfaces and Fisherfaces (which use the original image intensity values as features). Furthermore, the GFC, which applied the Enhanced Fisher linear discriminant Model (EFM) to the compressed augmented Gabor feature vector, achieved better performance than both PCA and LDA with the Gabor feature vector. The use of EFM helps improve discrimination and generalization and alleviates the small sample problem of FLD/LDA.

Ahonen *et al.* presented the first application of the local binary pattern (LBP) texture descriptor to face recognition [4]. Specifically, they used the uniform patterns extension

of LBP (*i.e.,* every circular pattern with at most two bitwise transitions contributes to its own bin in the histogram and all other non-uniform patterns contribute to a single bin). One major contribution of the Ahonen *et al.*'s LBP face representation was the *spatially enhanced histogram*. To incorporate regional and global properties in combination with the local features from the LBPs themselves, they placed a grid over the face, extracted a histogram of LBP for each grid location, and concatenated the results to form the final feature vector of the face. Because of this representation, different weights can be assigned to the grid locations to be used with the weighted Chi squared distance measure. Patches that contribute more to discriminating between identities (*e.g.,* the eyes) can be given more weight. In comparison with other local descriptors, Ahonen *et al.* [4] provided experiments to show that LBP representation typically demonstrated the best performance on subsets of the FERET database; likely due to the monotonic gray-scale invariance of LBP compared to the other local descriptors.

Recently, a few extremely high-dimensional local representations (with efficient dimensionality reduction techniques) have shown impressive performance on the LFW database. For example, high-dimensional features (sampled at multiple scales on dense landmarks detected by Cao *et al.* [29]) with Joint Bayesian classification [33] achieve 93-95% accuracy on the LFW database for LBP, Gabor, HOG, SIFT, and LE descriptors trained on WDRef database (99,773 images of 2,995 subjects) [34]. Most of the initial local representation methods have now been categorized as methods based on "handcrafted" or "engineered" features because image filters are pre-defined and performance typically depends on a fine tuning of the radius and scales of sampling.

1.2.4 Learned Representation

Motivated by the drawbacks of handcrafted local descriptors such as LBP, Gabor, SIFT etc., Cao *et al.* proposed a learning-based (LE) descriptor [30]. LE descriptors are extracted by sampling a ring-based pattern from the neighborhood of each pixel to form a low-level feature

vector which is then normalized to unit length. Cao *et al.* applied unsupervised learning methods (K-means, PCA tree, or random-projection tree) to encode feature vectors into discrete codes. Face images are represented as "code images," and histograms of LE codes can be extracted from grid locations and concatenated to form the final face representation. Cao *et al.* showed that the distribution of the LE descriptors is more uniform across face images than LBP and histogram of oriented gradients (HOG) and is therefore more informative, discriminative, and compact [30]. The LE descriptors are combined with a pose-adaptive matching method which aligns and matches nine components of the face separately, combines their similarity scores, and delegates the verification decision to a linear SVM classifier that has been trained on the two poses most similar to the input face images. Experimental results on the LFW (84.45% accuracy) and Multi-PIE (95.19% accuracy) databases show that the LE descriptors with pose-adaptive matching performs better than other methods trained in the same manner and is competitive with methods trained using additional information [30].

1.2.4.1 Deep ConvNets

More recently, deep neural networks have achieved impressive results for many visual recognition tasks [75], including face recognition. Neural networks are not new (*e.g.,* perceptrons were first developed in the 1950s); however, network models with many hidden layers (deep structures) can be trained due to better regularization strategies and availability of large face databases and processing capabilities. Again, rather than handcrafted features, face representations are learned by deep convolutional neural networks (ConvNets) trained to classify identities (or verify pairs of face images) from large-scale training sets of face images. The dimensionality of the feature representation is hierarchically reduced due to the structure of convolutional and pooling layers; both low-level features and global features are learned in a cascaded manner. Commonly, the output of the last hidden layer (prior to the classification layer) has been shown to have learned a highly robust face representation for new face images in testing [123, 128]. While the specific architectures of the networks in [123, 128, 146] are all

different, their high recognition performance can generally be attributed to a few common properties: better regularization strategies for learning very deep structures (4-11 layers), availability of large-scale training databases (*e.g.,* $>$ 4 million images [128]), and access to faster and cheaper computational resources.

However, the success of these deep ConvNets approaches is not due to sophisticated learning and large-scale training sets alone; many of these methods also include additional preprocessing and/or post-processing steps that further boost performance. For example, Taigman *et al.* directly feed raw RGB pixel values as input to their deep ConvNet under the assumption that their 3D face frontalization is successful [128]. This strong assumption may not have been possible a few years ago when 3D frontalization capabilities from unconstrained 2D images were not accurate and robust. Sun *et al.* train multiple deep ConvNets on various face patches at different scales [121], and DeepID [123] (and its variants) utilize the Joint Bayesian classification method [33] on their deep representation; Joint Bayesian [33] is a supervised subspace learning approach that has achieved high accuracies with other face representations as well (*e.g.,* [34, 85]).

1.3 Video Face Recognition

Face recognition in video is becoming increasingly important due to the abundance of video data captured by surveillance cameras and mobile devices, uploaded to the Internet, etc. Given the aggregate of facial information contained in a video (*i.e.,* a sequence of face images or frames), video-based face recognition solutions can potentially alleviate classic challenges caused by variations in pose, illumination, and expression. A summary of the common public domain databases used to evaluate video-based face recognition algorithms can be found in Table 1.2. Of particular interest for these databases is the number of subjects available, and whether or not the activities of the subjects were constrained or unconstrained (*e.g.,* subjects were directed to move in certain ways vs. subjects act naturally in an environment). Notably,

Table 1.2 Characteristics of popular face video databases in the public domain.

Database	Acquisition Conditions	Subjects	Videos	Accuracy
Motion of Body (MoBo) [52]	Treadmill walking: slowly, quickly, on incline, or with a ball	25	150	98.8% [89]
Face in Action (FIA) [49]	Variations in expressions and orientations; indoor/outdoor	221	n/a	99% [101][a]
1st Honda/UCSD [79]	Staged head rotations and expressions	20	75	99% [131]
MBGC [103]	Walking, activity, conversation; standard and high resolutions	821	3,764	see [103]
YouTube Celebrity [69]	Unconstrained, many same-subject tracks from the same video	47	1,910	78.9% [145]
YouTube Faces [141]	Unconstrained	1,595	3,425	54.8% [128][b]
IJB-A [72]	Unconstrained	500	2,085	40.6% [72][b]

[a] Authors used an indoor subset of FIA [b] TAR @ 1.0% FAR

Figure 1.10 Example images from face tracks of two subjects in the YouTube Faces (YTF) database. The top two and bottom two rows are face tracks from the same subject.

the YouTube Faces (YTF) database [141] contains the largest number of subjects and the faces in the video tracks are relatively more unconstrained than other face video databases. The MBGC video data [103] also has strong relevance to unconstrained faces in video, but the YTF database is more widely used due to the following reasons: (i) it contains the largest number of subjects, (ii) the actions of the subjects are naturally varied (as opposed to performing prescribed actions), (iii) the YTF database is easier to acquire (thus allowing the baselines to be used by the research community at large), and (iv) all subjects in the YTF database also have still images available in the LFW database [62] (thus allowing baselines to be compared to the video-to-still image matching scenario). The IJB-A database [72] also contains unconstrained face videos but with fuller pose variations and lower quality faces than the YTF database (faces in YTF were detected by Viola-Jones detector which can miss faces at extreme poses, while faces in IJB-A were manually annotated by humans).

Video-based face recognition approaches have been organized into the following two categories [10] based on how they leverage the multitude of information available in a video sequence: (i) sequence-based, and (ii) set-based. At a high-level, what most distinguishes these two approaches is whether or not they utilize temporal information. Sequence-based approaches consider all detected faces based on their temporal ordering. For example, Zhou *et al.* combined both face tracking and face recognition into a single framework, which allowed the inter-frame dynamics to be exploited during the recognition process [152]. See [10] for more details about sequence-based methods.

Set-based approaches to video-based face recognition consider all the available frames of a subject's face as an unordered set. Such methods have been further organized into approaches that fuse the available information prior to matching, and those that fuse information after performing matching [10]. Methods that fuse information prior to matching will generally output either a feature vector representation or a single face image. For example, manifold-based methods project the set of face images onto a manifold within a feature space, which in turn facilitates matching within the feature space [80, 138]. Manifold methods are similar

to sequence-based methods in that they require specialized matching algorithms. Both super resolution methods [6] and 3D modeling-based methods [101] output a single face image that in turn can be matched with an existing face recognition system. Thus, while such synthesis-based methods attempt to solve a difficult generative modeling task, these methods are compatible with existing face recognition engines. A few commercial solutions are available for such synthesis methods, though they are only semi-automated and hence more relevant to forensic applications.

Finally, set-based methods that fuse information after the face matching process seek to combine the comparison scores from static face matchers into a single similarity score. For example, Taigman *et al.* [128] randomly selected 100 pairs of frames from two videos and used the mean of the pairwise similarity scores as the similarity score between two videos; this simple extension of their static image-based method (*i.e.,* DeepFace) achieves 91.4% accuracy on the YTF database. Yi *et al.* also applied their deep ConvNet approach to video data in a similar manner (randomly select 15 frames from each video) and also achieve high accuracy (92.2%) on the YTF database. Like the LFW database, deep ConvNet methods are currently outperforming all other methods on the YTF database.

1.4 Face Image Quality

Face recognition system errors are often due to quality issues at the time of acquisition of the face image. In constrained and controlled capture environments (*e.g.,* passport and mugshot photos), low quality face images are typically due to operator issues or uncooperative users. Many users of face recognition systems are unaware of the sensitivity of automatic face recognition systems to illumination, facial pose, expression, eyeglasses, etc., or subjects may be uncooperative (*e.g.,* for mugshot photos). In unconstrained scenarios, ranging from surveillance imagery to face images available on the internet, low quality face images are unavoidable due to the nature of the applications. Available face images are either not col-

lected for use with identification documents and face recognition purposes, or face images are captured covertly where subjects are unaware of acquisition or purposely do not want a good quality face image to be acquired.

Following the accepted definition of biometric sample quality, a face quality measure should be predictive of automatic face recognition performance [5, 23, 56]. Hence, a face image determined to be of low (poor) quality should result in low genuine and high impostor similarity scores, and a high (good) quality face image should result in high genuine and low impostor similarity scores. The benefits of an automatic measure of face quality are similar to the benefits of automatic quality measures for any other biometric trait (*e.g.,* fingerprint or iris) [56]. Some examples include the following:

- To assist with the integrity of enrollment face databases, automatic quality measures could be integrated into face image acquisition protocols, where the process cannot be completed until a face image of desired quality has been acquired. The quality measures could also be applied retroactively to legacy face databases to "flag" low quality images which have been previously enrolled.

- Similarly, an automatic quality check could be incorporated at the time of verification or identification in controlled and constrained scenarios where capture of additional face images is possible if necessary. Rather than returning a false match or false non-match, where the operator (or user) would need to ascertain whether to attempt the process again, the system could have a "reject option" where no decision is given unless the query face image is of sufficient quality. If the acquired face image does not pass a quality check, the user can be prompted to provide a better quality face image.

- A face quality measure can be used to weight face image samples for fusion of different biometric traits (*e.g.,* face and iris) or of multiple face images (and/or video frames) in media collection scenarios such as those explored in Chapter 2.

- Automatic invocation of adaptive recognition systems based on the quality (*e.g.,* fusion

of multiple matchers if face quality is poor may boost the performance, but fusion could be avoided for high quality samples where additional computation is unnecessary or fusion may even degrade performance). Hence, it may be useful to have both matcher-independent and matcher-dependent quality measures.

Bharadwaj *et al.* [23] and Alonso-Fernandez [5] provide recent reviews of biometric sample quality for fingerprint, face, and iris. The most widely used biometric sample quality has without a doubt been the use of NIST Fingerprint Image Quality (NFIQ v1.0 [125] and NFIQ v2.0 [124]); with wide acceptance, it is now the de facto standard for assessing fingerprint image quality for many important applications, including the US-VISIT program [96]. NFIQ is an integer value of 1 to 5 (1 being the highest quality) that predicts the expected performance of fingerprint matching algorithms on a given fingerprint image. In comparison, face image quality has been studied in the literature (*e.g.,* [22, 35, 42]), but no satisfactory solutions are yet available from either the research community or commercial vendors, to the best of our knowledge.

1.5 Facial Aging

Because of the natural process of aging, appearance changes that affect both facial shape and texture are inevitable. Hence, the permanence/persistence of the face as a biometric tends to be lower than that of the other primary biometric traits (*i.e.,* fingerprint and iris). Unlike other factors such as pose and illumination, aging variations cannot be controlled; facial aging is a challenge that spans both constrained and unconstrained applications of face recognition. A common approach to handle the issue of faces changing over time is "template update" where subjects' enrolled samples are periodically updated. For example, driver's licenses and passport photos must be renewed every so many years. While template update is effective, there are many applications where template update is not a viable solution (*e.g.,* de-duplication, identification of missing persons, surveillance and watch-list scenarios).

To solve this problem, there have been two primary approaches in the literature: (i) age simulation/progression of face images prior to feature extraction and matching (*e.g.,* [78, 102]) and (ii) recognition methods which utilize "age-invariant" features and/or subspaces (*e.g.,* [67, 83, 87]). Ramanathan *et al.* provide a survey of approaches related to facial aging [111].

1.6 Benchmarking State of the Art

Progress in face recognition research has largely been driven by systematic large-scale evaluations of current methods, which not only encourage competition, but also help to identify future areas of research. While the research community attempts to benchmark published methods against each other, public access to large operational databases has been limited. Hence, third-party evaluations done by the National Institute of Standards and Technology (NIST) are invaluable for knowledge of current state-of-the-art algorithms; NIST has access to large operational databases and conducts extensive testing of multiple algorithms on protocols that mimic operational scenarios.[13] In particular, commercial vendors, whose algorithms are typically proprietary, submit their algorithms for the NIST evaluations. The research community should pay close attention to the results of these tests; actual state-of-the-art methods are different than "home-brewed" algorithms evaluated on small "in-house" or lab-collected databases.

To measure progress in face recognition, we can track the results of the various NIST evaluations, which began in September 1993 with the FERET program [107]. At that time, face recognition systems were limited to prototypes from research labs and universities, and few were fully automatic. Commercial systems have since been evaluated in multiple Face Recognition Vendor Tests (FRVTs). Table 1.3 shows that the FRVTs (and the MBE [57]) have documented continuously increasing TARs at 0.1% FAR on frontal constrained face images from 2000 to 2013; a decrease in error rates of approximately three orders of

[13] http://www.nist.gov/itl/iad/ig/face.cfm

Table 1.3 Face recognition performance on frontal, constrained face images as reported over the years in NIST evaluations.

Evaluation	Year	Rank-1 Accuracy	Gallery Size	TAR @ 0.1% FAR
FERET [107]	1993/94	78%	316	21%
FERET [107]	1996/97	95%	831	46%
FRVT [106]	2002	73%	37,437	80%
FRVT [108]	2006	n.a.	n.a.	99%
FRGC v2.0 [105]	2005	n.a.	16,028	99%
MBE [57]	2010	92%	1.6M	>99%
FRVT [55]	2014	96%	1.6M	n.a.

magnitude has been observed.

The most recent NIST evaluation, FRVT 2013, focused on large-scale face identification, both closed-set and open-set [55]. While closed-set accuracies of the top six commercial vendors were quite high (best was 4.1% rank-1 miss rate), open-set accuracies decreased significantly for a FAR of 0.2% (best was 7.5% rank-1 detection and identification miss rate). These evaluations have also experimented on face images captured in less ideal conditions (*i.e.,* non-uniform lighting, lower resolution, non-frontal); the FRGC and FRVT evaluations identified pose, illumination, and outdoor imagery as especially challenging for algorithms. However, with the exception of the webcam face images used in FRVT 2013, most of these evaluations on other factors have been on databases with staged variations (*i.e.,* lab collected).

1.6.1 Unconstrained Face Recognition

Current state-of-the-art methods for unconstrained face recognition have been benchmarked by the LFW database protocol since its release in 2007 [62]. Numerous methods have evaluated on the LFW protocol (almost 60 publications on the LFW website[14] at the time of writing). Recently, the LFW protocol has been dominated by convolutional neural network approaches with reported accuracies of 97–99% (*e.g.,* [121, 128, 146]). As previously dis-

[14]http://vis-www.cs.umass.edu/lfw/results.html

cussed, these high accuracies are largely due to the use of large-scale training databases external to LFW; methods which leverage outside training data (ConvNet methods already mentioned, as well as *e.g.,* [30, 34, 90]) have proven to achieve much higher accuracies than methods that only train on LFW face images (current best accuracies are 95.89% [7] and 88.97% [81]). The public availability of the LFW database has greatly contributed to advancements in the development of face recognition techniques that are robust to variations in pose, illumination, expression, etc. by facilitating competition amongst research teams, as well as the goal to outperform humans [76]. However, there are a few limitations of the LFW protocol as discussed in the next section.

1.6.1.1 Drawbacks of the LFW Protocol

The LFW database protocol was designed for the *classification* task of determining whether a pair of face images is the same (*i.e.,* genuine) or different (*i.e.,* impostor). Hence, the LFW protocol is specifically an evaluation of face verification. While face verification is a real-world biometric scenario, the LFW protocol suffers from the following limitations:

- Many methods that use the LFW protocol only report the accuracy of their final classifier that determines same vs. not-same face pairs. However, in a biometric verification system, we typically do not require a classifier to make a binary decision. A face recognition system will be deployed, and the system administrators will determine the threshold at which they wish to operate the system (depending on security and usability requirements of the application domain). Hence, a full receiver operating characteristic (ROC) curve should be reported to demonstrate performance across different thresholds.

- Because of the above point, biometric systems should especially be tested at low false accept rates (FARs) as this is typically where most applications operate (*e.g.,* FARs well below 1%). The LFW protocol, which contains only 300 impostor scores per cross-validation fold, does not allow for evaluating at these low FARs (*e.g.,* $1/300 = 0.3\%$

and is not statistically reliable).

- Many unconstrained face recognition scenarios require face identification, rather than verification, tasks. While verification and identification are related, DeCann and Ross show that a good verification system does not necessarily imply a good identification system (and vice versa) [40]. Hence, unconstrained face recognition methods should also be evaluated in identification modes (both closed-set and open-set).

Because of these drawbacks, the LFW protocol design has received recent criticisms [85, 129, 151], and research focus is beginning to shift towards evaluation in more realistic biometric settings. In 2014, Liao *et al.* released a new unconstrained face recognition protocol: Benchmark of Large-scale Unconstrained Face Recognition (BLUFR) [85]. The protocol is still 10-fold cross-validation but exploits the large number of face images available in the LFW database; BLUFR has both verification and open-set identification protocols consisting of about 157,000 genuine scores and 47 million impostor scores per fold.[15]

Liao *et al.* [85] provide results on the BLUFR protocol for some benchmark methods, including Chen *et al.*'s high-dimensional LBP with Joint Bayesian classification [33, 34]; while Chen *et al.*'s approach achieves 95% accuracy on the LFW protocol, the accuracies significantly drop for the more challenging BLUFR protocol (see Table 1.4). Similarly, deep neural network approaches (*e.g.,* Yi *et al.* [146] and Wang *et al.* [136]) achieve ∼98% accuracy on the LFW protocol, but only 90% and 56% accuracies on the BLUFR protocol. These results demonstrate that accuracies of ∼99% on the LFW protocol are misleading; there is still room for improvement in scenarios more representative of real-world (*i.e.,* large-scale) biometric applications.

The YTF database protocol (*i.e.,* 10-fold cross-validation of 250 same and 250 not-same pairs per fold [141]) is the video-equivalent of the LFW protocol and contains the same drawbacks. Additionally, another issue with these two databases is that web-collected data can easily contain labeling errors. Because the LFW and YTF protocols contain so few face

[15] http://www.cbsr.ia.ac.cn/users/scliao/projects/blufr/

Table 1.4 Comparison of performance on the LFW [62] vs. BLUFR [85] protocols.

	LFW Protocol	**BLUFR Protocol**	
Method	Accuracy (%)	TAR @ 0.1% FAR	DIR @ 1% FAR
HighDimLBP + JointBayes [34]*	95.17	41.66	18.07
Yi *et al.* [146]	97.73	80.26	28.90
Wang *et al.* [136]	98.20	89.80	55.90

*Performance here for [34] on BLUFR protocol was reported by [85]

pairs, these errors may be significant. By studying human performance (via crowdsourcing on Amazon Mechanical Turk), we discovered 111 errors out of the 2,500 genuine pairs in the YTF protocol [16]. Some of the errors were due to the difficult task of verifying ground truth because of the temporal aspect of videos; the person of interest may not appear in the video until a few or many frames into the sequence. Databases that have been reliably annotated with ground truth labels prior to release, such as the IJB-A database [72], are invaluable to the research community.

An additional limitation of the LFW and YTF unconstrained face databases is that they were both compiled using a commodity face detector, namely, an implementation of the Viola-Jones algorithm [135]. While automatic face detection facilitates the collection of large-scale face databases, this property immediately puts a *constraint* on the collected face images which are supposed to be *un*constrained. The constraint being that Viola-Jones based algorithms (and most other existing face detectors) perform best on near-frontal face images [36]. Additionally, poor illumination, extreme expression, and occlusions can also cause face detection to fail. Hence, current research efforts in unconstrained face recognition have been optimizing automatic face recognition only for those faces which can be detected by these commodity detectors.

For the reasons mentioned above, the IARPA Janus program released a new unconstrained face database, IARPA Janus Benchmark A (IJB-A), which is a joint face detection and recognition database [72]. IJB-A contains 500 subjects with an average of 11.4 face im-

ages and 4.2 videos per subject. All faces in both images and video frames were annotated manually via sophisticated methods using Amazon Mechanical Turk [126]. Because all faces are detected by humans (rather than automatically detected by a Viola-Jones face detector), the IJB-A database contains larger ranges of variations (particularly facial pose) that degrade performance of current face detection and recognition approaches. The IJB-A face recognition challenge managed by NIST[16] is a "template-based" matching scenario where each sample is a composite of still images and video frames of the same subject; the goal is to leverage complementary information that may be available in multiple unconstrained faces. The current leaderboard[17] accuracies for the IJB-A challenge are the following: 82% TAR @ 1% FAR (1:1 verification), 88% rank-1 accuracy (1:N closed-set), and 53% TPIR @ 1% FPIR (1:N open-set).

1.6.2 Age-Invariant Face Recognition

State-of-the-art age-invariant face recognition systems (in the literature) are currently benchmarked by the FG-NET [78] and MORPH [113] databases; a number of methods claim to improve the "age-invariance" of face recognition by reporting overall performance on FG-NET and/or MORPH. For example, [50] reports rank-1 identification accuracies of 69.0% and 91.1% on the FG-NET and MORPH-II databases, respectively. Using the periocular region, Xu *et al.* [67] reported 100% rank-1 accuracy and 98% TAR at 0.1% FAR on FG-NET. However, an overall performance improvement on a specific database does not necessarily indicate a good solution to the facial aging problem. Klare and Jain demonstrate that methods developed (trained) for age-invariance may actually decrease performance in non-aging scenarios [70]. Furthermore, simply stating accuracies on the entire longitudinal database does not provide any information/quantification regarding facial aging as a covariate to face recognition (*i.e.,* how much impact specific ages or time lapses have on comparison scores and/or accuracies).

[16] https://www.nist.gov/programs-projects/face-challenges

[17] IJB-A reports are periodically updated. Leaderboard results reported here are from Nov. 2016.

To further study facial aging, most researchers divide the database into partitions (of age groups or elapsed times) and report performance for each partition. Performance trends across increasing age group or elapsed time are then evaluated. While this approach provides empirical notions of how facial aging affects the performance of systems, covariate analysis is needed to account for the effects of other factors (*e.g.,* pose, image quality) that also play a role in performance. In particular, the FG-NET database contains a number of other variations that can make recognition difficult, in addition to those related to facial aging (see Fig. 1.11).

Longitudinal databases are difficult to acquire because images of the same subjects need to be collected over time. A database for studying facial aging should consist of both a large number of subjects and a large number of images per subject that have been collected over time. While the FG-NET and MORPH databases have primarily been the only publicly available databases for studying facial aging, they are not ideal for longitudinal study due to the following reasons:

- FG-NET contains only 82 subjects in total, and 48% of the 1,002 total face images are younger than 13 years old. Even with small elapsed times, face recognition of children is still an open research problem; the FRVT 2013 [55] reported that all of the top six commercial algorithms suffered an especially large decrease in performance for all age groups less than 13 years old.

- While the largest commercial version of the MORPH database has about 20,000 subjects, there are only an average of 4 face images per subject. Additionally, there are only 317 subjects with more than 5 face images collected over at least 5 years.

Hence, if we wish to study how facial changes of individuals affects face recognition performance over time, we need to leverage a database that is both fairly constrained with respect to other covariates and contains a large number of images per subject acquired over periods of time which are long enough for facial changes due to aging to occur.

(a) (b)

Figure 1.11 Face images of two example subjects from the FG-NET database [78]: (a) female at ages 3–38 years and (b) male at ages 19–63 years. As shown in these examples, the FG-NET database contains a significant amount of variations (pose, illumination, inter-pupillary distances, image quality, etc.), in addition to intrinsic variations due to facial aging.

Figure 1.12 Face images and corresponding ages (in years) of three example subjects from the MORPH database [113]. The largest commercial version of MORPH has 78,207 face images of 20,569 subjects. However, there are only 317 subjects with at least 5 images acquired over at least 5 years (these are three of the 317).

1.7 Contributions

Automatic face recognition has been an extensively studied topic for more than two decades. Significant advancements in the technology have been realized in numerous subtasks needed for robust recognition (face detection, alignment, feature extraction, matching). However, as the technology moves from research problems to real-world deployment systems, it is imperative that the research be driven by requirements of these real-world scenarios. In summary, this introduction has highlighted a few limitations of current research in unconstrained face recognition and studies on facial aging, particularly with respect to how these two challenging problems are benchmarked and evaluated.

The perceived contributions of this thesis are the following:

1. Experimental protocols are developed for *identification* of unconstrained face images. Baseline results using a state-of-the-art COTS face matcher and a separate 3D face modeler are provided for both closed-set and open-set scenarios.

2. A framework is provided for matching a collection of face media (image(s), video(s), 3D model(s), demographic data, and sketch) to mitigate the challenges associated with unconstrained face recognition (uncooperative subjects, unconstrained imaging conditions) and to boost recognition accuracy in scenarios where multiple instances of the face may be available (*e.g.,* persons of interest on a watch list).

3. An automatic measure of face image quality is proposed which can be used to reject low-quality face images prior to matching and rank a collection of face images in order of quality (*e.g.,* to determine which face image to put in the gallery or which face images to use to build a 3D face model).

4. The largest (to date) longitudinal study of face recognition performance is conducted to determine the state-of-the-art robustness to facial aging. The study involves two operational mugshot databases consisting of (i) 147,784 images of 18,007 subjects and (ii)

31,852 images of 5,636 subjects; each subject has a minimum of 4 mugshots collected over an average of 8.5 and 5.8 years for the two databases, respectively. Mixed-effects regression models are used to analyze trends in genuine scores over time (*i.e.,* as subjects age) and quantify the subject-specific variability. As such, estimates are provided for how many years of aging are tolerated by face matchers, *e.g.,* before 95% of the population's genuine scores will drop below the threshold at 0.1% FAR. The effects of demographics (age, gender, race) and face image quality are also analyzed.

1.8 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 focuses on utilizing a face media collection to improve unconstrained face recognition accuracy. Chapter 3 investigates human assessments of the quality of a large database of unconstrained face images and proposes an automatic measure of face image quality. Chapter 3 provides a longitudinal study on automatic face recognition which utilizes multilevel statistical models for a covariate analysis of elapsed time and other factors. Chapter 4 concludes this thesis with a summary of contributions and future work.

Chapter 2

Face Recognition with Media Collection

2.1 Introduction

As face recognition applications progress from constrained imaging and cooperative subjects (*e.g.,* identity card deduplication) to unconstrained imaging scenarios with uncooperative subjects (*e.g.,* watch list monitoring), a lack of guidance exists with respect to optimal approaches for integrating face recognition algorithms into large-scale applications of interest. In this work we explore the problem of identifying a person of interest given a variety of information sources about the person (face image, surveillance video, face sketch, 3D face model, and demographic information) in both closed-set and open-set identification modes.

Identifying a person based on unconstrained face images is an increasingly prevalent task for law enforcement and intelligence agencies. In general, these applications seek to determine the identity of a subject based on one or more probe images or videos, where a top-100 ranked list retrieved from the gallery (for example) may suffice for analysts (or forensic examiners) to identify the subject [64]. In many cases, such a forensic identification

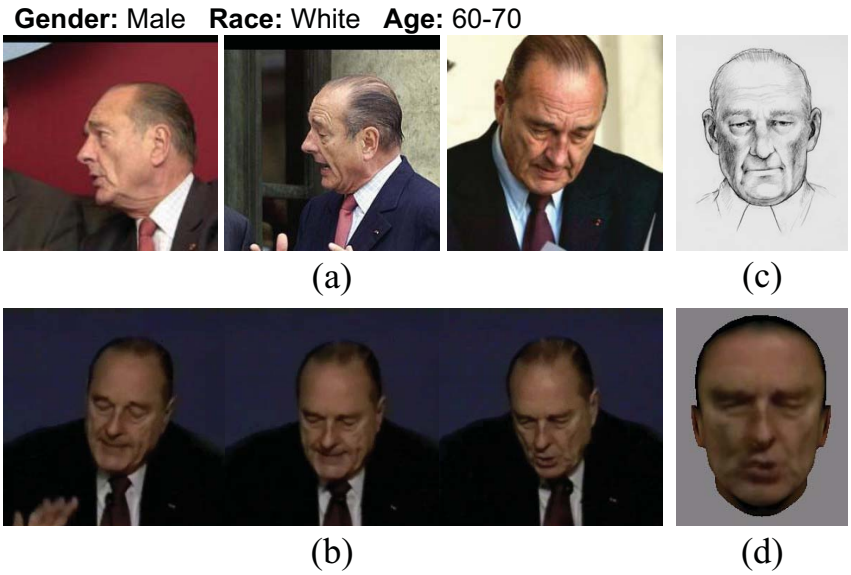


Figure 2.1 A collection of face media for a particular subject may consist of (a) multiple still images, (b) a face track from a video, (c) a forensic sketch, (d) a 3D face model of the subject derived from (a) and/or (b), and demographic information (*e.g.,* gender, race, and age). The images and video track shown here are from [62, 141]. The sketch was drawn by a forensic sketch artist after viewing the face video. In other applications, sketches could be drawn by an artist based on verbal description of the person of interest.

is performed when multiple face images and/or a face track (*i.e.,* a sequence of cropped face images which can be assumed to be of the same person) from a video of a person of interest are available (see Fig. 2.1). For example, in investigative scenarios, multiple face images of an unknown subject often arise from an initial clustering of visual evidence, such as a network of surveillance cameras, the contents of a seized hard drive, or from open source intelligence (*e.g.,* social networks). In turn, these probe images are searched against large-scale face repositories, such as mug shot or identity card databases.

High profile crimes such as the Boston Marathon bombings often rely on data extracted by significant manual effort to identify the person of interest:

> `"It's our intention to go through every frame of every video [from the marathon bombings]," Boston Police Commissioner Ed Davis`[1]

[1] http://www.washingtonpost.com/world/national-security/boston-marathon-bombings-investigators-sifting-through-images-debris-for-clues/2013/04/16/1cabb4d4-a6c4-11e2-b029-8fb7e977ef71_story.html

While other routine, but high value, crimes such as armed robberies, kidnappings, and acts of violence require similar identifications, only a fraction of the manual resources are available to solve these crimes. Thus, it is paramount for face recognition researchers and practitioners to have a firm understanding of optimal strategies for combining multiple sources of face information, collectively called *face media*, available to identify the person of interest.

While forensic identification is focused on human-driven queries, several emerging applications of face recognition technology exist where it is neither practical nor economical for a human to have a high degree of intervention with the automatic face recognition system. One such example is watch list identification from surveillance cameras, where a list of persons of interest are continuously searched against streaming videos. Termed as *open-set recognition*, these challenging applications will likely have better success as unconstrained face recognition algorithms continue to develop and mature [28]. While a closed-set identification system deals with the scenario where the person of interest is assumed to be present in the gallery, and always returns a non-empty candidate list, an open-set identification system allows for the scenario where the person of interest is not enrolled in the gallery, and so can return a possibly empty candidate list [82]. We provide experimental protocols, recognition accuracies on these protocols using COTS face recognition and 3D face modeling algorithms, and an analysis of the integration strategies to improve operational scenarios involving open-set recognition.

2.1.1 Overview

In forensic investigations, manual examination of a suspect's face image against a mug shot database with millions of face images is prohibitive. Thus, automatic face recognition techniques are utilized to generate a candidate suspect list. As shown in Fig. 2.2, forensic investigations using face images typically involve six stages: obtaining face media, preprocessing, automatic face matching, generating a suspect list, human or forensic analysis, and

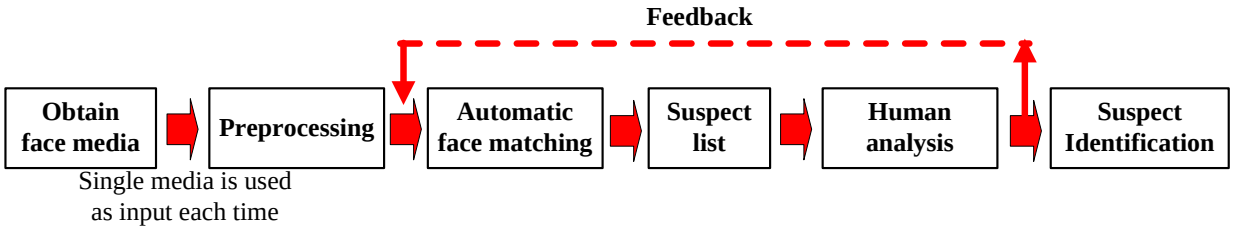


Figure 2.2 Forensic investigations by law enforcement agencies using face images typically involve six main stages: obtaining face media, preprocessing, automatic face matching, generating a suspect list, human analysis, and suspect identification. Feedback occurs after human analysis reveals that, for example, additional preprocessing of the input image (*e.g.,* illumination correction and/or manual eye locations), demographic filtering of the gallery, and/or a different face sample from the media collection is necessary.

suspect identification.[2] The available forensic data or media of the suspect may include still face image(s), video track(s), a face sketch, and demographic information (*e.g.,* age, gender, and race) as shown in Fig. 2.3. While traditional face matching methods take a single media (*i.e.,* a still face image, video track, or face sketch) as probe to generate a suspect list, a media collection is expected to provide more identifiable information about a suspect. The proposed approach contributes to forensic investigations by taking into account the entire media collection of the suspect to perform face matching. This approach generates a single candidate suspect list (rather than a separate list for each face sample in the collection), thereby reducing the amount of human analysis needed.

In this work, we examine the use of commercial off the shelf (COTS) face recognition systems with respect to the aforementioned challenges in large-scale unconstrained face recognition scenarios. First, the efficacy of forensic identification is explored by combining two public-domain unconstrained face databases, Labeled Faces in the Wild (LFW) [62] and YouTube Faces (YTF) [141], to create sets of multiple probe images and videos to be matched against a gallery consisting of a single image for each subject. To replicate forensic identification scenarios, we further populate our gallery with one million operational mug shot images from the Pinellas County Sheriff's Office (PCSO) database.[3] Using this data,

[2] A more detailed description of this forensic investigation process can be found at: http://www.justice.gov/criminal/cybercrime/docs/forensics_chart.pdf

[3] http://biometrics.org/bc2010/presentations/DHS/mccallum-DHS-Future-Opportunities.pdf

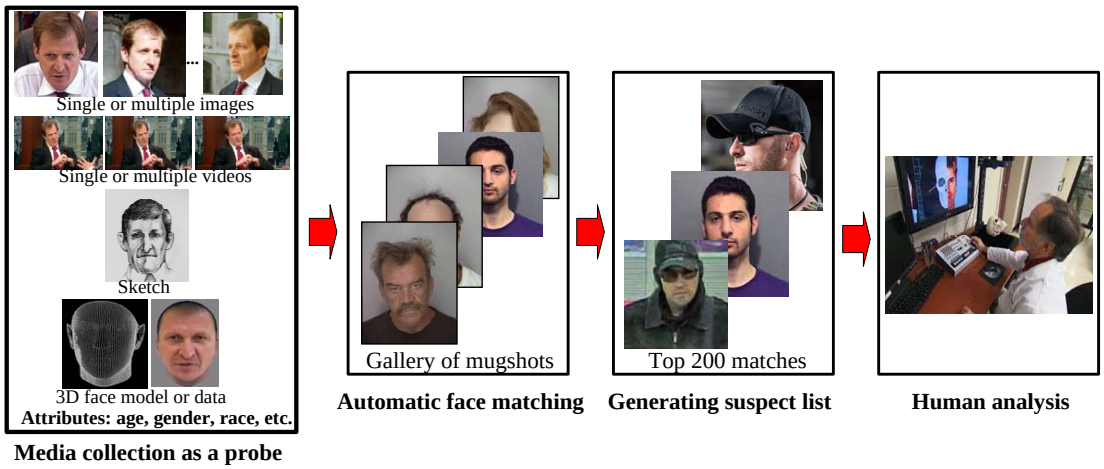


Figure 2.3 Schematic diagram of a person identification task given a face media collection as input.

we are able to examine how to boost the likelihood of face identification through different fusion schemes, incorporation of 3D face models and hand drawn sketches, and methods for selecting the highest quality video frames. Researchers interested in improving forensic identification accuracy can use this competitive baseline (on public-domain databases LFW and YTF) to provide more objectivity towards such goals.

Most of the work on unconstrained face recognition using the LFW and YTF databases has been reported in verification scenarios [98, 137]. However, in forensic investigations, it is the identification mode that is of interest, especially the open-set identification scenario where the person of interest may not be present in legacy face databases.

The contributions of this work are summarized as follows:

- We show, for the first time, how a collection of face media (image(s), video(s), 3D model(s), demographic data, and sketch) can be used to mitigate the challenges associated with unconstrained face recognition (uncooperative subjects, unconstrained imaging conditions) and boost recognition accuracy.

- Unlike previous studies that report results in verification mode, we present results for both open-set and closed-set identifications which are the norm in identifying persons of interest in forensic and watch list scenarios.

- We present effective face quality measures to determine when the fusion of information sources will help boost identification accuracy. The quality measures are also used to assign weights to different media sources in fusion schemes.

- To demonstrate the effectiveness of media-as-input for the difficult problem of unconstrained face recognition, we utilize a state of the art COTS face matcher and a separate COTS 3D face modeler, namely the Aureus 3D SDK provided by CyberExtruder[4]. Face sketches were drawn by forensic sketch artists who generated the sketch after viewing low quality videos. In the absence of demographic data for LFW and YTF databases, we used crowdsourcing to obtain the estimates of gender and race. The above strategy allows us to show the contribution of various media components as we incrementally add them as input to the face matching system.

- Pose-corrected versions of all face images in the LFW database, pose-corrected video frames from the YTF database, forensic sketches, and experimental protocols used in this work have been made publicly available.[5]

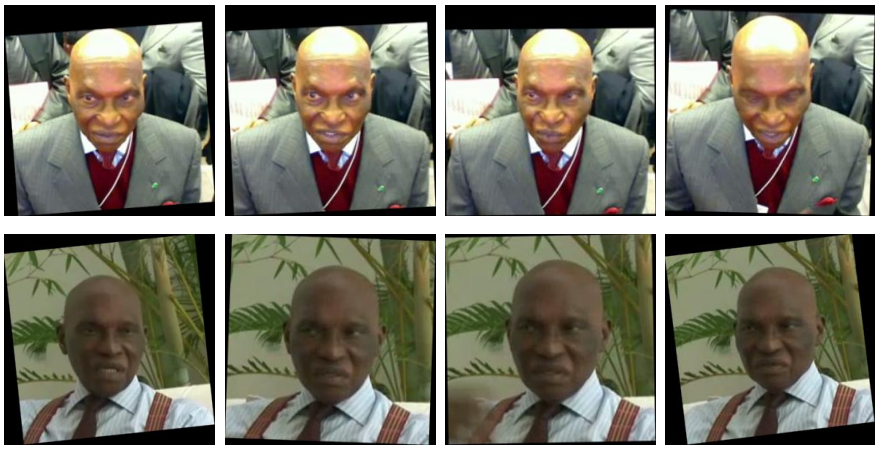
The remainder of this chapter is organized as follows. In Section 2.2, we briefly review published methods related to unconstrained face recognition. We detail the proposed face media collection as input and media fusion method in Sections 2.3 and 2.4, respectively. Experimental setup and protocols are given in Section 2.5, and experimental results are presented in Section 2.6. We conclude this work in Section 2.7.

2.2 Related Work

The release of the public-domain database Labeled Faces in the Wild[6] (LFW) in 2007 spurred interest and progress in unconstrained face recognition. The LFW database is a collection

[4]http://cyberextruder.com/products/aureus-3d-sdk/
[5]http://biometrics.cse.msu.edu/pubs/databases.html
[6]http://vis-www.cs.umass.edu/lfw/

(a) LFW face images



(b) YTF face video tracks

Figure 2.4 Example (a) face images from the LFW database and (b) face video tracks from the YTF database. All faces shown are of the same subject.

of 13,233 face images, downloaded from the Internet, of 5,749 different individuals such as celebrities, public figures, etc. [62]. These images were selected since they meet the criterion that faces can be successfully detected by the Viola-Jones face detector [135]. Despite this property, the LFW database contains significant variations in facial pose, illumination, and expression, and many of the face images are occluded. The LFW protocol consists of face verification based on ten-fold cross-validation, each fold containing 300 "same face" and 300 "not-same face" image pairs.

The YouTube Faces[7] (YTF) database, released in 2011, is the video-equivalent to LFW for unconstrained face matching in videos. The YTF database contains 3,425 videos of 1,595 individuals. The individuals in the YTF database are a subset of those in the LFW database. Faces in the YTF database were also detected with the Viola-Jones face detector at 24 fps, and face tracks were included in the database if there were at least 48 consecutive frames of that individual's face. Similar to the LFW protocol, the YTF face verification protocol consists of ten-fold cross-validation, each fold containing 250 "same face" and 250 "not-same face" track pairs. Figure 2.4 shows example face images and video tracks from the LFW and YTF databases for one particular subject. In this work, we combine these two databases to evaluate the performance of face recognition on unconstrained face media collections.

We provide a summary of related work on unconstrained face recognition, focusing on various face media matching scenarios in Table 2.1. We emphasize that most prior work has evaluated unconstrained face recognition methods in the verification mode. While fully automated face recognition systems are able to achieve ∼99% True Accept Rate (TAR) at 0.1% False Accept Rate (FAR) in constrained imagery and cooperative subject conditions, face recognition in unconstrained environments remains a challenging problem [97]. However, face verification accuracies on the LFW protocol have recently seen drastic improvements. When utilizing outside training data, recent works have achieved TARs greater than 94% at

[7] http://www.cs.tau.ac.il/~wolf/ytfaces/

1% FAR and classification accuracies over 97% (*e.g.,* [122, 128]). However, at 1% FAR, the LFW protocol only contains three impostor scores per fold, so these saturated accuracies may overestimate the abilities of FR systems on unconstrained faces. Liao *et al.* propose a new benchmark for LFW which allows for evaluation at lower FARs; out of three features and seven learning algorithms, they find the best performance is 42% and 66% at 0.1% and 1% FAR, respectively [85]. Open-set identification performance is even lower at 18% for Rank-1 and 1% FAR [85].

Unconstrained face recognition methods can be grouped into two main categories: single face media based methods and face media collection based methods. Single media based methods focus on the scenario where both the query and target instances contain only one type of face media, such as a still image(s), video track(s), or 3D image(s) or model(s). However, the query and target instances can be different media types, such as single image vs. single video. These methods can be effective for unconstrained illumination and expression variations but can only handle limited pose variations. For example, while ∼97% TAR at 0.1% FAR has been reported in MBGCv2.0 unconstrained vs. unconstrained face matching, under large pose variations, this performance drops to ∼17% TAR in MBGCv2.0 non-frontal vs. frontal face matching (see Table 2.1). Such challenges were also observed in single image vs. single image face matching in LFW, and single video vs. single video face matching in YTF and MBGCv2.0 walking vs. walking databases.

These observations suggest that in unconstrained scenarios, a single face media probe, especially of "low quality", may not be able to provide a sufficient description of a face. This motivates the use of a face media collection which utilizes any source of information that is available for a probe (or query) instance of a face. One preliminary study in this direction is the FRGCv2.0 Exp. 3 where (i) a single 3D face image and (ii) a collection of single 3D image and a single 2D face image were used as queries. Results show that 2D face image and 3D face image did improve the face matching performance (79% TAR for 3D face and 2D face vs. 53% TAR for just the 3D face at 0.1% FAR) in unconstrained conditions. It

Table 2.1 A summary of published methods on unconstrained face recognition (UFR). Performance is reported as True Accept Rate (TAR) at a fixed False Accept Rate (FAR) of 0.1% or 1%, unless otherwise noted.

	Dataset	Query Type (size) vs. Target Type (size)	Accuracy (TAR @ FAR)	Source
UFR on Single Media	FRGC v2.0 Exp. 4 unconstrained vs. constrained	Single image (8,014) vs. single image (16,028)	12% @ 0.1%	[97]
	MBGC v2.0 unconstrained vs. unconstrained	Single image (10,687) vs. single image (8,014)	97% @ 0.1%	[97]
	MBGC v2.0 non-frontal vs. frontal	Single image (3,097) vs. single image (16,028)	17% @ 0.1%	[97]
	MBGC v2.0 unconstrained vs. HD video	Single image (1,785) vs. single HD video (512)	94% @ 0.1%	[97]
	MBGC v2.0 walking vs. walking	Notre Dame: Single video (976) vs. single video (976) UT Dallas: Single video (487) vs. single video (487)	Notre Dame: 46% @ 0.1% UT Dallas: 65% @ 0.1%	[97]
	FRGC v2.0 Exp. 3 3D vs. 3D	Single 3D image (4,007) vs. single 3D image (4,007)	53% @ 0.1%	[97]
	LFW Image-Unrestricted Protocol (w/ outside training data)	300 genuine and 300 impostor pairs per fold	88% @ 1% 94% @ 1% 95% @ 1%	[34] [128] [122]
	LFW BLUFR Protocol	4,249 subjects and 9,708 images per fold	90% @ 0.1%	[136]
	YouTube Celebritites	1,500 video clips of 35 celebrities	79% Rank-1 Acc.	[145]
	YouTube Faces	250 genuine and 250 impostors per fold	55% @ 1% 63% @ 1%	[128] [19]
UFR on Media Collection	FRGC v2.0 Exp. 3	Single image & single 3D image (8,014) vs. single 3D image (943)	79% @ 0.1%	[97]
	MBGC v2.0 unconstrained face and iris vs. NIR& HD videos	Single face & iris (14,115) vs. single NIR & single HD (562)	97% @ 0.1%	[97]
	LFW & YouTube Faces (plus 3D face models & demographic information)	Single image vs. single image	56.7%	This work[1]
		Multiple images vs. single image	72.0%	
		Single video vs. single image	31.3%	
		Multiple videos vs. single image	44.0%	
		Multiple images & multiple videos vs. single image	77.5%	
		Multiple images, multiple videos, & 3D model vs. single image	83.0%	
		Multiple images, multiple videos, 3D model, & demographics vs. single image	84.9%	

[1]Performance measures reported here for scenarios considered in this work are Rank-1 identification accuracies.

is, therefore, important to determine how we can improve the face matching accuracy when presented with a collection of face media of different types, albeit of different qualities, as probe.

2.3 Media-as-Input

A face media collection can consist of still images, video tracks, a 3D model, a forensic sketch, and demographic information. In this section, we discuss how we use face "media-as-input" as probe and our approach to media fusion.

2.3.1 Still Image and Video Track

Still image and video track are two of the most widely used sources of media in face recognition systems [82]. Given multiple still images and videos, we use the method reported in [19] to match all still images and video frames available for a subject of interest to the gallery mugshot (frontal pose) images using a COTS face matcher. The resulting match scores are then fused to get a single match score for either multiple probe images or video(s).

2.3.2 3D Face Models

One of the main challenges in unconstrained face recognition is large variations in facial pose [47, 94]. In particular, out-of-plane rotations drastically change the 2D appearance of a face, as they cause portions of the face to be occluded. A common approach to mitigate the effects of pose variations is to build a 3D face model from a 2D image(s) so that synthetic 2D face images can then be rendered at designated poses (*e.g.,* [8, 63, 86]).

In this work, we use a state of the art COTS 3D face modeling SDK, namely CyberExtruder's Aureus 3D SDK, to build 3D models from 2D unconstrained face images.[8] We input eye locations (extracted automatically by [34] for LFW images and the COTS face matcher

[8] http://www.cyberextruder.com/aureus-3d-sdk

for YTF video frames) to the SDK to help with model robustness. The entire 3D face modeling process is fully automatic. The 3D face model is then used to render a "pose corrected" (*i.e.,* frontal facing) image of the unconstrained probe face images. The pose corrected image can then be matched against a frontal gallery. We also pose correct "frontal" gallery images because even the gallery images can have variations in pose as well. Experimental results show that including pose corrected gallery images indeed improves the identification performance.

Given the original and pose corrected probe and gallery images, there are four matching scores that can be computed between any pair of probe and gallery face images (see Fig. 2.5). We use the score s_1 as the baseline to determine whether including scores s_2, s_3, s_4, or their fusion can improve the performance of a COTS face matcher. A face in a video frame can be pose corrected in the same manner. The Aureus SDK also summarizes faces from multiple frames in a video track as a "consolidated" 3D face model (see Fig. 2.6).

2.3.3 Demographic Attributes

In many law enforcement and government applications, it is customary to collect ancillary information like age, gender, race, height, and eye color from the subjects during enrollment. We explore how to best utilize demographic data to boost the recognition accuracy. Demographic information such as age, gender and race becomes even more important in complementing identity information provided by face images and videos in unconstrained face recognition due to the difficulty of the face matching task.

In this work, we take gender and race attributes of each subject in the LFW and YTF face databases as one type of media. Since this demographic information is not available for the subjects in the LFW and YTF face databases, we utilized the Amazon Mechanical Turk (MTurk) crowdsourcing service[9] to obtain the "ground-truth" gender, and race of the 596 subjects that are common in LFW and YTF datasets. Most studies on automatic

[9]www.mturk.com/mturk/

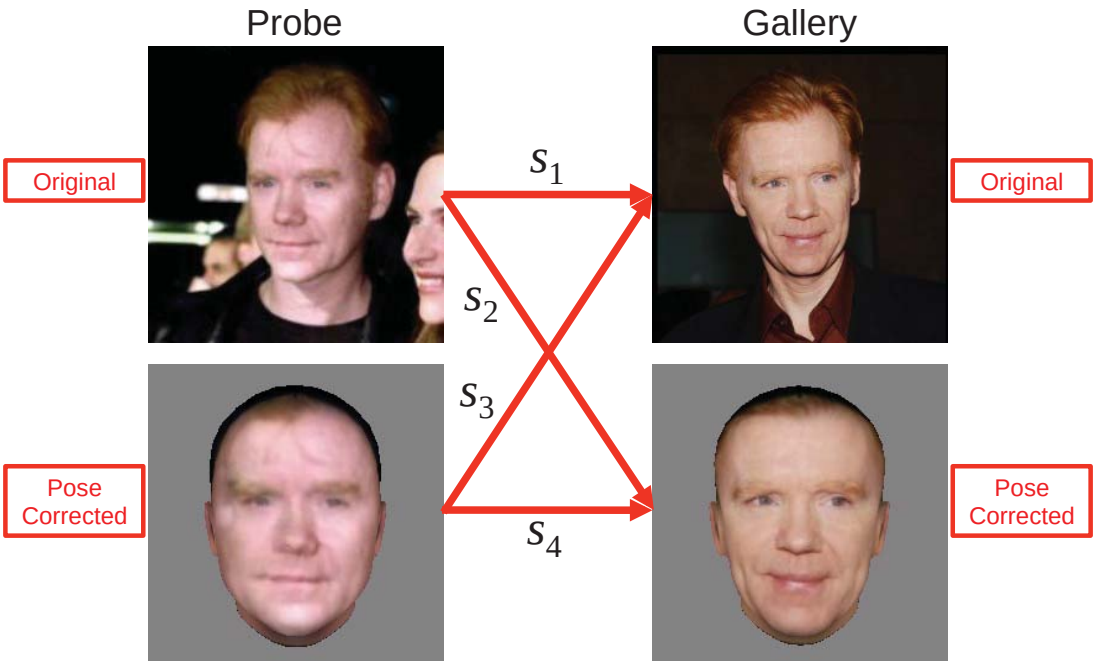


Figure 2.5 Pose correction of probe (left) and gallery (right) face images using CyberExtruder's Aureus 3D SDK. We consider the fusion of four different match scores (s_1, s_2, s_3, and s_4) between the original probe and gallery images (top) and synthetic pose corrected probe and gallery images (bottom).

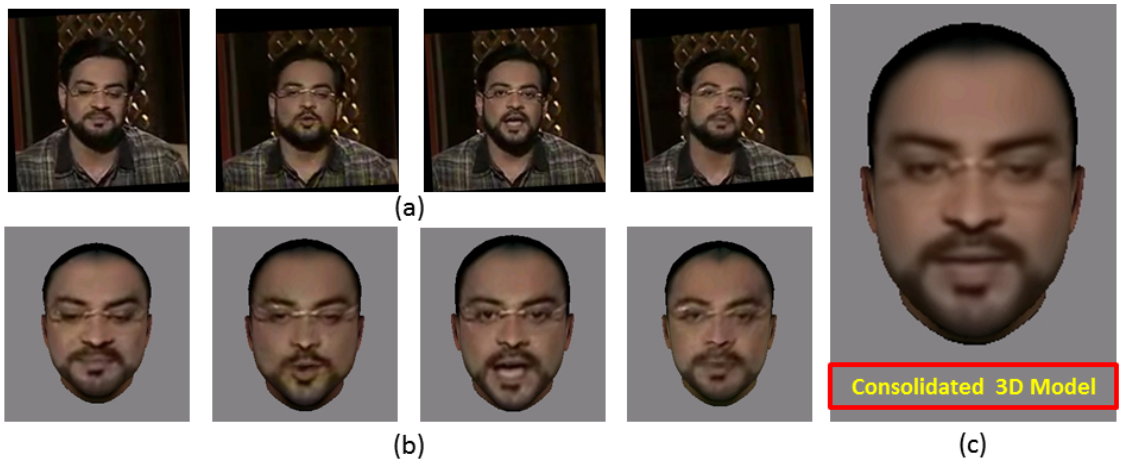


Figure 2.6 Pose corrected faces (b) in a video track (a) and the resulting "consolidated" 3D face model (c). The consolidated 3D face model is a summarization of all frames in the video track.

demographic estimation are limited to frontal face images [59]; demographic estimation from unconstrained face images (*e.g.,* the LFW database) is challenging [76]. For gender and race estimation tasks, we submitted 5,749 (*i.e.,* the number of subjects in LFW) Human Intelligence Tasks (HITs), with ten human workers per HIT, at a cost of 2 cents per HIT. Finally, a majority voting scheme (among the responses) was utilized to determine the gender (Female or Male) and race (Black, White, Asian or Unknown) of each subject. We did not consider age in this work due to large variations in age estimates by crowd workers.

2.3.4 Forensic Sketches

Face sketch based identification dates back to the 19th century [130], where the paradigm for identifying subjects using face sketches relied on human examination. Recent studies on automated sketch based identification systems show that sketches can also be helpful to law-enforcement agencies to identify the person of interest from mugshot databases [58, 74]. In situations where the suspect's photo or video is not available, expertise of forensic sketch artists are utilized to draw a suspect's sketch based on a verbal description provided by an eyewitness or victim. In some situations, even when a photo or video of a suspect is available, the quality of this media can be poor. In this situation also, a forensic sketch artist can be called in to draw a face sketch based on the low-quality face photo or video. For this reason, we also include the face sketch in a face media collection.

We manually selected 21 low-quality (large pose variations, shadow, blur, etc.) videos (one video per subject) from the YTF database (for three subjects, we also included a low quality still image from LFW). We then asked two forensic sketch artists to draw a face sketch for each subject in these videos (10 subjects were drawn by one forensic sketch artist, and 11 subjects by the other). Our current experiments are limited to sketches of 21 subjects due to the high cost of hiring a sketch artist. Examples of these sketches and their corresponding low-quality videos are shown in Figs. 2.7 and 2.15.

(a) Video (b) Cropped face image from video (c) Forensic Sketch

Figure 2.7 An example of a sketch drawn by a forensic artist by looking at a low-quality video. (a) Video shown to the forensic artists, (b) facial region cropped from the video frames, and (c) sketch drawn by the forensic artist. Here, no verbal description of the person of interest is available.

2.4 Media Fusion

Given a face media collection as probe, there are various schemes to integrate the identity information provided by each individual media component, such as score level, rank level, and decision level fusion [114]. Among these approaches, score level fusion is the most commonly adopted. Some COTS matchers do not output a meaningful match score (to prevent hill-climbing attacks [133]). Thus, in these situations, rank level or decision level fusion is typically adopted.

In this work, we match each face media (image, video, 3D model, sketch, or demographic information) of a probe collection to the gallery and combine the scores using score level fusion. Specifically, score level fusion takes place in two different layers: (i) fusion within one type of media, and (ii) fusion across different types of media. The first fusion layer generates a single score from each media type if multiple instances are available. For example, matching scores from multiple images or multiple video frames can be fused to get a single score. Additionally, if multiple video clips are available, matching scores of individual video clips can also be fused. Score fusion within the ith face media can generally be formulated as

$$s_i = \mathfrak{F}(s_{i,1}, s_{i,2}, \cdots, s_{i,n}), \tag{2.1}$$

where s_i is a single match score based on n instances of the ith face media type; $\mathfrak{F}(\cdot)$ is a score level fusion rule; we use the *sum* rule, *e.g.,* $s = \frac{1}{n}\sum s_{i,n}$, which has been found to be quite effective in practice [19]. Note that the *sum* and *mean* rules are equivalent, but we use the terms *mean* and *sum* for situations when normalization by the number of scores is and is not necessary, respectively. Given a match score for each face media type, the next fusion step involves fusing the scores across different types of face media. Again, the *sum* rule is used and found to work very well in our experiments; however, as shown in Fig. 2.8, face media for a person of interest can be of different quality. For example, a 3D face model can be corrupted due to inaccurate localization of facial landmarks. As a result, match scores calculated from individual media sources may have different degrees of confidence.

We take into account the quality of individual media type by designing a quality based fusion. Specifically, let $\mathbf{S} = [s_1, s_2, \cdots, s_m]^T$ be a vector of the match scores between n different media types in a collection of probe and gallery, and $\mathbf{Q} = [q_1, q_2, \cdots, q_m]^T$ be a vector of quality values for the corresponding input media. Match scores from the COTS matcher are normalized with *z-score* normalization. The quality values are normalized to the range $[0, 1]$. The final match score between a probe and a gallery image is calculated by a weighted *sum* rule fusion,

$$s = \frac{1}{m}\sum_{i=1}^{m} q_i s_i = \mathbf{Q}^T\mathbf{S}. \tag{2.2}$$

Note that the quality based across-media fusion in (2.2) can also be applied to score level fusion within a particular face media type (*e.g.,* 2D video frames).

In this work, we have considered five types of media in a collection: 2D face image, video, 3D face model, sketch, and demographic information. However, since sketches of only 21 persons (out of 596 persons that are common in LFW and YTF databases) are available, in most of the experiments, we perform quality-based fusion in (2.2) based on only four types of media ($m = 4$). The quality measures for individual media type are defined as follows.

- **Image and video:** For a probe image, the COTS matcher assigns a face confidence

value in the range of $[0, 1]$, which is used as the quality value. For each video frame, the same face confidence value measure is used. The average face confidence value across all frames is used as the quality value for a video track.

- **3D face model:** The Aureus 3D SDK used to build a 3D face model from image(s) or video frame(s) does not output a confidence score. We define the quality of a 3D face model based on the pose corrected 2D face image generated from it. Given a pose corrected face image, we calculate its structural similarity (SSIM) [139] to a set of predefined reference images (manually selected frontal face images). Let $\mathbf{I}_{PC}$ be a pose corrected face image (from the 3D model), and $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_t\}$ be the set of t reference face images. The quality value of a 3D model based on SSIM is defined as

$$\begin{aligned} q(\mathbf{I}_{PC}) &= \tfrac{1}{t} \sum_{i=1}^{t} \text{SSIM}(\mathbf{I}_{PC}, \mathbf{R}_i) \\ &= \tfrac{1}{t} \sum_{i=1}^{t} l(\mathbf{I}_{PC}, \mathbf{R}_i)^{\alpha} \cdot c(\mathbf{I}_{PC}, \mathbf{R}_i)^{\beta} \cdot s(\mathbf{I}_{PC}, \mathbf{R}_i)^{\gamma} \end{aligned} \tag{2.3}$$

  where $l(\cdot)$, $c(\cdot)$, and $s(\cdot)$ are luminance, contrast, and structure comparison functions [139], respectively; α, β, and γ are parameters used to adjust the relative importance of the three components. We use the recommended parameters $\alpha = \beta = \gamma = 1$ in [139]. The quality value is in the range of $[0, 1]$.

- **Demographic information:** As stated earlier, we collected demographic attributes (gender and race) of each face image using the MTurk crowdsourcing service with ten MTurk workers per task. Hence, the quality of demographic information can be measured by the degree of consistency among the ten MTurk workers. Let $\mathbf{E} = [e_1, e_2, \cdots, e_k]^T$ be the collection of estimates of one specific demographic attribute (gender or race) by k (here, $k = 10$) MTurk workers. The quality value of this demographic attribute can be calculated as

$$q(\mathbf{E}) = \frac{1}{k} \max_{i=1,2,\cdots,c} \{\sum (\mathbf{E} == i)\}, \tag{2.4}$$

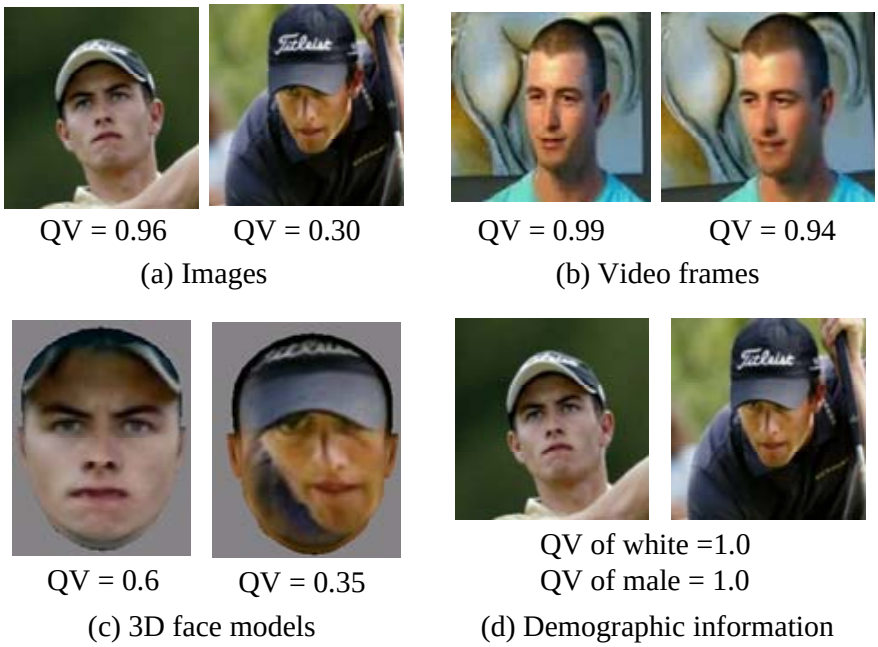


Figure 2.8 Examples of different face media types with varying quality values (QV) of one subject: (a) images, (b) video frames, (c) 3D face models, and (d) demographic information. The range of QV is [0,1].

where c is the total number of classes for one demographic attribute. Here, $c = 2$ for gender (Male and Female); while $c = 4$ for race (Black, White, Asian, and Unknown). The notation $\sum(\mathbf{E} == i)$ denotes the number of estimates that are labeled as class i. The quality value range in (2.4) is in $[0, 1]$.

Quality values for different face media of one subject are shown in Fig. 2.8. We note that the proposed quality measures give reasonable quality assessments for different input media.

2.5 Experimental Setup

The 596 subjects who have at least two images in the LFW database and at least one video track in the YTF database (subjects in YTF are a subset of those in LFW) are used to evaluate the performance of face identification on media-as-input in both closed-set and open-set scenarios. The state of the art COTS face matcher used in our experiments was one of the top performers in the 2010 NIST Multi-Biometric Evaluation [97]. Though the

Table 2.2 Number of probe face images (from the LFW database) and video tracks (from the YTF database) available for the 596 subjects that are common in the two databases.

# images/videos per subj.	1	2	3	4	5	6	7+
# subjects (LFW images)	238	110	78	57	25	12	76
# subjects (YTF videos)	204	190	122	60	18	2	0

COTS face matcher is designed for matching still images, we apply it to video-to-still face matching via multi-frame fusion to obtain a single score for the video track [19]. In all cases where video tracks are part of the face media collection, we use the *mean* rule for multi-frame fusion (the *max* fusion rule performed comparably [19]).

2.5.1 Closed Set Identification

In closed-set identification experiments, one frontal LFW image per subject is placed in the gallery (one with the highest frontal score from the COTS matcher), and the remaining LFW images are used as probes. All YTF video tracks for the 596 subjects are used as probes. Table 2.2 shows the distribution of number of probe images and videos per subject. The average number of images, video tracks, and total media instances per subject is 5.3, 2.2, and 7.4, respectively. We further extend the gallery size with an additional $3,653$ LFW images (of subjects with only a single image in LFW). In total, the size of the gallery is $4,249$.

We evaluate five different scenarios depending on the contents of the probe set: (i) single image as probe, (ii) single video track as probe, (iii) multiple images as probe, (iv) multiple video tracks as probe, and (v) multiple images and video tracks as probe. We also take into account the 3D face models and demographic information in the five scenarios. To better simulate the scenarios in real-world forensic investigations, we also provide a case study on the Boston Marathon bomber to determine the efficacy of using media, and the generalization ability of our system to a large gallery with one million background face images.

For all closed-set experiments involving still images from LFW, we input automatically

extracted eye locations (from [34]) to the COTS face matcher to help with enrollment because the COTS matcher sometimes enrolls a background face in the LFW image that is not the subject of interest. Against a gallery of approximately $5,000$ LFW frontal images, we observed a 2–3% increase in accuracy for Rank-20 and higher by inputting the automatically extracted eye locations from [34]. Note that for the YTF video tracks, there are no available ground-truth eye locations for faces in each frame. Recall from Section 2.3.2 that we input eye locations from [34] and the COTS face matcher to build the 3D models for LFW images and YTF video frames, respectively; hence, the entire 3D face modeling process is fully automatic. We report closed-set identification results as Cumulative Match Characteristic (CMC) curves.

2.5.2 Open Set Identification

Here, we consider the case when the person of interest in the probe image or video track may not have a true mate in the gallery. This is representative of a watch list scenario. The gallery (watch list) consists of 596 subjects with at least two images in the LFW database and at least one video in the YTF database. To evaluate performance in the open-set scenario, we construct two probe sets: (i) a *genuine probe set* that contains faces matching gallery subjects, and (ii) an *impostor probe set* that does not contain faces matching gallery subjects.

We conduct two separate experiments: (i) randomly select one LFW image per watch list subject as the genuine probe set and use the remaining LFW images of subjects not on the watchlist as the impostor probe set (596 gallery subjects, 596 genuine probe images, and $9,494$ impostor probe images), and (ii) use one YTF video per watch list subject as the genuine probe set, and the remaining YTF videos which do not contain watch list subjects as the impostor probe set (596 gallery subjects, 596 genuine probe videos, and $2,064$ impostor probe videos). For each of these experiments, we evaluate three scenarios for the gallery: (i) single image, (ii) multiple images, and (iii) multiple images and videos.

Open-set identification can be considered a two step process: (i) decide whether or not to

reject a probe image as not in the watchlist, and (ii) if probe is in the watchlist, recognize the person. Hence the performance is evaluated based on (i) Rank-1 detection and identification rate (DIR), which is the fraction of genuine probes matched correctly at Rank-1, and not rejected at a given threshold, and (ii) the false alarm rate (FAR) of the rejection step (i.e. the fraction of impostor probe images which are not rejected). We report the DIR vs. FAR curve describing the tradeoff between true Rank-1 identifications and false alarms.

2.6 Experimental Results

2.6.1 Pose Correction

We first investigate whether using a COTS 3D face modeling SDK to pose correct a 2D face image prior to matching improves the identification accuracy. The closed-set experiments in this section consist of a gallery of $4,249$ frontal LFW images and a probe set of $3,143$ LFW images or $1,292$ YTF videos. Table 2.3 (a) shows that the COTS face matcher performs better on face images that have been pose corrected using the Aureus 3D SDK. Matching the original gallery images to the pose corrected probe images (*i.e.,* match score s_3) performs the best out of all four match scores, achieving a 7.25% improvement in Rank-1 accuracy over the baseline (*i.e.,* match score s_1). Furthermore, fusion of all four scores (s_1, s_2, s_3, and s_4) with the simple *sum* rule provides an additional 2.6% improvement at Rank-1. Consistent with the results for still images, match scores s_3 and $sum(s_1, s_2, s_3, s_4)$ also provide significant increases in identification accuracy over using match score s_1 alone for matching frames of a video track (Table 2.3 (b)). We note that s_4 likely performs lower than s_3 because the gallery images are already fairly frontal. If both the gallery and the probe face images are unconstrained then s_4 may perform better.

Next, we investigate whether the Aureus SDK consolidated 3D models (*i.e.,* n frames of a video track summarized as a single 3D face model rendered at frontal pose) can achieve comparable accuracy to matching all n frames. Table 2.4(a) shows that the accuracy of

Table 2.3 Closed-set identification accuracies (%) for pose corrected gallery and/or probe face images using 3D model. The gallery consists of 4,249 LFW frontal images and the probe sets are (*a*) 3,143 LFW images and (*b*) 1,292 YTF video tracks. Performance is shown as Rank retrieval results at Rank-1, 20, 100, and 200. Computation of match scores s_1, s_2, s_3, and s_4 are shown in Fig. 2.5.

LFW Images				
	R-1	R-20	R-100	R-200
s_1	56.7	78.1	87.1	90.2
s_2	57.7	77.6	86.0	89.9
s_3	**63.9**	**83.4**	**90.7**	**93.6**
s_4	55.6	78.8	88.0	91.9
sum	**66.5**	**85.9**	**92.4**	**95.1**

(a)

YTF Video Tracks				
	R-1	R-20	R-100	R-200
s_1	31.3	54.2	68.0	74.5
s_2	32.3	55.3	67.8	73.9
s_3	**36.3**	**58.8**	**71.3**	**77.2**
s_4	31.7	54.4	68.7	76.5
sum	**38.8**	**61.4**	**73.6**	**79.0**

(b)

Table 2.4 Closed-set identification accuracies (%) for matching consolidated 3D face models built from (*a*) all frames of a video track or (*b*) a subset of high quality (HQ) video frames.

Consolidated 3D Model: All Frames				
	R-1	R-20	R-100	R-200
s_3	33.1	54.1	67.3	72.8
s_4	29.4	51.7	64.8	71.1
sum	34.6	56.4	68.2	74.1

(a)

Consolidated 3D Model: Frame Selection				
	R-1	R-20	R-100	R-200
s_3	34.4	56.6	67.8	73.4
s_4	29.8	52.4	66.5	72.7
sum	35.9	58.3	69.9	75.1

(b)

$sum(s_3, s_4)$ (*i.e.,* consolidated 3D models matched to original and pose corrected gallery images) provides the same accuracy as matching all n original frames (*i.e.,* score s_1 in Table 2.3 (b)). However, the accuracy of the consolidated 3D model is slightly lower ($\sim 5\%$) than *mean* fusion over all n pose corrected frames (*i.e.,* score s_3 in Table 2.3 (b)). Hence, the consolidated 3D model built from a video track is not able to retain all discriminatory information contained in the collection of n pose-corrected frames.

2.6.2 Forensic Identification: Media-as-Input

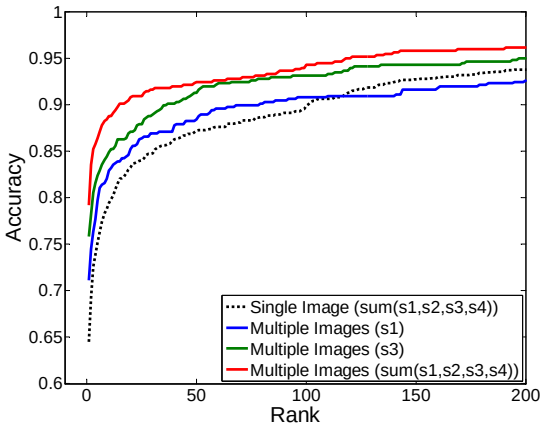
A summary of results for the various media-as-input scenarios is shown in Fig. 2.9. For all scenarios that involved multiple probe instances (*i.e.,* multiple images and/or videos), the

mean fusion method gave the best result. For brevity, all CMC curves and results that involve multiple probe instances are also obtained via *mean* fusion. We also investigated the performance of rank-level fusion; the highest-rank fusion performed similar to score-level fusion, while the Borda count method [114] performed worse.

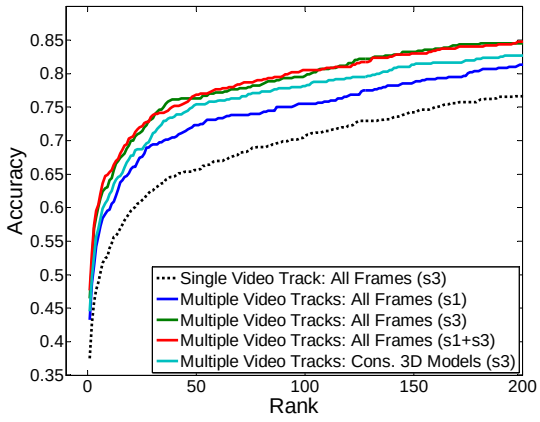
As observed in the previous section, pose correction with the Aureus 3D SDK to obtain scores s_3 or $sum(s_1, s_2, s_3, s_4)$ achieves better accuracies than score s_1. This is also observed in Figs. 2.9(a) and 2.9(b) where scores $sum(s_1, s_2, s_3, s_4)$ and s_3 provide approximately a 5% increase in accuracy over score s_1 for multiple images and multiple videos, respectively. This improvement is also observed in Fig. 2.9(c) for matching media that includes both still images and videos, but the improvement is mostly at low ranks ($<$ Rank-50).

Figure 2.9 shows that (i) multiple probe images and multiple probe videos perform better than their single instance counterparts, but (ii) multiple probe videos actually perform worse than single probe image (see Figs. 2.9(a) and 2.9(b)). This is likely due in part to videos in the YTF database being of lower quality than the still images in the LFW database. However, we note that though multiple videos perform poorly compared to still images, there are still cases where the fusion of multiple videos with the still images does improve the identification performance. This is shown in Fig. 2.9(c); the best result for multiple images is plotted as a baseline to show that the addition of videos to the media collection improves identification accuracy. An example of this is shown in Fig. 2.10. For this particular subject, there is only a single probe image available that exhibits extreme pose. The additional information provided by the 3D model and video track improves the true match from Rank-438 to Rank-8. In fact, the performance improvement of media (*i.e.,* multiple images and videos) over multiple images alone can mostly be attributed to cases where there is only a single probe image with large pose, illumination, and expression variations.

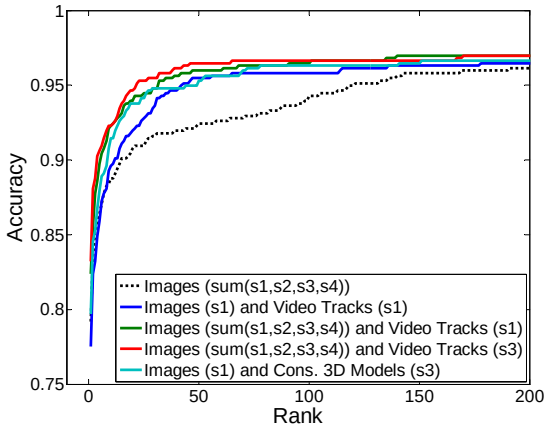
While Fig. 2.9 shows that including additional media to a probe collection improves identification accuracies on average, there are cases where matching the entire media collection can degrade the matching performance. An example is shown in Fig. 2.11. Due to the



(a) Images



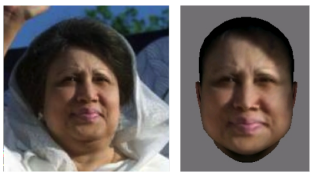
(b) Video Tracks



(c) Media Collection

Figure 2.9 Closed-set identification results for different probe sets: (a) multiple still face images, (b) multiple face video tracks, and (c) face media collection (images, videos and 3D face models). Single face image and video track results are plotted in (a) and (b) for comparison. Note that the ordinate scales are different in (a), (b), and (c) to accentuate the difference among the plots.

(a) Probe media collection



(b) Gallery true mate

Figure 2.10 A collection of face media for a subject (a) consisting of a single still image, 3D model, and video track improves the retrieval rank of the true mate in the gallery (b). Against a gallery of 4,249 frontal images, the single still image was matched at Rank-438 with the true mate. Including the 3D model along with the still image improved the match to Rank-118, while the entire probe media collection was matched to the true mate at Rank-8.

(a) Probe image and 3D model

(b) Gallery image and 3D model

(c) Probe video tracks

Figure 2.11 Additional face media does not always improve the identification accuracy. In this example, the probe image with its 3D model (a) was matched at Rank-5 against a gallery of 4,249 frontal images. Inclusion of three video tracks of the subject (c) to the probe set degraded the true match to Rank-216.

fairly low quality of the video tracks, the entire media collection for this subject is matched at Rank-216 against the gallery of $4,249$ images, while the single probe image and pose corrected image (from the 3D model) are matched at Rank-5. This necessitates the use of quality measures to assign a degree of confidence to each media.

We evaluated the face verification performance (see Fig. 2.12) using the same database as the closed-set identification protocol (*i.e.,* gallery (target) of 4,249 images and probe (query) media collections of 596 subjects). We found that score s_3 still outperforms s_1, s_2, and s_4 for still images and videos frames. In investigating why s_3 performs better than s_4, we found that s_4 provides a better genuine score distribution than s_3, but the impostor distribution of s_4 has a longer tail. We believe this is partially due to similarities in the contours of two pose-corrected images. However, we find that multiple images with their 3D models (sum(s_1,s_2,s_3,s_4)) perform better than a media collection of multiple images (s_1) and video frames (s_1 or consolidated 3D model), whereas in closed-set identification, these media collections perform better than the multiple images and 3D models alone. In both identification and verification modes, the best performance is a collection of images with their 3D models and video frames. Image and video scores were normalized with *z-score* normalization.

2.6.3 Quality-based Media Fusion

In this section, we evaluate the proposed quality measures and quality-based face media fusion. As discussed in Section 2.4, quality measures and quality-based face media fusion can be applied at both within-media layer and across-media layer.

Tables 2.5 (a) and (b) show the closed-set identification accuracies of quality-based fusion of match scores $(s_1, \cdots, s_4)$ of single image per probe and multiple images per probe, respectively. The performance with *sum* rule fusion is also provided for comparison. Our results indicate that the proposed quality measures and quality based fusion are able to improve the matching accuracies in both scenarios. Examples where the quality-based fusion

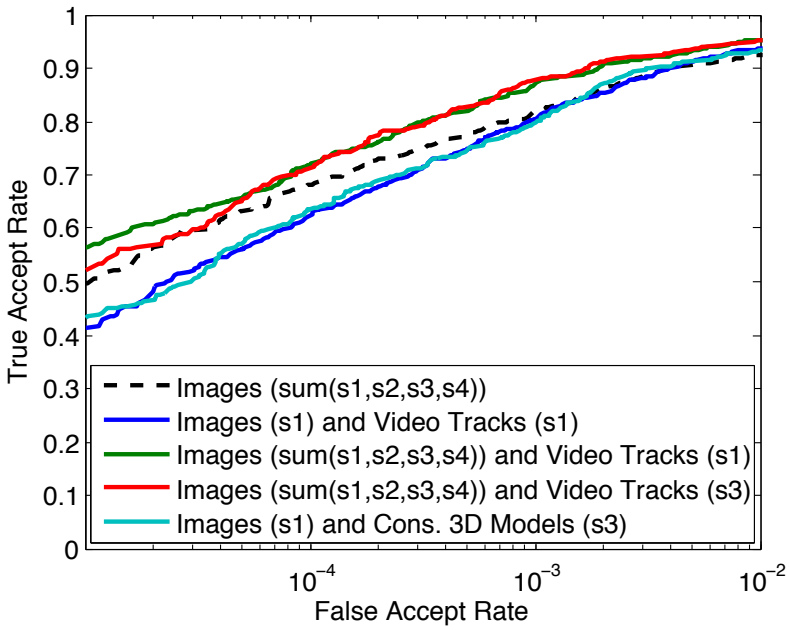


Figure 2.12 Face verification performance of a gallery of 4,249 frontal LFW images and probe media collections of 596 subjects.

Table 2.5 Closed-set identification accuracies (%) for quality based fusion (QBF) (*a*) within a single image, and (*b*) across multiple images.

QBF within a single image

	R-1	R-20	R-100	R-200
sum	65.7	83.2	90.1	93.5
QBF	66.5	85.9	92.6	95.3

(a)

QBF across multiple images

	R-1	R-20	R-100	R-200
sum	79.4	91.1	94.5	96.5
QBF	80.0	91.8	94.5	96.5

(b)

performs better than *sum* rule fusion are shown in Fig. 2.13 (a). Although in some cases the quality-based fusion may perform worse than *sum* rule fusion (see Fig. 2.13 (b)), overall, it still improves the matching performance (see Table 2.5).

We have also applied the proposed quality measure for 3D face model to select high-quality frames that are used to build a consolidated 3D face model for a video clip. Figure 2.14 (a) shows two examples where the consolidated 3D models using frame selection with SSIM quality measure (see Sec. 2.4) gets better retrieval ranks than using all frames. Although, a single value, *e.g.,* the SSIM based quality measure, may not always be reliable to describe the quality of a face image (see Fig. 2.14 (b)), frame selection still slightly improves the identification accuracy of the consolidated 3D face models at low ranks (see Table 2.4).

2.6.4 Forensic Sketch Experiments

In this experiment, we study the effectiveness of forensic sketches in a media collection. For each subject with a forensic sketch, we input the forensic sketch to the COTS matcher to obtain a retrieval rank. Among the 21 subjects for whom we have a sketch, sketches of 12 subjects are observed to perform significantly better than the corresponding low-quality videos. Additionally, when demographic filtering using gender and race is applied, we can further improve the retrieval ranks. Figure 2.15 shows three examples where the face sketches significantly improved the retrieval ranks compared to low quality videos. The retrieval ranks

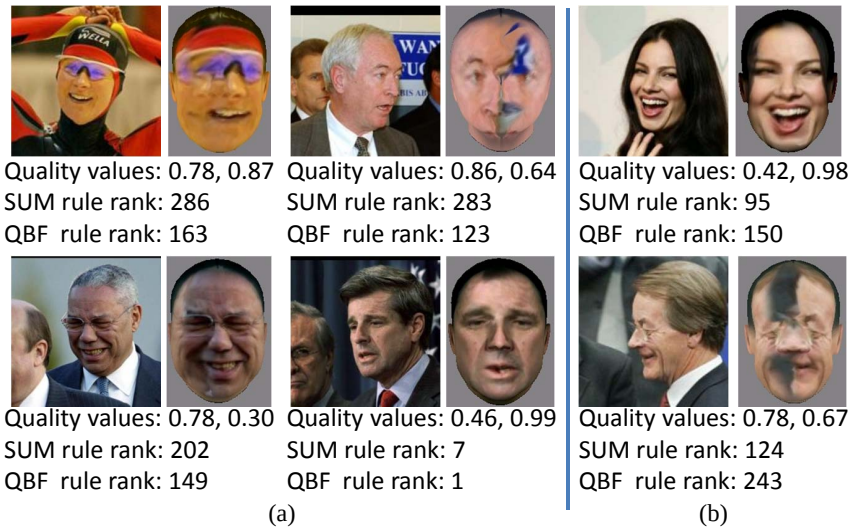


Figure 2.13 A comparison of quality based fusion (QBF) vs. simple *sum* rule fusion (SUM). (a) Examples where quality based fusion provides better identification accuracy than *sum* fusion; (b) Examples where quality based fusion leads to lower identification accuracy compared with *sum* fusion.

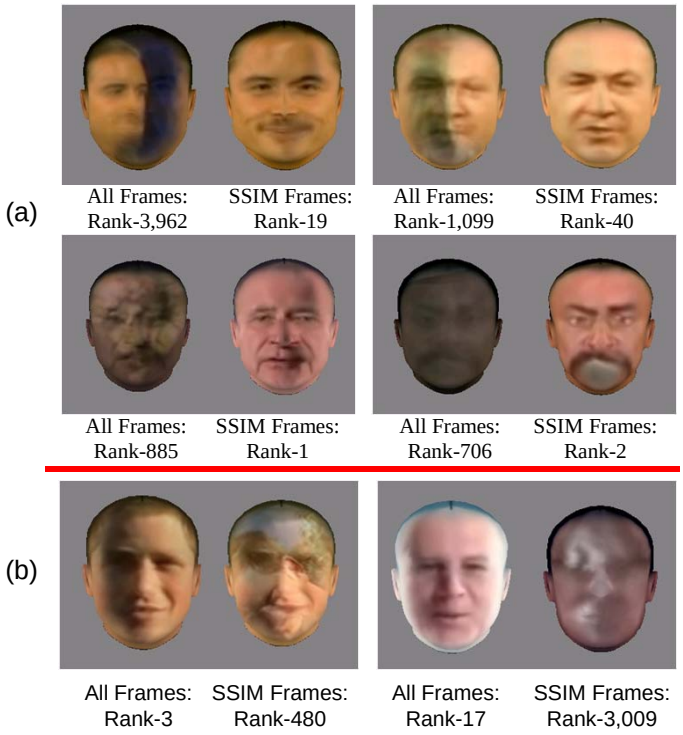


Figure 2.14 Retrieval ranks using consolidated 3D face models (built from video tracks). Frame selection with SSIM quality measure (see Sec. 2.4) prior to building the consolidated 3D face model (a) improves and (b) degrades the identification accuracy. However, overall, frame selection using the proposed quality measure based on SSIM improves the COTS matcher's performance by an average of 1.43% for low ranks 1 to 50.

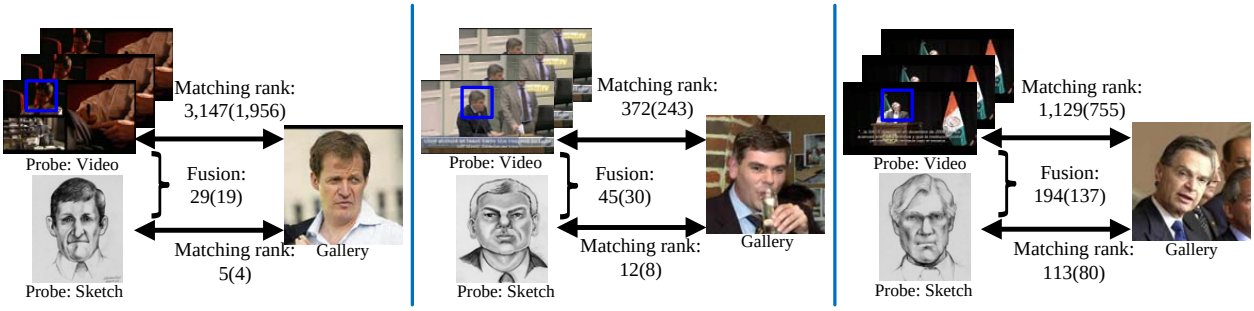


Figure 2.15 Three examples where the face sketches drawn by a forensic artist after viewing the low-quality videos improve the retrieval rank. The retrieval ranks without and with combining the demographic information (gender and race) are given in the form of #(#).

of sketch and low-quality video fusion are also reported in Fig. 2.15.

To further demonstrate the efficacy of forensic sketch, we focus on identification of Tamerlan Tsarnaev, the older brother involved in the 2013 Boston Marathon bombing. In an earlier study Klontz and Jain [73] showed that while the younger brother, Dzhokhar Tsarnaev, could be identified at Rank-1 based on his probe images released by the authorities, the older brother could only be identified at Rank-12,446 (from a gallery of one million images with no demographic filtering). Figure 2.16 shows three gallery face images of Tamerlan Tsarnaev (*1x*, *1y*, and *1z* [73]) and two probe face images (*1a* and *1b*) which were released by the FBI during the investigation.[10] Because the probe images of Tamerlan Tsarnaev are of poor quality, particularly due to wearing of sunglasses and a hat, we also asked a sketch artist to draw a sketch of Tamerlan Tsarnaev (*1c* in Fig. 2.16) while viewing the two probe images.[11]

To simulate a large-scale forensic investigation, the three gallery images of Tamerlan Tsarnaev were added to a background set of one million mugshot images of 324,696 unique subjects from the PCSO database. Particularly due to the occlusion of eyes, the probe images are difficult for the COTS face matcher to identify (though they can be enrolled with

[10] http://www.fbi.gov/news/updates-on-investigation-into-multiple-explosions-in-boston

[11] *"I was living in Costa Rica at the time that event took place and while I saw some news coverage, I didn't see much and I don't know what he actually looks like. The composite I am working on is* 100% *derived from what I am able to see and draw from the images you sent. I can't make up information that I can't see, so I left his hat on and I can only hint at eye placement."* - Jane Wankmiller, forensic sketch artist, Michigan State Police.

Race: White
Gender: Male
Age: 20 to 30

1a *1b* *1c*

1x *1y* *1z*

Figure 2.16 Face images used in our case study on identification of Tamerlan Tsarnaev, one of the two suspects of the 2013 Boston Marathon bombings. Probe (*1a*, *1b*) and gallery ($1x$, $1y$, and $1z$) face images are shown. *1c* is a face sketch drawn by a forensic sketch artist after viewing *1a* and *1b*, and a low quality video frame from a surveillance video.

Table 2.6 Retrieval ranks for probe images (*1a, 1b)* and sketch (*1c*) matched against gallery images *1x, 1y,* and *1z* with an extended set of one million mug shots (*a*) without and (*b*) with demographic filtering. Rows max and mean denote score fusion of multiple images of this suspect in the gallery; columns max and sum are score fusion of the three probes.

(*a*) Without Demographic Filtering

	1a	*1b*	*1c*	max	sum
1x	117,322	475,769	**8,285**	18,710	27,673
1y	12,444	440,870	63,313	38,298	28,169
1z	87,803	237,704	53,771	143,389	55,712
max	9,409	117,623	**6,259**	14,977	**6,281**
mean	13,658	125,117	8,019	20,614	8,986

(*b*) With Demographic Filtering (white male, 20-30)

	1a	*1b*	*1c*	max	sum
1x	5,432	27,617	**112**	114	353
1y	518	25,780	1,409	1,656	686
1z	3,958	14,670	1,142	2,627	1,416
max	374	6,153	94	109	106
mean	424	5,790	**71**	109	**82**

manually marked eye locations), as shown in Table 2.6. However, the retrieval rank for the sketch (*1c* in Fig. 2.16) is much better compared to the two probe images (*1a* and *1b* in Fig. 2.16), with the best match at Rank-6,259 for max fusion of multiple images of Tamerlan Tsarnaev (*1x*, *1y*, and *1z*) in the gallery. With demographic filtering [71] (white male in the age range of 20 to 30 filters the gallery to $54,638$ images of $13,884$ subjects), the sketch is identified with gallery image *1x* (a mugshot)[12] in Fig. 2.16 at Rank-112. Again, score fusion of multiple images per subject in the gallery further lowers the retrieval to Rank-71. The entire media collection (here, *1a*, *1b*, and *1c* in Fig. 2.16) is matched at Rank-82 against the demographic-filtered and multiple image-fused gallery.

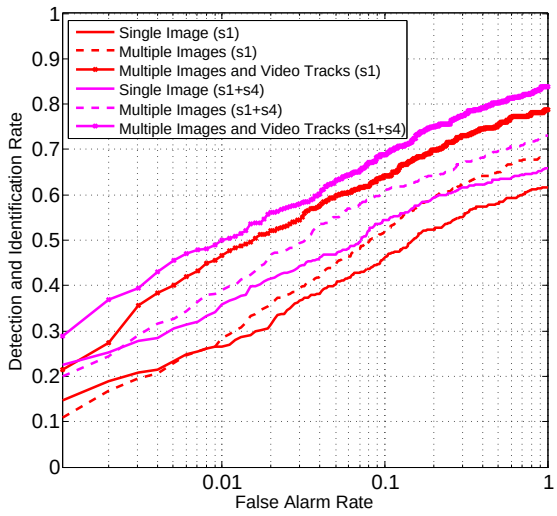
2.6.5 Watch List Scenario: Open Set Identification

We report the DIR vs. FAR curves of open-set identification in Figs. 2.17 (a) and (b). With a single image or single video per subject in the gallery, the DIR values at 1% FAR are about 25% and 10% for still image probe and video clip probe, respectively. This suggests that a large percentage of probe images or video clips that are matched to their gallery true mates at a low rank in a closed-set identification scenario, can no longer be successfully matched in an open-set scenario. Of course, this comes at the benefit of much lower false alarms than in the closed-set identification. The proposed face media collection based matching still shows improvement over single media based matching. For example, at 1% FAR, face media collection based matching leads to about 20% and 15% higher DIRs for still image and video clip probes, respectively.

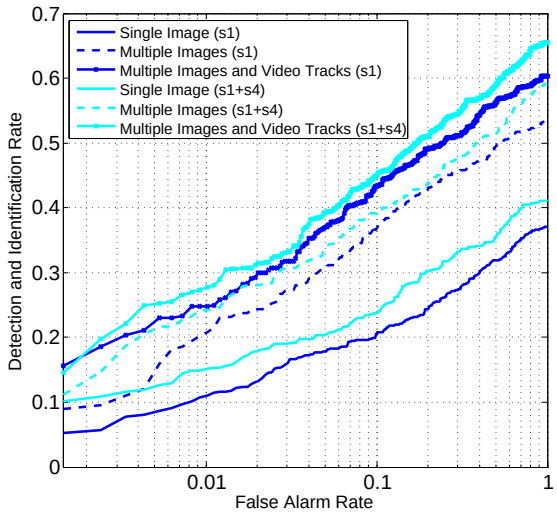
2.6.6 Large Gallery Results

In order to simulate the large-scale nature of operational face identification, we extend the size of our gallery by including *one million* face images from the PCSO database. We

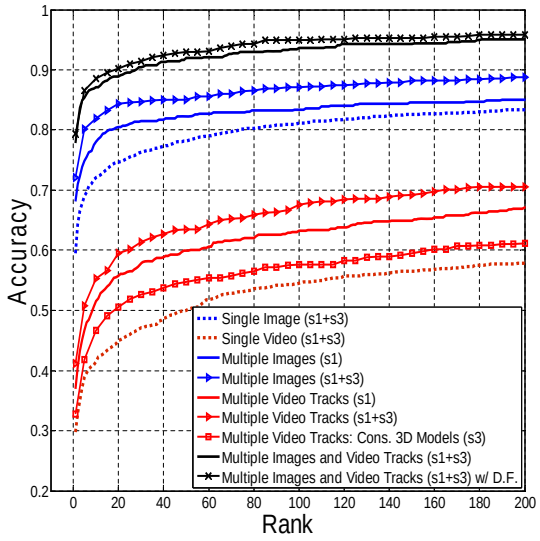
[12] http://usnews.nbcnews.com/_news/2013/05/06/18086503-funeral-director-in-boston-bombing-case-used-to-serving-the-unwanted?lite



(a) Probe: Single Image



(b) Probe: Single Video Track



(c) Large Gallery

Figure 2.17 Scenarios of open-set and closed-set identifications. Open-set identification with (a) a single face image as the probe and various media collections as the gallery and (b) a single face video track as the probe and various media collections as the gallery; the legend denotes the gallery media collection in (a) and (b). Closed-set identification of (c) various media collections as probe against a large gallery set with one million background face images from the PCSO database; the legend denotes the probe media collection; the black curve denoted with "D.F." indicates that demographic information (gender and race) is also fused with the other face media. Note that the ordinate scales are different in (a) and (b) to accentuate the difference among the plots.

acknowledge that there may be a bias towards matching between LFW probe and LFW gallery images versus matching LFW probe with PCSO gallery images. This bias is likely due to the fact that the gallery face images in LFW are not necessarily frontal with controlled illumination, expression, etc., while the background face images from PCSO are mugshots of generally cooperative subjects. The extended gallery set with $1M$ face images makes the face identification problem more challenging. Figure 2.17(c) gives the media collection based face identification accuracies with $1M$ background face images. A comparison between Fig. 2.17 (c) and Fig. 2.9 shows that the proposed face media collection based matching generalizes well to a large gallery set.

2.7 Conclusions

We studied face identification of persons of interest in unconstrained imaging scenarios with uncooperative subjects. Given a face media collection of a person of interest (*i.e.,* face images and video clips, 3D face models built from image(s) or video frame(s), face sketch, and demographic information), we have demonstrated an incremental improvement in the identification accuracy of a COTS face matching system. We believe this is of great value to forensic investigations and "lights out" watch list operations, as matching the entire probe collection outputs a *single* ranked list of candidate identities, rather than a ranked list for each face media sample. Evaluations are provided in the scenarios of closed-set identification, open-set identification, closed-set identification with a large gallery, and verification. Our contributions can be summarized as follows:

1. A collection of face media, such as image, video, 3D face model, face sketch, and demographic information, on a person of interest improves identification accuracies, on average, particularly when individual face samples are of low quality.

2. Pose correction of unconstrained 2D face images and video frames (via 3D face modeling) prior to matching improves the accuracy of a state of the art COTS face matcher.

This improvement is especially significant when match scores from rendered pose corrected images are fused with match scores from original face imagery.

3. A single consolidated 3D face model summarizes the entire video track to a single representation, but score-level fusion of the multiple pose corrected frames from the video track performs better than the consolidated model.

4. Quality based fusion of match scores of different media types performs better than fusion without incorporating the quality.

5. The value of forensic sketch drawn based on low quality videos or low quality images of the suspect is demonstrated in the context of one of the Boston bombing suspects and YTF video tracks.

While the LFW and YTF databases contain variations in pose, illumination, expression, occlusion, resolution, etc., matching a face media collection may not boost the performance if there are long elapsed times between the probe face samples and the true mate in the gallery. Figure 2.18 shows an example of two age-separated face images of the same subject in the LFW database. This type of scenario is difficult to analyze because the LFW and YTF databases do not contain age information for the images.

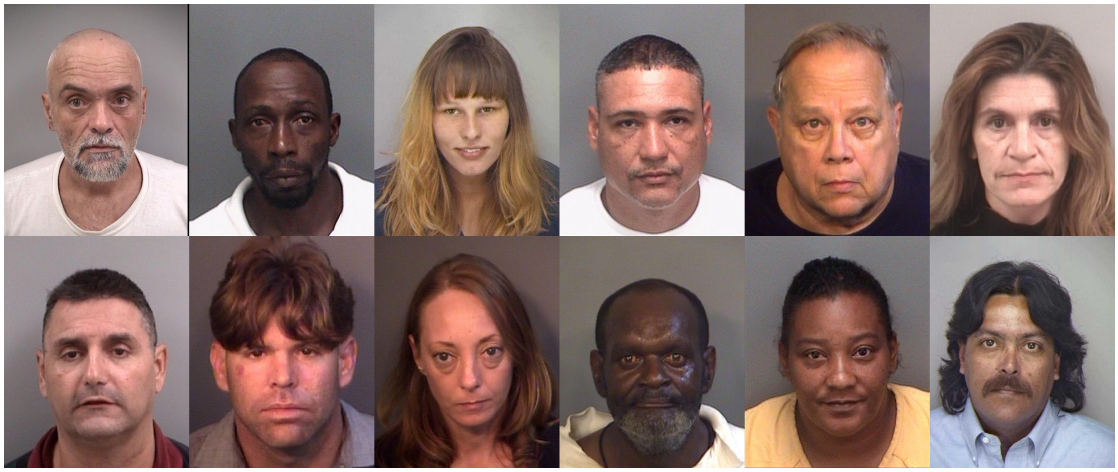
Figure 2.18 An example of two face images of the same subject in the LFW database where facial aging has occurred.

Chapter 3

Automatic Face Image Quality

The performance of automatic face recognition systems largely depends on the quality of the face images acquired for comparison. Under controlled image acquisition conditions (*e.g.,* mugshot photos) with uniform lighting, frontal pose, neutral expressions, and standard image resolution, face recognition systems can achieve extremely high accuracies (*e.g.,* >99% TAR at 0.1% FAR [57]). The system errors still present here are often caused by a relatively small portion"poor" quality face images. This could be due to uncooperative subjects or operator negligence during the acquisition of a mugshot, for example (see Fig. 3.1). There are many emerging applications of face recognition which seek to operate on face images captured in less than ideal conditions (*e.g.,* surveillance). In such cases where large intra-subject facial variations are more prevalent, or even the norm, the accuracy of face recognition degrades. The 2014 large-scale evaluation conducted by NIST demonstrated that mugshot-to-mugshot recognition error rates more than doubled for the top six commercial algorithms when comparing a mugshot gallery to lower quality webcam face images [55].

The performance of biometric recognition, in general, is driven by the quality of biometric samples (*e.g.,* fingerprint, iris, and face images) [5,24,56]. Biometric sample quality is defined as *a measure of a sample's utility to automatic matching* [5, 24, 56]. A desirable property of a biometric quality measurement is that it should be indicative of recognition performance



(a)

(b)

Figure 3.1 Examples of (a) high and (b) low quality mugshots from the PCSO database.

and be correlated with error rates such as false non-match rate (FNMR), false match rate (FMR), or identification miss rates. If a system can automatically determine the quality of a biometric sample defined in this way, it can be useful for several practical applications.

- Negative identification systems - *e.g.,* automated security checkpoints at airports to compare passengers against watch list photos. If passengers are purposely trying to evade detection, automatic face quality assessment can flag their attempt and/or deny entry through the checkpoint.

- Quality-based fusion: multiple face images (*e.g.,* sequence of video frames), multi-biometric fusion [109] (*e.g.,* face and fingerprint), or 3D face modeling from collection of face images.

- Dynamic assignment of comparisons to different matching algorithms. High quality face images can be assigned to high-throughput algorithms, while low quality face images could be assigned to slower, but more robust, algorithms.

A biometric quality measure able to detect "bad" quality samples can subsequently process them accordingly (*e.g.,* reject poor quality samples, request a better sample from the user, employ a slower but more robust matching algorithm, etc.). Additionally, a quality measure can be used to rank a collection of biometric samples which is particularly useful when multiple samples of a subject are available (*e.g.,* frames from a video track, see Fig. 3.2).

Because a biometric sample's quality is specific to automatic recognition performance, human visual perception of the sample's quality may not be well correlated with recognition performance [24, 56]. Particularly, given a fingerprint or iris image, it is difficult for a human to assess the quality in the context of recognition because humans (excluding forensic experts) do not naturally use fingerprints or iris textures for person recognition. However, the human visual system is extremely advanced when it comes to recognizing the faces of individuals, a routine daily task. In fact, it was recently shown that humans surpass the performance of current state-of-the-art automated systems on recognition of very challenging,

(a)

(b)

Figure 3.2 (a) Video frames from a sample video in the IJB-A [72] unconstrained face database and (b) corresponding cropped faces sorted from high to low quality by the proposed approach.

low quality, face images [25]. To the best of our knowledge, very few studies have actually investigated face image quality assessment by humans. Adler and Dembinsky [2] found very low correlation between human and algorithm measurements of face image quality (98 mugshots of 29 subjects, 8 human evaluators), while Hsu *et al.* [60] found some consistency between human perception and recognition-based measures of face image quality (frontal and controlled illumination face images, 2 human evaluators).

Face recognition performance is highly sensitive to factors such as pose, illumination, expression, occlusion, resolution, and other intrinsic or extrinsic properties of face images. The primary goal of face recognition research is to develop systems which are more robust to these factors. Recent works on automatic face recognition have devoted efforts towards recognition of *unconstrained* facial imagery [136] where facial variations of any kind can be simultaneously present (*e.g.,* face images from surveillance cameras). While much prior work has been conducted in face image quality, it has primarily focused on the quality of lab-collected face image databases where facial variations such as pose and illumination are synthetic/staged/simulated in order to isolate and facilitate evaluation of the different factors. In this work, we focus on automatic face image quality of unconstrained face images using the Labeled Faces in the Wild (LFW) [62] and IARPA Janus Benchmark A (IJB-A) [72] unconstrained face datasets. The contributions of this work are summarized as follows:

- Collection of human ratings of face image quality for a large database of unconstrained face images (namely, LFW [62]) by crowdsourcing a small set of pairwise comparisons of face images and inferring the complete ratings with matrix completion.

- Investigation of the utility of face image quality assessment by humans in the context of automatic face recognition performance. This is the first study on human quality assessment of face images that exhibit a wide range of quality factors (*i.e., unconstrained* face images).

- Comparison of two methods for "ground truth" labeling the quality of face images in

a database: (i) human quality ratings and (ii) quality labels computed from similarity scores from COTS matchers. The latter serves as an "oracle" for a face quality measure that is correlated with recognition performance.

- Automatic prediction of the face image quality of an unseen image using image features from a deep neural network.

Our experimental evaluation follows the methodology advocated by Grother and Tabassi [56] where a biometric quality measurement is tested by "relating quality values to empirical matching results." Our evaluation focuses on two primary uses of the proposed face image quality measure: (i) for ranking a collection of face images, and (ii) to reject low quality face images to improve error rates (*e.g.,* FNMR) of automatic face recognition systems.

3.1 Related Work

A number of studies (*e.g.,* [1, 20, 21]) have offered in depth analyses of the performance of automatic face recognition systems with respect to different *covariates*. These studies have identified key areas of research and have guided the community to develop algorithms that are more robust to the multitude of variations in face images. The covariates studied include *image-based*, such as pose, illumination, expression, resolution, and focus, as well as *subject-based*, such as gender, race, age, and facial accessories (*e.g.,* eyeglasses). In general, it is typically shown that face recognition performance degrades due to these different sources of variability. Intuitively, the magnitude of degradation is algorithm-specific.

Prior works have proposed face image quality as some measure of the similarity to reference face images (typically frontal pose, uniform illumination, neutral expression). For example, [116] uses luminance distortion from a high quality reference image for adaptive fusion of two face representations. Wong *et al.* [142] propose probabilistic similarity to a reference model of "ideal" face images for selecting high quality frames in video-to-video verification, and Best-Rowden *et al.* [17] investigated structural similarity (SSIM) for quality-

Table 3.1 Summary of Related Work on Automatic Methods for Face Image Quality

Study (year)	**Database: Num. imgs. (subj.)**	**Target Quality Value**	**Learning Approach**	**Evaluation**
Hsu *et al.* [60] (2006)	FRGC: 1,886 (n/a) passports: 2,000 (n/a) mugshots: 1,996 (n/a)	Continuous (genuine score)	Neural network to combine 27 quality measures (exposure, focus, pose, illumination, etc.) for prediction of genuine scores	ROC curves for different levels of quality (FaceIt algorithm by Identix)
Aggarwal *et al.* [3] (2011)	Multi-PIE: 6,740 (337)* FacePix: 1,830 (30)	Continuous (genuine score) or Binary (algorithm success vs. failure, requires matching prior to quality)	MDS to learn a mapping from illumination features to genuine scores. Predicted genuine score compared to algorithm score to predict algorithm success or failure	Prediction accuracy of algorithm success vs. failure, ROC curves for predicted, actual, 95% and 99% retained (SIFT-based and PittPatt algorithms)
Phillips *et al.* [104] (2013)	PaSC: 4,688 (n/a) GU†: 4,340 (437)	Binary (low vs. high)	PCA + LDA classifier	Error vs. Reject curve for FNMR vs. percent of images removed
Bharadwaj *et al.* [22] (2013)	CAS-PEAL: n/a (1,040) SCFace: n/a (130)	Quality bins (poor, fair, good, excellent)	SVM on GIST and HOG features	ROC curves, rank-1 accuracy, EER, % histogram overlap (COTS algorithm)
Abaza *et al.* [1] (2014)	GU†: 4,340 (437)	Binary (good vs. ugly)	Neural network (1-layer) to combine contrast, brightness, sharpness, focus, and illumination measures	Rank-1 identification for blind vs. quality-selective fusion
Dutta *et al.* [42] (2014)	Multi-PIE: 3,370 (337)‡	Continuous (false reject rate)	Probability density functions (PDFs) model interaction between image quality (deviations from frontal and uniform lighting) and recognition performance	Predicted vs. actual verification performance for different clusters of quality (FaceVACS algorithm)
Kim *et al.* [68]	FRGC: 10,448 (322)	Binary (low vs. high) or Continuous (confidence of the binary classifier)	Objective (pose, blurriness, brightness) and Relative (color mismatch between train and test images) face image quality measures as features fed into AdaBoost binary classifier	Identification rate w.r.t. fraction of images removed, ROC curve with and without low quality images (SRC face recognition algorithm)
Chen *et al.* [35] (2015)	SCFace: 2,080 (130) (trained with FERET, FRGC, LFW, and non-face images)	0 – 100 (rank-based quality score)	A ranking function is learned by assuming images from different databases are of different quality and images from same database are of equal quality	Visual quality-based rankings, Identification rate
Proposed Approach	LFW: 13,233 (5,749) IJB-A: 5,399 (500)	Continuous (human quality ratings or normalized comparison scores)	Support vector regression with image features from a deep convolutional neural network [136]	Error vs. Reject curves, visual quality-based ranking

Note: n/a *indicates that the authors did not report the number (an unknown subset of the database may have been used)*

*Only the illumination subset of Multi-PIE

†GU denotes the Good and Ugly partitions of the Good, Bad, and Ugly (GBU) face database

‡Only neutral expressions from Multi-PIE

based fusion within a collection of face media. Reference-based approaches are dependent on the face images used as reference and may not generalize well to different databases or face images with multiple quality factors present.

More recently, especially with the influx of unconstrained face images, interest has peaked in automatic measures for face image quality that can encompass multiple quality factors, and hence, determine the degree of suitability for automatic matching of an arbitrary face image. Table 3.1 summarizes related works in automatic face image quality which are learning-based approaches. These methods are related in that they all define some target quality which is related to automatic recognition performance. The target quality value can be a prediction of the genuine score (*e.g.,* [3, 60]), a bin indicating that an image is poor, fair, or good for matching (*e.g.,* [22]), or a binary value of low vs. high quality image (*e.g.,* [1, 68, 104]). For example, Bharadwaj *et al.* fuse similarity scores from two COTS matchers, define quality bins based on CDFs of images that were matched correctly and incorrectly, and use a support vector machine (SVM) trained on holistic image features to classify a test image as poor, fair, good, or excellent quality [22]. Rather than defining target quality values for a training database of face images, Chen *et al.* propose a "learning to rank" framework which assumes a rank-ordering of a set of databases (*e.g.,* non-face images $<$ unconstrained face images $<$ ID card face images) where face images from the same database have equal quality; rank weights from multiple types of features are learned and then mapped to a quality score 0$\sim$100 [35].

In our approach, we annotate a large database of unconstrained face images with target quality values (defined as either human quality ratings or score-based values from a COTS matcher), extract image features using a deep convNet [136], and learn a model for prediction of face quality from the deep convNet features using support vector regression. The target quality values in this work are continuous and allow for a fine-tuned quality-based ranking of a collection of face images.

Table 3.2 Performance of Face Recognition Algorithms on the BLUFR Protocol [85]

Algorithm	TAR @ 0.1% FAR	DIR @ 1% FAR
HDLBP + JointBayes [34]*	41.66	18.07
Yi *et al.* [146]	80.26	28.90
DCNN *et al.* [136]	89.80	55.90
COTS-A	88.14	76.28
COTS-B	76.01	53.21

*Performance here for [34] was reported by [85]

3.2 Face Image Databases and COTS Matchers

In this work, we utilize two unconstrained face databases: Labeled Faces in the Wild (LFW) [62] and IARPA Janus Benchmark A (IJB-A) [72]. Both LFW and IJB-A contain face images with unconstrained facial variations that affect the performance of face recognition systems (*e.g.,* pose, expression, illumination, occlusion, resolution, etc.). The LFW database consists of 13,233 images of 5,749 subjects, while the IJB-A database consists of 5,712 images and 2,085 videos of 500 subjects. Face images in the LFW database were detected by the Viola-Jones face detector [135] so the pose variations are limited by the pose tolerance of the Viola-Jones detector. Face images in IJB-A were manually located, so the database is considered more challenging than LFW due to full pose variations [72]. See Fig. 3.3 for sample face images from the two databases.

Because face image quality needs to be evaluated in the context of automatic face recognition performance, we make use of two commercial face matchers, denoted as COTS-A and COTS-B. Table 3.2 shows that COTS-A and COTS-B are competitive algorithms on the BLUFR protocol [85] for the LFW database. Performance is also reported for the deep learning-based matcher proposed by Wang *et al.* [136] as DCNN. The feature representation from [136] is used in this work to predict face image quality.

(a) LFW

(b) IJB-A

Figure 3.3 Sample face images from the (a) LFW [62] and (b) IJB-A [72] unconstrained face databases.

3.3 Face Image Quality Labels

Biometrics and computer vision heavily rely on supervised learning techniques when training sets of *labeled* data are available. When the aim is to develop an automatic method for face image quality, compiling a quality-labeled face image database is not straightforward. The definition of face image quality (*i.e.,* a predictor of automatic matching performance) does not lend itself to explicit labels of face image quality, unlike labels of facial identity or face vs. non-face labels for face recognition and detection methods, respectively. Possible approaches for generating quality labels of face images include:

1. Combine various measurements of image quality factors into a single value which indicates the overall face quality.

2. Human annotations of perceived image quality.

3. Based on comparison scores (or performance measures) from automatic face recognition matchers.

The issues with 1) are that it is an *"ad-hoc"*/heuristic approach and, thus far, has not achieved much success (*e.g.,* [104]). The issue with 2) is that human perception of quality may not be indicative of automatic recognition performance; previous works [22, 56] have stated this consensus but, to our knowledge, the only studies to investigate these statements were conducted on constrained face images (*e.g.,* mugshots) [2,60]. The issue with 3) is that comparison scores are obtained from a *pair* of images, so labeling single images based on comparison scores (or performance) can be problematic. However, this approach achieved some success for fingerprint [56, 125], and only few studies [22, 104] have considered it for face quality. In this work, we investigate both methods 2) and 3), detailed in the remainder of this section.

3.3.1 Human Ratings of Face Image Quality

Because of the inherent ambiguity in the definition of face image quality, framing an appropriate prompt to request a human to label the quality of a face image is challenging. If asked to rate a face image on a scale of 1 to 5, for example, there are no notions as to the meaning of the different levels. Additionally, some prior exposure to the variability in the face images that the human will encounter may be necessary so that they know what kinds of "quality" to expect in face images (*i.e.,* a baseline) before beginning the quality rating task.

In this work, we choose to only collect quality labels for relative pairwise comparisons of face images by asking the following question: "Which face (left or right) has better quality?" Crowdsourcing literature [148] has demonstrated that ordinal (comparison-based) tasks are generally easier and take less time than cardinal (score-based) tasks. Ordinal tasks additionally avoid calibration efforts needed for cardinal responses from raters inherently using different ranges for decision making (*i.e.,* biased ratings, inflated vs. conservative ratings, meaning of absolute ratings changes with exposure to more data).

To obtain absolute quality ratings for individual face images, we make use of a matrix completion approach [148] to infer the quality rating matrix from the pairwise comparisons. Because it is infeasible to have multiple persons manually assess and label the qualities of *all* face images in a large database, this approach is desirable in that it only requires a small set of quality labels from each human rater in order to infer the quality ratings for the entire database. The details of data collection and the matrix completion approach are discussed in the remainder of this section.

3.3.1.1 Crowdsourcing Comparisons of Face Quality

Amazon Mechanical Turk (MTurk)[1] was utilized to facilitate collection of pairwise comparisons of face image quality from multiple human raters (*i.e.,* MTurk "workers"). Given a pair of face images, displayed side by side, our Human Intelligence Task (HIT) was to select a

[1]https://www.mturk.com

response to the prompt "Indicate which face has better quality" out of the following options: (i) left face is much better, (ii) left face is slightly better, (iii) both faces are similar, (iv) right face is slightly better, and (v) right face is much better. Fig. 3.4 shows the interface used to collect the responses.[2]

Our HIT requested each worker to provide responses to a total of 1,001 face image pairs, made up of 6 tutorial pairs, 974 random pairs, and 21 consistency check pairs. The tutorial pairs were pre-selected from the LFW database where the quality of one image was clearly better than the quality of the other (Fig. 3.5 shows the sets of images used). Because these pairs had "correct" responses, they allowed us to ensure that the worker had completed the tutorial introduction and understood the goal of the task. The next 974 pairs of images were chosen randomly from the LFW database, while the final 21 pairs were selected from the set of 974 as repeats to test the consistency of the worker's responses. MTurk workers who attempted our HIT were only allowed to complete it if they passed the tutorial pairs, and we only accepted the submitted responses from workers who were consistent on at least 10 out of the 21 consistency check pairs.

In order to be eligible to attempt our HIT for assessment of face image quality, MTurk workers had to have previously completed at least 10,000 HITs from other MTurk "requesters" with an approval rate of at least 99%. These stringent qualifications helped to ensure that only experienced and reliable workers (in terms of MTurk standards) participated in our data collection.[3] A total of 435 MTurk workers began our HIT. After removing 245 workers who did not complete the full set of 1,001 pairwise comparisons and 4 workers who failed the consistency check (inconsistent response for 10 or more of the 21 repeated pairs), a total of 194 workers were each compensated US $5.00 through the MTurk crowdsourcing service.

[2]The tool is available at http://cse.msu.edu/~bestrow1/FaceOFF/.

[3]The MTurk worker qualifications are managed by the MTurk website.

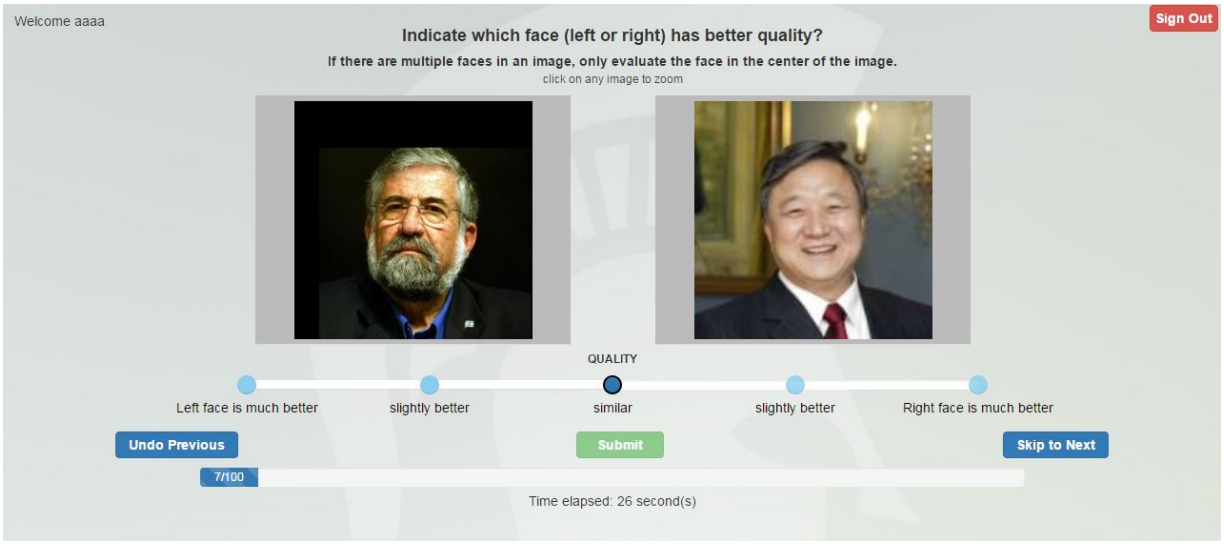


Figure 3.4 The interface used to collect responses for pairwise comparisons of face image quality from MTurk workers.

Figure 3.5 Face images (from the LFW database) used for the 6 tutorial pairs used to check whether MTurk workers understood the task before completing the pairwise comparisons used in our study of face image quality. For each of the tutorial pairs, one image was selected from the top row (high quality images) and one image was selected from the bottom row (low quality images), so the pairwise comparison of face quality had an unambiguous answer.

3.3.1.2 Matrix Completion

After collecting random sets of pairwise comparisons of face image quality from 194 workers via MTurk, we use the matrix completion approach proposed by Yi *et al.* [148] to infer a complete set of quality ratings for each worker on the entire LFW database (13,233 total face images). The aim is to infer $\hat{F} \in \mathbb{R}^{m \times n}$, the worker-rating matrix for face image qualities, where n is the number of workers and m is the number of face images.

Yi *et al.* [148] show that only $O(r \log m)$ pairwise queries are needed to infer the full ranking list of a worker for all m items (face images), where r is the rank of the unknown rating matrix ($r \ll m$). The maximum possible rank of the unknown rating matrix is $r = n = 194$ workers, $O(194 \log 13,233) \approx 800$; hence, the 974 random pairs per worker collected in our study are sufficient to do the matrix completion, especially since we expect $r < n$ (*i.e.,* the quality ratings from the n workers are not all independent).

While relative pairwise comparisons are often preferred in crowd-based tasks [148] because they avoid the biases from raters' tendencies to give conservative or inflated responses when using an absolute scale (*e.g.,* quality levels 1 to 5), we still observed a bias after the matrix completion where the bias is from a tendency to respond "Similar". Fig. 3.6 shows an inverse relationship between the number of pairs that a worker marked "Similar" and the resulting range of quality ratings for that worker (after matrix completion). Note that this bias is not due to the coarse levels of left image is "much better" vs. "slightly better" because prior to matrix completion we combine these responses to simply "left is better". Because of this observation, *min-max* normalization was performed on each worker's quality ratings to transform them to the same range (0 to 1).

After matrix completion, there are face image quality ratings from 194 different workers for each face image in the LFW database. With the aim of obtaining a single quality rating per face image in the LFW database, we simply take the *median* value from all 194 workers to reduce the $194 \times 13,233$ matrix of quality ratings to a $1 \times 13,233$ vector of quality ratings (one per image in LFW). We empirically tested other heuristics (mean, min, max) but found

that median seemed to result in the best quality ratings.

3.3.2 Recognition-based Face Image Quality Labels

Target quality labels acquired from similarity scores serve as an "oracle" for a quality measure that is highly correlated with automatic recognition performance. For example, if the goal is to detect and remove low-quality face images to improve the FNMR, then face images could be removed from a database in the order of their genuine comparison scores. Previous works on biometric quality (fingerprint [56,125] and face [22]) have defined "ground truth" or "target" quality labels as a measure of the separation between the sample's genuine score and its impostor distribution when compared to a gallery of enrollment samples. A normalized comparison score for the jth query sample of subject i can be defined as,

$$z_{ij} = (s_{ij}^G - \mu_{ij}^I)/\sigma_{ij}^I, \tag{3.1}$$

where s_{ij}^G is the genuine score and μ_{ij}^I and σ_{ij}^I are the mean and standard deviation, respectively, of the impostor scores for the query compared to the gallery. Previous works then bin the normalized comparison scores into quality bins based on the cumulative distribution functions (CDFs) of sets of correctly and incorrectly matched samples [22, 56, 125]. Instead, we propose to directly predict the z_{ij} for a given face image to obtain a continuous measure of face image quality.

Target quality values defined based on comparison scores are confounded by the fact that a comparison score is computed from *two* face images, but we are trying to label the quality of a *single* face image. A simplifying assumption can be made if it can be assumed that the quality of the enrollment samples is at least as good as the quality of the probe samples; because comparison scores are typically governed by the low quality samples [56], the quality value can be assigned to the probe image.

To allow for this simplifying assumption, we manually selected the best quality image

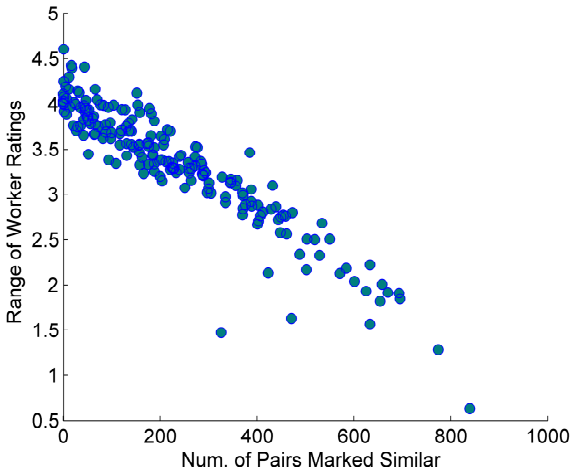


Figure 3.6 The resulting range of the face quality values (after matrix completion) for a particular worker inversely depends on the number of pairs that the worker marked "Similar" quality. Although collection of *relative* responses avoids bias present when workers are asked to rate individual images on an *absolute* scale, bias is still present from tendency to respond "Similar". This indicates that normalization is required to transform the quality ratings from each worker to the same scale.

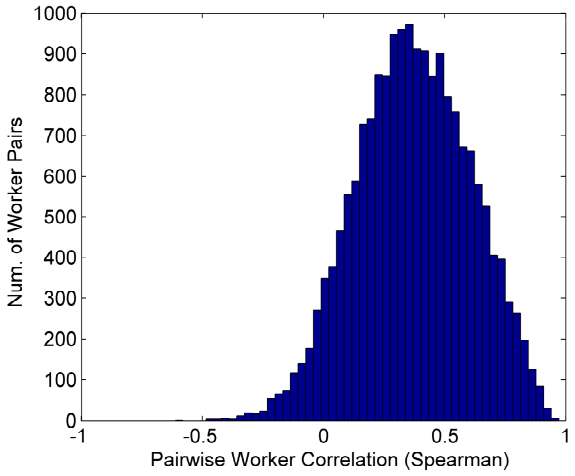


Figure 3.7 Histogram of rank correlations between the face image quality ratings of all pairs of MTurk workers ($\binom{194}{2} = 18,721$ total pairs of workers). The quality ratings are those obtained after matrix completion. The degree of concordance between workers is 0.37, on average.

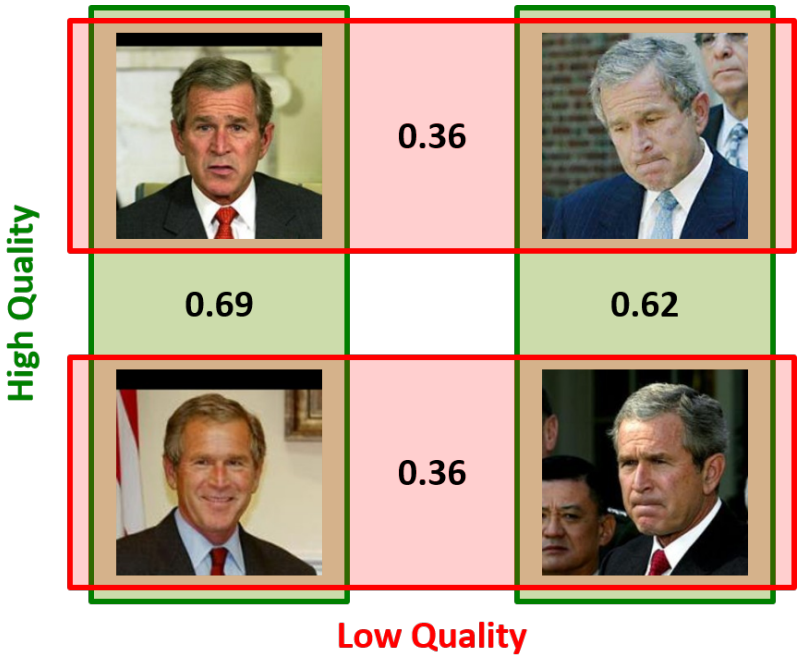


Figure 3.8 Illustration of the pairwise quality issue. Images in the left and right columns are individually of high and low qualities, respectively. However, when compared with the other images, they can produce both high and low similarity scores. (Similarity scores are from COTS-A with range of $[0, 1]$.)

for every subject in the LFW database. There are 1,680 subjects in LFW with at least two face images. The best image selected by us is placed in the gallery (1,680 images, one per subject), while the remaining 7,484 images of these subjects are used as the probe set. The additional 4,069 images in the LFW database (subjects with only a single image) are used to extend the size of the gallery. Normalized comparison scores are computed using Eqn. (3.1) for the 7,484 probe images for each of the face matchers (COTS-A, COTS-B, and DCNN) and are used as score-based target face quality values.

3.4 Automatic Prediction of Face Quality

Given that we have obtained face image quality labels for the LFW database, we now wish to train a model to automatically predict the quality of an unseen face image. Ideally, we would compile a set of automatically extracted image features that are measurements of

known quality factors that affect face recognition performance, such as pose, illumination, expression, occlusion, contrast, focus, etc. Rather than trying to handcraft a set of image features for our task of predicting face image quality, we make use of features extracted from a deep convolutional neural network which was trained for recognition purposes by Wang *et al.* [136]. The features are 320-dimensional, so we refer to them as *Deep-320* features. The deep network in [136] was trained on the CASIA WebFaces database [147]. We additionally consider a 5-dimensional feature set, referred to as *Vishnu-5*, which includes a face alignment score, number of occluded landmarks (out of 68 total), and measures of facial pose (*i.e.,* yaw, pitch, and roll). Using either the Deep-320 or Vishnu-5 image features, we then train a support vector regression (SVR) [31] model with radial basis kernel function to predict either the normalized comparison scores (z_{ij}) from a commercial matcher or the human quality ratings. The parameters for SVR are determined via grid search on a validation set of face images.

3.5 Experimental Evaluation

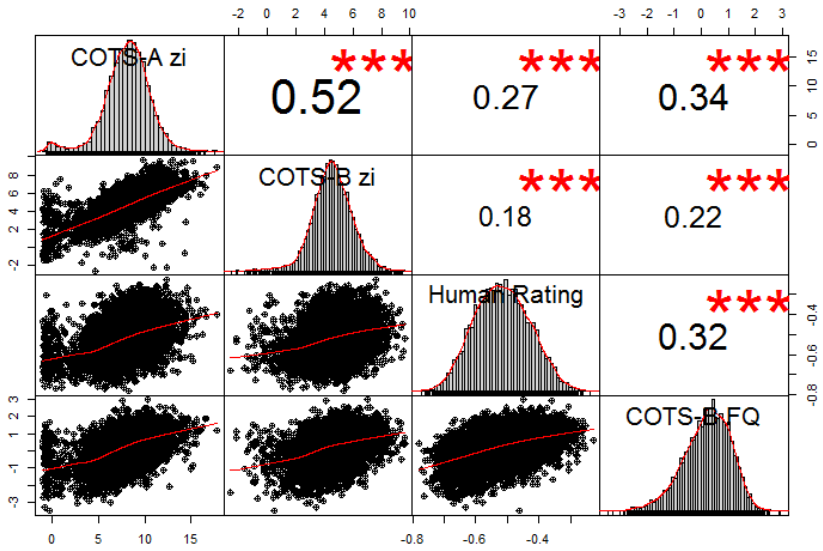
The aim of this work is twofold:

1. Label the target, or "ground truth", quality values of a face image database.

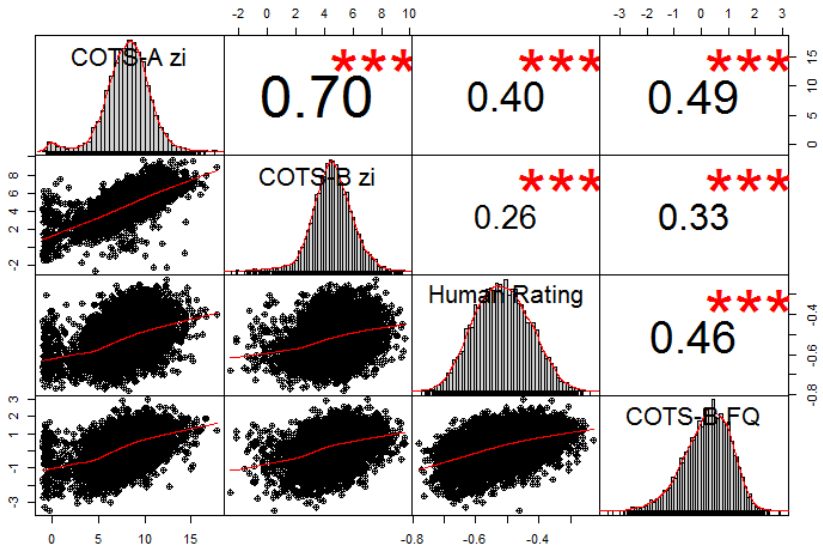
2. Train a model to automatically predict the target quality values using features automatically extracted from an unseen test face image (prior to matching).

Hence, in Sec. 3.5.1, we first evaluate the *target* quality values to determine their utility for automatic recognition. In Sec. 3.5.2 we then evaluate how well the target quality values can be predicted by the proposed model for automatic face image quality. Following the methodology advocated by Grother and Tabassi [56], we evaluate the face quality measures using the following performance metrics.

- *Error versus Reject (EvR) curve* evaluates how efficiently rejection of low quality samples results in decreased error rates. The EvR curve plots an error rate (FNMR or



(a) Kendall's tau



(b) Spearman

Figure 3.9 Rank correlations between the different target face quality values considered in this work. COTS-B FQ is a face quality measure output by COTS-B (black-box method to us, included for comparison). Three red asterisks indicate that the correlations are statistically significant at $\alpha = 0.001$. The score-based measures of face quality (z_{ij}) from COTS-A and COTS-B have the strongest correlation, while the human quality ratings have the weakest correlation with the other quality measures.

FMR) versus the fraction of images removed/rejected, where the error rates are re-computed using a fixed threshold (*e.g.,* overall FMR = 0.01%) after a fraction of the images have been removed.

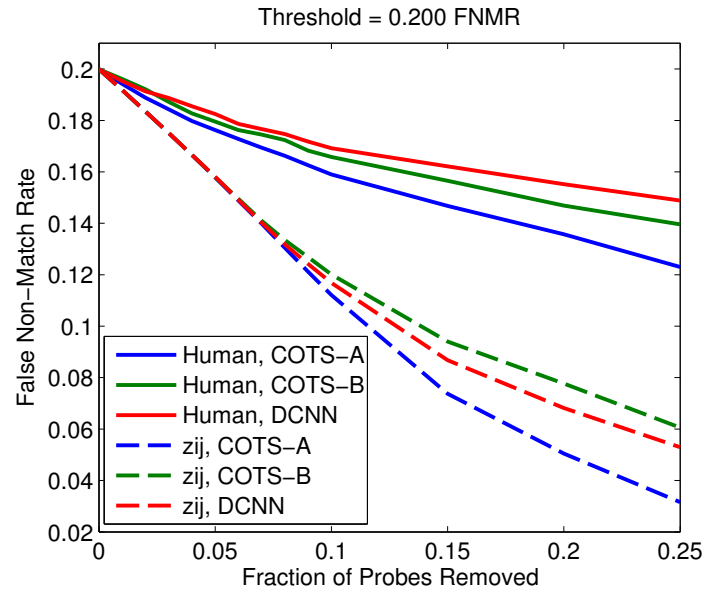
We additionally provide visual inspections of face images rank-ordered by the proposed face image quality.

3.5.1 Target Face Image Quality Values

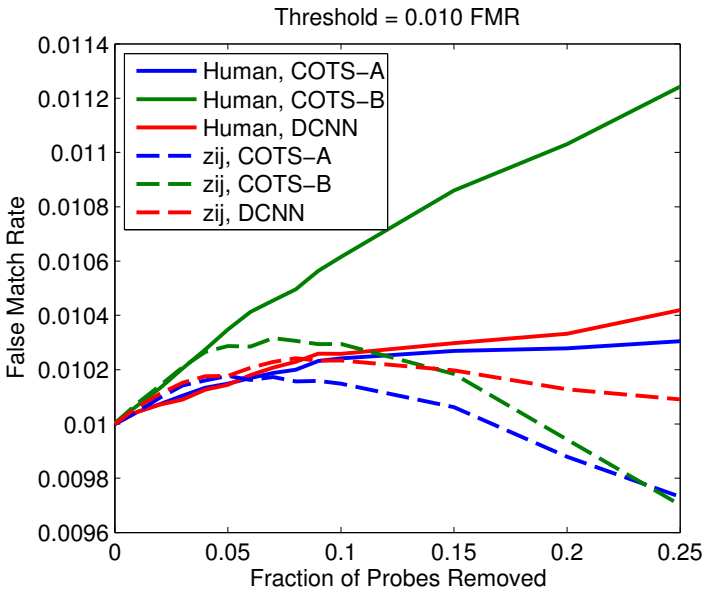
First, the face images in the LFW database are "ground truth" labeled with the methods discussed in Section 3.3. We refer to these quality values as *target* quality values and the ones predicted by our model as *predicted*. Fig. 3.9 shows the distributions of the target labels for COTS-A z_{ij}, COTS-B z_{ij}, and the human ratings after matrix completion, as well as a measure of quality output by the COTS-B matcher (for comparison). Fig. 3.9 also shows that the rank correlation is fairly low between the human ratings of quality and the score-based quality values, while the score-based quality values from the two matchers are highly correlated.

We evaluate the target quality values using the same gallery/probe setup of the LFW database that was used to compute the normalized comparison scores (z_{ij}). This allows for comparison of the human quality ratings and the score-based quality values. Fig. 3.10 plots EvR curves for both methods, evaluated for three different face matchers (COTS-A, COTS-B, and DCNN [136]). Fig. 3.10(a) shows that removing probe images in order of human quality ratings does decrease FNMR for all three matchers. So, human quality ratings are correlated with recognition performance; however, the score-based quality values are much more efficient in reducing FNMR. This is expected because the score-based target quality values are computed from the same comparison scores used to compute the FNMR. Again, the score-based quality values here somewhat serve as an "oracle" for a desirable quality measure.

The utility of the target quality values in terms of reducing FMR in Fig. 3.10(b) is not



(a) FNMR



(b) FMR

Figure 3.10 Error vs. Reject curves for (a) FNMR and (b) FMR on the LFW database (5,749 gallery and 7,484 probe images). Probe images were rejected in order of *target* (*i.e.*, "ground truth") quality values of human quality ratings or score-based quality values (z_{ij}). Thresholds are fixed at (a) 0.2 FNMR and (b) 0.01 FMR for comparison of the three face matchers (COTS-A, COTS-B, and DCNN [136]).

as apparent; in fact, removing low quality images based on human quality ratings clearly *increases* FMR for COTS-B (though the magnitude of the increase is quite small). The relation between face quality and impostor scores (*i.e.,* FMR) is generally less of a concern. For biometric quality, in general, we desire *high* quality samples to produce *low* impostor similarity scores, but *low* quality samples may also produce *low* (or even lower) impostor scores. If this is the case, low quality face images may be beneficial to FMR for empirical evaluation, but still undesirable operationally.

3.5.2 Predicted Face Image Quality Values

The proposed framework for automatic prediction of face image quality (both human ratings and score-based quality values) is used to predict the quality of face images from the LFW [62] and IJB-A [72] databases. The prediction models for both databases are trained using LFW face images and the following experimental protocols.

3.5.2.1 Train, Validate, and Test on LFW:

We first divide 7,484 face images of the 1,680 subjects with two or more images in LFW into 10 random splits for training and testing data, where the subjects are randomly split into 2/3 and 1/3 for training and testing, respectively. For each split, we then conduct 5-fold cross-validation within the training set to determine the parameters (via grid-search) for the support vector regression model. The selected set of parameters is then applied to the full training set to result in a single model for each of the 10 splits, which are then used to predict the quality labels of the images in each of the 10 test sets. This framework ensures subject-disjoint training and testing sets, and parameter selection is conducted within a validation set, not optimized for the test sets.

Table 3.3 gives the rank correlation (mean and standard deviation over the 10 splits) between the target and predicted quality values for human quality ratings and score-based quality values (for COTS-A and COTS-B). The first observation is that the Deep-320 features

Table 3.3 Rank Correlation, (a) Kendall's tau and (b) Spearman, Between Target and Predicted Quality Labels (Mean ± Standard Deviation Over 10 Random Splits of LFW Images)

(a)

	Face Quality Label		
	COTS-A z_{ij}	COTS-B z_{ij}	Human Rating
Deep-320	0.395 ± 0.018	0.305 ± 0.019	0.412 ± 0.016
Vishnu-5	0.232 ± 0.031	0.202 ± 0.018	0.295 ± 0.018

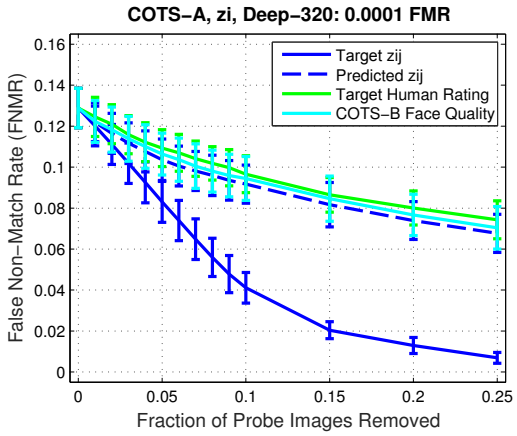
(b)

	Face Quality Label		
	COTS-A z_{ij}	COTS-B z_{ij}	Human Rating
Deep-320	0.558 ± 0.023	0.442 ± 0.026	0.585 ± 0.019
Vishnu-5	0.340 ± 0.042	0.297 ± 0.026	0.431 ± 0.025

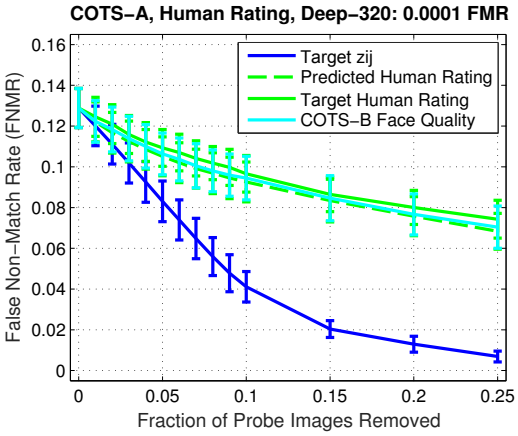
better predict all three quality measures than the Vishnu-5 features. Additionally, prediction of human quality ratings is more accurate than prediction of score-based quality from either COTS-A or COTS-B, likely due to the difficulty in predicting particular nuances of each matcher.

To further investigate the resulting face quality predictions, we computed the Spearman rank correlation between the target and predicted values separately for the multiple images of each subject; *i.e.,* given multiple face images of a subject, we rank them based on the target and the quality values, and compute the correlation between the two ranking lists. Figs. 3.12 and 3.13 show examples of *strong* correlation between target and predicted human quality ratings, while Figs. 3.14 and 3.15 show examples of *weak* or even *negative* correlation. Figs. 3.16 and 3.17 show examples of negative correlation between target and predicted score-based quality for COTS-A. It appears that weak correlation is observed when the multiple images of a subject are of similar quality; it is difficult to achieve a consistent fine-tuned ranking of face images when all of the qualities are similar.

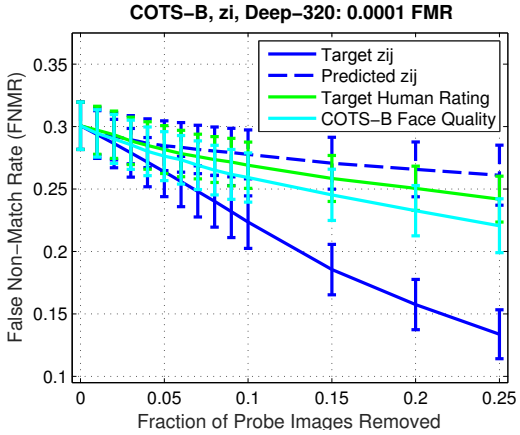
To evaluate the quality values in the context of automatic face recognition performance, error vs. reject curves (for FNMR) are plotted in Fig. 3.11 for both target and predicted quality values. The figures demonstrate that rejecting low quality face images based on predicted z_{ij}, predicted human ratings, or the COTS-B measure of face quality, results in



(a) COTS-A, z_{ij}



(b) COTS-A, Human



(c) COTS-B, z_{ij}

Figure 3.11 Error vs. Reject curves for target and predicted face image quality values. The curves show the efficiency of rejecting low quality face images in reducing FNMR at a fixed FMR of 0.001%. The model used for the face quality predictions in (a)-(c) are support vector regression on the deep-320 features from the deep convNet in [136].

comparable efficiency in reducing FNMR (*e.g.,* removal of 5% of probe images lowers FNMR by ∼2%). However, none of the methods are near as efficient as rejecting images based on the target z_{ij} values, which serve as an oracle for a predicted face quality measure that is highly correlated with the recognition performance.

3.5.2.2 Train and Validate on LFW, Test on IJB-A:

In this framework, we conduct 5-fold cross-validation over the 7,484 LFW images (folds are subject-disjoint) to determine the parameters for the support vector regression model via grid search. We then apply the selected set of parameters to all of the LFW training images. This model trained on LFW face images is then used to predict the quality of face images in the IJB-A database. The Deep-320 image features [136] are used here.

We currently do not have any ground truth quality labels for IJB-A face images because we did not collect human annotations for this database, and we do not have a recognition protocol set up with a higher quality gallery. Initial efforts to construct a high quality gallery (faces with frontal pose, neutral expression, no occlusion, etc.) for IJB-A indicated that this is not possible for all the subjects. Hence, current evaluation entails visual inspection of the rank-ordering of face images based on the predicted quality values. Figs. 3.18–3.20 shows that the proposed automatic face quality measure does a fairly good job at sorting face images (and video frames in Fig. 3.21) in order of face quality. Figs. 3.18–3.20 also show face images sorted by the Rank-based Quality Score (RQS) of Chen *et al.* [35] for comparison. Though it is difficult to compare the two methods without recognition experiments, there are a few cases where the top highest quality faces predicted by our method appear to be better than the RQS ranking (*e.g.,* top row in Fig. 3.20).

Ranked by Target Human Quality Ratings

Ranked by Predicted Human Quality Ratings

Figure 3.12 Face images from a subject in LFW are rank-ordered by target (left) and predicted (right) human quality ratings, in order of increasing face quality. The Spearman correlation between the target and predicted rank orderings for this subject is 0.72.

Ranked by Target Human Quality Ratings **Ranked by Predicted Human Quality Ratings**

Figure 3.13 Face images from LFW are rank-ordered by target (left) and predicted (right) human quality ratings, in order of increasing quality. Examples shown have *positive* rank correlation between target and predicted rankings. For each of the three example subjects, the Spearman correlation between the target and predicted rank orderings are 0.94, 0.90, and 0.50 (top to bottom).

Ranked by Target Human Quality Ratings **Ranked by Predicted Human Quality Ratings**

Figure 3.14 Face images from LFW rank-ordered by target (left) and predicted (right) human quality ratings, in order of increasing quality. Examples shown have *negative* (or zero) rank correlation between target and predicted rankings. For each of the example subjects, the Spearman correlation between the target and predicted rank orderings are -0.50, and 0.00 (top to bottom).

Ranked by Target Human Quality Ratings **Ranked by Predicted Human Quality Ratings**

Figure 3.15 Face images from LFW rank-ordered by target (left) and predicted (right) human quality ratings, in order of increasing quality. Examples shown have strong *negative* rank correlation between target and predicted rankings. For each of the example subjects, the Spearman correlation between the target and predicted rank orderings are -0.90, and -0.70 (top to bottom).

Ranked by Target COTS-A z_{ij} **Ranked by Predicted COTS-A z_{ij}**

Figure 3.16 Face images from LFW rank-ordered by target (left) and predicted (right) score-based quality values (COTS-A z_{ij}), in order of increasing quality. Examples shown have *negative* rank correlation between target and predicted rankings. For each of the example subjects, the Spearman correlation between the target and predicted rank orderings are -0.33 and -0.37 (top to bottom).

Ranked by Target COTS-A z_{ij} **Ranked by Predicted COTS-A z_{ij}**

Figure 3.17 Face images from LFW rank-ordered by target (left) and predicted (right) score-based quality values (COTS-A z_{ij}), in order of increasing quality. Examples shown have *negative* rank correlation between target and predicted rankings. For each of the three example subjects, the Spearman correlation between the target and predicted rank orderings are -1.00, -0.20, and -0.31 (top to bottom).

Ranked by Predicted Human Rating

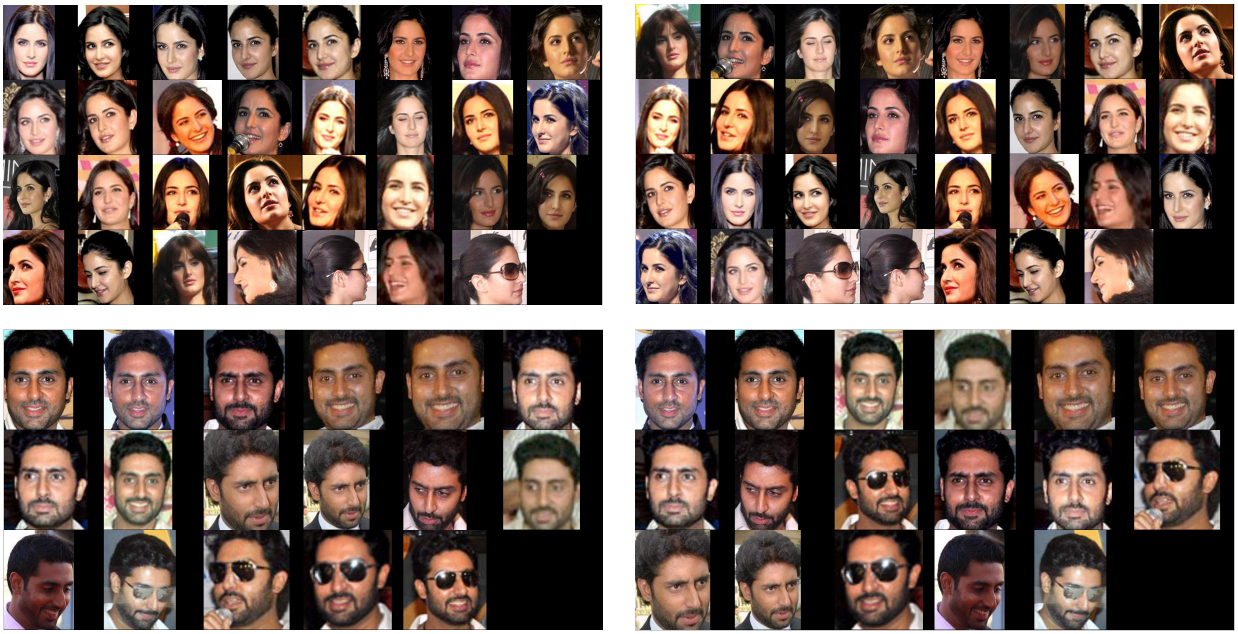
Ranked by RQS [35]

Figure 3.18 Face images from IJB-A [72] sorted by face image quality (best to worst). The face image qualities were automatically predicted by (left) the proposed approach (SVR model on Deep-320 image features [136]) and human quality ratings from the LFW database) and (right) Rank-based Quality Score (RQS) [35] for comparison.

Ranked by Predicted Human Rating

Ranked by RQS [35]

Figure 3.19 Face images from IJB-A [72] sorted by face image quality (best to worst). The face image qualities were automatically predicted by (left) the proposed approach (SVR model on Deep-320 image features [136]) and human quality ratings from the LFW database) and (right) Rank-based Quality Score (RQS) [35] for comparison.



Ranked by Predicted Human Rating **Ranked by RQS [35]**

Figure 3.20 Face images from two subjects in IJB-A [72] sorted by face image quality (best to worst). The face image qualities were automatically predicted by (left) the proposed approach (SVR model on Deep-320 image features [136]) and human quality ratings from the LFW database) and (right) Rank-based Quality Score (RQS) [35] for comparison.

Figure 3.21 Face images from the videos of example subjects in IJB-A [72] sorted by face image quality (best to worst) which was automatically predicted by the proposed approach using a model (SVR on Deep-320 image features [136]) trained on human quality ratings from the LFW database.

3.6 Conclusion

Automatic face image quality assessment is a challenging problem with important operational applications. Automatic detection of low quality face images would be beneficial in maintaining the integrity of enrollment databases, reacquisition prompts, quality-based fusion, and adaptive recognition approaches. In this work, we have investigated two methods for assigning target face image quality values to a large database of face images to be used for training, and proposed a model for automatic prediction of face image quality using only image features extracted prior to matching. The conclusions and contributions can be summarized as follows:

- Human ratings of face image quality (obtained from crowdsourcing and matrix completion) are correlated with automatic recognition performance for unconstrained face images. Rejection of 5% of the lowest quality face images (based on human quality ratings) in the LFW database resulted in $\sim 2\%$ reduction in FNMR.

- Human quality ratings are not as correlated with recognition performance as are target face quality values obtained from similarity scores (matcher-specific). This was as expected since score-based quality serves as an oracle for an ideal quality measure (performance is directly computed from the same similarity scores), whereas human quality ratings are solely based on single images.

- Automatic prediction of human quality ratings is more accurate than prediction of score-based face quality values. It is difficult to predict the score-based quality because of nuances of specific matchers and pairwise quality factors (*i.e.,* comparison scores are a function of *two* face images, but we are using the scores to label the quality of a *single* face image).

- Visual inspection of face images rank-ordered by the proposed automatic face quality measures (both human ratings and score-based quality) are promising, even for cross-

database prediction (*i.e.,* model trained on LFW [62] and tested on IJB-A [72] face images).

Chapter 4

Longitudinal Study of Automatic Face Recognition

4.1 Introduction

Technological advancements in automatic face recognition have progressively tackled challenges caused by variations in facial pose, illumination, and expression (collectively called PIE variations). Current efforts (*e.g.,* [128,136]) are breaking ground on robustness to "faces in the wild" (*e.g.,* images posted on the web) to account for PIE, occlusion, and partial face images. Comparatively, aging variations (*i.e.,* large time lapse between pairs of images being compared) have received considerably less attention in the face recognition community.

Published studies on facial aging in the context of automatic face recognition have primarily employed *cross-sectional* techniques where a population of individuals who differ in age are analyzed according to differences between age groups [15, 55, 70, 87, 99]. However, cross-sectional analysis cannot adequately explore age-related effects because assumptions of independent observations require that there be only one measurement per individual in the study (see Fig. 4.6. Past and future measurements are either not considered or are

(a) Ages 30.5 and 39.6 (0.423)

(b) Ages 32.2 and 40.3 (0.433)

(c) Ages 29.5 and 38.3 (0.498)

(d) Ages 39.2 and 48.6 (0.500)

Figure 4.1 Face image pairs of four subjects from the PCSO_LS mugshot database which are age-separated by eight to ten years. Similarity scores from a state-of-the-art face matcher (COTS-A) are shown in parentheses (score range is [0.0, 1.0]). The thresholds at 0.01% and 0.1% FAR are 0.533 and 0.454, respectively. Hence, all of these genuine pairs would be falsely rejected at 0.01% FAR, while the two female subjects, (a) and (b), would also be rejected at 0.1% FAR.

summarized into a single measurement which loses information; trends of individuals over time are not analyzed. Hypotheses about facial aging are, instead, *longitudinal* by nature and require multiple measurements of the same individuals over time to reveal trends in comparison scores with respect to facial aging.

To what extent facial aging affects the performance of automatic face recognition systems is of more than academic concern. Because the appearance of the face changes throughout a person's life, most identity documents containing face images expire after a designated period of time; U.S. passports are only valid for five years for minors and ten years for adults, while U.S. driver's licenses typically require renewal every five years. Additionally, to our knowledge, ensuring that a new (more recent) photo has been submitted for renewal is not verified, especially for renewals by mail or online. Validity periods of such identity documents may be too long if these photos are to be used with state-of-the-art face matching

systems. Fig. 4.1 shows that elapsed times of eight to ten years between two face images can cause false non-match errors. Studying how the actual comparison scores change over time is important for understanding the implications of operating with a global threshold[1] (*e.g.,* de-duplication and other open-set scenarios) on face recognition accuracy.

While longitudinal studies for automatic iris recognition [54] and fingerprint recognition [149] have been published, to our knowledge, no large-scale longitudinal study of automatic face recognition performance has been reported in the literature. We aim to fill this gap by addressing the following question: *How robust are state-of-the-art automatic face recognition systems to facial aging?* In this chapter, we conduct a longitudinal analysis of the performance of state-of-the-art COTS face matchers on two longitudinal face image databases consisting of repeat criminal offenders (mugshots) from two different law enforcement agencies (see Table 4.2). The COTS matchers used here are among the top-ranked performers in the FRVT 2013 face recognition evaluation [55]. The contributions of this chapter can be summarized as follows:

1. *Longitudinal analysis of two of the largest longitudinal databases studied to date.* LEO_LS contains $31,852$ images of $5,636$ subjects, and PCSO_LS contains $147,784$ images of $18,007$ subjects, where the average time span between a subject's multiple image acquisitions is 6.1 and 8.5 years, respectively. Such large-scale databases allow for evaluation of performance at low FAR values (*e.g.,* 0.01% and 0.1%). Previous studies (*e.g.,* [70, 99]) evaluated at 1% FAR and higher.

2. *Determine the age-invariant properties of current state-of-the-art face matchers.* Rates of change over time in genuine comparison scores are analyzed using mixed-effects regression models, which are appropriate for longitudinal data. In doing so, we quantify (i) the population-mean rate of change in genuine scores over time and (ii) the variability in subject-specific longitudinal trends (*i.e.,* how closely individuals in the

[1] A biometric system operating with a global threshold uses the same decision threshold for all subjects across all comparisons.

population follow the population-mean trend). We also investigate the influence of age at enrollment, sex, race, and face image quality.

3. *Methodology and analysis tools for advancing the development and evaluation of age-invariant face recognition algorithms.* The analysis conducted in this chapter can be applied to any matcher and any database. Periodic reevaluation will be necessary as face recognition technology evolves to better address facial aging.[2]

Our previous longitudinal analysis of automatic face recognition was first published in [18]. The present work extends and refines our previous study in significant ways. The primary differences are as follows. (i) We study longitudinal effects of both *aging* (elapsed time) and *age* (biological age); [18] only studied elapsed time. (ii) Genuine scores are computed to represent a scenario where the youngest image of each subject is enrolled in a gallery (a subject with n_i total images has $n_i - 1$ scores, whereas [18] computed all $\binom{n_i}{2}$ genuine scores). Comparing query images to an enrollment image (a fixed point in time) simplifies the complex correlation structure that is present for all pairwise comparisons. (iii) We analyze an additional longitudinal face database (namely, LEO_LS) from a different law enforcement agency than the PCSO_LS database used in [18], and a different COTS matcher is used to obtain genuine scores for LEO_LS. Still, longitudinal analysis shows similar results for both databases and matchers.

The remainder of this chapter is organized as follows. Section 4.2 highlights related work on facial aging as it pertains to automatic face recognition. Section 4.3 details the two longitudinal face databases used in this study. Section 4.4 explains the methodology used for longitudinal analysis. Section 4.5 gives results for both the PCSO_LS and LEO_LS face databases. Section 4.6 summarizes our observations about the current longitudinal capabilities of automatic face recognition.

[2]To facilitate longitudinal study on other face datasets and matchers, the code of our longitudinal analysis will be made publicly available at http://biometrics.cse.msu.edu/.

4.2 Related Work

Almost all of the published studies that investigate the effects of facial aging on automatic face recognition performance adopt the following approach: (i) divide the database (face pairs) into partitions depending on age group or time lapse, (ii) report summary performance measures (*e.g.,* TAR at fixed FAR) for each partition independently, and then (iii) draw conclusions from the differences in performance across the partitions. Such an approach has led to the following general conjectures [91]: (i) Face recognition performance decreases as the time elapsed between two images of the same person increases (*e.g.,* [70, 87, 99]). (ii) Faces of older individuals are easier to recognize/discriminate than faces of younger individuals (*e.g.,* [55, 87]). See Table 4.1 for a summary of these studies.[3]

Partitioning of data (images or subjects) based on age group or time lapse is often arbitrary and varies from one study to another. Erbilek and Fairhurst show that different age group partitionings result in different performance trends for both iris and signature modalities [43]. Furthermore, this cohort-based analysis with summary statistics cannot address whether age-related performance trends are due to changes in genuine (same subject) comparison scores, impostor (different subjects) comparison scores, or both.

Multilevel (hierarchical or mixed-effects) statistical models have been used for determining important factors (covariates) to explain the performance of face recognition systems. Beveridge *et al.* [20] apply generalized linear mixed models to verification decisions (accept or reject) made by three algorithms in the FRGC Exp. 4 evaluation. In addition to eight levels of FAR as a covariate, they analyze gender, race, image focus, eye distances, age, and elapsed time. The limitations of this study include (i) the maximum elapsed time between face images of the same subject is less than one year, and (ii) it only involves 351 subjects. Poh *et al.* [110] utilized regression models to estimate subject-specific biometric (face and speech) performance trends over time, but the database used only contains 150 subjects and

[3]Studies that address developing age-invariant face recognition algorithms (*e.g.,* [50,67]) are beyond the scope of this work.

Table 4.1 Table of related work on the effects of facial aging on face recognition performance.

Study	Database	Age or Elapsed Time Partitions	Summary of Findings
Ling *et al.* [87]	Passports (private)	4–11 years elapsed time	Degradation in EER saturates after 4 years elapsed time.
	FG-NET	0–8, 8–18, and 18+ years old	Verification accuracies increase with increasing age group.
Klare and Jain [70]	PCSO (200,000 mugshots, 64,000 subjects)	0-1, 1-5, 5-10, 10+ years elapsed time	TARs at 1% FAR are 96.3%, 94.3%, 88.6%, and 80.5% for the listed elapsed time partitions. Training/testing on different aging partitions decreases performance in some non-aging scenarios.
Otto *et al.* [99]	MORPH-II	0–1, 1–5 years elapsed time	TARs at 1% FAR are 97% and 95% for the listed elapsed time partitions. The nose is the most stable facial component over time.
Bereta *et al.* [15]	FG-NET	0–5, 6–10, 11–15, 16–20, 21–30, and 30+ years elapsed time; 23–30, 31–40, 41–50, and 50+ years old	Identification accuracies of local descriptors (*e.g.,* variants of LBP) when combined with Gabor wavelet magnitudes become relatively consistent across absolute ages and age gap groups, but accuracies are still fairly low for a small gallery.
NIST FRVT [55]	Visa images (19,972 subjects)	baby, kid, pre-teen, teen, young, parents, older	Error rates (for open-set identification) are higher for younger age groups when the same threshold is used for all age groups.

EER = equal error rate; TAR = true accept rate; FAR = false accept rate

the elapsed times are less than two years. The longitudinal study on face recognition in this work follows the general methodology of linear mixed-effects statistical models outlined in [54] for iris recognition and [149] for fingerprint recognition.

The two main databases used for research on facial aging, including automatic age estimation, age progression, and age-invariant face recognition, are FG-NET [78] and MORPH [113]. Panis *et al.* [100] provide a recent overview of research that has utilized the FG-NET database. While the public release of these databases greatly encouraged progress in these areas, the databases are not suitable for longitudinal analysis because (i) FG-NET contains only 82 subjects in total, and (ii) MORPH contains only a small number of subjects with multiple images over time (only 317 subjects have at least 5 images over at least 5 years).[4] The Cross-Age Celebrity Dataset (CACD) [32] was recently released, containing $163,446$ images of $2,000$ celebrities across 10 years. However, because the images were downloaded from the web (via Google search), the unconstrained quality makes it difficult to statistically model

[4]Images in FG-NET are relatively unconstrained (scanned from personal photo collections), while the MORPH databases are mugshots, similar to LEO_LS and PCSO_LS used in this work but with different database properties (see Table 4.2).

Table 4.2 Facial Aging Databases

Database	**Num. Subjects**	**Num. Imgs**	**Num. Imgs per Subject**	**Age Range (years)**
FG-NET [78]	82	1,002	6–18 (avg. 12)	0–69 (avg. 16)
MORPH-II [113]	13,000	55,134	2–53 (avg. 4)	16–77 (avg. 42)
MORPH-II commercial [113][a]	20,569	78,207	1–76 (avg. 4)	15–77 (avg. 33)
CACD [32]	2,000	163,446	n.a. (avg. 81)	16–62 (n.a.)
LEO_LS[b]	5,636	31,852	4–20 (avg. 6)	12–69 (avg. 31)
PCSO_LS[b]	18,007	147,784	5–60 (avg. 8)	18–83 (avg. 35)

[a] This largest version of MORPH-II only has 317 subjects with at least 5 images acquired over at least 5 years.
[b] The longitudinal face image databases used in this study (details in Sec. 4.3).

the effects of facial aging. Variations in pose, illumination, expression, etc., may largely influence the trends in similarity scores. Such covariates are difficult to quantify in order to "tease out" these effects from the longitudinal effects, so standardized imaging (near-frontal, neutral expression, uniform illumination) is preferable for the longitudinal study conducted in this work. Relatively constrained images, such as mugshots, help to ensure that other effects, such as PIE variations, are captured in the noise term in the statistical models. For the above reasons, our longitudinal analysis utilizes two new longitudinal face databases, detailed in Section 4.3.

4.3 Longitudinal Face Databases

Operational face image datasets maintained by government and law enforcement agencies can contain longitudinal records of individuals of magnitudes that are infeasible to collect in laboratory settings (*e.g.,* elapsed times over 10+ years). These agencies routinely collect face images of the same individuals over time and have been doing so for relatively long

durations, primarily for applications involving driver's licenses, visa and passport applications/renewals, frequent travelers, and multiple arrests of repeat criminal offenders. The sources of face images in our longitudinal analysis are mugshot bookings. While we acknowledge that lifestyle factors (*e.g.,* drug[5] and alcohol use, trauma, etc.) may increase aging rates for some individuals in this population (adult repeat criminal offenders), these accelerated agers are expected to be outliers in the statistical models in our analysis; the overall trends should be relatively robust to this factor. Additionally, we were not able to access any other longitudinal face data. We did attempt to use longitudinal face images from the State Department visa databases. However, we discovered that roughly 5% of genuine face images were duplicate photo submissions (*e.g.,* an individual reuses the same photo for a visa renewal application), so the corresponding inaccurate age information rendered it unsuitable for longitudinal study.

The two databases used in this longitudinal study (LS), denoted LEO_LS and PCSO_LS, are subsets of subjects and images from two larger mugshot databases initially consisting of 3.7 and 1.5 million images, respectively. The following criteria were used to compile the subsets: (i) Each subject has at least 4 (LEO_LS) or 5 (PCSO_LS) face images that were (ii) acquired over at least a 5 year time span, and (iii) each pair of consecutive images is time-separated by at least one month. Database statistics are shown in Fig. 4.2.

The facial variations in the PCSO_S and LEO_LS databases are well-controlled because the mugshots adhere to standards similar to those detailed in the ANSI/NIST-ITL 2011 face image standards.[6] The standards specify that mugshots should be captured at frontal pose, with neutral expression, uniform illumination, and a background set to 18% gray, for examples. Because these databases are both from operational sources, some confounding factors are still present, such as minor pose and expression variations (see Fig. 4.5). We also observed rare occurrences of facial occlusions or injury, as shown in Fig. 4.4, but have retained such images in this study.

[5]See Yadav *et al.* [143] for work specifically on the effects of drug abuse on face recognition performance.
[6]https://www.nist.gov/itl/iad/image-group/ansinist-itl-standard-history

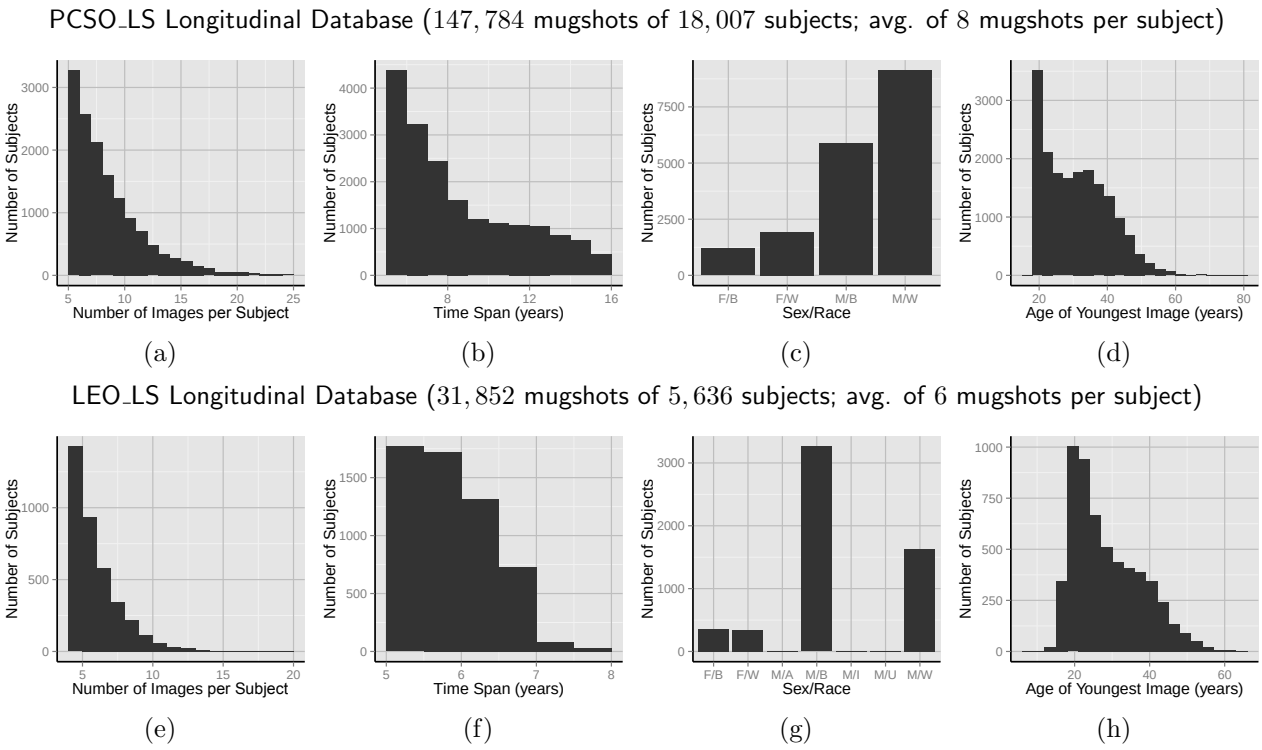


Figure 4.2 Statistics of the two longitudinal face image databases (PCSO_LS and LEO_LS) used in this study. (a) and (e) Number of face images per subject, (b) and (f) the time span of each subject (*i.e.,* the number of years between a subject's youngest and oldest face image acquisitions), (c) and (g) demographic distributions of sex (male, female) and race (white, black, Asian, Indian, unknown), and (d) and (h) the age of the youngest image of each subject (in years).

For both databases, we only include white and black race subjects in this study because there are too few subjects of other races to do a meaningful statistical analysis. Since human labeling errors pertaining to demographic attributes and subject ID can be inadvertently introduced in large-scale legacy databases, we determine the sex, race, and date of birth of a subject as the majority vote from each subject's records to ensure consistent labels within each subject. Identifying all such errors was not feasible due to the large size of these databases, but a cursory examination of the PCSO_LS database revealed 134 subject records that contained multiple identities (Fig. 4.3). These subject records were removed from our study.

(a) (b) (c)

Figure 4.3 Three examples of labeling errors in the PCSO_LS face database. All pairs show two different subjects who are labeled with the same subject ID number in the database.

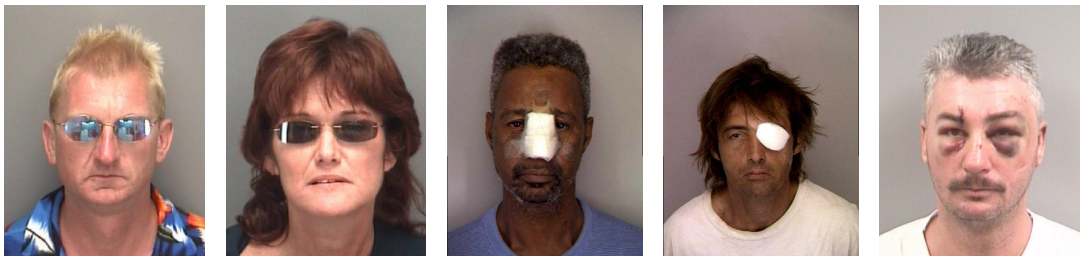


Figure 4.4 Examples of facial occlusions (sunglasses, bandages, and bruises) in the PCSO_LS face database.

4.3.1 LEO_LS Face Database

The LEO_LS database contains 31,852 images of 5,636 subjects from an operational dataset of law enforcement images. Each subject has an average of 6 images over an average time span of 5.8 years (maximum of 8 years). Demographic makeup of the LEO_LS database includes 2,009 white and 3,627 black subjects where 4,922 subjects are males and 714 are females. Subjects in LEO_LS are primarily adults, but there are 656 images of 369 subjects that are younger than 18 years-old; these may be juvenile[7] arrests or they could be data entry errors. Due to privacy considerations, we only have access to the comparison scores (both genuine and impostor), so we cannot show face images from this database.

4.3.2 PCSO_LS Face Database

The PCSO_LS database consists of 147,784 operational mugshots of 18,007 repeat criminal offenders booked by the Pinellas County Sheriff's Office (PCSO) from 1994 to 2010. Each

[7]In the United States, a juvenile is typically under the age of 17.

Table 4.3 Overall true accept rates (TARs) at fixed false accept rates (FARs) for various face matchers on the PCSO_LS and LEO_LS databases.

		0.01% FAR	0.1% FAR	1% FAR
PCSO_LS	**COTS-A**	**94.98**	**97.83**	**99.14**
	PittPatt	41.54	58.65	78.30
LEO_LS	**COTS-B**	**99.35**	**99.66**	**99.84**
	COTS-2	90.62	94.96	97.92
	COTS-3	78.97	86.87	93.49
	COTS-4	96.68	98.47	99.31

subject has an average of 8 images over an average time span of 8.5 years (maximum of 16 years). Demographic makeup of the PCSO_LS database includes 11,002 white and 7,004 black subjects where 14,882 subjects are males and 3,124 are females. Example face images from PCSO_LS are shown in Fig. 4.5. Each booking record in PCSO_LS contains both the date of birth and the date of arrest (actual dates were unavailable for LEO_LS, only the ages were provided to us).

4.3.3 Face Comparison Scores

Face comparison scores (similarities) were obtained from various commercial face matchers with the aim of evaluating current state-of-the-art longitudinal performance. Two matchers were applied to the PCSO_LS database, and comparison scores were obtained from four different matchers for the LEO_LS database.[8] As shown in Table 4.3, COTS-A and COTS-B were the overall most accurate matchers. Due to space limitations, longitudinal results are only reported for COTS-A and COTS-B throughout the remainder of the chapter. COTS-A and COTS-B were both among the top-3 performers in the FRVT 2013 [55].

The original mugshot images were input to each COTS matcher, and a total of 26,216 and 129,773 genuine scores were computed for the LEO_LS and PCSO_LS databases, respectively, under the scenario where each subject's set of face images are compared to his/her enrollment

[8]Comparison scores and ancillary information (sex, race, age) for the LEO_LS face image database were provided by the Image Group, National Institute of Standards and Technology (NIST), http://www.nist.gov/itl/iad/ig/.

image. Genuine comparison scores, s_{ij}, between the enrollment and jth face images of subject i were standardized so $y_{ij} = (s_{ij} - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the genuine scores from all subjects. This standardized response, y_{ij}, is in terms of standard deviations from the mean of the genuine distribution, which allows interpretation of coefficients from mixed-effects regression models as quantifying the change in genuine scores as β standard deviations per year. Fig. 4.8 shows the distributions of COTS-A and COTS-B standardized genuine scores.

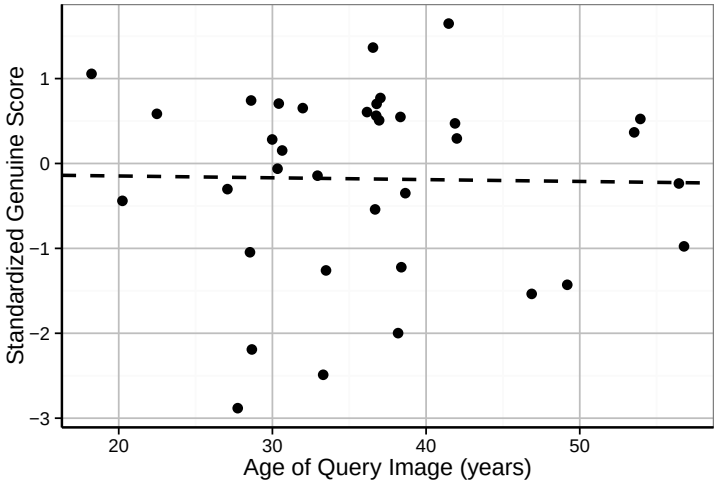
The response variable for all mixed-effects models in this study are standardized *genuine* comparison scores. However, to evaluate face recognition performance, trends in genuine scores should be considered in context with an impostor distribution. For both the LEO_LS and PCSO_LS databases, we computed all possible impostor scores (5.5 million and 11.1 billion, respectively) to calculate thresholds at different fixed FAR values. The threshold at 0.01% FAR, for example, is used to determine when genuine scores drop below the threshold, causing false rejection errors.

4.4 Mixed-Effects Models

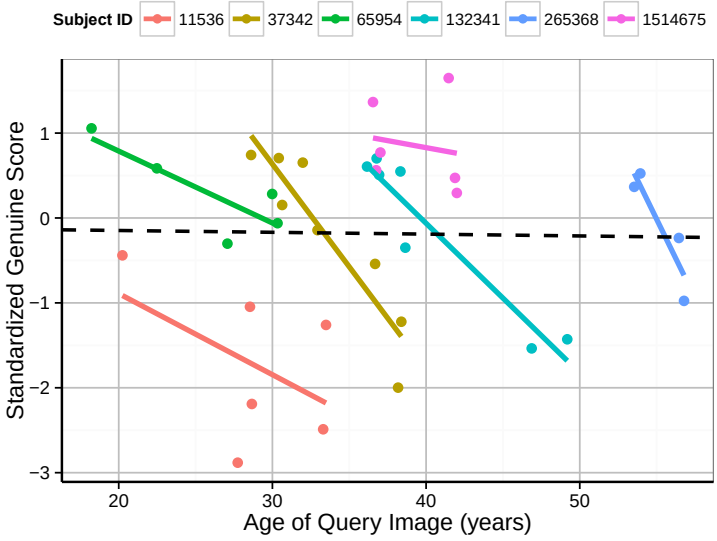
Mixed-effects models (also known as random-effects, multilevel, and hierarchical models) are widely used in various scientific disciplines for studying data that is hierarchically structured, including longitudinal data of repeated observations over time [44, 118]. In our case, face images are grouped by subject because we have repeated observations of each individual in our study. When data is structured in such a manner, responses from the same cluster/group/individual are correlated with each other and across time (for longitudinal data). Mixed-effects models enable analysis of variation in the response (here, standardized face comparison scores) that occurs at different levels of the data hierarchy.

Ideally, longitudinal data collection would observe all individuals in the study following the exact same schedule over the entire duration of interest. However, longitudinal data is

Figure 4.5 Face images of six example subjects from the PCSO_LS database. The enrollment face image (leftmost column) is the youngest image of each subject, and all query images are in order of increasing age. In this study, genuine similarity scores are computed by comparing the query images of each subject to his/her enrollment image.



(a)



(b)

Figure 4.6 An example of cross-sectional vs. longitudinal analysis. In (a), a cross-sectional approach (ordinary least squares (OLS) linear regression) is applied, which incorrectly assumes that all the scores are independent. In (b), OLS is instead applied six times, separately to each subject's set of scores (subjects shown in Fig. 4.5). The slope estimated by cross-sectional analysis (black dotted line) is much flatter than the slopes of subject-specific trends in (solid colored lines in (c)). The longitudinal analysis in this work utilizes mixed-effects models, which provide "shrunken" OLS estimates for each subject, where the OLS trends shrink towards a population-mean trend [44,118], further accounting for the correlation that exists between scores from the same subject.

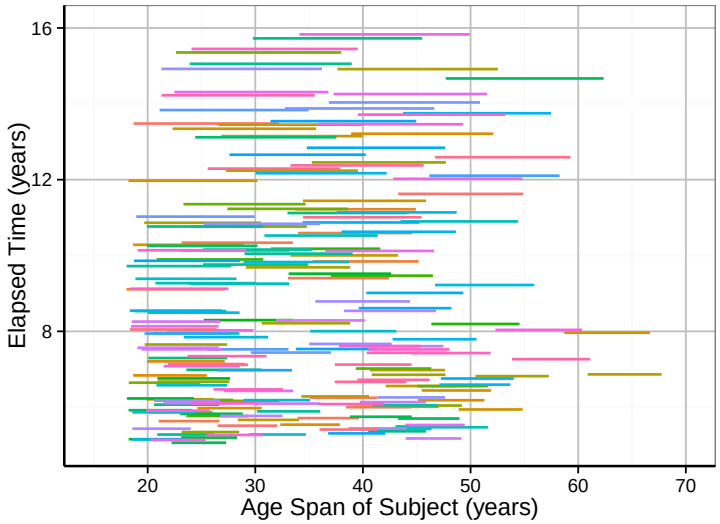


Figure 4.7 Age distribution of a random sample of 200 subjects from the PCSO_LS database. Each line denotes the age span of a subject (*i.e.,* age of youngest image to age of the oldest image), separated along the y-axis by the elapsed time for each subject (*i.e.,* the length of the age span).

typically not this nicely structured because it is difficult (and expensive) to collect, or it must be analyzed retrospectively, as is the case with the mugshot databases used in this study. Instead, longitudinal data is most often *time-unstructured* and *unbalanced*, meaning individuals in the study population are observed at different schedules and have different numbers of observations. For the mugshot databases, this translates to different rates of recidivism for each subject. Fig. 4.2 shows that subjects in the LEO_LS and PCSO_LS databases have anywhere from 4 to more than 20 mugshots, and Fig. 4.7 shows that the age spans of the subjects are highly unstructured.

Mixed-effects models can handle imbalanced and time-unstructured data and are preferable over other approaches because they model both the mean response (fixed effects define the population-mean trend), as well as the covariance structure (random effects allow deviations of individuals from the population-mean). In longitudinal data, this covariance structure has a complicated form which stems from the fact that error terms are *not* independent (as is assumed in standard linear regression). The remainder of this section provides details of the models and covariates of interest.

Table 4.4 Mixed-Effects Model Formulations

Model	Level-1 Model	Level-2 Model: Intercept	Level-2 Model: Slope
A	$y_{ij} = \varphi_{0i} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + b_{0i}$	
BT	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
CT	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
CA	$y_{ij} = \varphi_{0i} + \varphi_{1i}AGE_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
D	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + \beta_{02}AGE_{ie}^2 + b_{0i}$	$\varphi_{1i} = \beta_{10} + \beta_{11}AGE_{ie} + b_{1i}$
E	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + \beta_{02}AGE_{ie}^2 + \beta_{03}M_i + \beta_{04}B_i + b_{0i}$	$\varphi_{1i} = \beta_{10} + \beta_{11}AGE_{ie} + \beta_{12}M_i + \beta_{13}B_i + b_{1i}$
Q	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varphi_{2i}Q_{ij} + \varphi_{3i}Q_{ij}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}Q_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + \beta_{11}Q_{ie} + b_{1i}$, $\varphi_{2i} = \beta_{20} + \beta_{21}Q_{ie} + b_{2i}$, $\varphi_{3i} = \beta_{30}$

$\triangle T_{ij}$: elapsed time (years) between the enrollment and jth face image of subject i;
AGE_{ie}: age (years) of subject i in her enrollment face image;
AGE_{ij}: age (years) of subject i in her jth face image;
M_i: binary indicator of subject sex ($M_i = 1$ if male, 0 if female);
B_i: binary indicator of subject race ($B_i = 1$ if black, 0 if white)
Q_{ie}: quality (*e.g.,* frontalness or interpupillary distance) of the enrollment image of subject i;
Q_{ij}: quality (*e.g.,* frontalness or interpupillary distance) of the jth query image of subject i

4.4.1 Model Formulations

Given n_i face images of subject i, let AGE_{ij} denote the absolute age of the ith individual for the jth face image, where $AGE_{ij} < AGE_{ik}$ for $j = 0, \ldots, n_i - 2$ and $k = j + 1, \ldots, n_i - 1$ (*i.e.,* the n_i images are ordered by increasing age). To begin with, assume that the youngest image (first acquisition) of each subject is enrolled in the gallery, and let $AGE_{ie} = AGE_{i0}$ denote the age of individual i at enrollment where $AGE_{ie} < AGE_{ij}$ for $j = 1, \ldots, n_i - 1$. We can compute $m_i = n_i - 1$ genuine comparison scores by comparing every other image to the enrollment image. Hence, in this scenario, y_{ij} ($j = 1, \ldots, m_i$) is the comparison score between the jth face image of individual i and his/her enrollment image. AGE_{ij} is the age of the jth query/probe image of subject i, so the elapsed time between enrollment and query image is $\triangle T_{ij} = AGE_{ij} - AGE_{ie}$.

When studying age-related effects on automatic face recognition performance, there are two different, albeit closely related, time-varying covariates which are of primary interest: (i) the *elapsed time* between image acquisitions and (ii) the *absolute ages* of the subject in

the two face images being compared. Below, we discuss mixed-effects models which include these and other covariates.

4.4.1.1 Function of Elapsed Time

The simplest notion of face recognition performance over time is a function of the elapsed time between a subject's enrollment and query face images, $f(\triangle T_{ij})$. A linear mixed-effects model with two levels (to account for subject-specific trends) and a single covariate for elapsed time can be formulated as follows. At level-1, the comparison score y_{ij} between the enrollment and jth query image of subject i can be modeled as a linear function of $\triangle T_{ij}$:

$$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}, \tag{4.1}$$

where the ith individual's intercept, φ_{0i}, and slope, φ_{1i}, are

$$\begin{aligned} \varphi_{0i} &= \beta_{00} + b_{0i}, \\ \varphi_{1i} &= \beta_{10} + b_{1i}. \end{aligned} \tag{4.2}$$

The *level-1* equation in (4.1) models *within-subject* longitudinal change in y_{ij} where a subject's scores can vary around his/her linear trend by ε_{ij} (level-1 residual variation). The *level-2* model in (4.2) accounts for *between-subject* variation in comparison scores because each subject's intercept and slope parameters, φ_{0i} and φ_{1i}, respectively, are modeled as a combination of fixed and random effects. The *fixed effects*, β_{00} and β_{10}, are the grand means of the population intercepts and slopes, respectively, and define the overall *population-mean trend*, while the *random effects*, b_{0i} and b_{1i}, are subject-specific deviations from the population-mean parameters. Since each subject can have his/her own intercept and slope parameters, mixed-effects models are flexible in handling/allowing for biometric zoo effects [41,144] (some subjects generally have higher or lower scores). Fig. 4.5 shows six example subjects from the PCSO_LS database at different ages, with their subject-specific trends in genuine scores

over time shown in Fig. 4.6(b).

The random structure of the above two-level model includes the level-1 residuals, $\{\varepsilon_{ij}\}$, as well as the random effects, b_{0i} and b_{1i}, which can be thought of as level-2 residuals. The distributional assumptions of these two error terms are:

$$\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon}^2) \tag{4.3}$$

and

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right), \tag{4.4}$$

where $N(.,.)$ denotes a Gaussian distribution.

Substituting the level-2 equations for subject-specific intercepts and slopes into the level-1 model in (4.1), the *composite form* of the two-level mixed-effects model is:

$$y_{ij} = \big[\beta_{00} + b_{0i}\big] + \big[\beta_{10} + b_{1i}\big] \triangle T_{ij} + \varepsilon_{ij}. \tag{4.5}$$

Here, the model terms inside the two brackets in (4.5) correspond to all coefficients for the intercept and slope terms.

When the error terms are equal to their assumed means of zero, (6) reduces to the population-mean trend of $y_{ij} = \beta_{00} + \beta_{10} \triangle T_{ij}$. The grand mean intercept β_{00} quantifies the expected *marginal* mean comparison score when $\triangle T_{ij} = 0$. Note that this intercept is not particularly meaningful, as our data does not contain any same-day comparisons. However, interpretation of β_{00} does give us some notion of differences in subject's comparison scores at a projected baseline of zero years elapsed time. The primary coefficient we are interested in is β_{10} which quantifies the expected change in mean comparison score per one-year increase in elapsed time since enrollment. Because this model, as well as all others considered in this work, include random terms for both intercepts and slopes (b_{0i} and b_{1i}), we can also analyze the variation in the population parameters (*i.e.,* differences in the trends of individuals in

the population).

4.4.1.2 Function of Elapsed Time and Age at Enrollment

If rates of change in comparison scores are steeper or flatter throughout an individual's lifetime, then face recognition performance may also be a function of absolute age. If we add the age of the enrollment image to (4.5):

$$y_{ij} = \big[\beta_{00} + \beta_{01} AGE_{ie} + b_{0i}\big] + \big[\beta_{10} + b_{1i}\big] \triangle T_{ij} + \varepsilon_{ij}. \tag{4.6}$$

Because AGE_{ie} is a *fixed* effect for each subject (*time-invariant*), the above composite model actually has a two-level specification with the same level-1 model in (4.1). Hence, AGE_{ie} cannot improve the model fit at level-1 (within-subject); it can only influence the level-2 subject-specific variations.[9] The population-mean trend for (4.6) is:

$$\begin{aligned} \mathbf{E}(y_{ij}) &= \beta_{00} + \beta_{01} AGE_{ie} + \beta_{10} \triangle T_{ij} \\ &= \beta_{00} + \beta_{01} AGE_{ie} + \beta_{10} (AGE_{ij} - AGE_{ie}). \end{aligned} \tag{4.7}$$

By definition, $\triangle T_{ij}$ is a *centered* version of AGE_{ij}, where the centering term (AGE_{ie}) is subject-specific. Hence, the model for aging as a function of elapsed time and age at enrollment, $f(\triangle T_{ij}, AGE_{ie})$, is mathematically equivalent to a model for aging as a function of the age of the query image and age at enrollment, $f(AGE_{ij}, AGE_{ie})$:

$$\mathbf{E}(y_{ij}) = \beta_{00} + \beta_{01} AGE_{ie} + \beta_{10} AGE_{ij}. \tag{4.8}$$

The two models in (4.7) and (4.8) will result in the same estimate for longitudinal change, β_{10}. What distinguishes them is the interpretation of the coefficient β_{01} quantifying the effect of AGE_{ie}. Note the relationship between the two models: $\beta_{01}^{(4.8)} = \beta_{01}^{(4.7)} - \beta_{10}^{(4.7)}$. Hence, $\beta_{01}^{(4.8)}$

[9] Comparing all images of a given subject to her fixed enrollment image means that AGE_{ij} and $\triangle T_{ij}$ are perfectly correlated at level-1 (within-subject) of the model. Hence, we cannot include both of these covariates; the effect of age must be added as a level-2 covariate.

is the "contextual" effect that models the *difference* between the within- and between-subject effects of aging [14].[10] The significance of subject age at enrollment in (4.8) is tested with the null hypothesis of $H_0 : \beta_{01} = 0$, whereas *restricted* inference is needed to test significance in (4.7) because the null hypothesis must instead be $H_0 : \beta_{01} = \beta_{10}$.

The relationship between these two models (CT and CA) is similar to common approaches for decoupling the longitudinal and cross-sectional effects of a time-varying covariate. A time-varying covariate at level-1 (*e.g.,* age or elapsed time) exhibits variability *within*, but also *between* individuals; models which assume that the within- and between-individual effects are equal do not properly estimate either of these effects [12, 14, 44, 95]. Typically, the time-varying covariate is "centered" on subject-specific means, so as to remove between-subject variation at level-1 of the model.

4.4.2 Model Comparison and Evaluation

The goal of statistical modeling is to find a model that includes substantive predictors and excludes unnecessary ones (parsimony). A common approach is to fit increasingly complex models to successively evaluate the impact of adding different covariates [118]. Models can be compared using goodness-of-fit measures based on log-likelihood statistics: deviance, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Deviance quantifies how much worse the current model is compared to the (hypothetical) saturated model that includes all possible covariates to perfectly fit the data. Because the log-likelihood (LL) of the saturated model is zero,

$$\text{Deviance} = -2[LL_{\text{current}} - LL_{\text{saturated}}] = -2LL_{\text{current}}. \tag{4.9}$$

Deviance can be used to compare nested models (*i.e.,* the more complex model can be reduced to the simpler model by placing constraints on its parameters) that are fit to the

[10] The equality $\beta_{01}^{(4.8)} = \beta_{01}^{(4.7)} - \beta_{10}^{(4.7)}$ holds for mixed-effects models with random intercepts, and is approximately true for models with both random intercepts and random slopes.

same data. To compare non-nested models, AIC and BIC penalize the log-likelihood based on the complexity of the models[11] and the sample size. Smaller values indicate better fit for all three goodness-of-fit measures.[12]

Further comparisons of models depend on whether the successive model has included a time-invariant (*e.g.,* sex, race) or time-varying (*e.g.,* face image quality) covariate to the baseline model. For both cases, pseudo-R^2 statistics can be used to measure the proportional reduction in level-2 variance (σ_0^2, σ_1^2) and level-1 residual variance (σ_ε^2) attributable to inclusion of time-invariant and time-variant covariates, respectively.

4.5 Results

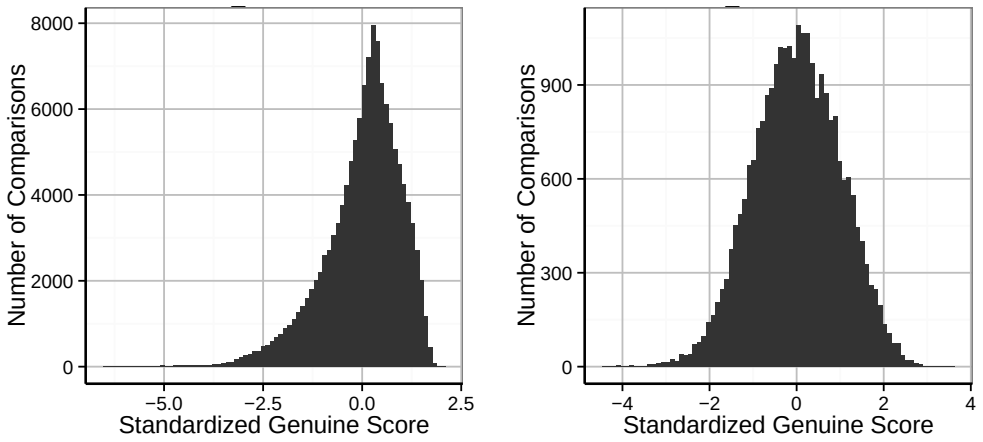
We first focus on analysis of the PCSO_LS database, starting with simpler models (*i.e.,* Model A and BT) and progressing to more complex models including covariates for subject sex/race and face image quality. We then present results for the LEO_LS database. Recall that models are discussed in Section 4.4 and equations are provided in Table 4.4. All models in our analysis are fit with full maximum likelihood (ML) estimation via iterative generalized least-squares (GLS) using the **lme4** package (v1.1-9) [11] for R (v3.2.2).

4.5.1 Model Assumptions

While mixed-effects models are capable of handling non-Gaussian response distributions (*e.g.,* COTS-A genuine scores in Fig. 4.8(a)), the error terms must follow Gaussian distribution. Fig. 4.9(a) shows normal probability plots of the level-1 residuals, ε_{ij}, from fitting Model BT to genuine scores from the PCSO_LS database. Since significant departure from linearity is observed at the tails, we cannot verify that the model assumptions hold; normal probability plots of random effects, b_{0i} and b_{1i}, also depart from linearity (Figs. 4.9(b),

[11]For full ML estimation, the number of parameters includes both the fixed effects and the variance components.

[12]For AIC and BIC, the magnitude of the reduction in model fit is difficult to interpret.



(a) PCSO_LS (b) LEO_LS

Figure 4.8 Distributions of standardized genuine comparison scores from the two longitudinal face databases used in this study: (a) COTS-A on PCSO_LS and (b) COTS-B on LEO_LS. There are a total of 129,773 and 26,216 genuine scores in (a) and (b), respectively.

4.9(c)). This behavior was observed for other models as well, precluding the use of standard errors for formal hypothesis tests of parameters [134].

When parametric model assumptions are violated, it is common to resort to non-parametric bootstrap to establish confidence intervals for the parameter estimates, as followed in Yoon and Jain [149]. Hence, for the PCSO_LS database, we conduct a non-parametric bootstrap by *case resampling* [134]; 1,000 bootstrap replicates are generated by sampling 18,007 subjects with replacement. Multilevel models are fit to each bootstrap replicate, and the mean parameter estimates over all 1,000 bootstraps are reported. Tests for fixed effects parameters can be conducted by examining the bootstrap confidence intervals.[13] Table 4.5 gives the bootstrap parameter estimates (with 95% confidence intervals), variance components, and goodness-of-fit for the models in Table 4.4.

[13]The null hypothesis of the parameter equal to 0 can be rejected at significance of 0.05 if the 95% confidence interval does not contain 0.

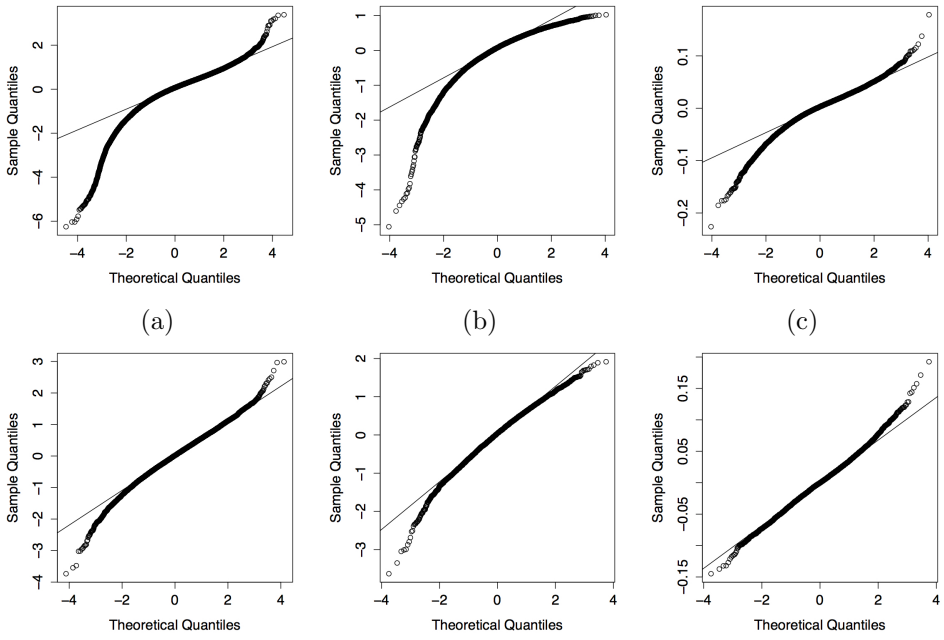


Figure 4.9 Normal probability plots of ((a) and (d)) level-1 residuals, ε_{ij}, and level-2 random effects for ((b) and (e)) intercepts, b_{0i}, and ((c) and (f)) slopes, b_{1i}, from Model BT on the PCSO_LS and LEO_LS databases (top and bottom rows, respectively). Departure from normality at the tails of the distributions is likely due to low quality face images or errors in subject IDs.

Table 4.5 Bootstrap results for mixed-effects models on the PCSO_LS database and COTS-A genuine scores.

		Model A	Model BT	Model CT	Model D
Fixed Effects (95% confidence intervals):					
Intercept	β_{00}	0.0274	0.6734	0.7226	0.5158
		$(0.0171, 0.0376)$	$(0.6624, 0.6849)$	$(0.6905, 0.7556)$	$(0.4073, 0.6239)$
Time	β_{10}		-0.1364	-0.1364	-0.1372
			$(-0.1379, -0.1349)$	$(-0.1379, -0.1349)$	$(-0.1426, -0.1316)$
Age Group	β_{01}			-0.0016	0.0120
				$(-0.0027, -0.0006)$	$(0.0047, 0.0189)$
Age Group × Time	β_{11}				$0.0000^{\#}$
					$(-0.0002, 0.0002)$
Age Group^2	β_{02}				-0.0002
					$(-0.0003, -0.0001)$
Variance Components:[a]					
Level-1 Residual	σ^2_ε	0.6076	0.3912	0.3912	0.3912
Random Intercepts	σ^2_0	0.3841	0.3243	0.3239	0.3231
Random Slopes	σ^2_1		0.0028	0.0028	0.0028
Covariance	σ_{01}		-0.0039	-0.0039	-0.0038
Goodness-of-Fit:[b]					
	AIC	333433	287016	287006	286985
	BIC	333462	287074	287075	287073
	Deviance	333427	287004	286992	286967

[a] Confidence intervals for variance components have been omitted due to space limitations.

[b] Goodness-of-fit values are the mean values of the $1,000$ bootstrap samples.

4.5.2 Unconditional Means Model (Model A)

The simplest mixed-effects model is the unconditional means model, which partitions the total variation in comparison scores by subject. Denoted Model A in Table 4.4, and with composite form of $y_{ij} = \beta_{00} + b_{0i} + \varepsilon_{ij}$, b_{0i} is the *subject-specific mean* and β_{00} is the *grand mean*. Similar to analysis of variance (ANOVA), Model A provides initial estimates of the within-subject variance σ^2_ε (*i.e.,* deviations around each subject's own mean comparison score) and the between-subject variance σ^2_0 (*i.e.,* deviations of subject-specific means around the grand mean). The intraclass correlation coefficient (ICC) quantifies the proportion of between-subject variation in the response, $\rho = \sigma^2_0/(\sigma^2_0 + \sigma^2_\varepsilon)$. Variance components for Model A shown in Table 4.5 indicate that between-subject differences in genuine scores (*i.e.,* biometric zoo) account for 38.7% ($\rho = 0.3873$) of the total variation in genuine scores from the PCSO_LS database. Baseline goodness-of-fit measures are also shown in Table 4.5.

4.5.3 Unconditional Growth Model (Model BT)

The next model to consider in longitudinal analysis is the unconditional growth model that includes the time-related covariate. In our case, we add elapsed time, $\triangle T_{ij}$, as well as random effects for slopes, b_{1i}, to Model A, resulting in Model BT. Table 4.5 shows that Model BT estimates that PCSO_LS genuine scores decrease by 0.1364 standard deviations per one-year increase in elapsed time (see solid black line in Fig. 4.10). Comparing the level-1 residual variation of Models A and BT, elapsed time explains 35.6% of the variation in a given subject's genuine scores around his/her own average genuine score.[14]

Longitudinal change estimated by Model BT implies that the *population-mean trend* will drop below the thresholds for 0.01% and 0.1% FAR after 19.1 and 24.0 years elapsed time, respectively, but this only provides insight into performance on subjects in the population with average (or higher) genuine scores over time. A reliable face recognition system must be able to recognize much more than just 50% of the population it encounters, so we are also

[14]Using pseudo-$R^2 = (\sigma^2_\varepsilon(A) - \sigma^2_\varepsilon(BT))/\sigma^2_\varepsilon(A)$.

interested in the spread of the population around the population-mean trend. Do all subjects closely follow the population-mean trend, or is there large variability between subjects? Do biometric zoo effects extend to rates of change over time?

Using the estimated variance components for slopes and intercepts (σ_0^2, σ_1^2, and σ_{01}), we compute a 2D confidence ellipse (random effects are assumed to be 2D Gaussian distributed) to define a region that contains, for example, 95% of the estimated subject-specific parameters. In order to translate from the 2D space of intercepts and slopes to obtain a confidence region for genuine scores versus elapsed time, we sample 100 combinations of intercept and slope parameters along the contour of the confidence ellipse, compute the predicted genuine scores for each of the 100 trends, and define the confidence region as between the minimum and maximum predicted scores for different values of elapsed time. Results are shown in Fig. 4.10.

From the confidence bands of subject variations in Fig. 4.10, we infer that genuine scores for 99% of the population will remain above the threshold at 0.01% FAR for up to approximately 5.5 years elapsed time, which reduces to 95% of the population after 7 years (*i.e.,* false reject errors would occur, on average, for 5% of subjects after 7 years since enrollment). Similarly, at a higher FAR of 0.1%, 99% of subjects can be recognized up to 8.5 years elapsed time, which reduces to 95% after 10.5 years. Fig. 4.11 shows face images from six example outlier subjects whose estimated trends lie outside the 99% region of confidence due to extreme intercepts and/or slopes; subjects significantly deviate from the population spread due to alignment errors, face quality issues (illumination, facial occlusion), and changes to facial hair, for example.

4.5.4 Age at Enrollment (Models CT and D)

We next investigate whether the population-mean trends in genuine scores over time depend on a subject's absolute age (*i.e.,* whether variation in subject-specific trends observed in Model BT can be explained by differences in subject age). The significance of the AGE_{ie}

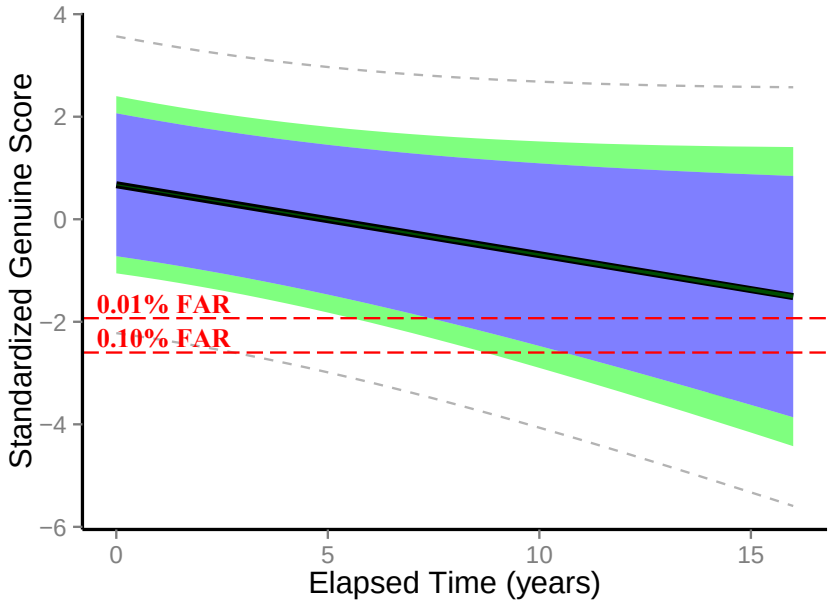


Figure 4.10 Results from Model BT on COTS-A genuine scores from the PCSO_LS database. The bootstrap-estimated population-mean trend is shown in black (bootstrap confidence intervals are too small to be visible). The blue and green bands plot regions of 95% and 99% confidence, respectively, for subject-specific variations around the population-mean trend. Grey dotted lines additionally add one standard deviation of estimated residual variation, σ_ε. Hence, Model BT estimates that 95% and 99% of the subject trends fall within the blue and green bands, but scores can vary around their trends, extending to the grey dotted lines. Thresholds at 0.01% and 0.1% FAR for COTS-A are shown as dashed red lines.

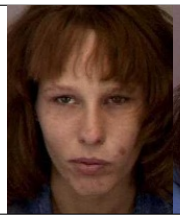
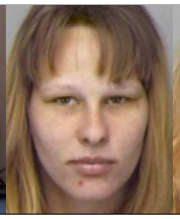
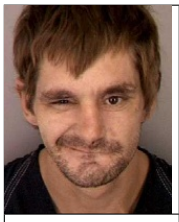
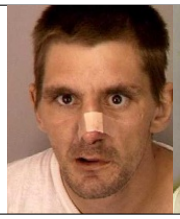
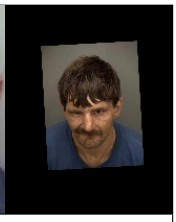
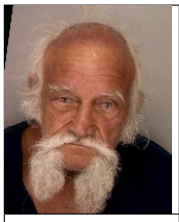
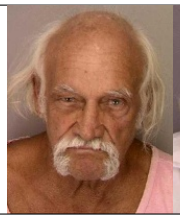
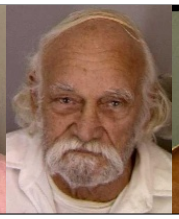
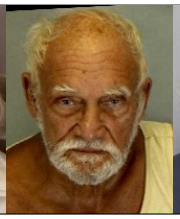
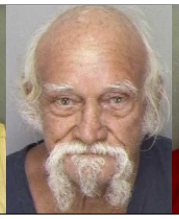
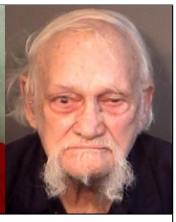
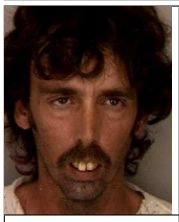
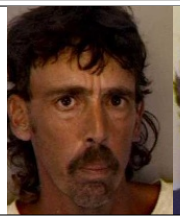
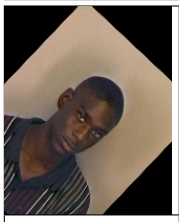
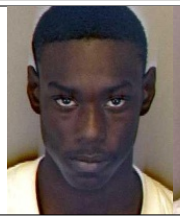
Enrollment Image	**Query Images** (in order of increasing age in years old)					
						
19.9	21.7	22.0	24.4	24.7	25.1	25.3
						
31.1	31.3	34.5	34.7	34.8	35.7	37.0
						
69.1	69.3	70.8	72.4	73.2	73.4	82.3
						
31.5	33.2	34.6	35.7	38.2	43.5	45.3
						
20.6	20.8	21.3	22.7	23.1	23.5	27.9

Figure 4.11 Example outlier subjects, *i.e.,* subjects whose subject-specific trends, estimated by Model BT, significantly deviate from the spread of the population in the PCSO_LS database. All images were aligned using COTS-A eye locations.

term in Model CT suggests a *negative* linear relationship between age at enrollment and genuine scores, but the magnitude of β_{01} is relatively small.

To further test the complexity of the effects of age at enrollment, we add additional terms associated with AGE_{ie}, resulting in Model D (see Table 4.4). The hypotheses of interest are 1) older subjects are easier to recognize than younger subjects, and 2) younger subjects age at faster rates than older subjects. These two hypotheses manifest in younger subjects having lower genuine scores, on average, and steeper negative rates of change. Table 4.5 shows that the interaction term $AGE_{ie} \times \triangle T_{ij}$ in Model D is not significantly different from zero because the 95% confidence interval for β_{11} contains zero; hence, we cannot conclude that subject enrollment age has a linear effect on rates of change in COTS-A genuine scores. The statistically significant β_{02} coefficient indicates a quadratic relationship between subject enrollment age and intercepts, and goodness-of-fit measures are lower compared to Model BT. However, further comparing to Model BT, level-2 variation in random effects for intercepts (σ_0^2) is only reduced by 0.4% after including AGE_{ie} terms. The differences between scores for different ages at enrollment are marginal compared to the change in scores due to elapsed time; the change in score between a 20 year-old and a 30 or 50 year-old (at enrollment) is equivalent to only 7 and 5 *months* of elapsed time (within-subject longitudinal change), respectively.

4.5.5 Sex and Race (Model E)

Model E in Table 4.4 is used to test the effects of subject sex and race. First, we observed that Model E results in better model fit than Model D (deviance for Model E is $285,712$ compared to $286,967$ for Model D). The main effect of subject sex is statistically non-zero at significance level of 0.05, but the main effect of subject race is not (the 95% bootstrap confidence interval contains 0). Male genuine scores at baseline ($\triangle T_{ij} = 0$ years) are 0.3987 standard deviations higher than female scores. Significant interactions with elapsed time indicate that rates of change in genuine scores depend on both sex and race; population-mean slopes are

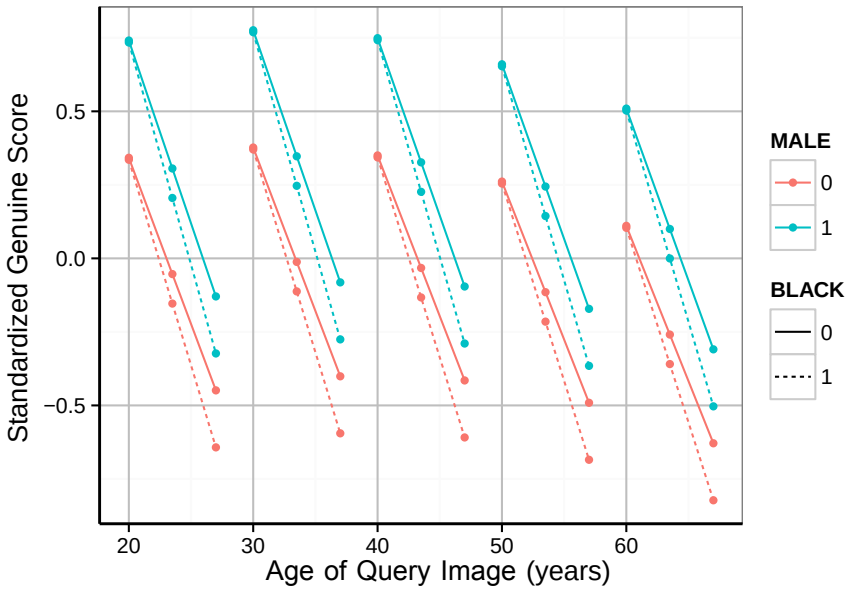


Figure 4.12 Model E fit to COTS-A genuine scores from the PCSO_LS database. Population-mean trends are plotted by subject demographics of sex and race. Each trend line represents seven years of elapsed time since enrollment at five different ages (20–60 years old). For example, the solid blue line beginning at $AGE_{ij} = 20$ years represents the average decrease in genuine scores for white males enrolled at age 20 with query images until age 27.

-0.0113 and -0.0267 standard deviations steeper for males and black subjects, respectively. Population-mean trends separated by subject demographics are shown in Fig. 4.12 for different ages at enrollment. while male genuine scores decrease at slightly faster rates than female scores, males are clearly easier to recognize with higher genuine scores overall. Fig. 4.12 also shows that the differences between subject race are minor compared to differences between males and females.

4.5.6 Face Image Quality (Model Q)

Adding level-2 covariates (*i.e.,* time-invariant values for each subject, such as AGE_{ie}) cannot improve the fit of the model at level-1 (within-subject). Table 4.5 shows that the level-1 residual variation σ^2_ε (*i.e.,* deviation of scores around each subject's own linear trend) is quite large when time is the only level-1 covariate for all models considered thus far. One standard deviation of level-1 residual variation estimated by Model BT (and similarly Models CT and

D) is equivalent to 4.6 years of elapsed time (calculated as $\sqrt{\sigma_\varepsilon^2}/\beta_{10} = \sqrt{0.3912}/-0.1372$). This is visually shown by the dotted grey lines in Fig. 4.10.

Level-1 residual variation can only be reduced by level-1 time-varying covariates (*i.e.,* image-specific); in this section we investigate whether face image quality measures can be used to improve the model fit. The quality measures considered are interpupillary distance (IPD) and a "frontal" score, both of which are output by COTS-A. While higher frontalness indicates better quality, the range of the frontal score has little meaning, since its computation is proprietary. We standardize (z-score) the frontalness score so we can interpret model parameters as standard deviations from the mean of the frontalness scores from all images in PCSO_LS.

After finding that neither of the quality measures alone explain variation in genuine scores as well as Model BT with only elapsed time as covariate (details are omitted due to space limitations), we then added the quality measures to Model BT, resulting in Model Q in Table 4.4. Table 4.6 gives estimated level-1 residual variation and goodness-of-fit for models with frontalness, IPD, and both frontalness and IPD (Model QF, QI, and QFI, respectively). Model QF has a better overall fit than Model QI. Table 4.7 gives the elapsed times for when population-mean scores cross thresholds at 0.001% and 0.01% FAR for different values of frontalness and IPD. Note how changing frontalness has a greater impact on when population-mean genuine scores cross the thresholds than changes in IPD. Model QFI with both measures of quality further reduces both the level-1 residual variation and values of goodness-of-fit values.

The values of 100 and 120 pixels for IPD in Table 4.7 were chosen because we observed systematic changes in IPDs over time (see Fig. 4.13); in particular, mean IPD varies around 100 pixels from 1994–2002 but increases to a consistent $\sim$120 pixels starting in 2003. This observation, along with correspondence with Pinellas County Sheriff's Office, suggests that booking agencies began to adhere to imaging standards around this time. To investigate whether this aspect of the data confounds the estimation of longitudinal effects (face images

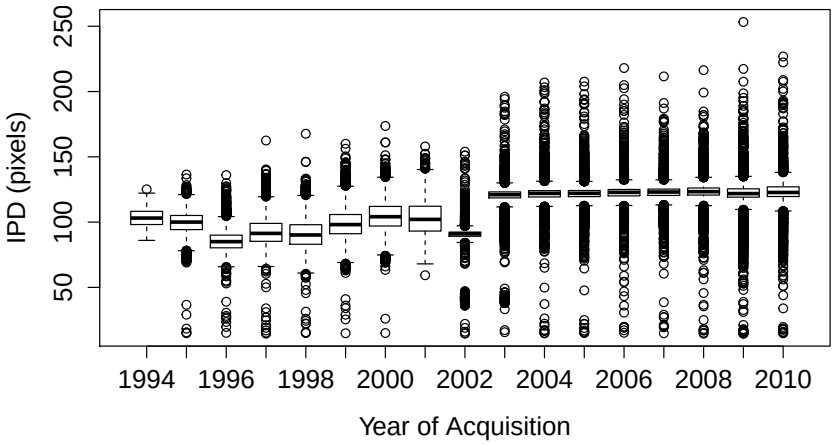


Figure 4.13 A boxplot of interpupillary distances (IPDs) versus year of acquisition shows that mean IPDs systematically changed over time for the PCSO_LS database, likely due to booking stations adhering to face imaging standards only in more recent years.

Table 4.6 Bootstrap results for mixed-effects models with elapsed time and face quality covariates for the PCSO_LS database and COTS-A genuine scores.

	Model QF	Model QI	Model QFI
σ_ε^2	0.3302	0.3539	0.3218
AIC	275108	281296	273643
BIC	275283	281471	273848
Deviance	275072	281260	273601

Table 4.7 Elapsed times (in years) for when population-mean trends in genuine scores drop below the decision thresholds at 0.001% and 0.01% FAR for different measures related to face quality (frontalness and IPD) of the enrollment image Q_{ie} and the query image Q_{ij}.

	Q_{ie}	Q_{ij}	0.001% FAR	0.01% FAR
Frontal	-1σ	-1σ	10.9	15.6
	μ	μ	13.0	18.4
	1σ	1σ	16.8	23.0
IPD	100 pixels	100 pixels	13.8	19.4
	100 pixels	120 pixels	14.0	20.0
	120 pixels	120 pixels	13.0	18.4

Table 4.8 Mixed-effects model results for the LEO_LS database and COTS-B genuine scores.

		Model A	Model BT	Model CT	Model D
Fixed Effects (standard errors):					
(Intercept)	β_{00}	0.0037	0.5395	0.5468	0.0894
		(0.0098)	(0.0127)	(0.0325)	(0.1057)
Time	β_{10}		−0.1699	−0.1699	−0.1980
			(0.0023)	(0.0023)	(0.0076)
Age Group	β_{01}			−0.0003	0.0346
				(0.0011)	(0.0068)
Age Group × Time	β_{11}				0.0010
					(0.0003)
Age Group^2	β_{02}				−0.0006
					(0.0001)
Variance Components:					
Level-1 Residual	σ^2_ε	0.5985	0.4276	0.4276	0.4275
Intercepts	σ^2_0	0.4009	0.5543	0.5542	0.5516
Slopes	σ^2_1		0.0059	0.0058	0.0058
Covariance	σ_{01}		−0.0317	−0.0317	−0.0316
Goodness-of-Fit:					
	AIC	68705	62647	62649	62606
	BIC	68730	62697	62707	62679
	Deviance	68699	62635	62635	62588

in later years may be of higher quality), we also tested for a difference in slope prior to 2003 versus after 2003 by using a piecewise linear formulation for the mixed-effects model (with a breakpoint at 2003). We found that slope after 2003 was significantly flatter (less negative).

Additional face quality factors known to cause changes in face recognition performance are illumination, expression, and occlusions. However, there are no widely accepted methods for quantifying such variations in face images and doing so is beyond the scope of this work.

4.5.7 LEO_LS Database

Table 4.8 gives results for the models in Table 4.4 fit to COTS-B genuine scores from the LEO_LS database. Fixed-effects parameter estimates are given with standard errors; bootstrapping was not conducted for LEO_LS models because the error terms better follow Gaus-

sian distributions (see Fig. 4.9). Model results are summarized as follows.

Model A estimates that 40% of the total variation in genuine scores is due to between-subject differences. The longitudinal change in genuine scores estimated by both Model BT and Model CT indicates that a one year increase in elapsed time decreases genuine scores by $\beta_{10} = -0.1699$ standard deviations. From the confidence bands of subject variations in Fig. 4.14 (estimated by Model BT), we infer that genuine scores for 99% of the population will remain above the threshold at 0.01% FAR for up to approximately 6.5 years elapsed time, which reduces to 95% of the population after 8.5 years (*i.e.,* false reject errors would occur, on average, for 5% of subjects after 8.5 years since enrollment). Similarly, at a higher FAR of 0.1%, 99% of subjects can be recognized up to 8.0 years, which reduces to 95% after 9.5 years elapsed time.

Although the between-subject effect of age at enrollment (β_{01}) is significantly different from β_{10} in Model CT, the effect is not significantly different from zero, indicating that there is no linear relationship between subject enrollment age and average genuine scores. However, additional terms involving AGE_{ie} result in significant effects of enrollment age in Model D. The significant β_{02} coefficient indicates a downward quadratic relationship between age at enrollment and average genuine scores (similar to COTS-A on PCSO_LS). Furthermore, the significant interaction term $AGE_{ie} \times \triangle T_{ij}$ indicates that longitudinal change in scores tends to vary with subject's age at enrollment; a 10-year increase in subject age results in a longitudinal slope that is $\beta_{11} = -0.0098$ standard deviations steeper. Population-mean rates of change range from -0.1784 to -0.1490 standard deviations per year for subjects with age at enrollment of 20 to 50 years (calculated as $\beta_{10} + \beta_{11} AGE_{ie}$). Recall that age at enrollment had no effect on rates of change for COTS-A on PCSO_LS.

Model E results indicate that intercepts are 0.0565 and 0.4238 standard deviations higher for black and male subjects, respectively (so, black-male subjects have intercepts that are 0.4803 standard deviations higher than white-female subjects). Slopes are not statistically different for black and white subjects, but the population-mean slope for males is *steeper*

(*i.e.,* more negative) than for females. These population-mean trends are shown in Fig. 4.15 for different ages at enrollment. Fig. 4.15 also shows that the differences between subject race are minor compared to differences between males and females, as was also the case for COTS-A on the PCSO_LS database.

4.6 Conclusions

We presented a longitudinal study of automatic face recognition, utilizing two large operational databases of mugshots, PCSO_LS (147,784 images of 18,007 subjects, avg. 8 images per subject over avg. 8.5 years) and LEO_LS (31,852 images of 5,636 subjects, avg. 6 images per subject over avg. 5.8 years), where each subject has at least four face images acquired over at least a five-year time span. Linear mixed-effects regression models were used to analyze variation in genuine scores due to elapsed time, age, sex, and race, as well as subject-specific differences in scores (*i.e.,* biometric zoo effects). Face similarity scores were obtained from state-of-the-art COTS matchers for both the PCSO_LS and LEO_LS databases. Based on our analysis, we make the following observations (statements apply to both databases and matchers):

✦ Population-mean trends indicate that genuine scores significantly decrease with increasing elapsed time between enrollment (gallery) and query (probe) images, as expected. However, population-mean trends (average genuine scores) do not fall below thresholds at 0.01% FAR until after 15 years elapsed time. This suggests that in a practical application, an average individual's genuine scores decrease at a rate that will not affect the recognition accuracy at 0.01% FAR until more than 15 years since enrollment.

✦ Significant subject-specific variability around the population-mean trends is observed; genuine scores for some subjects decline at much faster rates than the population-mean. Analysis of the estimated variance in subject-specific parameters (intercepts and slopes) allowed for estimation of subject-based accuracies (*i.e.,* how many subjects are estimated to

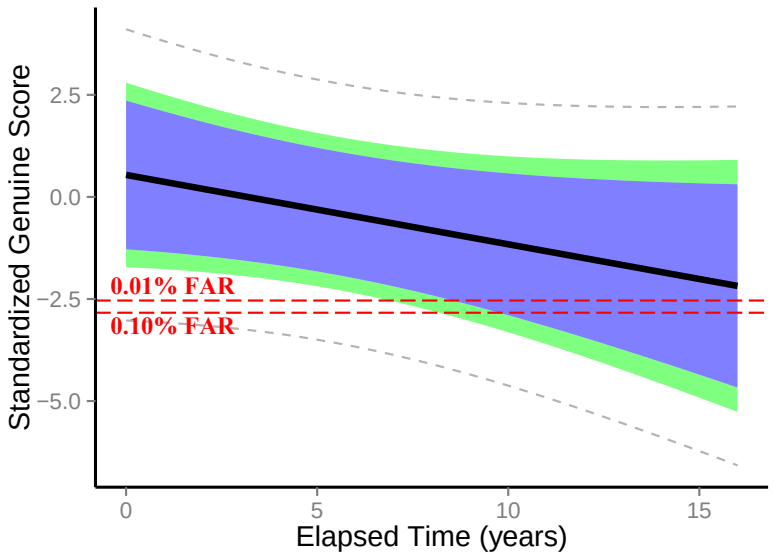


Figure 4.14 Results from Model BT on COTS-B genuine scores from the LEO_LS database. The population-mean trend is shown in black. The blue and green bands plot regions of 95% and 99% confidence, respectively, for subject-specific variations around the population-mean trend. Grey dotted lines additionally add one standard deviation of estimated residual variation, σ_ε. Hence, Model BT estimates that 95% and 99% of the subject trends fall within the blue and green bands, but scores can vary around their trends, extending to the grey dotted lines. Thresholds at 0.01% and 0.1% FAR for COTS-B are shown as dashed red lines.

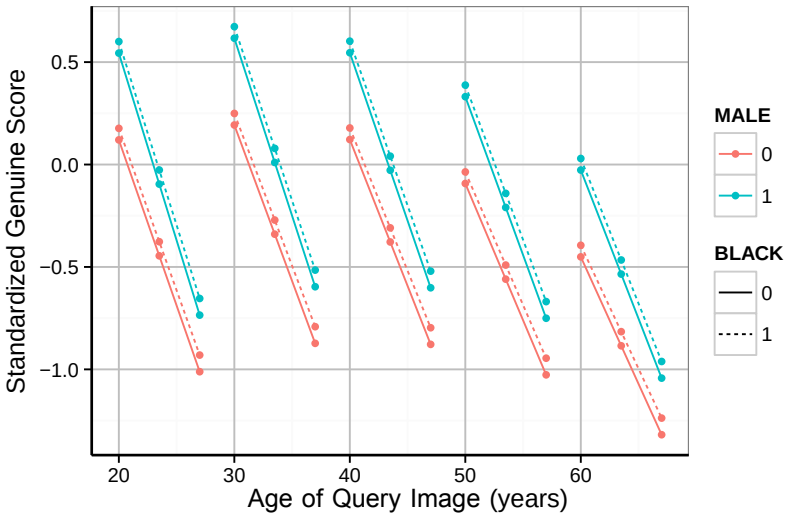


Figure 4.15 Model E for COTS-B genuine scores from the LEO_LS database. Population-mean trends are plotted by subject demographics of sex and race, in addition to five different ages at enrollment (20 to 60 years). Each trend line represents seven years of elapsed time since enrollment. For example, the solid blue line beginning at $AGE_{ij} = 20$ years represents the average decrease in genuine scores for white males enrolled at age 20 with query images until age 27.

be falsely rejected, rather than standard image-based accuracy calculations). For example, the models estimate that genuine scores for 99% of the population will remain above the threshold at 0.01% FAR until 6.5 years elapsed time for PCSO_LS and 5.5 years for LEO_LS. Other calculations (e.g. 95% of the population) are also within approximately one year for both databases.

✦ Subject-specific variance in rates of change (*i.e.,* linear slopes) is only marginally attributable to subject age at enrollment, sex, and race. Subject sex was the most significant factor for between-subject differences in genuine scores, with males having significantly higher genuine scores than females. The magnitude of the difference suggests that false reject errors may occur approximately two years earlier for females than for males (assuming that a global threshold is used operationally).

✦ While the model fit improved for more complex models incorporating simple measures of face quality (for the PCSO_LS database), the models are still limited for *prediction* purposes. The within-subject variability (*i.e.,* level-1 residual variance) is still quite large. All models considered in this study indicate that one standard deviation in genuine scores due to short-term variations (*e.g.,* illumination, hairstyle, etc.) is approximately equivalent to the change in genuine scores due to ± 4 years of elapsed time (for these particular databases and matchers).

Longitudinal analysis, in general, is an important, yet very difficult, problem. To the best of our knowledge, no proper statistical analysis has yet been conducted for studying face recognition performance on a large population over periods of time longer than five years. In this work, we attempted to analyze the covariates of interest that were available to us (elapsed time, age, sex, race, some measures of quality), but there are additional covariates that cannot be accounted for because we do not have the information (*e.g.,* camera characteristics, IPD for the LEO_LS database, expression variations, etc.). Despite this, the longitudinal study on automatic face recognition presented here utilizes two of the largest, deepest, and longest (in terms of number of subjects, number of images per subject, and

time spans of subject images, respectively) face image databases studied to date, and the COTS matchers are representative of current state-of-the-art. Given that the performance of face recognition systems continues to improve, longitudinal analysis should be conducted periodically to reevaluate robustness to facial aging (and other covariates).

Chapter 5

Summary and Future Work

This thesis has addressed some of the important challenges associated with automatic face recognition. The primary contributions involve the role of quality covariates present in unconstrained face images and the effect of facial aging on face recognition performance.

5.1 Contributions

In Chapter 2, we studied operational scenarios for recognition of unconstrained face media. The contributions include:

- A framework for matching a collection of face media (*i.e.,* images, videos, 3D models, demographics) was provided for scenarios where multiple instances of a subject's face are available (*e.g.,* to identify a person of interest). This is particularly of value to forensic investigations, as matching the collection of face media outputs a single candidate list for a human operator to review, rather than multiple candidate lists (one for each of the face samples available on the person of interest). This work is one of the first baselines provided for "template-based" matching which is rapidly gaining interest (*e.g.,* the NIST IJB-A protocol [72]).

Table 5.1 Published works which have reported results using the experimental protocols introduced in Chapter 2 for the LFW database [62] (single-image matching). COTS results were reported in Chapter 2.

Method	**Rank-1 Accuracy (%)**	**DIR (%) @ 1% FAR**
COTS-A	56.7	27.0
COTS-A (s1+s4)	66.5	36.0
DeepFace [128]	64.9	44.5
WST Fusion [129]	82.5	61.9
DeepID2+ [123]	95.0	80.7
DeepID3+ [120]	96.0	81.4

- Evaluation protocols introduced in Chapter 2 were publicly released[1] for closed-set and open-set *identification* of unconstrained face images and videos in the LFW [62] and YTF [141] databases. While our work focused on matching a collection of unconstrained face media, we also reported baseline results for single-image matching. At the time of release, identification protocols for unconstrained face images were lacking in the research community, as efforts were focused on maximizing performance on the LFW *verification* protocol [62] (which has some limitations, see Chapter 1). Table 5.1 shows that our evaluation protocols introduced in Chapter 2 have since been adopted by other published works for comparisons and have encouraged competition, particularly for the more challenging open-set identification problem.[2]

Chapter 3 focused on the important and challenging problem of automatic face image quality. This chapter offers the following contributions:

- The first study on human assessments of *unconstrained* face image quality. To the best of our knowledge, there have been no other work on human assessment of face quality since preliminary studies on mugshot quality by Adler *et al.* [2] and Hsu *et al.* [60]

[1]Evaluation protocols are available at: http://biometrics.cse.msu.edu/pub/databases.html

[2]The BLUFR protocol [85] was released around the same time and is also a valuable benchmark for unconstrained face recognition algorithms.

in 2006. Relative pairwise comparisons of face image quality (*i.e.,* "Which face image has better quality?") were collected via crowdsourcing on Amazon Mechanical Turk[3]. With a relatively small number of pairwise responses per "worker" (<1,000 pairs), a matrix completion approach [148] was utilized to obtain face quality ratings from each worker for all 13,233 images in the LFW database. The resulting human quality ratings were shown to be correlated with automatic face recognition performance.

- An automatic method was proposed to predict either (i) human face quality rating or (ii) similarity score-based face quality value. The proposed method uses image features extracted prior to matching and does not require any comparisons to reference high quality images.

- Evaluation of the proposed automatic face image quality measure showed efficiency in reducing false non-match errors by removing low quality face images from a database (*i.e.,* operational reject option).

- Visual inspections of face images rank-ordered by the predicted quality values demonstrated the effectiveness of the approach in separating high quality face images (*e.g.,* frontal, uniform illumination, no occlusion) from low quality face images (*e.g.,* out-of-plane rotation, low resolution, occluded facial regions).

Lastly, the contributions of the longitudinal study on automatic face recognition in Chapter 4 are summarized as follows:

- First large-scale statistical analysis of the longitudinal effects of facial aging on the performance of automatic face recognition. The study involved two operational mugshot databases consisting of (i) 147,784 images of 18,007 subjects and (ii) 31,852 images of 5,636 subjects with a minimum of 4 mugshots per subject collected over an average of 8.5 and 5.8 years for the two databases, respectively.

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

- Mixed-effects regression models were used to analyze trends in genuine scores over time (*i.e.,* as subjects age) and quantify the subject-specific variability in the longitudinal trends of a large population of subjects. As such, estimates were provided for how many years of aging are tolerated by commercial face matchers before recognition errors are attributable to be expected. For example, we showed that a state-of-the-art face matcher operating at a threshold of 0.1% FAR can recognize 95% of the population until 10.5 years elapsed time between enrollment and query face images.

- Demographics (age, gender, race) and face image quality were shown to only marginally affect the longitudinal trends in genuine scores.

- A methodology for the longitudinal evaluation of face recognition performance was detailed which will ideally be conducted periodically to reevaluate state-of-the-art systems as robustness to facial aging continues to evolve.

5.2 Future Work

In conducting the studies on face recognition included in this dissertation, a number of areas for future work have been realized. This section concludes the dissertation by suggesting extensions to the work presented in the previous chapters that can be explored by researchers in automatic face recognition.

Template-based matching is still an open research problem, indicated by the recognition accuracies reported in Chapter 2 for matching collections of face media, as well as the current leaderboard[4] accuracies for the IJB-A face challenge [72]. Score-level fusion of all face samples in the collection, as was explored in Chapter 2, is not computationally efficient, especially for 1:N matching scenarios. Face representations which can extract information from multiple face samples to result in a single template are preferable. This template-to-template matching reduces comparisons to the same complexity as image-to-image matching

[4]https://www.nist.gov/programs-projects/face-challenges

while still leveraging the multiple face samples of a subject.

Research in automatic face image quality assessment is still in its infancy. While a very challenging problem due to the large facial variations that are possible, particularly in unconstrained scenarios, face image quality has many important operational applications. The work presented in Chapter 3 suggests the following next steps for face image quality.

- Face quality may need to be distinguished as three scenarios: (i) determining face vs. non-face (flagging face detection failures), (ii) assessment of the accuracy of face alignment, and (iii) given an aligned face image, now what is the quality? These three modules of a face image quality algorithm may allow for the integration of face matcher-dependent properties (*e.g.,* IPD, alignment errors) with more generalizable face image quality measures.

- A hierarchical prediction approach may improve the prediction accuracy. For example, face quality of an image could first be classified as low, medium, or high (where the bins are defined to be highly correlated with recognition performance), followed by regression within each bin for a fine-tuned ranking (useful for visual purposes and other ranking applications).

- The current image features extracted from a deep convNet [136] show promising results for face image quality. However, the deep convNet in [136] was trained for face *recognition* purposes, so the representation should ideally be robust to face quality factors. It would be desirable to retrain a deep convNet for prediction of face image quality, rather than identity.

- More extensive evaluation of face image quality measures in the context of face recognition performance are needed. A methodical evaluation of the pairwise quality factor may offer new insights.

Persistence (or permanence) of a biometric trait is one of two fundamental premises of biometrics (uniqueness being the other) [82]. Our systematic longitudinal study in Chapter 4

offered significant insights about the persistence property of automatic face recognition systems. The following related avenues of research could be pursued in future. (i) Development of a single face quality measure for mugshot images would be beneficial for longitudinal study. Incorporating individual face quality factors (*e.g.,* IPD and pose) into the mixed-effects regression model quickly increases the complexity and interpretation of the results. (ii) Longitudinal analysis could be conducted on different face cropping (particularly, pre-cropped images to exclude most of the hair region) to investigate the impact of changing hairstyle over time. (iii) The longitudinal capabilities of face recognition for children (0–18 years old) is still relatively unknown. Operational mugshot databases do not contain this population, and longitudinal face images of young children are difficult to obtain. Recognition of child face images is an important application for law enforcement agencies seeking to analyze digital media containing faces of exploited children (see the Child Exploitation Image Analysis (CHEXIA) face challenge[5]). (iv) The stability of impostor scores should be investigated, as recognition errors can also manifest in increased impostor similarity scores. A longitudinal study of impostor scores over time will help to quantitatively address questions related to the second fundamental premise of uniqueness, such as: Does the probability of false acceptance depend on the age of the two subjects in question? The hypothesis is that younger individuals are more likely to falsely match to other younger individuals because distinctive characteristics such as wrinkles and spots have not yet formed. Mixed-effects regression models applied to impostor scores may additionally be useful for location of duplicate identities in a large operational database of subjects. Lastly, the methodology detailed in Chapter 4 can and should be used to periodically reevaluate the longitudinal robustness of state-of-the-art face recognition systems.

[5]https://www.nist.gov/programs-projects/chexia-face-recognition

BIBLIOGRAPHY

Bibliography

[1] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross. Design and evaluation of photometric image quality measures for effective face recognition. *IET Biometrics*, 3(4):314–324, Dec. 2014.

[2] A. Adler and T. Dembinsky. Human vs. automatic measurement of biometric sample quality. In *Canadian Conf. on Electrical and Computer Engineering (CCECE)*, 2006.

[3] G. Aggarwal, S. Biswas, P. J. Flynn, and K. W. Bowyer. Predicting performance of face recognition systems: An image characterization approach. In *Proc. CVPR Workshops*, 2011.

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

[5] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. Quality measures in biometric systems. *IEEE Security Privacy*, 10(6):52–62, Nov. 2012.

[6] O. Arandjelovic and R. Cipolla. A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. In *Proc. ICCV*, 2007.

[7] S. R. Arashloo and J. Kittler. Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features. *IEEE Trans. Information Forensics and Security (TIFS)*, 9:2100–2109, Dec. 2014.

[8] A. Asthana, T.K. Marks, M.J. Jones, K.H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3D pose normalization. In *Proc. ICCV*, 2011.

[9] M. Ballantyne, R. S. Boyer, and L. Hines. Woody Bledsoe: His life and legacy. *AI Magazine*, 17(1):7–20, Spr. 1996.

[10] J. H. Barr, K. W. Boyer, P. J. Flynn, and S. Biswas. Face recognition from video: A review. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 26(05), 2012.

[11] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[12] M. D. Begg and M. K. Parides. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, 22(16):2591–2602, Aug. 2003.

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

[14] A. Bell and K. Jones. Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1):133–153, Jan. 2015.

[15] M. Bereta, P. Karczmarek, W. Pedrycz, and M. Reformat. Local descriptors in application to the aging problem in face recognition. *Pattern Recognition*, 46(10):2634–2646, Oct. 2013.

[16] L. Best-Rowden, S. Bisht, J. Klontz, and A. K. Jain. Unconstrained face recognition: Establishing baseline human performance via crowdsourcing. In *Proc. IJCB*, 2014.

[17] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Trans. Information Forensics and Security (TIFS)*, 9(12):2144–2157, Dec. 2014.

[18] L. Best-Rowden and Anil K. Jain. A longitudinal study of automatic face recognition. In *Proc. ICB*, 2015.

[19] L. Best-Rowden, B. Klare, J. Klontz, and A. K. Jain. Video-to-video face matching: Establishing a baseline for unconstrained face recognition. In *Proc. BTAS*, 2013.

[20] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding (CVIU)*, 113:750–762, 2009.

[21] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, and Y. M. Lui. FRVT 2006: Quo vadis face quality. *Image and Vision Computing*, 28(5):732–743, May 2010.

[22] S. Bharadwaj, M. Vatsa, and R. Singh. Can holistic representations be used for face biometric quality assessment? In *Proc. ICIP*, 2013.

[23] S. Bharadwaj, M. Vatsa, and R. Singh. Biometric quality: a review of fingerprint, iris, and face. *EURASIP Journal on Image and Video Processing*, 34, Jul. 2014.

[24] S. Bharadwaj, M. Vatsa, and R. Singh. Biometric quality: A review of fingerprint, iris, and face. *EURASIP Journal on Image and Video Processing*, 34(1), 2014.

[25] A. Blanton, K. C. Allen, T. Miller, N. D. Kalka, and A. K. Jain. A comparison of human and automated face verification accuracy on unconstrained image sets. In *Proc. CVPR Workshops*, 2016.

[26] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, 1999.

[27] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Analysis Machine Intelligence (PAMI)*, 25:1063–1074, Sep. 2003.

[28] M. Burge. IARPA Broad Agency Announcement: BAA-13-07, Janus Program. http://www.iarpa.gov/index.php/research-programs/janus/baa, Nov. 2013.

[29] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Proc. CVPR*, 2012.

[30] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Proc. CVPR*, 2010.

[31] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[32] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimedia*, 17(6):804–815, Apr. 2015.

[33] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. ECCV*, 2012.

[34] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proc. CVPR*, 2013.

[35] J. Chen, Y. Deng, G. Bai, and G. Su. Face image quality assessment based on learning to rank. *IEEE Signal Processing Letters*, 22(1):90–94, 2015.

[36] J. Cheney, B. Klein, A. K. Jain, and B. F. Klare. Unconstrained face detection: State of the art baseline and challenges. In *Proc. ICB*, 2015.

[37] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 23(6):681–685, Jun. 2000.

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

[39] Z. Cui, Wen Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *Proc. CVPR*, 2013.

[40] B. DeCann and A. Ross. Can a "poor" verification system be a "good" identification system? a preliminary study. In *Proc. WIFS*, 2012.

[41] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proc. ICSLP*, 1998.

[42] A. Dutta, R. Veldhuis, and L. Spreeuwers. A bayesian model for predicting face recognition performance using image quality. In *Proc. IJCB*, 2014.

[43] M. Erbilek and M. Fairhurst. A methodological framework for investigating age factors on the performance of biometric systems. In *Proc. Multimedia and Security*, 2012.

[44] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition, 2011.

[45] A. P. Founds, N. Orlans, W. Genevieve, and Craig I. Watson. NIST special database 32 - multiple encounter dataset II (MEDS-II). NIST Interagency Report 7807, Jul. 2011.

[46] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, Aug. 1997.

[47] X. Ge, J. Yang, Z. Zheng, and F. Li. Multi-view based face chin contour extraction. *Eng. Appl. Artif. Intel.*, 19(5):545–555, Aug. 2006.

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

[49] R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In *Proc. AMFG*, pages 255–263, 2005.

[50] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Proc. ICCV*, 2013.

[51] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *Proc. FGR*, 2008.

[52] R. Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA, June 2001.

[53] P. Grother. Face recognition vendor test 2002: Supplemental report. NIST Interagency Report 7083, Feb. 2004.

[54] P. Grother, J. R. Matey, E. Tabassi, G. W. Quinn, and M. Chumakov. IREX VI: Temporal stability of iris recognition accuracy. NIST Interagency Report 7948, Jul. 2013.

[55] P. Grother and M. Ngan. FRVT: Performance of face identification algorithms. NIST Interagency Report 8009, May 2014.

[56] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 29(4):531–543, Apr. 2007.

[57] P. J. Grother, G. W. Quinn, and P. J. Phillips. Multiple biometric evaluation (MBE) 2010: Report on the evaluation of 2D still-image face recognition algorithms. Interagency report 7709, NIST, 2010.

[58] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Trans. Information Forensics and Security (TIFS)*, 8(1):191–204, Jan. 2013.

[59] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *Proc. ICB*, 2013.

[60] R.-L. Hsu, J. Shah, and B. Martin. Quality assessment of facial images. In *Biometrics Symposium: Special Issue on Research at the Biometric Consortium Conference (BCC)*, 2006.

[61] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Proc. CVPR*, 2014.

[62] 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. Tech. Report 07-49, Univ. of Mass., Amherst, Oct. 2007.

[63] C. P. Huynh, A. Robles-Kelly, and E. R. Hancock. Shape and refractive index from single-view spectro-polarimetric images. *Int J. Comput. Vis.*, 101(1):64–94, 2013.

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

[65] Anil K. Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, Jan. 2016.

[66] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *Proc. ICCV*, 2015.

[67] F. Juefei-Xu, K. Luu, M. Savvides, T. D. Bui, and C. Y. Suen. Investigating age invariant face recognition based on periocular biometrics. In *Proc. IJCB*, 2011.

[68] H. I. Kim, S. H. Lee, and Y. M. Ro. Face image assessment learned with objective and relative face image qualities for improved face recognition. In *IEEE International Conference on Image Processing (ICIP)*, Sep. 2015.

[69] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Proc. CVPR*, 2008.

[70] B. Klare and A. K. Jain. Face recognition across time lapse: On learning feature subspaces. In *Proc. IJCB*, 2011.

[71] B. F. Klare, M.J. Burge, J.C. Klontz, R.W. Vorder Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Trans. Information Forensics and Security (TIFS)*, 7(6):1789–1801, Dec. 2012.

[72] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In *Proc. CVPR*, 2015.

[73] J. C. Klontz and A. K. Jain. A case study on unconstrained facial recognition using the boston marathon bombing suspects. Tech. Report MSU-CSE-13-4, Michigan State Univ., May 2013.

[74] S. Klum, H. Han, A. K. Jain, and B. Klare. Sketch based face recognition: Forensic vs. composite sketches. In *Proc. ICB*, 2013.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[76] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. ICCV*, 2009.

[77] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42:300–311, 1993.

[78] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 24(4), Apr. 2002.

[79] K.-C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *CVIU*, 99(3):303–331, 2005.

[80] K.-C. Lee and D. Kriegman. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In *Proc. CVPR*, volume 1, pages 852–859, 2005.

[81] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-PEP for video face recognition. In *Proc. ACCV*, 2014.

[82] S. Z. Li and A. K. Jain, editors. *Handbook of Face Recognition*. New York: Springer, 2 edition, 2011.

[83] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE Trans. Information Forensics and Security (TIFS)*, 6(3):1028–1037, Sep. 2011.

[84] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 35:1193–1205, May 2013.

[85] S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study on large-scale unconstrained face recognition. In *Proc. IJCB*, 2014.

[86] Y. Lin, G. Medioni, and J. Choi. Accurate 3D face reconstruction from weakly calibrated wide baseline images with profile contours. In *Proc. CVPR*, 2010.

[87] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs. Face verification across age progression using discriminative methods. *IEEE Trans. Information Forensics and Security (TIFS)*, 5(1):82–91, Mar. 2010.

[88] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Processing*, 11(4):467–476, Aug. 2002.

[89] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *Proc. CVPR*, pages 340–345, 2003.

[90] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with gaussianface. http://arxiv.org/abs/1404.3840, Apr. 2014.

[91] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips. A meta-analysis of face recognition covariates. In *Proc. BTAS*, 2009.

[92] A. M. Martinez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center, University of Barcelona, 1998.

[93] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):123–164, Nov. 2004.

[94] E. Mostafa, A. Ali, N. Alajlan, and A. Farag. Pose invariant approach for face recognition at distance. In *Proc. ECCV*, 2012.

[95] J. M. Neuhaus and J. D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645, Jun. 1998.

[96] U.S. Department of Homeland Security. Biometric standards requirements for us-visit: Version 1.0. https://www.dhs.gov/xlibrary/assets/usvisit/usvisit_biometric_standards.pdf, Mar. 2010.

[97] National Institute of Standards and Technology (NIST). Face homepage. http://face.nist.gov, Jun. 2013.

[98] E. G. Ortiz and B. C. Becker. Face recognition for web-scale datasets. *Comput. Vis. Image Und.*, 118(0):153 – 170, Jan. 2013.

[99] C. Otto, H. Han, and A. K. Jain. How does aging affect facial components? In *ECCV WIAF Workshop*, 2012.

[100] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. An overview of research on facial aging using the FG-NET aging database. *IET Biometrics*, May 2015.

[101] U. Park and A. K. Jain. Face recognition in video: Adaptive fusion of multiple matchers. In *Proc. CVPR*, pages 1–8, 2007.

[102] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 32(5), May 2010.

[103] J. Phillips. Video challenge problem multiple biometric grand challenge preliminary results of version 2. In *MBGC 3rd Workshop*, December 2009.

[104] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, S. Cheng, M. N. Teli, and H. Zhang. On the existence of face quality measures. In *Proc. BTAS*, pages 1–8, Sep. 2013.

[105] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, and W. Worek. Preliminary face recognition grand challenge results. In *Proc. FG*, 2006.

[106] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002: Evaluation report. NIST Interagency Report 6965, Mar. 2003.

[107] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 22(10), Oct. 2000.

[108] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 32:831–846, 2010.

[109] N. Poh and J. Kittler. A unified framework for biometric expert fusion incorporating quality measures. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 34(1):3–18, Jan 2012.

[110] N. Poh, J. Kittler, C.-H. Chan, and M. Pandit. Algorithm to estimate biometric performance change over time. *IET Biometrics*, 4(4):236–245, Dec. 2015.

[111] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages and Computing*, 20:131–144, 2009.

[112] H. T.F. Rhodes. *Alphonse Bertillon, father of scientific detection*. London: George G. Harrap & Co., 1956.

[113] K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Proc. FGR*, 2006.

[114] A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. New York: Springer, 2006.

[115] J. Roth, Y. Tong, and X. Liu. Unconstrained 3D face reconstruction. In *Proc. CVPR*, 2015.

[116] H. Sellahewa and S. A. Jassim. Image-quality-based adaptive face recognition. *IEEE Transactions on Instrumentation and Measurement*, 59(4):805–813, Apr. 2010.

[117] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, Dec. 2003.

[118] J. D. Singer and J. B. Willett, editors. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford Univ. Press, Inc., 2003.

[119] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, Mar. 1987.

[120] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. https://arxiv.org/abs/1502.00873, Feb. 2015.

[121] Y. Sun, X. Wang, and X Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014.

[122] Y. Sun, X Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014.

[123] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proc. CVPR*, 2015.

[124] E. Tabassi, M. A. Olsen, A. Makarov, and C. Busch. Towards NFIQ II lite: Self-organizing maps for fingerprint image quality assessment. NIST Interagency Report 7973, Dec. 2013.

[125] Elham Tabassi and Charles L. Wilson. A novel approach to fingerprint image quality. In *IEEE International Conference on Image Processing (ICIP)*, 2005.

[126] E. Taborsky, K. Allen, A. Blanton, A. K. Jain, and B. F. Klare. Annotating unconstrained face imagery: A scalable approach. In *Proc. ICB*, 2015.

[127] Y. Taigman and L. Wolf. Leveraging billions of faces to overcome performance barriers in unconstrained face recognition. http://arxiv.org/abs/1108.1122, Aug. 2011.

[128] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014.

[129] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *Proc. CVPR*, 2015.

[130] K. T. Taylor. *Forensic Art and Illustration*. Boca Raton, FL: CRC Press, 2000.

[131] D. Thomas, K. W. Bowyer, and P. J. Flynn. Multi-frame approaches to improve face recognition. In *IEEE Workshop on Motion and Video Computing*, pages 19–19. IEEE, 2007.

[132] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, Winter 1991.

[133] U. Uludag and A. K. Jain. Attacks on biometric systems: A case study in fingerprints. In *Proc. SPIE*, 2004.

[134] R. van der Leeden, F. Busing, and E. Meijer. Bootstrap methods for two-level models. In *Multilevel Conf.*, 1997.

[135] P. Viola and M. J. Jones. Robust real-time face detection. *Int J. Computer Vision*, 57(2):137–154, May 2004.

[136] D. Wang, C. Otto, and A. K. Jain. Face search at scale: 80 million gallery. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, PP(99), Jun. 2016.

[137] H. Wang, B. Kang, and D. Kim. PFW: A face database in the wild for studying face identification and verification in uncontrolled environment. In *Proc. ACPR*, 2013.

[138] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *Proc. CVPR*, pages 1–8, 2008.

[139] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, Apr. 2004.

[140] C. I. Watson. NIST mugshot identification database. http://www.nist.gov/srd/nistsd18.htm, 1994.

[141] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. CVPR*, 2011.

[142] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *Proc. CVPR Workshops*, pages 74–81, Jun. 2011.

[143] D. Yadav, N. Kohli, P. Pandey, R. Singh, M. Vatsa, and A. Noore. Effect of illicit drug abuse on face recognition. In *Proc. WACV*, 2016.

[144] Neil Yager and Ted Dunstone. The biometric menagerie. *IEEE IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 32(2):220–230, Feb. 2010.

[145] M. Yang, P. Zhu, L. Van Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *Proc. FG*, 2013.

[146] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. http://arxiv.org/abs/1411.7923, Nov. 2014.

[147] D. Yi, S. Liao, and S. Z. Li. Learning face representation from scratch. arXiv:1411.7923v1, 2014.

[148] J. Yi, R. Jin, S. Jain, and A. K. Jain. Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach. *First AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2013.

[149] S. Yoon and A. K. Jain. Longitudinal study of fingerprint recognition. *Proc. National Academy of Sciences (PNAS)*, 112(28):8555–8560, Jul. 2015.

[150] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Tech. Report MSR-TR-2010-66, Microsoft Research, Jun. 2010.

[151] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? Tech. report, Face++, Megvii Inc., Jan. 2015.

[152] S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Trans. Image Processing*, 13(11):1491–1506, 2004.

[153] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. CVPR*, 2012.