Computational Statistics and Data Analysis 67 (2013) 236–247

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Introduction to face recognition and evaluation of algorithm performance

G.H. Givens [a,*], J.R. Beveridge [b], P.J. Phillips [c], B. Draper [b], Y.M. Lui [b], D. Bolme [d]

[a] Department of Statistics, Colorado State University, Fort Collins, CO, 80523, USA
[b] Department of Computer Science, Colorado State University, Fort Collins, CO, 80523, USA
[c] National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD, 20899, USA
[d] Oak Ridge National Laboratory, PO Box 2008 MS6075, Oak Ridge, TN, 37831-6075, USA

ARTICLE INFO

Article history:
Received 12 July 2012
Received in revised form 27 February 2013
Accepted 31 May 2013
Available online 6 June 2013

Keywords:
Face recognition
Biometrics
Computer vision
Generalized linear mixed model
GLMM

ABSTRACT

The field of biometric face recognition blends methods from computer science, engineering and statistics, however statistical reasoning has been applied predominantly in the design of recognition algorithms. A new opportunity for the application of statistical methods is driven by growing interest in biometric performance evaluation. Methods for performance evaluation seek to identify, compare and interpret how characteristics of subjects, the environment and images are associated with the performance of recognition algorithms. Some central topics in face recognition are reviewed for background and several examples of recognition algorithms are given. One approach to the evaluation problem is then illustrated with a generalized linear mixed model analysis of the Good, Bad, and Ugly Face Challenge, a pre-eminent face recognition dataset used to test state-of-the-art still-image face recognition algorithms. Findings include that (i) between-subject variation is the dominant source of verification heterogeneity when algorithm performance is good, and (ii) many covariate effects on verification performance are 'universal' across easy, medium and hard verification tasks. Although the design and evaluation of face recognition algorithms draw upon some familiar statistical ideas in multivariate statistics, dimension reduction, classification, clustering, binary response data, generalized linear models and random effects, the field also presents some unique features and challenges. Opportunities abound for innovative statistical work in this new field.

1. Introduction

Over the past twenty years, face recognition has matured from a nascent multidisciplinary scientific research topic to a widely deployed commercial technology. Applications range from checking identities at international borders and searching mugshots in national criminal databases to tagging faces in photos on social media websites. Face recognition was considered important enough to be addressed in a major report from the National Academy of Sciences (NAS) on biometrics, which summarized the state of the field, its importance, and the critical open questions (Pato et al., 2010).

The performance of face recognition systems depends on the conditions under which the face images are taken. In the most recent US Government evaluation conducted in 2010, the identification rate was 93% for the best commercial systems on a database of 1.6 million mugshots (Grother et al., 2010). This represents a significant improvement of the technology over the past two decades. US Government evaluation of face recognition technology began in 1993 (Phillips et al., 2000),

* Corresponding author. Tel.: +1 970 491 6402; fax: +1 970 491 7895.
E-mail address: geof@lamar.colostate.edu (G.H. Givens).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.025

and over that time the error rate has decreased by a factor of 272 (a reduction in the false reject rate from 0.79 to 0.0029, at a false positive rate of 1 in 1000) (Phillips et al., 2012). In Moore's law terms, the false reject rate has halved every 2 years. This dramatic decrease in error rates, however, is limited to frontal face images of motionless subjects acquired in either a mobile studio or as mugshots. Recognition performance is much worse at the other end of the spectrum of imaging conditions, such as for personal photos taken with cell phone cameras or video footage from low cost security systems.

The statistical and scientific challenge is to understand what factors improve or degrade the performance of face recognition algorithms, and to be able to predict performance on a novel set of images. The Good, Bad, and Ugly Face Challenge Problem (GBU) presented by NIST in 2009 encapsulates a wide range of performance under varying imaging conditions (Phillips et al., 2012). The GBU was created to encourage the development of robust still-image face recognition algorithms and to study the reasons for the rapid drop off in performance for frontal images acquired under difficult conditions. The GBU consists of three strata, or partitions (Good, Bad, and Ugly). The challenge problem asks: given two face images, are they images of the same face? On the Good partition, the benchmark algorithm detects same-face pairs at 98% with a false alarm rate (FAR), also called a false positive rate, of 1 in a 1000. The corresponding rate for the Bad partition is 80%, and 15% on the Ugly partition. Table 1 lists the performance of six additional algorithms on the GBU data.

Because all images were acquired with a high-end consumer camera and the faces are nominal frontal and large sized, one might presume that verification performance would be relatively consistent and strong. However, the results in Table 1 highlight the challenges of developing robust face recognition algorithms. One of the keys to advancing automatic face recognition is understanding the sources of variability that impact face recognition performance.

In face recognition, research directions for advancing face recognition are based on the prevailing conventional wisdom that a small set of covariates substantially explains algorithm performance. Some of the covariates, like illumination direction, are based on formal theories. For others, such as subject age and facial expression, the effects have been quantified from repeated experimentation. Finally, some have been discovered from practical experience during field implementations.

Over the past decade we have investigated over 60 of the most commonly proposed covariates using a variety of statistical techniques and found convincing evidence that changes in these factors can affect algorithm performance (Givens et al., 2004; Beveridge et al., 2005b, 2009b; Lui et al., 2009b). We include an analysis here that reconfirms and consolidates these findings. However, our work identifies three inadequacies in the face recognition community's understanding of how the critical factors relate to recognition performance and how to estimate such effects.

The first problem stems from our key new finding, presented below, that most of these covariate effects are universal in the sense that they do not interact much with image 'difficulty'. Thus, the fundamental assumption that if a covariate is predictive of performance, then it should be able to separate difficulty strata (as for GBU) is not valid. Moreover, efforts to develop an effective measure of image 'quality' have thus far failed – despite enormous effort – to find a way to accurately predict the difficulty of recognizing a particular image, e.g., predicting the GBU partition membership. If it were possible to identify and keep only the 'highest quality' images, one would have a high performing system. Unfortunately, existing quality measures, including ours, do not work.

Second, contrary to conventional wisdom, the covariate factors do not explain a large portion of the observed variation in algorithm performance, and the existence or direction of effects of certain covariates can be surprising. Most variation in performance remains unexplained. The current analysis confirms this result.

Third, there are methodological challenges inherent in the statistical analysis of recognition data. For example, although performance is best modeled by studying image pairs rather than individual images, face recognition cannot be modeled as a forced choice categorization/classification task. As we explain in the next section, every person would be their own category, and we would have at most a few samples of each—often only one.

Having found that key components of the existing conventional wisdom have little empirical support or are wrong, now is the time for the research community, including statisticians, to build a new theory. A better understanding of the factors that affect performance and the most effective methods for determining those effects is essential. Face recognition straddles both scientific and societal issues, thus successful use of face recognition systems and development of the next generation of face recognition algorithms has fundamental scientific merit and broad societal impact.

2. Motivation

The National Academy of Sciences report on biometrics (Pato et al., 2010) specifically points to the importance of biometric evaluation, noting that testing and evaluation is one of "the unsolved fundamental problems and research opportunities related to biometric systems" (p. 116). It also stresses that

> Methods used successfully for the study and improvement of systems in other fields (for example, controlled observation and experimentation on operational systems guided by scientific principles and statistical design and monitoring) should be used in developing, maintaining, assessing, and improving biometric systems (p. 123).

The recognition of this need is longstanding, and the progress to date for face recognition is notable but not sufficient. With respect to experimental design, significant contributions include the development of common datasets for performance comparison (Phillips et al., 2000) including sequestered data used to evaluate vendor supplied algorithms (Phillips et al., 2002, 2010), along with standardized performance measures (Phillips et al., 2000) and baseline algorithms (Phillips et al., 2005). Notwithstanding these advances, let us specifically highlight the call for better statistical design and monitoring in

Table 1
Verification rates (percent) at 0.1% false accept rate for seven algorithms on the GBU dataset partitions. From top to bottom, citations for these methods are: (Phillips et al., 2012; Lui et al., 2012; Pinto et al., 2009a; Baudat and Anouar, 2000; Phillips et al., 2012; Ahonen et al., 2006; Beveridge et al., 2005a).

Algorithm	Good	Bad	Ugly
Fusion of algorithms from face recognition vendor test 2006	98	80	15
Cohort linear discriminant analysis	84	48	11
Gabor wavelets ("V1-like")	73	24	6
Kernel generalized discriminant analysis	69	29	5
Local region principal component analysis	64	24	7
Local binary pattern	51	5	2
Elastic bunch graph matching	50	8	2

the context of controlled and experimental operational systems, and consider possible reasons why it has proven difficult to apply standard statistical methods to characterize face recognition performance.

One reason for this difficulty is that face recognition is not a traditional classification problem. The first difficulty that arises if one equates face recognition to traditional classification is the nature of the class label set. While it is true that a face recognition system may be asked to return a name given an image of a person, it is equally true that to be of practical value, the system must be able to rapidly extend the set of people it can recognize. If face recognition is solved using a classifier to map images to people's names, then each new person introduced to the system requires introducing a new class label and subsequent reconstruction of the classifier. This difficulty is compounded by the standard protocol for adding a new person to a face recognition system, which is to provide one face image for each person. Traditional approaches to classifier construction have little to offer when the label set, i.e., the number of people, can be many thousands and only a single example of each class is available for construction of the classifier.

For the reasons just given, as well as others, face recognition is formalized as the task of supplying a similarity score between pairs of images where higher similarity implies a greater confidence that the two images are pictures of the same person. Great attention is paid to the means by which this similarity score is computed, almost certainly involving training over large sets of labeled images of people. However, once constructed, the algorithm for computing the similarity score between pairs of face images is intended to be of general value over all people; not merely those over which it was trained. The requirement to avoid specialization to a limited set of people goes even farther in most government sponsored evaluation protocols (Phillips et al., 2002, 2010, 2012), where training a face recognition system on the same set of people on which it is tested is forbidden.

Just as face recognition may not be expressed as a classification problem over a fixed, closed set of people, it is not amenable to cluster analysis either. Even in the unlikely event that clusters correspond to individual people, no amount of clustering over a fixed training set is going to enable a system to recognize a new person based upon a single image of that new person.

Thus, many statistical methods designed for analysis and performance evaluation of traditional classification or clustering tasks do not apply directly to face recognition, at least so long as the labels are assumed to represent people. However, it is important to understand that we can recast the face recognition task in a manner where it resembles a binary classification task. Specifically, similarity scores between *pairs* of images are thresholded and divided into two groups: those said by the face recognition system to belong to the same person and those said to belong to different people.

Although our re-posing of the problem in this manner may seem elementary, it is in fact an important change that complements one of the most common formalizations of face recognition as a verification task (defined below) and the formulation lies at the heart of both our past work (Givens et al., 2004; Beveridge et al., 2009a,b) and the analysis presented here. However, there are complications with this approach. For example, image pairs are typically correlated since the same people (or even the same images) often appear in multiple pairings. Algorithm training and performance evaluation (including data generation and collection) must also be consistent with this viewpoint. In summary, this paper will show how we organize performance and covariate data around the binary classification of image pairs and then adapt a generalized linear mixed model to extract information from experimental data.

3. Face recognition: a review

Here we define basic concepts in face recognition and describe a very simple algorithm. More advanced techniques are also briefly discussed in order to provide a glimpse at the variety of approaches that have been proposed.

3.1. Algorithm fundamentals

A face recognition algorithm A may be defined as a mapping from a pair of face images to a real number s_A representing some measure of the similarity (equivalently, distance) between those images. Formally,

$$A : \mathbb{I}_T \times \mathbb{I}_Q \to \Re \tag{1}$$

where it is common to describe the first image as being drawn from a set of target images $\mathcal{I}_T$ and the second from a set of query images $\mathcal{I}_Q$. This definition gives rise to a data structure called the similarity matrix:

$$S_A = \{A(t, q)\ \forall t \in \mathcal{I}_T, q \in \mathcal{I}_Q\} = [s_A(t, q)]. \tag{2}$$

A similarity score $s_A(t, q)$ is said to be a match score when the same person is pictured in images q and t and a non-match score when the pictures are of different people. For most typical evaluation tasks, an algorithm is run on all pairwise combinations of images from $\mathcal{I}_T$ and $\mathcal{I}_Q$, and all subsequent questions about algorithm performance are addressed by studying the similarity matrices.

Recognition, and hence performance evaluation, may be expressed in terms of two distinct tasks: identification and verification. In an identification task, a query image $q \in \mathcal{I}_Q$ is compared to a set of (target) images in a gallery $\mathcal{I}_T$. The gallery is then sorted by similarity with respect to q and either the most similar or a small set of the most similar target images are returned as a presumed match or match ranking for q.

In a verification task, a person presents themselves to a system that already has a stored (target) image of his/her face. The system acquires a new (query) image of the person and compares the resulting similarity $s_A(t, q)$ to an acceptance threshold. If the similarity exceeds that threshold then the system confirms that the person is who s/he claims; otherwise the person's claimed identity is rejected. The acceptance threshold is chosen using a training dataset where true identities are known. The threshold is usually set to achieve a pre-determined false verification (i.e., false acceptance) rate. The rate of correct verification will depend on the extent to which the distributions of match and non-match scores overlap.

Most of the major face recognition evaluations carried out in the past decade have concentrated on the verification task. One reason is that the outcome of a verification test depends only on the similarity score $s_A(t, q)$ and the threshold for acceptance. In contrast, identification performance depends upon the other people and images in the gallery. Consequently, results for small galleries typically do not scale to problems involving more people and images. Accordingly, the development of sophisticated statistical methods for evaluation of identification performance is an open field for new research.

3.2. Algorithm basics and a modern benchmark

The algorithm commonly credited for setting off an explosion of activity in the early 1990s was developed by Turk and Pentland (1991) using principal component analysis (PCA) applied to face images in a manner first suggested by Sirovich and Kirby (1987) as well as O'Toole et al. (1988). Because of its seminal importance, we begin with this algorithm in Section 3.2.2.

3.2.1. Face localization and image preprocessing

All face recognition algorithms must first establish a spatial correspondence between the faces in two images before they can proceed to measure similarity. This is commonly done by first locating the eyes in both images. Then the image may be adjusted by positioning, scaling, and rotating the face so that the eyes always fall at exactly the same position. The result of localization is typically a new smaller image, called a face image chip, created by re-sampling pixels in the original, such that eyes, mouth, and so forth appear at approximately the same pixel coordinates in every face chip.

Beyond this geometric localization, a variety of additional preprocessing steps may be introduced that modify pixel values. For example, it is common to zero out pixels that fall outside an oval defined by the shape of the face. This step removes background clutter, hair, etc. Illumination normalization is another common step. Consider that lighting can have a much larger effect on pixel values than the identity of the person being imaged. For example, an image of a face lit from the left will produce very different pixel values compared to the same face lit from the right. A variety of algorithms exist for illumination normalization, such as the Self Quotient Image approach which approximates the albedo of the face and adjust pixel values accordingly (Moses et al., 1994). This greatly reduces harsh lighting artifacts.

It is important to understand that the relative success of preprocessing profoundly influences the recognition performance of face recognition algorithms. Clearly, if the face is not found in the image, no amount of cleverness in the comparison step will ever overcome this initial failure. The impacts of preprocessing are critical in less extreme cases, too. If factors such as lighting make images of the same person sufficiently different in appearance and preprocessing is unable to suppress this difference, then face recognition will perform poorly.

3.2.2. A canonical PCA algorithm

Having completed geometric normalization and image preprocessing, the simplest approach begins with a standard PCA decomposition. Consider m gray scale images where each pixel corresponds to a scalar between 0 and 255. Let the ith image be expressed as a column vector x_i representing the 'unrolled' face chip, defined as the vector obtained by concatenating pixel rows. Construct a data matrix X with one column for each training image:

$$X = \begin{bmatrix} x_1 - \mu & x_2 - \mu & \cdots & x_m - \mu \end{bmatrix} \quad \text{where } \mu = \frac{1}{k}\sum_{i=1}^{m} x_i. \tag{3}$$

The principal components are eigenvectors of the sample covariance matrix $X^T X/(m-1)$. The PCA recognition algorithm selects a subset of eigenvectors, say the first k, to create a linear subspace in which to compare images. Typically, k might be chosen in the range of 50–500, while an image chip vector might contain upwards of one hundred thousand pixels or more.

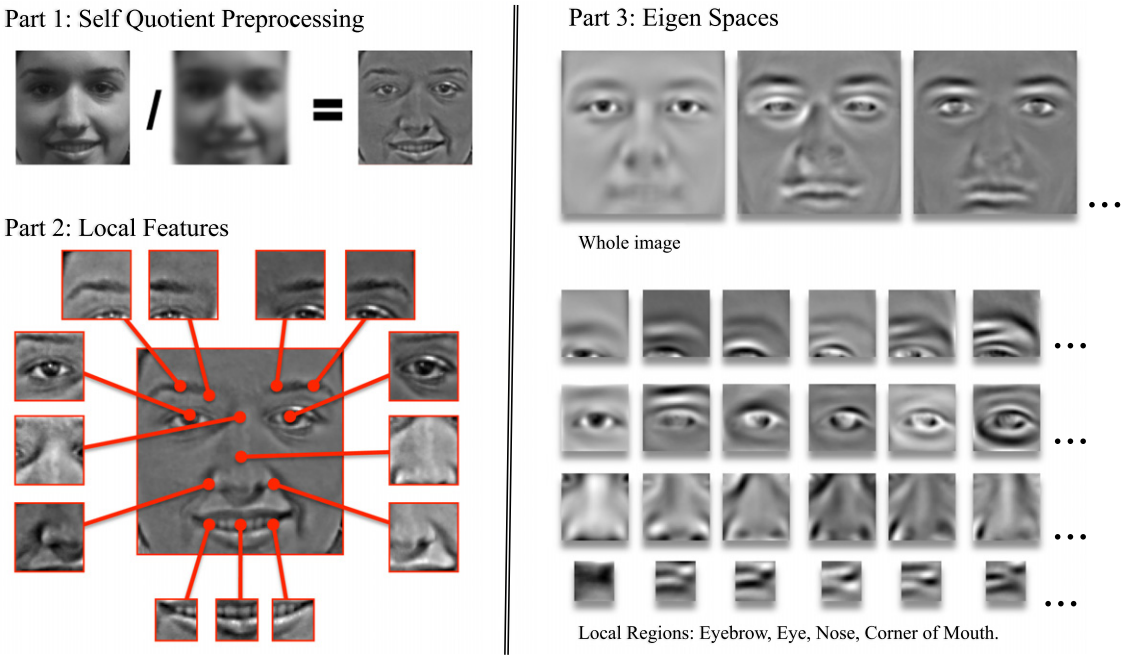

Fig. 1. Visual overview of key steps in the Local Region PCA algorithm. In part 1, images are normalized for illumination variations using a sequence of processing techniques including Self Quotient Image preprocessing (Moses et al., 1994). In part 2, the image is divided into 14 regions including one that is the entire face. In part 3, PCA is applied to each region. In the figure, the first 6 principal components are shown for 5 regions (rows) including the whole face.

Define M to be a matrix whose rows contain the first k principal components. One variant of the canonical PCA algorithm defines similarity between a target image t and query image q as the negative Euclidean distance between the two as measured in the PCA subspace:

$$s_A(t,q) := -d(t,q), \qquad d(t,q) = \|\acute{t} - \acute{q}\|, \quad \acute{t} = M(t-\mu), \acute{q} = M(q-\mu). \tag{4}$$

A host of refinements have been proposed for this basic technique, including alternative measures of similarity, more complex ways of selecting PCA dimensions, and so forth. In general, the performance of the canonical PCA algorithm is not competitive with more modern approaches to face recognition.

3.2.3. Local region PCA

Local Region PCA (LRPCA), a refined implementation of PCA, is among the best performing algorithms in the public domain (i.e., not proprietary or commercially developed) (Phillips et al., 2012). (See Table 1 for performance on GBU.) This approach attempts to extracting feature over local regions of the face (Pentland et al., 1994; Wiskott et al., 1997; Ahonen et al., 2006). Like other well-performing algorithms, LRPCA requires significant image preprocessing and tuning to be successful. The three major parts of the LRPCA algorithm described here are illustrated in Fig. 1, including image preprocessing, local regions definition, and PCA decomposition.

After preprocessing, the algorithm illustrated in Fig. 1 splits the face into 13 local regions at the expected locations of the major facial landmarks, such as the eyes, eyebrows, nose, and mouth, with the entire face treated as a 14th 'region'. These regions are then processed independently to produce a set of eigenvectors for each region. Setting $k = 250$ for each of the 14 regions reduces whole images to vectors with only $14 \times 250 = 3500$ elements. Although this seems large, consider that the full face image chip may contain $1024 \times 1024 = 1,048,576$ pixels.

The final step is to re-weight the values in the resultant vector to emphasize the more discriminative features. In general, it has been found that whitening the covariance matrix for each region using the variance estimates obtained as part of the PCA analyses improves the performance of PCA based algorithms. LRPCA goes a step further by reweighting each of the 3500 elements by the ratio of between-class to within-class variance, which emphasizes values that better discriminate between people. Once the values have been re-weighted, similarity between vectors is defined as Pearson's correlation.

3.3. Some advanced approaches

One of the most ubiquitous extensions to PCA is the adoption of linear discriminant analysis (Fisher, 1936; Etemad and Chellappa, 1997). Yet another approach to finding a linear subspace in which recognition performs best is independent component analysis (Draper et al., 2003).

Beyond linear classifiers, in the past decade there has been attention paid to support vector machine classifiers in conjunction with the reproducing kernel Hilbert space for face recognition. As a result, many of the more recent algorithms combine kernel methods, generalized linear discriminants and SVMs (Liu, 2006; Pinto et al., 2009b).

It is also common to transform image pixels into an alternative feature space where recognition will proceed more easily. Alternative transformations proposed include Gabor wavelets (Wiskott et al., 1997; Liu, 2006; Pinto et al., 2009b), local binary patterns (Ahonen et al., 2006), correlation filters (Kumar et al., 2006) and Grassmann manifolds (Lui and Beveridge, 2008).

Presentation of many algorithms from very early ideas to state-of-the-art is beyond the scope of this paper. Below, however, we illustrate the diversity of modern methods by describing an approach that is far removed from the PCA-related methods reviewed above.

One of the major challenges in face recognition is illumination variation. Fortunately, the theory of illumination and its connection with non-linear manifolds is well developed. From the illumination cone principle (Belhumeur and Kriegman, 1998; Basri and Jacobs, 2003), it is known that a set of convex objects under a fixed pose with a Lambertian reflectance surface forms a convex polyhedral in $\mathbb{R}^n$. An image can then be relighted using a Bayesian model under the Lambertian constraint (Lui et al., 2009a). Formally, a generic image can be represented from the illumination basis as $I = Bs + e(s)$ where I is the image, B is the illumination basis, s is the illumination coefficient, and $e(s)$ is the error term. The relighted images can be estimated using maximum a posterior estimation as:

$$B = \left(\frac{\hat{I} - \mu_B \hat{s} - \mu_e}{\sigma_e^2 + \hat{s}^T C_B \hat{s}} \right) C_B \hat{s} + \mu_B \tag{5}$$

where $\hat{I}$ is a novel image, $\hat{s}$ is the estimated lighting coefficient, μ_e is the mean error from $\hat{s}$, σ_e is the standard derivation from $\hat{s}$, and μ_B and C_B are the mean basis and covariance basis from a training set.

The relighted images B form the basis of an illumination cone. Since illumination cones reside in a vector space, they endow a geometric structure on a Stiefel manifold (Edelman et al., 1999). A Stiefel manifold $\mathcal{V}_{n,p}$ consists of a set of p-dimensional subspace of $\mathbb{R}^n$, i.e. $\{Y \in \mathbb{R}^{n \times p} : Y^T Y = I_p\}$. As such, one may form a tangent space centered at the relighted images on the Stiefel manifold. The Stiefel manifold is also endowed with a smoothly varying inner product on a tangent space in which the inner product g is a canonical metric such that $g_x : T_x \mathcal{V}_{n,p} \times T_x \mathcal{V}_{n,p} \longrightarrow \mathbb{R}$ at any given point x in $\mathcal{V}_{n,p}$. It is consequently possible to build a face recognition algorithm that takes a single query image q and target image t, relights each in the manner just described, and defines similarity $s_A(t, q)$ in terms of projections in the tangent spaces associated with the Stiefel manifold (Lui et al., 2009a). Such an algorithm performs well on some known benchmarks, but so far its computational demands leave it at the fringe of current practice. It is mentioned here to provide some feel for extent to which face recognition algorithms are starting to incorporate sophisticated mathematical modeling.

For more details on automatic face recognition, see the surveys by Zhao et al. (2003) and Chellappa et al. (2010).

4. Performance evaluation

Independent government-sponsored evaluation efforts play a major role in benchmarking and characterizing biometrics technology in general. Often, these evaluations are run by the National Institute of Standards and Technology (NIST). These include major efforts in the recognition of faces, fingerprints, speaker/voice, and more recently irises (Greenberg and Martin, 2009; Grother et al., 2012; Phillips et al., 2010). The most prominent of these large-scale face recognition studies since 1994 are 'challenge' problems (Phillips et al., 2000; Grother et al., 2003; Phillips et al., 2006, 2010, 2009). Two of the most common challenges are 'open challenge' problems and more formal tests using sequestered data. In open challenge problems, participants have open access to the data and typically carry out experiments following a common protocol. The common protocol allows for comparisons across research groups. In contrast, formal test participants submit self-contained systems that are run independently on data not seen by the system developers themselves. Together, these studies document a three order of magnitude improvement in face recognition technology, and provided in-depth analysis of performance to those responsible for making decisions about the deployment of face recognition technology.

4.1. The Good, the Bad and the Ugly face challenge

NIST has released a new challenge problem: the Good, the Bad and the Ugly Challenge Problem (GBU) (Phillips et al., 2012), available through the NIST Face and Ocular Challenge Series website http://www.nist.gov/itl/iad/ig/focs.cfm. One motivation for the release of the GBU Challenge Problem was that reliable recognition of cooperative people standing in front of, and looking at, a high-quality camera was, and still is, a difficult problem under some circumstances. Indeed, the purpose of the study presented below was to begin the process of better understanding what characterizes these difficult situations.

The 6340 face image data for GBU were acquired using a high quality consumer-level digital camera. The photographs were posed with a person standing and looking at a camera mounted at eye level. Images were acquired with ambient lighting in hallways and outdoors. The outdoor images were acquired with a variety of lighting conditions and backdrops. All images were acquired during a 9 month period.

The GBU image data was divided into three partitions constructed to present a graduated level of difficulty. In each partition, there are 2170 images of 437 people. There is the same number of images of each person in each partition. Because of the partition design controls, the differences in performance among the three partitions are a result of how a face is presented in each image. Also, changes in pose and face aging do not affect performance.

The images in each partition were further split into disjoint target and query sets. Algorithm performance was expressed in terms of similarity scores for all pairs of query and target images. The partitions were constructed from a larger initial set of images, using similarity scores constructed from fusing three top-performing algorithms from the Face Recognition Vendor Test (FRVT) 2006 (Phillips et al., 2010, 2012). At a false accept rate of 0.001, the FRVT 2006 fusion algorithm achieved a verification rate of 0.98, 0.80, and 0.15 on the Good, the Bad, and the Ugly partitions respectively.

5. GBU performance evaluation

We limit consideration here to three questions: Under what conditions does an algorithm perform well/poorly, how much performance variation is attributable to specific covariates, and what are the covariate effects? Fundamentally, this is a predictor/response investigation.

5.1. Outcomes

Our unit of analysis was a match pair, and our response variable was $I(s_A(q, t) > \tau)$ where τ is the threshold yielding a false acceptance rate of 0.001 (i.e., the 0.999 percentile of the non-match scores) and $I(X)$ is 1 if X is true and 0 otherwise. The alternative response variable was the similarity score, s_A.

For modern face recognition algorithm, this is a poor option. Some algorithms produce similarity scores that are a mix of ratings and continuous numbers: for example there might be categories of 'terrible', 'bad', 'possible', and numerical scores above that. Second, similarity scores have very different distributions depending on whether the image pair comprises two images of the same person (a 'match score') or two different people (a 'non-match score'). Because of these reasons, similarity scores between algorithms are incomparable. Non-match scores tend to have broad and very heavy-tailed distributions compared to match scores. Moreover, distributions of both types of scores can be markedly skewed, especially for non-matches. Third, because the calculation of similarity scores can be highly esoteric, one should have no confidence that some sort of score normalization would produce values for which incremental changes scale in any sensible way with algorithm performance. Our approach of limiting consideration to trials on match pairs discards the vast majority of similarity scores; however, it allows us to pose interpretations in terms of the estimated probability of successful verification, which is directly relevant for algorithm evaluators and end users.

5.2. Covariates

Two sets of covariates are associated with the similarity matrix. The first set contains measurements associated with a single image, such as imaging setting (an environment variable), gender (subject), or focus (image). The second set contains covariates associated with image pairs, such as setting change (e.g., between outdoor and indoor), time lapse between images, and comparative focus. The covariates available in our dataset are listed in Table 2.

The lighting variable is an estimate of how the face is illuminated using models that connect the theory of illumination of convex objects with the empirical representation of non-linear manifolds (Beveridge et al., 2010). Location is the exact location where images were acquired: one of 10 possible sites used by the University of Notre Dame to collect these data. Edge density is the average edge magnitude of a Sobel operator applied to the chip; this and the focus measure are reviewed by Beveridge et al. (2009b, 2010). The remaining covariates are mostly self-explanatory.

5.3. The Good, Bad and Ugly partition labels

As noted above, the dataset is partitioned into three portions. In a rough sense, the image pairs are grouped in a manner related to ease of verification. This may be a concern because the partition label is a covariate that seems to be a virtual recoding of the response variable. There are three reasons for proceeding despite this concern.

The first is pragmatic. The biometric algorithm development community needs to improve performance on difficult problems while at the same time not sacrificing performance on easier problems for which performance is already good. This is best done with a graduated challenge problem. The second is also pragmatic, but having more directly to do with choosing between two modeling options. Concern about the GBU covariate could be alleviated by breaking the analysis apart and treating each partition as its own unique study. While cleaner, such an approach denies us the opportunity to directly quantify interactions between the partitions and other covariates.

A final reason is based on how GBU should be interpreted. For each person separately, the person's image pairs were sorted by similarity score, with the very best being placed in the Good partition, the very worst in the Ugly partition, and a part of the remainder placed in the Bad partition. It is critical to note that the GBU factor level for a match pair is a person-specific measure of the extent to which that image pair approaches the subject's maximum potential matchability. Thus, a

Table 2
Covariates available for analysis. Except for the asterisked items, all covariates are measured separately for the probe and gallery image and therefore permit probe/gallery combination variables like 'indoor probe with outdoor gallery'. The additional covariate of GBU is described in the text.

Subject	Environment	Image
Person ID*	Lighting	Eye location
Image ID*	Setting	Resolution
Gender*	Location	Tilt
Race*		Edge density
Initial subject age*		Focus
Facial expression		
Glasses		
Elapsed time*		

good image pair for a person represents an imaging occasion where that individual 'presented' data containing nearly the greatest effective information content as s/he could. An ugly image pair represents the case when this person was 'having a particularly bad day' with respect to the match information contained in his/her images. Due to between-subject variation, similarity scores and verification rates for Good image pairs for one person might be inferior to Bad scores and rates for another person.

5.4. Model structure and selection

We fit a generalized linear mixed model (GLMM) using the residual mean pseudo-likelihood estimation approach (Wolfinger and O'Connell, 1993; Breslow and Clayton, 1993; SAS, 2006). The Kenward–Roger adjustment for residual degrees of freedom was used when testing fixed effects (Kenward and Roger, 1997).

Exploratory data analysis, experience with past analyses, and conventional wisdom in the face recognition field motivated consideration of a very broad collection of potential model terms including two-way interactions, interactions with GBU labels, interactions between target and query features, and cubic orthogonal polynomials in target/query values for focus, resolution, and edge density.

We considered two sources of overdispersion modeled with random effects. First, our models included a subject-specific random effect. There were between 3 and 48 match scores per subject. Second, the location variables contributed random effects. Three of the ten locations were outside, and understanding the effects of indoor and outdoor settings was very important. However, the particular locations within each setting were unimportant and incidental. To complicate matters, the target and query locations for each match pair could differ. We considered the merits of two potential approaches. The first option was to model the location effects as additive. In this case, the target location and query location would contribute separate random effects, and there would be 20 potential combinations. The second option for location effects was to permit a separate random effect for each target/query location combination. Not all location pairings were sampled; this option would have yielded 93 different combinations.

We adopted the additive random effects structure for location. A strong reason for this choice was its simplicity. This model is consistent with an assumption that the designations of 'target' and 'query' are interchangeable and hence S_A is symmetric. In practice, surprisingly, this assumption is not always true due to cohort normalization (Rosenberg et al., 1992; Phillips et al., 2010), meaning that a query image is compared to a set of reserved images kept with the system in order to assess the relative difficult of a query image and subsequently adjust its score accordingly.

The final model had 63 parameters for fixed effects and 457 random effects (the 437 subjects and 20 location effects). Although an objective, automated model selection technique would have been helpful here, log likelihood and AIC values could not be compared between models due to the pseudo-likelihood approach needed for estimation. Model selection was therefore conducted using a non-automated blocked stepwise approach biased toward forward selection. The blocks of subject, setting, and image covariates were considered separately in that order (including the polynomial and interaction terms). Between-block interactions and further stepwise selection were then addressed. Contributions to F statistics and effect sizes were used to guide variable selection. In the latter case, model terms were required to cause an effect on estimate verification probability of at least 2 in 100, which is the minimal effect size considered scientifically relevant in application. This minimal effect magnitude criterion was more useful than reliance on p-values because the number of image pairs is so large here that all our model effects had very small p-values.

5.5. Results and discussion

5.5.1. Random effects

We can quantify the importance of the random effects by looking in the log odds space. The marginal mean log odds for verification are 5.14, 1.33, and −1.85 for Good, Bad, and Ugly respectively. The standard deviations of the random effects – which are defined on the log odds scale – are 0.97 and 0.79 for subjects and locations, respectively. The ratios of random effect standard errors to marginal log odds are shown in Table 3. These numbers are akin to CVs for the log odds of verification.

Table 3
Magnitude of random effects on log odds scale, expressed as the ratio of random effect standard error to mean log odds of verification.

	Ugly	Bad	Good
Subject	0.72	0.72	0.97
Location	0.59	0.60	0.15

The ratios shown in Table 3 associated with subjects are greater than that for locations. Both types of random effect are notably large, indicating that subject- and location-specific variation are both very important factors affecting algorithm performance. For algorithm developers, both factors clearly should be targets for additional research. Location effects are particularly troubling since face chips mask the background and are image normalized before any match effort occurs. We hypothesize that location variation is attributable to lighting effects not currently accounted for in our lighting model and/or tested algorithms. However, this hypothesis is weakened because we included a model term indicating whether locations matched. Effects of location somehow apply within-location rather than merely between them. Thus our results suggest that it is difficult to develop generic algorithms that perform well on new people and in novel locations.

Another result in Table 3 is evident from a comparison between the GBU partitions. Although the total extra variation is roughly equivalent across partitions, the subject/location ratios are markedly different. Unlike the Ugly and Bad images, for Good images, between-subject variation is overwhelmingly dominant. We believe that this finding suggests that, for Good images, the challenge of verification across locations is nearly solved. If there is any practical improvement to performance to be made under Good circumstances, developers should focus on controlling subject-specific effects.

For end users planning to implement a recognition algorithm, these results are cause for hesitation. In the real world, images are usually obtained in non-ideal and uncontrolled circumstances where Ugly and Bad are the norm. In this situation, it will be insufficient to alter the acquisition protocol somehow to reduce location variation. Subject effects, which are wholly uncontrollable, would remain a substantial impediment to good performance.

After fitting this model which incorporates random effects for subject and location, the Pearson chi-square per degree of freedom goodness-of-fit statistic does not exceed 1. This suggests that we have not overlooked any major source of additional overdispersion.

5.5.2. *Fixed effects*

Fig. 2 shows population-weighted least-squares mean probabilities of successful verification from our model for various subject factor covariate cells, marginalized across remaining covariates. The Good (green), Bad (blue) and Ugly (red) partitions are visible as three strata in this plot (from top to bottom in black and white). Uncertainty estimates were obtained by bootstrapping subjects. Quantile intervals of bootstrap distributions are shown, with the quantile being Bonferroni-adjusted to maintain 95% joint confidence within each block of covariate $\times$ GBU partition results. Marginal results are also shown for comparison. Within each G/B/U $\times$ factor block, effects that are substantially greater than at least one other level in that block are plotted with heavier lines.

Among the fixed covariate effects, GBU is of course dominant, and because of its construction we interpret only its interactions. Also, performance within the Good partition is nearly perfect so we consider it no further.

Where not self explanatory, the factor labels used in Fig. 2 are as follows. For Expression and Glasses, labels are paired to indicate the state of the target and then the query image. Thus, for Expression, Smile/Neutral indicates a smile in the target image and a neutral expression in the query image. For Age, Younger subjects were no older than 23, Older subjects were at least 30, and between these were Typical ages. Lighting indicates the estimated lighting direction for the target and query image, with Front indicating a fully illuminated face lit from the front, Side indicating strong shading across part of the face consistent with illumination from the side, Shade indicating shadowing consistent with a face illuminated from above.

A few important observations can be made from this figure. One key point is that Asian subjects are easier to verify than other races. This is consistent with past findings (Lui et al., 2009b). There are several possible explanations for this effect. Since only 25% of the people in the dataset are Asian, it may be that recognition is easier because there are fewer possible confusions among Asians because there are fewer Asians total relative to Caucasians in this dataset. Alternatively, it is possible that something about Asian faces makes them intrinsically easier to recognize. Additional studies will be needed to clarify if and how recognition difficulty varies between races.

Verification performance is improved when target and query facial expressions are the same. However, an interaction with GBU indicates that matching two neutral images clearly is easier than two smiling images in the Ugly data, but not so in the Bad data. These results are consistent with past findings (Lui et al., 2009b) and have important operational implications, the most obvious being policies to promote consistency of expression. For example, the rules for Canadian passport photos include: "... taken with a neutral facial expression (eyes open and clearly visible, mouth closed, no smiling)" (Canada, 2012).[1] Such policies are well supported in terms of promoting consistency. They are problematic if supported by arguments that neutral expressions are intrinsically superior to smiling expressions, something not consistently true in our findings here.

[1] http://www.ppt.gc.ca/cdn/photos.aspx.

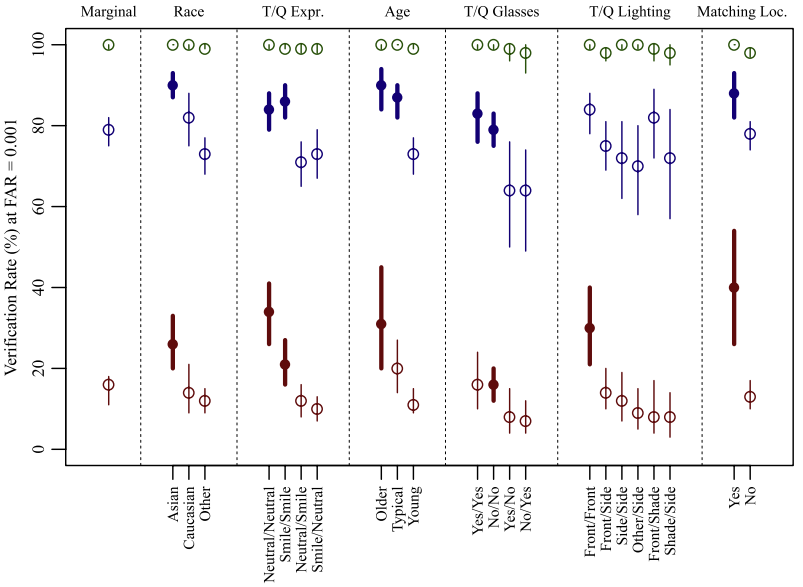

Fig. 2. Population-weighted least squares means and 95% bootstrap intervals for verification rate when the false accept rate is set at 1 in 1000. The Good (green), Bad (blue) and Ugly (red) partitions are visible as three strata in this plot, from top to bottom. The intervals have been Bonferroni-corrected within each G/B/U × factor block, and effects that are significantly greater than at least one other level in that block are plotted with heavier lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Verification performance improves with increasing subject age. This finding also corroborates previous work (Lui et al., 2009b), and indeed this age effect is perhaps the most consistently observed across multiple studies. There is an emerging view that this result may be due to increased and more detailed idiosyncratic facial features associated with aging, such as freckles and wrinkles.

As with expression, taking on or off a pair of glasses substantially degrades recognition performance. However, this finding suggests no reason to support always wearing one's glasses versus never wearing one's glasses. This finding too has important policy implications. If one considers the scenario where a person is photographed every day, perhaps while entering a secure facility, there is probably no good reason for asking them to constantly take off their glasses. However, in a different context, specifically law enforcement applications, it is probably best to favor a no glasses policy.

With respect to imaging environment effects, it is surprising how little effect lighting appears to have. However, the one place where the lighting model clearly identifies superior matching is Front–Front on the Ugly partition. This may be indicating that lighting, as captured by our lighting model, only begins to seriously influence recognition on the most challenging data, and only in so much as it indicates the single most favorable circumstance.

The location effect is also somewhat surprising. Algorithms are not supposed to be influenced by externalities such as where a photo is taken. Yet, the results show a substantial improvement in all three GBU partitions when query and target images are acquired at the same location as opposed to two different locations. Further, the effect is huge for the Ugly case. In the short term, this suggests that where it is possible to engineer a system to consistently acquire images at a single location, current algorithms will perform much better. Longer term, this finding indicates a critical and under-appreciated weakness in algorithms that the research community needs to better understand and reduce.

Fig. 3 illustrates some findings related to image covariates. These graphs show multiplicative effects on the odds of verification as a function of target and query focus (top) and edge density (bottom) for the Ugly data partition. Contours are masked to include only the portion of the space reasonably well-represented in the dataset to avoid extrapolation. The figures show that increased focus and edge density are associated with higher verification odds. Moreover, the contours suggest that the focus for query and target images should be roughly equal to improve performance.

Finally, variable selection led to the notable finding that aside from limited effects for lighting and expression, most predictors did not interact with the GBU partition labels. This suggests that most features of subjects and images tend towards 'universality' in that their effect on recognition is similar whether the subject presents excellent or poor match opportunities. This finding is important for many implementation protocols where imaging is opportunistic and uncontrolled. In these cases, face recognition algorithms must do as well as they can with whatever images they get.

5.6. Validity and generalizability

Table 1 illustrated how verification performance on the GBU dataset can differ markedly between algorithms. It is also true that performance of a single algorithm can vary substantially across datasets. Before considering the implications of this, it is worth repeating that the purpose of our analysis is to estimate the effects of covariate factors, not absolute performance

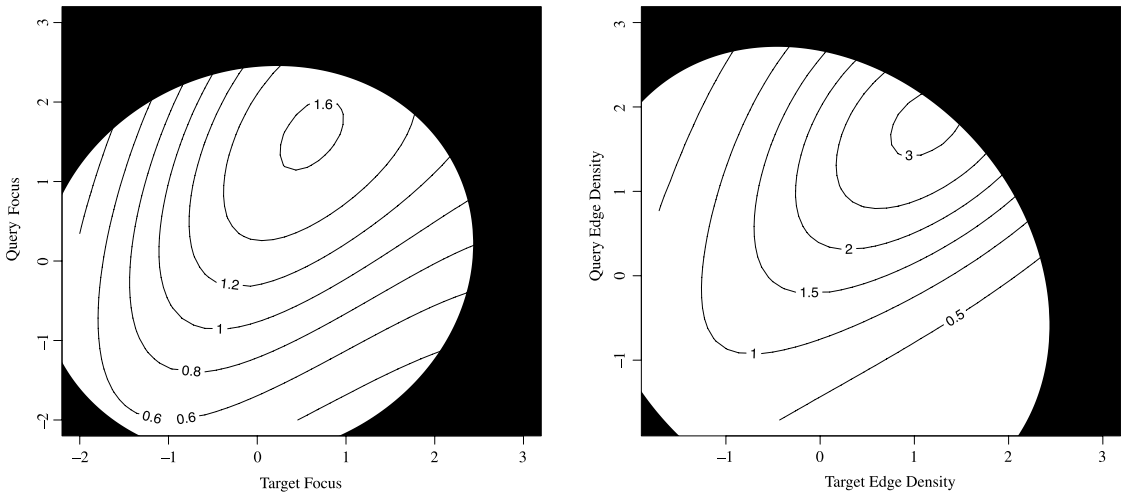

Fig. 3. Multiplicative effect on verification odds as a function of target and query focus (left) and edge density (right) for the Ugly data partition.

levels per se. This intent renders irrelevant the confounding question about whether poor (or good) algorithm performance is attributable to the algorithm or the dataset.

Of course, the constant ('intercept') term in the link-linear predictor of the GLMM establishes a baseline verification rate that is specific to the algorithm and dataset analyzed. Using our model parameter estimates, prediction of verification rates on easier (or harder) datasets would be inaccurate, as would predictions for different algorithms. If empirical performance prediction was a goal, one could fit a model like ours to data from the new application.

However, we are interested in finding covariates whose effects are consistent across diverse datasets and, ideally, across various recognition algorithms. This addresses a question of greater relevance to algorithm developers and opens the door to more profound questions about the role and utility of biometric identification.

It is difficult to assess the extent to which our estimated covariate effects may be 'universal'. Note, however, that the GBU partitions might be viewed as three independent datasets. Our finding (Section 5.5.2) that most covariates do not interact with the GBU partition label therefore supports the hypothesis that these effects are generalizable across datasets. Also, we have completed similar analyses on performance data from the Face Recognition Grand Challenge (FRGC) (Beveridge et al., 2009a) and Face Recognition Vendor Test 2006 (FRVT) (Beveridge et al., 2009b) and a consistent pattern is emerging for some covariates. Specifically, for all three studies (i.e., GBU, FRGC and FRVT) the effects of Race, Age and Glasses are consistent. Expression was not part of the FRVT study, but for GBU and FRGC, the dominant Expression result is consistent: recognition is easier when expressions (smiling vs. neutral) match. Finally, results for image-related covariates such as focus are more varied across algorithms and datasets. Our prior work on FRVT has suggested that image covariate effects depend upon the recognition algorithm being used.

6. Conclusion

The results presented here illustrate how statistical models can be applied in a manner to help algorithm developers and users understand – and hopefully improve – recognition performance. It is clear that the statistical toolbox contains many more applicable methods for development and evaluation. Performance evaluation is a particularly fertile topic for statistical attention due to burgeoning interest from the biometric community and the fact that a predictor–response formulation is one natural way to pose key questions using available datasets. More broadly, opportunities for statistical research in face recognition are abundant because many aspects of recognition algorithms, the structure of the identification and verification tasks they face, and their performance have only recently begun to be posed explicitly in a statistical context.

Acknowledgments

This work was supported by the Technical Support Working Group (TSWG) under Task SC-AS-3181C. P. Jonathon Phillips thanks the Federal Bureau of Investigation (FBI) for their support of this work. The identification of any commercial product or trade name does not imply endorsement or recommendation by Colorado State University and the National Institute of Standards.

References

Ahonen, T., Hadid, A., Pietikainen, M., 2006. Face description with local binary patterns: application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12), 2037–2041.
Basri, R., Jacobs, D., 2003. Lambertian reflectance and linear subspaces. PAMI 25 (2), 218–233.

Baudat, G., Anouar, F., 2000. Generalized discriminant analysis using a kernel approach. Neural Computation 12 (10), 2385–2404.
Belhumeur, P., Kriegman, D., 1998. What is the set of images of an object under all possible illumination conditions. IJCV 28 (3), 245–260.
Beveridge, J.R., Bolme, D.S., Draper, B.A., Givens, G.H., Lui, Y.M., Phillips, P.J., 2010. Quantifying how lighting and focus affect face recognition performance. In: IEEE CVPR 2010 Workshop on Biometrics. pp. 74–81.
Beveridge, J.R., Bolme, D., Draper, B.A., Teixeira, M., 2005a. The CSU face identification evaluation system. Machine Vision and Applications 16 (2), 128–138. URL http://dx.doi.org/10.1007/s00138-004-0144-7.
Beveridge, J.R., Draper, B.A., Givens, G.H., Fisher, W., 2005b. Introduction to the statistical evaluation of face recognition algorithms. In: Zhao, W., Chellappa, R. (Eds.), Face Processing: Advanced Modeling and Methods. Elsevier, pp. 87–124.
Beveridge, J.R., Givens, G.H., Phillips, P.J., Draper, B.A., 2009a. Factors that influence algorithm performance in the face recognition grand challenge. Computer Vision and Image Understanding 113 (6), 750–762. URL http://www.sciencedirect.com/science/article/pii/S1077314209000022.
Beveridge, J.R., Givens, G.H., Phillips, P.J., Draper, B.A., Bolme, D.S., Lui, Y.M., 2009b. FRVT 2006: Quo vadis face quality. Image and Vision Computing 28 (5), 732–743. URL http://www.sciencedirect.com/science/article/B6V09-4XC57FT-1/2/9238086cd217effa81ec8cac8a84e70a.
Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 8, 9–25.
Canada. 2012. Rules for Canadian passport photos. May. URL http://www.ppt.gc.ca/cdn/photos.aspx.
Chellappa, R., Sinha, P., Phillips, P.J., 2010. Face recognition by computers and humans. IEEE Computer 46–55.
Draper, B.A., Baek, K., Bartlett, M.S., Beveridge, J., 2003. Recognizing faces with pca and ica. Computer Vision and Image Understanding 91 (1–2), 115–137. Special Issue on Face Recognition. URL http://www.sciencedirect.com/science/article/pii/S1077314203000778.
Edelman, A., Arias, R., Smith, S., 1999. The geometry of algorithms with orthogonal constraints. SIAM Journal on Matrix Analysis and Applications 20 (2), 303–353.
Etemad, K., Chellappa, R., 1997. Discriminant analysis for recognition of human face images. Journal of the Optical Society of America 14, 1724–1733.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 17, 179–188.
Givens, G.H., Beveridge, J.R., Draper, B.A., Grother, P., Phillips, P.J., 2004. How features of the human face affect recognition: a statistical comparison of three face recognition algorithms. In: Proceedings: IEEE Computer Vision and Pattern Recognition 2004. pp. 381–388.
Greenberg, C.S., Martin, A.F., 2009. NIST speaker recognition evaluations 1996–2008. In: SPIE Proceedings. vol. 7324. pp. 732411-1–732411-12.
Grother, P., Micheals, R.J., Phillips, P.J., 2003. Face recognition vendor test 2002 performance metrics. In: Kittler, J., Nixon, M.S. (Eds.), Audio-and Video-Based Biometric Person Authentication. In: LNCS, vol. 2688. Springer, pp. 937–945.
Grother, P., Quinn, G.W., Matey, J.R., Ngan, M., Salamon, W., Fiumara, G., Watson, C., 2012. IREX III performance of iris identification algorithms. Tech. Rep. NISTIR 7836. National Institute of Standards and Technology.
Grother, P.J., Quinn, G.W., Phillips, P.J., 2010. Mbe 2010: report on the evaluation of 2d still-image face recognition algorithms. Tech. Rep. NISTIR 7709. National Institute of Standards and Technology.
Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53, 983–997.
Kumar, B., Savvides, M., Xie, C., 2006. Correlation pattern recognition for face recognition. Proceedings of the IEEE 94 (11), 1963–1976.
Liu, C., 2006. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5), 725–737.
Lui, Y.M., Beveridge, J.R., 2008. Grassmann registration manifolds for face recognition. In: Forsyth, D., Torr, P., Zisserman, A. (Eds.), ECCV 2008 Proceedings. Springer Verlag, pp. 44–57.
Lui, Y.M., Beveridge, J.R., Kirby, M., 2009a. Canonical stiefel quotient and its application to generic face recognition in illumination spaces. In: IEEE International Conference on Biometrics: Theory, Applications and Systems. Washington, DC, pp. 431–438.
Lui, Y.M., Bolme, D., Draper, B.A., Beveridge, J.R., Givens, G., Phillips, P.J., 2009b. A meta-analysis of face recognition covariates. In: IEEE International Conference on Biometrics: Theory, Applications, and Systems. pp. 139–146.
Lui, Y.M., Bolme, D., Phillips, P., Beveridge, J., Draper, B., 2012. Preliminary studies on the Good, the Bad, and the Ugly face recognition challenge problem. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2012. pp. 9–16.
Moses, Y., Adini, Y., Ullman, S., 1994. Face recognition: the problem of compensating for changes in illumination direction. In: European Conference on Computer Vision. Marseille. pp. 286–296.
O'Toole, A.J., Millward, R.B., Anderson, J.A., 1988. A physical system approach to regognition memory for spatially transformed faces. Neural Networks 1, 179–199.
Pato, J., Committee, W.B., Millet, L., Council, N.R., 2010. Biometric Recognition: Challenges and Opportunities. National Academies Press.
Pentland, A., Moghaddam, B., Starner, T., 1994. View-based and modular eignspaces for face recognition. In: Proceedings Computer Vision and Pattern Recognition 94. pp. 84–91.
Phillips, P.J., Beveridge, J.R., Draper, B.A., Givens, G., O'Toole, A.J., Bolme, D., Dunlop, J., Lui, Y.M., Sahibzada, H., Weimer, S., 2012. The Good, the Bad, and the Ugly face challenge problem. In: Best of Automatic Face and Gesture Recognition 2011. Image and Vision Computing 30 (3), 177–185. URL http://www.sciencedirect.com/science/article/pii/S0262885612000091.
Phillips, P.J., Flynn, P.J., Beveridge, J.R., Scruggs, W.T., O'Toole, A.J., David, B., Bowyer, K.W., Draper, B.A., Givens, G.H., Lui, Y.M., Sahibzada, H., Scallan, Iii, J.A., Weimer, S., 2009. Overview of the multiple biometrics grand challenge. In: Proceedings of the Third International Conference on Advances in Biometrics. ICB'09. Springer-Verlag, Berlin, Heidelberg, pp. 705–714. URL http://dx.doi.org/10.1007/978-3-642-01793-3_72.
Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W., 2005. Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 947–954.
Phillips, P.J., Flynn, P.J., Scruggs, W.T., Bowyer, K.W., Worek, W., 2006. Preliminary face recognition grand challenge results. In: Seventh International Conference on Automatic Face and Gesture Recognition. pp. 15–24.
Phillips, P., Grother, P., Micheals, R., Blackburn, D., Tabassi, E., Bone, J., 2002. FRVT 2002: Overview and Summary. Tech. Rep., Face Recognition Vendor Test 2002. www.frvt.org.
Phillips, P., Moon, H., Rizvi, S., Rauss, P., 2000. The FERET evaluation methodology for face recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10), 1090–1104.
Phillips, P.J., Scruggs, W.T., O'Toole, A.J., Flynn, P.J., Bowyer, K.W., Schott, C.L., Sharpe, M., 2010. FRVT 2006 and ICE 2006 large-scale results. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (5), 831–846.
Pinto, N., DiCarlo, J.J., Cox, D., 2009a. How far can you get with a modern face recognition test set using only simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. p. (online only).
Pinto, N., DiCarlo, J.J., Cox, D.D., 2009b. How far can you get with a modern face recognition test set using only simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2591–2598.
Rosenberg, A., Delong, J., Lee, C., Juang, B.-H., Soong, F., 1992. The use of cohort normalized scores for speaker recognition. International Conference on Spoken Language Processing. Banff, Canada.
SAS 2006. The GLIMMIX Procedure. SAS Institute Inc., Cary, NC.
Sirovich, L., Kirby, M., 1987. A low-dimensional procedure for the characterization of human faces. The Journal of the Optical Society of America 4, 519–524.
Turk, M., Pentland, A., 1991. Eigenfaces for recognition. Journal of Cognitive Neuroscienc 3 (1), 71–86.
Wiskott, L., Fellous, J.-M., Kruger, N., von der Malsburg, C., 1997. Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (7), 775–779.
Wolfinger, R., O'Connell, M., 1993. Generalized linear models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48, 233–243.
Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J., 2003. Face recognition: a literature survey. ACM Computer Surveys 35, 399–458.