

Preliminary Studies on the Good, the Bad, and the Ugly Face Recognition Challenge Problem

Yui Man Lui^a, David Bolme^a, P. Jonathon Phillips^b, J. Ross Beveridge^a, and Bruce A. Draper^a
Department of Computer Science, Colorado State University, Fort Collins, CO 80523, USA^a
The National Institute of Standards and Technology, Gaithersburg, MD 20899, USA^b
{lui,bolme,ross,draper}@cs.colostate.edu^a, jonathon.phillips@nist.gov^b

Abstract

Face recognition has made significant advances over the last twenty years. State-of-the-art algorithms push the performance envelope to near perfect recognition rates on many face databases. Recently, the Good, the Bad, and the Ugly (GBU) face challenge problem has been introduced to focus on hard aspects of face recognition from still frontal images. In this paper, we introduce the CohortLDA baseline algorithm, which is an Linear Discriminant Analysis (LDA) algorithm with color spaces and cohort normalization. CohortLDA greatly outperforms some well known face recognition algorithms on the GBU challenge problem. The GBU protocol includes rules for creating training sets. We investigate the effect on performance of violating the rules for creating training sets. This analysis shows that violating the GBU protocol can substantially over estimate performance on the GBU challenge problem.

1. Introduction

Face recognition is a common biometric for recognizing people. Automated face recognition has been studied for more than two decades [12, 27], and the recognition performance has improved by several orders of magnitude [17]. For some face databases, the recognition rates have achieved nearly perfect results. While great progress has been made on frontal faces from controlled environments, face recognition of frontal images taken in ambient illumination remains a challenge. In particular, illumination, expression, and focus are the dominant factors for frontal view face recognition [15, 5]. To continue making strides on face recognition, we need to shift our attention to uncontrolled environments.

In recent years, there have been efforts in pushing face recognition images taken from mobile studio environments to ambient illumination environments. The prime examples are the Face Recognition Grand Challenge (FRGC) [20],

the Face Recognition Vendor Test 2006 (FRVT) [17], and Labeled Faces in the Wild (LFW) [22]. While progress has been made, the solution to face recognition remains unclear. To be more specific, whether an algorithm overfits some particular dataset or captures the idiosyncratic aspects of faces is generally unknown. This is a serious concern for algorithm development and needs to be addressed. One solution is testing on sequestered data; however, it is an expensive task.

An alternative may involve a database with a variety of difficulties with an experiment protocol that attempts to minimize the effects of overfitting. The recent introduction of the Good, the Bad, and the Ugly (GBU) face recognition challenge [18] focuses research on difficult aspects of face recognition; specifically, the current best verification on the Ugly partition is only 15% at the 1/1,000 false accept rate. This illustrates the challenge of frontal views of faces from uncontrolled environments. The goal of the GBU is to encourage the development of algorithms that improve performance on the difficult partitions (Bad and Ugly) while not sacrificing performance on the Good partition. In addition, the GBU protocol considers the danger of overfitting and does not allow any training on images of subjects in the GBU image set.

The GBU challenge problem provides a great platform for further face recognition advancement. The motivation of this work is twofold. First, we introduce the CohortLDA algorithm, a new variant of Fisher linear discriminant analysis-based face recognition algorithm. CohortLDA is designed to serve as an open source baseline software for the GBU challenge problem, which will likely advance further algorithm development and benchmarks. Second, we quantitatively show the effect on performance of violating the GBU training protocol. More specifically, when the training and testing sets contain the same people, the algorithm performance is inflated. To facilitate algorithm development, we distribute a valid training set called GBUx8 which ensures no subject identity overlap with the GBU images.

The remainder of this paper is organized as follows. The important characteristics of the proposed baseline algorithm are given in Section 2. The GBU data sets and evaluation protocol are described in Section 3. The effects of deviating from the evaluation protocol is discussed in Section 4. The details of our baseline algorithm are presented in Section 5. The baseline experimental results and analyses are reported in Section 6 and Section 7, respectively. Finally, the conclusion is provided in Section 8.

2. What Makes a Good Baseline Algorithm?

Establishing a good baseline is important for a challenge problem. A baseline provides a basis for performance comparison with both old and new algorithms. It also characterizes the difficulty of a data set.

A good baseline algorithm should be simple and easy to understand. In addition, the performance results should be reproducible and respectable. The baseline algorithm proposed in this paper exhibits the following seven attributes.

- Simple representation: Faces are represented holistically.
- Simple preprocessing: The preprocessing steps consist of standard pixel transforms.
- Simple algorithm: The face representation is based on Fisher's linear discriminant analysis.
- Simple post-processing: Raw matching scores are cohort normalized using an *independent* set of images—neither in the training or test sets.
- Reasonable speed: Training takes about an hour and running on the GBU takes about 10 minutes (performance is benchmarked on a high end iMac).
- Decent recognition results: The algorithm achieves respectable performance results.
- Availability: The source code and the sigset of GBUx8 are publicly available¹.

3. Data Sets and Evaluation Protocol

The Good, the Bad, and the Ugly (GBU) challenge problem [18] comprises three levels of difficulty the Good, the Bad, and the Ugly. The GBU Challenge Problem was constructed such that recognition difficulty varies markedly while at the same time factors such as the individual people or number of images per person remained the same. To gauge the relative difficulty associated with recognizing a pair of images, similarity scores were created by fusing scores from three of the top performing algorithms in the



Figure 1. Examples of match pairs from each partition: a good pair (left column), a challenging pair (middle column), and a very challenge pair (right column).

FRVT 2006 [17]. All images in the GBU data set were nominally frontal face images. The images were collected in ambient lighting both indoors and outdoors. Example images of the GBU partitions are given in Figure 1.

Each partition in the GBU is specified by two sets of images: a target set and a query set. Across all three partitions, all target and query sets consisted of 1,085 images from 437 subjects. All target and query sets contained the same number of images of each subject.

For each partition, an algorithm computes a similarity score between all pairs of images in that partition's target and query sets. A similarity score is a measure of the similarity between two faces. Higher similarity scores imply greater likelihood that the face images are of the same person. If an algorithm reports a distance measure, then a smaller distance measure implies greater likelihood that the face images are of the same person. Distances are converted to similarity scores by multiplying by negative one. The set of all similarity scores between a target and a query set is called a similarity matrix. A pair of face images of the same person is called a match pair, and a pair of face images of different people is called a non-match pair. From the similarity matrix, receiver operating characteristics (ROC) and other measures of performance can be computed.

The main points of the GBU protocol are summarized below:

- All training, model selection, and tuning need to be completed prior to computing performance on the GBU partitions.
- All algorithms are one-to-one matchers. Thus, all similarity scores between target and query images cannot depend in anyway on the target and query sets.
- Algorithms cannot be trained on images of subjects in the GBU image set.

¹<http://www.cs.colostate.edu/facerec/algorithms/baselines2011.php>

One of the most important aspects of the GBU protocol is that algorithms cannot be trained on images of subjects in the GBU image set. The same emphasis has also been adopted in the BANCA database [2]. This type of evaluation can help to develop algorithms that generalize and prevent over fitting the training set.

To compile with the GBU requirement, we searched the Multiple Biometric Grand Challenge (MBGC) [19] still image data for images that meet this requirement. In the MBGC still face data set, we found 345 subjects distinct from the 437 subjects in the GBU data set. We constructed a training set, called GBUx8, that contains up to 8 randomly selected images per subject. The GBUx8 training set contains 1,766 images from 345 subjects. All images in the GBUx8 training set were acquired under ambient lighting conditions from uncontrolled environments. Examples are provided in Figure 2.

4. Effects of Deviating from the Protocol

The GBU is very specific in stating the properties of a valid training set where people in the GBU image set cannot be used in training. Not following these rules can result in over estimation of performance on the GBU challenge problem. There are three basic ways to construct training sets that do not follow the protocol.

The first is to include images of subjects that are in the GBU image set. While this may seem to be a minor infraction, we show that constructing training sets in this manner does over estimate performance.

The second is for the training set to be one of the target sets. This method of constructing a training set is a version of cohort / gallery normalization. In cohort normalization, the similarity score between target and query images can depend on the target set. In some applications and experiment protocols, this is allowed. Cohort normalization has the potential to improve performance. However, because of the design of the GBU, cohort normalization using the same people from the target set is prohibited.

The third is for the training set to consist of the union of all three target sets. Our experiments, show that constructing the training set in this manner results in significant over estimation of performance.

It is imperative that researchers follow the GBU protocol; in particular, following the rules for constructing the training set. This allows for a fair comparison among algorithms; otherwise, the results may be over estimated. To demonstrate the severity, we provide quantitative results on the effect of training set violation in Section 7.

5. The CohortLDA Baseline Algorithm

The CohortLDA baseline algorithm introduced here is based on the Fisher’s Linear Discriminant Analysis

(LDA) [9]. LDA is a statistical learning technique which has been widely used in pattern recognition, particularly in the face recognition community [4, 8, 16]. CohortLDA is simple for face recognition, and is a good strawman for the GBU. The following subsections describe the components of CohortLDA including the preprocessing, training, and test phases.

5.1. Preprocessing Phase

Image preprocessing is an essential step for pattern recognition. In particular for face recognition, geometrical alignment and illumination normalization are the typical preprocessing mechanisms. The geometrical alignment is based on the eye coordinates so that the face region is scaled, rotated, and cropped to a predefined size. Specifically, all images are resized to 65×75 where the distance between the eyes is set to 32 pixels.

It is known that variation in illumination presents a significant challenge in face recognition. Over the years, many illumination normalization techniques have been proposed to transform face images to illumination insensitive representations. The exemplars of these methods include anisotropic Retinex model [10], self-quotient [28], and local-texture [26] representations. While these approaches have shown promise in face recognition, proper parameter choice is usually complicated.

Here, we seek a simpler form for lighting compensation. Among many illumination normalization techniques, the use of color spaces [11, 25, 13] seems the most direct and effective for face recognition. From our experience, the red from the RGB color space preserves the face structure in the presence of mild lighting variation and the *I* chrominance from the YIQ color space can compensate for severe illumination variation. Thus, both the red channel and the *I* chrominance serve as basis for our feature representations.

While the red can be directly extracted from the RGB color space, the *I* chrominance from the YIQ is the color difference computed by a linear combination of RGB. As such, the *I* chrominance is calculated as

$$0.596 * R - 0.275 * G - 0.321 * B.$$

Because the red channel is similar to the gray-scale image, it usually does not work well with large lighting variation. The next logical and simple step is to apply a logarithm to the red channel. In other words, we consider the pixel contrast instead of the pixel difference. Let *r* be an image derived from the red channel; the logarithm transformed image *x* is computed from $\log(r + 0.1)$ where 0.1 is added to avoid ill-conditioning. To further compensate for the absolute gain during acquisition, we apply a zero-mean and standard deviation (z-norm) normalization to both the logarithm transformed and the *I* chrominance images as follows:

$$\hat{x} = \frac{x - \bar{x}}{\sigma} \tag{1}$$



Figure 2. Example images from the GBUx8 training set.



Figure 3. Examples of preprocessed images: The original color image, the red channel image after log and z-norm, and the I chrominance image after z-norm are shown on the left, middle, right, respectively.

	Color Space	Log	Z-Norm
Training Set	Red	Yes	Yes
Cohort Set	Red	Yes	Yes
Query Set	Red	Yes	Yes
Target Set	Red	Yes	Yes
Training Set	I chrominance	No	Yes
Cohort Set	I chrominance	No	Yes
Query Set	I chrominance	No	Yes
Target Set	I chrominance	No	Yes

Table 1. Preprocessing Settings

where \bar{x} and σ are the mean and the standard deviation of the image x , respectively, and \hat{x} is the preprocessed image. A summary of our preprocessing settings is presented in Table 1 where the red channel is used with the logarithm transform. The examples of preprocessed images are depicted in Figure 3. We can see in the I chrominance image that much of the strong side lighting has been removed; on the other hand, it also eliminates some details of the face.

After the geometric and illumination normalization, we proceed with a training phase and a test phase.

5.2. Training Phase

Fisher’s LDA is a supervised learning method that seeks a linear projection of the training data by maximizing the distance of between-class samples while minimizing the distance of within-class samples. This criterion can be for-

mulated as:

$$J(W) = \operatorname{argmax}_W \frac{\operatorname{tr}\{W^T S_B W\}}{\operatorname{tr}\{W^T S_W W\}} \quad (2)$$

where S_B and S_W are the between-class scatter matrix and within-class scatter matrix defined as:

$$S_B = \sum_{k=1}^c p_i (u_i - u)(u_i - u)^T \quad (3)$$

$$S_W = \sum_{i=1}^p (\hat{x}_i - u_i)(\hat{x}_i - u_i)^T \quad (4)$$

where c and p are the number of classes and the number of samples in the training set; p_i is the number of samples in class i , u_i is the i th class mean, u is the global mean, and \hat{x}_i is the data sample.

Since Equation (2) can be recognized as the generalized Rayleigh quotient, the projection vectors W satisfies

$$S_B W = \lambda S_W W \quad (5)$$

and can be solved by a generalized eigenvalue decomposition where λ is the associated eigenvalues.

5.2.1 Dimensionality Reduction

Because of the curse of dimension, dimensionality reduction is often used with LDA. Before we solve the generalized eigenvalue problem in Equation (5), we perform dimensionality reduction using a linear projection obtained from Principal Component Analysis (PCA). Let X be a data sample matrix centered at zero in $\mathbb{R}^{n \times p}$; a linear projection U can be computed using the Singular Value Decomposition (SVD) as:

$$X = U S V^T \quad (6)$$

where U and V are the left singular and right singular vectors, respectively; S is a diagonal matrix containing the singular values. We further truncate the projection matrix U to

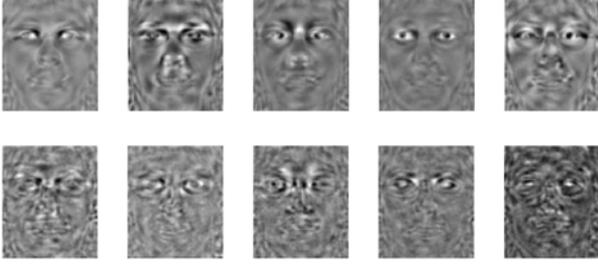


Figure 4. LDA faces: The top row shows the LDA faces acquired from the red channel after log and z-norm and the bottom row depicts the LDA faces obtained from the I chrominance after z-norm.

retain 98% of the energy. The retaining number of dimensions, z , is determined by

$$\operatorname{argmin}_m \left\{ \sum_{i=1}^m \operatorname{diag}(S)^2 - 0.98 \sum_{i=1}^p \operatorname{diag}(S)^2 \right\} \quad (7)$$

where m is the retaining dimension for the projection U , $\operatorname{diag}(S)$ is the diagonal elements of S , and p is the number of dimensions in $\operatorname{diag}(S)$. Then, we have

$$\hat{S}_B = U^T S_B U \quad (8)$$

$$\hat{S}_W = U^T S_W U \quad (9)$$

where \hat{S}_B and \hat{S}_W become $m \times m$ matrices ($m < n$).

Then, we solve the generalized eigenvalue problem described in Equation (5) using \hat{S}_B and \hat{S}_W and obtain the Fisher LDA projection W . The Fisher LDA projection vectors are organized according to the eigenvalues in a descending order. A PCA projection is then used to project W back to the space of \mathbb{R}^n described as:

$$\hat{W} = UW \quad (10)$$

An additional truncation is applied to reduce the dimension of the Fisher LDA projection such that the dimension z is the minimum of 128, the number of classes - 1, and the number of columns in \hat{W} . The truncated projection $\hat{W} \in \mathbb{R}^{n \times z}$ is the LDA projection. The first five LDA-faces for the red channel and I chrominance are exhibited in Figure 4. As Figure 4 shows, the LDA faces acquired from the red channel and I chrominance capture different details of the face.

5.3. Test Phase

5.3.1 Projection on the LDA Space

In supervised learning, the test phase performs pattern classification using the learned information from the training phase. In particular for LDA, all images are projected on the LDA space such that they are the spanning set of \hat{W} .

Let q and t be the query and target images, respectively; the same preprocessing steps described in Section 5.1 are applied to both q and t followed by a LDA projection and unit length normalization shown as:

$$\hat{q} = \frac{\hat{W}^T q}{\|\hat{W}^T q\|_2}, \quad \hat{t} = \frac{\hat{W}^T t}{\|\hat{W}^T t\|_2} \quad (11)$$

where \hat{q} and \hat{t} are the projected query and target images in \mathbb{R}^z , respectively.

5.3.2 Cohort Normalization

In face verification, we need a decision threshold to determine whether a pair of faces is a match or not. Because some images are harder than others, a fixed threshold may not adapt well from image to image. The use of cohort / gallery normalization [24] has been shown to enhance verification rates, particularly in face [14, 7] and fingerprint verifications [23]. A set of images called the cohort set is adopted to adjust the match distance. Since changing the distance is the same as adjusting the threshold, the verification threshold becomes adaptive when cohort normalization is applied.

While the traditional gallery normalization exploits the match scores for score normalization, multiple images with the same identities are needed. This is not possible in the GBU protocol. In this work, we employ an independent set of images as a cohort set whose identities do not overlap with the test set; therefore, it does not violate the GBU protocol. As such, the image identities are discarded and we normalize the similarity scores based on a set of non-matches.

In cohort normalization, the same preprocessing steps are first applied to all images in the cohort set, forming the red and I chrominance images. A subset of cohort images is selected for each query image and target image using the k nearest neighbor rule described as:

$$c_k^* = c_{k-1}^* \cup \operatorname{argmin}_{c_j \notin c_{k-1}^*} \|\hat{q} - \hat{c}_j\|_2 \quad (12)$$

$$c_k^+ = c_{k-1}^+ \cup \operatorname{argmin}_{c_j \notin c_{k-1}^+} \|\hat{t} - \hat{c}_j\|_2 \quad (13)$$

where $c_0^* = \{\}$, $c_0^+ = \{\}$, \hat{c}_i is the LDA projected pattern in the cohort set, and k is set to 100 in our experiments. The c^* and c^+ are the cohort subsets for query and target images. The cohort distance offsets can then be calculated as follows:

$$d_q = \frac{1}{k} \sum_{i=1}^k \|\hat{q} - c_i^*\|_2 \quad (14)$$

$$d_t = \frac{1}{k} \sum_{i=1}^k \|\hat{t} - c_i^+\|_2 \quad (15)$$

Method	Good	Bad	Ugly	Training Set
FRVT Fusion [18]	98%	80%	15%	Proprietary
<i>Our CohortLDA</i>	83.8%	48.2%	11.4%	GBUx8
V1-like [21]	73%	24.1%	5.8%	GBUx8
Kernel GDA [3]	69.1%	28.5%	5.1%	GBUx8
LRPCA [18]	64%	24%	7%	GBUx2
EBGM [6]	50%	8.1%	1.9%	FERET
LBP [1]	51.4%	5%	1.9%	None

Table 2. Valid Protocol: Verification rates at 0.1% FAR for the GBU data sets

where d_q and d_t are the query and target distance offsets.

5.3.3 Computing the Distance

Using the cohort normalization, the distance between a query image and a target image is computed as:

$$d = \|\hat{q} - \hat{t}\|_2 - \frac{d_q + d_t}{2} \quad (16)$$

The final distance d^* is obtained based on the red channel and I chrominance images using a simple sum rule:

$$d^* = d_R + d_I \quad (17)$$

where d_R and d_I are the distance computed from Equation (16) using the red channel and the I chrominance images, respectively.

6. Experimental Results

This section summarizes our empirical results and demonstrates the effectiveness of our CohortLDA on the GBU challenge problem. We assess the recognition performance as the verification accuracy given at a 0.1% false accept rate (FAR). The experimental results are reported in Table 2 and the associated bar chart is given in Figure 5.

The current best results on the GBU are the fusion of the top three algorithms in the FRVT 2006 [18] given in the first row of Table 2. Although the training set is proprietary, the test and target sets in the FRVT 2006 were sequestered. Thus, the results from the fusion are considered legitimate. As Table 2 shows, the fusion algorithm performs very well on the Good and the Bad data sets but not on the Ugly partition, achieving only a verification rate of 15%. The Ugly data set clearly shows that frontal face recognition remains a challenging problem.

Our CohortLDA results are reported in the second row of Table 2. The receiver operating characteristic (ROC) is given in Figure 6 where the verification rate at 0.1% FAR is highlighted. Here, we employ the GBUx8 as the training and cohort sets for CohortLDA; therefore, we do not have people’s identities overlapped between the training and test sets.

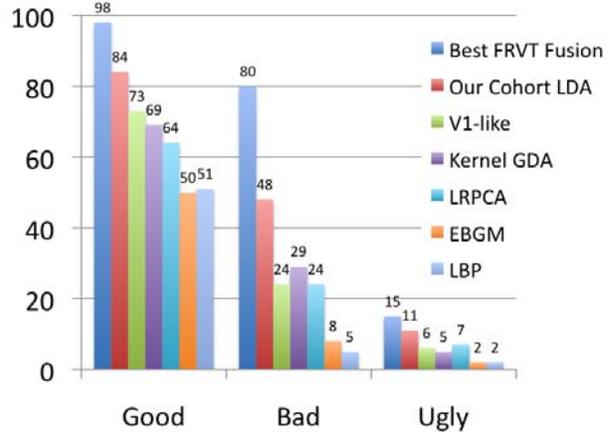


Figure 5. The verification results at 0.1% FAR on the GBU partitions where the Best-FRVT fusion is from the top three algorithms in FRVT 2006 [18]; the V1-like is the method simulating the V1-simple-cell like unit [21]; the kernel GDA is the kernel generalized discriminant analysis [3]; the LRPCA is the local region PCA in [18]; the EBGM is the elastic bunch graph matching [6]; and the LBP is the local binary patterns [1].

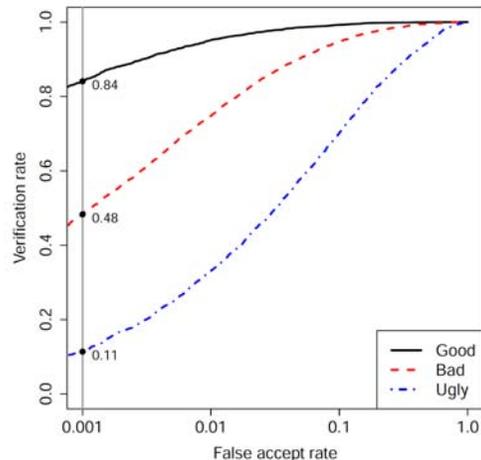


Figure 6. ROC for the GBU partitions trained by the GBUx8. The verification rates are highlighted by the vertical line at 0.1% FAR.

Compared to some known face recognition algorithms, CohortLDA outperforms the V1-like (A) representation [21], the Kernel Generalized Discriminant Analysis (Kernel GDA) [3], Local Region Principal Component Analysis (LRPCA) [18], Elastic Bunch Graph Matching (EBGM) [6], and Local Binary Patterns (LBP) [1] by a significant margin on all GBU partitions. While the performance of CohortLDA is not comparable to the FRVT 2006 fusion algorithm, it is competitive with all non-commercial published results at the time the paper was submitted. As a baseline algorithm, CohortLDA has fewer parameters to tune; thus it is simpler than other methods. CohortLDA

Good	Bad	Ugly	Training	Valid Protocol
83.8%	48.2%	11.4%	GBUx8	Yes
87.8%	52.6%	15%	FRGC	No
95.7%	48.8%	6.2%	Good-Target	No
84.9%	71.4%	15.1%	Bad-Target	No
79%	42%	21.1%	Ugly-Target	No
99%	88.8%	61.2%	All-Target	No

Table 3. Verification rates at 0.1% FAR for the GBU data sets

would benefit the face recognition community by offering a new baseline for performance comparison.

7. Discussion

Training sets play a vital role in supervised learning. One of the objectives of this paper is to demonstrate the effects of training sets and the consequences of violating different aspects of proper evaluation protocol. To do this, we construct five additional training sets. The first is the FRGC experiment 4 training set [20] consisting of 222 people with 12,776 images taken in studio and ambient illumination environments. More importantly, there are 91 subjects whose identities overlap between the FRGC training and the GBU test sets. The next three consist of images drawn directly from the individual Good, Bad, and Ugly target sets. Finally, the fifth training set combines all the Good, the Bad, and the Ugly target images in one training set.

While training and testing on the same people may be common in some face recognition experiments, the results in Table 3 reveal how doing so unrealistically inflates estimated performance. The verification rates shown in Table 3 are for the CohortLDA algorithm configured exactly as for the results shown in Table 2, but with the exception that training and cohort normalization is done using the problematic training sets involving people with the same identities.

The CohortLDA results reported in the first row of Table 3 is trained by the valid GBUx8 training set. When we employ the FRGC training set, the verification rates increase in all GBU partitions reported in the second row. This boosted performance is due to the partial overlapping subject identities between the training and test sets. In addition, the FRGC training set has a large number of images; so it covers a wider range of variation.

The results between the third and the fifth rows illustrate the effects when we train on each individual target set; in other words, the people in the training and test sets are identical. Subsequently, the diagonal entries between the third and the fifth rows exhibit strong recognition performance indicating that the distributions between the training set and the query set are similar. In contrast, poor results may still occur with identical subjects when the imaging conditions

between the training and query sets are significantly different. For example, the result on the Ugly partition is only 6.2% when LDA is trained on the Good target set.

In a more extreme scenario, we can combine all target images from the GBU partitions to form a training set. As a consequence, CohortLDA can outperform the FRVT 2006 fusion algorithm shown in the last row in Table 3. This performance enhancement is mainly due to two reasons: 1) Identical subject identities between the training and test sets; 2) Large amounts of training data. After all, one may expect to achieve better recognition results when the people in the training set and the test set are the same.

There are applications of face recognition where training and testing on the same people may be a reasonable thing to do, such as family photo libraries. However, it is entirely unacceptable for large scale deployed systems that must manage many enrolled people. For example, re-training a deployed system each time that a new person is enrolled is a logistical nightmare. Table 3 illustrates why new algorithms, proposed for large scale face recognition, should avoid presenting results where the subject identities between training and test sets overlapped.

8. Conclusions

We have presented the CohortLDA face recognition algorithm, and it has respectable performance on the GBU challenge problem. CohortLDA performs illumination compensation from color spaces and cohort normalization from a training set. The proposed CohortLDA algorithm provides a new baseline on the GBU where performance can be compared with other algorithms. We observe the GBU evaluation protocol where the subject identities do not overlap between the training and the test sets. One of the characteristics of CohortLDA is its simplicity, leaving plenty of room for improvement. In particular, kernel methods and local regions may be promising areas in boosting the performance.

We showed the effects of not following the GBU protocol with experiments on five different training sets. Experiments with these training sets show that recognition results can be inflated. We recommend that researchers publicly post their training sets. This will provide confidence in the veracity of the reported results.

9. Acknowledgments

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

References

- [1] T. Ahonen, A. Hadid, and M. Pietikinen. Face recognition with local binary patterns. In *European Conference on Computer Vision, Czech Republic*, 2004.
- [2] E. Bailly-Baillire, S. Bengio, F. Bimbot, J. K. Miroslav Hamouz, J. Mariethoz, J. Matas, V. P. Kieron Messer, F. Poree, B. Ruiz, and J.-P. Thiran. The banca database and evaluation protocol. In *International Conference on Audio and Video-Based Biometric Person Authentication*, 2003.
- [3] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.
- [4] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
- [5] R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, and P. J. Phillips. Quantifying how lighting and focus affect face recognition performance. In *IEEE Computer Society Workshop on Biometrics (in conjunction with CVPR), San Francisco*, 2010.
- [6] D. Bolme, R. Beveridge, M. Teixeira, and B. Draper. The csu face identification evaluation system: Its purpose, features and structure. In *International Conf. on Computer Vision Systems*, Graz, Austria, 2003.
- [7] C. H. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikainen. (multiscale) local phase quantisation histogram discriminant analysis with score normalisation for robust face recognition. In *ICCV Computer Vision Workshop, Kyoto, Japan*, 2009.
- [8] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *Journal of Optical Society of America A*, 14(8):1724–1733, 1997.
- [9] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 17(2):179–188, 1936.
- [10] R. Gross and V. Brajovic. An image pre-processing algorithm for illumination invariant face recognition. In *International Conference on Audio-and Video Based Biometric Person Authentication*, 2003.
- [11] S. Karungaru, M. Fukumi, and N. Akamatsu. Face recognition using genetic algorithm based template matching. In *ISCIT*, 2004.
- [12] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990.
- [13] Z. Liu and C. Liu. Fusing frequency, spatial and color features for face recognition. In *IEEE International Conference on Biometrics: Theory, Applications and Systems, Washington D.C.*, 2008.
- [14] Y. M. Lui, R. Beveridge, B. Draper, and M. Kirby. Image-set matching using a geodesic distance and cohort normalization. In *IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands*, 2008.
- [15] Y. M. Lui, D. Bolme, B. Draper, R. Beveridge, G. Givens, and P. J. Phillips. A meta-analysis of face recognition covariates. In *IEEE International Conference on Biometrics : Theory, Applications and Systems, Washington, D.C.*, 2009.
- [16] A. Martinez and A. Kak. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.
- [17] P. Phillips, W. Scruggs, A. O’Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 Large-Scale Results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5), 2010.
- [18] P. J. Phillips, J. R. Beveridge, B. Draper, G. Givens, A. J. O’Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *IEEE International Conference on Automatic Face and Gesture Recognition, Santa Barbara*, 2011.
- [19] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O’Toole, D. S. Bolme, K. W. Bowyer, B. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan, and S. Weimer. Overview of the multiple biometrics grand challenge. In *International Conference on Biometrics*, 2009.
- [20] P. J. Phillips, P. J. Flynn, W. T. Scruggs, K. W. Bowyer, J. Change, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [21] N. Pinto, J. DiCarlo, and D. Cox. 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*, 2009.
- [22] N. Pinto, J. J. DiCarlo, , and D. D. Cox. Establishing good benchmarks and baselines for face recognition. In *Faces in Real-Life Images Workshop (in conjunction with ECCV), Marseille, France*, 2008.
- [23] N. Poh, A. Merati, and J. Kittler. Adaptive client-impostor centric score normalization: A case study in fingerprint verification. In *IEEE Conference on Biometrics: Theory, Applications and Systems, Washington D.C.*, 2009.
- [24] A. Rosenberg, J. Delong, C. Lee, B.-H. Juang, and F. Soong. The use of cohort normalized scores for speaker recognition. In *International Conference on Spoken Language Processing, Banff, Canada*, 1992.
- [25] P. Shih and C. Liu. Comparative assessment of content-based face image retrieval in different color spaces. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(7):873–893, 2005.
- [26] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2007.
- [27] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.
- [28] H. Wang, S. Li, Y. Wang, and J. Zhang. Self quotient image for face recognition. In *International Conference on Image Processing*, 2004.