An Introduction to the Good, the Bad, & the Ugly Face Recognition Challenge Problem

P. Jonathon Phillips, J. Ross Beveridge, Bruce A. Draper, Geof Givens, Alice J. O'Toole,
David S. Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer

Abstract— **The Good, the Bad, & the Ugly Face Challenge Problem was created to encourage the development of algorithms that are robust to recognition across changes that occur in still frontal faces. The Good, the Bad, & the Ugly consists of three partitions. The Good partition contains pairs of images that are considered easy to recognize. On the Good partition, the base verification rate (VR) is 0.98 at a false accept rate (FAR) of 0.001. The Bad partition contains pairs of images of average difficulty to recognize. For the Bad partition, the VR is 0.80 at a FAR of 0.001. The Ugly partition contains pairs of images considered difficult to recognize, with a VR of 0.15 at a FAR of 0.001. The base performance is from fusing the output of three of the top performers in the FRVT 2006. The design of the Good, the Bad, & the Ugly controls for pose variation, subject aging, and subject "recognizability." Subject recognizability is controlled by having the same number of images of each subject in every partition. This implies that the differences in performance among the partitions are result of how a face is presented in each image.**

I. INTRODUCTION

Face recognition from still frontal images has made great strides over the last twenty years. Over this period, error rates have decreased by three orders of magnitude when recognizing frontal faces in still images taken with consistent controlled illumination in an environment similar to a studio [1], [2], [3], [4], [5], [6]. Under these conditions, error rates below 1% at a false accept rate of 1 in 1000 were reported in the Face Recognition Vendor Test (FRVT) 2006 and the Multiple Biometric Evaluation (MBE) 2010 [4], [6].

With this success, the focus of research is shifting to recognizing faces taken under less constrained conditions. Less constrained conditions include allowing greater variability in pose, ambient lighting, expression, size of the face, and distance from the camera. The trick in designing a face recognition challenge problem is selecting the degree to which the constraints are relaxed so that the resulting problem has the appropriate difficulty. The complexity of this task is compounded by the fact that it is not well understood how the above factors effect performance. The problem cannot be too easy that it is an exercise in tuning existing algorithms, nor so hard that progress cannot be made—the three bears problems [2].

Traditionally, a challenge problem is specified by the two sets of images that are to be compared. The difficulty of the problem is then characterized by the performance of a set of algorithms tasked with matching the two sets of face images. To create a problem of a desired level of difficulty, a set of algorithms could be one component in the image selection process. Others factors in the selection process include limiting the number of images per person and requiring that pairs of images of a person are collected on different days.

The Good, the Bad, and the Ugly (GBU) challenge problem consists of three partitions which are called the Good, the Bad, and the Ugly. The difficulty of each partition is based on the performance of three top performers in the FRVT 2006. The Good partition consists of pairs of face images of the same person that are easy to match; the Bad partition contains pairs of face images of a person that have average matching difficulty; and the Ugly partition concentrates on difficult to match face pairs. For the Good partition, the nominal performance based on the FRVT 2006 is a verification rate (VR) of 0.98 at a false accept rate (FAR) of 0.001. For the Bad and Ugly partitions, the corresponding VR at a FAR of 0.001 are 0.80 and 0.15. The performance range over the three partitions is roughly an order of magnitude. The three partitions capture the range of performance inherent in less constrained images[1].

There are numerous sources of variation, known and unknown, in face images that can effect performance. Four of these factors are explicitly controlled in the design of the GBU challenge problem: subject aging, pose, change in camera, and variations among faces. The data collection protocol eliminated or significantly reduced the impact of three of the factors. Changes in the appearance of a face due to aging is not a factor because all images were collected in the same academic year. However, the data set contains the natural variations in a person's appearance that would occur over an academic year. Because all the images were collected by the same model of camera, difference in performance cannot be attributable to changes in the camera. Changes in pose are not a factor because the data set consists of frontal face images.

One potential source of variability in performance is that

P. J. Phillips and H. Sahibzada are with the National Institute of Standards and Technology, 100 Bureau Dr., MS 8940 Gaithersburg MD 20899, USA (e-mail: jonathon@nist.gov). Please direct correspondence to P. J. Phillips.

J. R. Beveridge, B. A. Draper, D. S. Bolme, and Y-M Lui are with the Department of Computer Science, Colorado State U., Fort Collins, CO 46556, USA.

G. Givens is with the Department of Statistics, Colorado State U., Fort Collins, CO 46556, USA.

A. J. OToole, J. Dunlop, and S. Weimer are with the School of Behavioral and Brain Sciences, GR4.1 The University of Texas at Dallas Richardson, TX 75083-0688, USA

[1] Instructions for obtaining the complete GBU distribution can be found at http://face.nist.gov. Instructions for obtaining the LRPCA algorithm can be found at http://www.cs.colostate.edu/facerec.

people vary in their "recognizability." To control for this source of variability, the face images of the same people are in each partition. In addition, each partition has the same number of images of each person. Because the partition design controls for variation in the recognizability of faces, the differences in performance among the three partitions are a result of how a face is presented in each image, and with the pairs of faces that are matched.

II. Generation of the Good, the Bad, & the Ugly Partitions

The GBU partitions were constructed from the Notre Dame multi-biometric data set used in the FRVT 2006 [4]. The images for the partitions were selected from a superset of 9,307 images of 570 subjects. All the images in the superset are frontal still face images collected either outside or with ambient lighting in hallways. The images were acquired with a 6 Mega-pixel Nikon D70 camera. All photos were taken in the 2004-2005 academic year (Aug 2004 through May 2005).

Each partition in the GBU is specified by two sets of images, a target set and a query set. 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. A higher similarity scores implies 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. A distance measure is converted to a similarity score by multiplying by minus 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.

To construct the GBU Challenge Problem we sought to specify target and query sets for each of the three partitions such that recognition difficulty would vary markedly while at the same time factors such as the individual people involved 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 FRVT 2006; this fusion process is described more fully in the next section.

The following constraints were imposed when selecting the GBU partitions:

Distinct Images: An image can only be in one target or query set.

Balanced subject counts: The number of images per person are the same in all target and query sets.

Different days: The images in all match pairs were taken on different days.

After applying these constraints, and given the total number of images available, the number of images per person in the target and query sets was selected to fall between 1 and 4. This number depended upon the total availability of images for each person.

The selection criteria for the partition results in the following properties. An image is only in one partition. There are the same number of match face pairs in each partition and the same number of non-match pairs between any two subjects. This implies that any difference in performance between the partitions is not a result of different people. The difference in performance is a result of the different conditions under which the images were acquired. Figures 1, 2, and 3, are examples of matching face pairs from each of the partitions.

The images included in the GBU target and query sets were decided independently for each person. For each subject i, a subject-specific similarity matrix S_i is extracted from a larger matrix containing similarity scores from the FRVT 2006 fusion algorithm. Each subject-specific matrix contains all similarity scores between pairs of images of subject i. For the Good partition, a greedy selection algorithm iteratively added match face pairs for subject i that maximized the average similarity score for subject i; for the Ugly partition, match face pairs were selected to minimize the average similarity score for subject i; and for the Bad partition, face pairs for subject i were selected to maintain an approximately average similarity score. The selection process for each subject was repeated until the desired number of images were selected for that subject. Since the images for each subject are selected independently, the similarity score associated with a good face pair can vary from subject to subject (similarly for the Bad and Ugly partitions).

Each of the GBU target and query sets contains 1,085 images for 437 distinct people. The distribution of image counts per person in the target and query sets are 117 subjects with 1 image; 122 subjects with 2 images; 68 subjects with 3 images; and 130 subjects with 4 images. In each partition there is 3,297 match face pairs and 1,173,928 non-match face pairs. In the GBU image set 58% of the subjects were male and 42% female; and 69% of the subjects were Caucasian, 22% east Asian, 4% Hispanic, and the remaining 5% other groups; and 94% of the subjects were between 18 and 30 years old with the remaining 6% over 30 years old. For the images in the GBU, the average distance between the centers of the eyes is 175 pixels with a standard deviation of 36 pixels.

III. The FRVT 2006 Fusion Performance

Performance results for the GBU Challenge Problem are reported for the GBU FRVT 2006 fusion algorithm, which is a fusion of three of the top performers in the FRVT 2006. The algorithms were fused in a two-step process. In the first step, for each algorithm, the median and the median absolute deviation (MAD) were estimated from every 1 in 1023 similarity scores ($median_k$ and MAD_k are the median and MAD for algorithm k). The median and MAD were estimated from 1 in 1023 similarity scores to avoid over tuning the estimates to

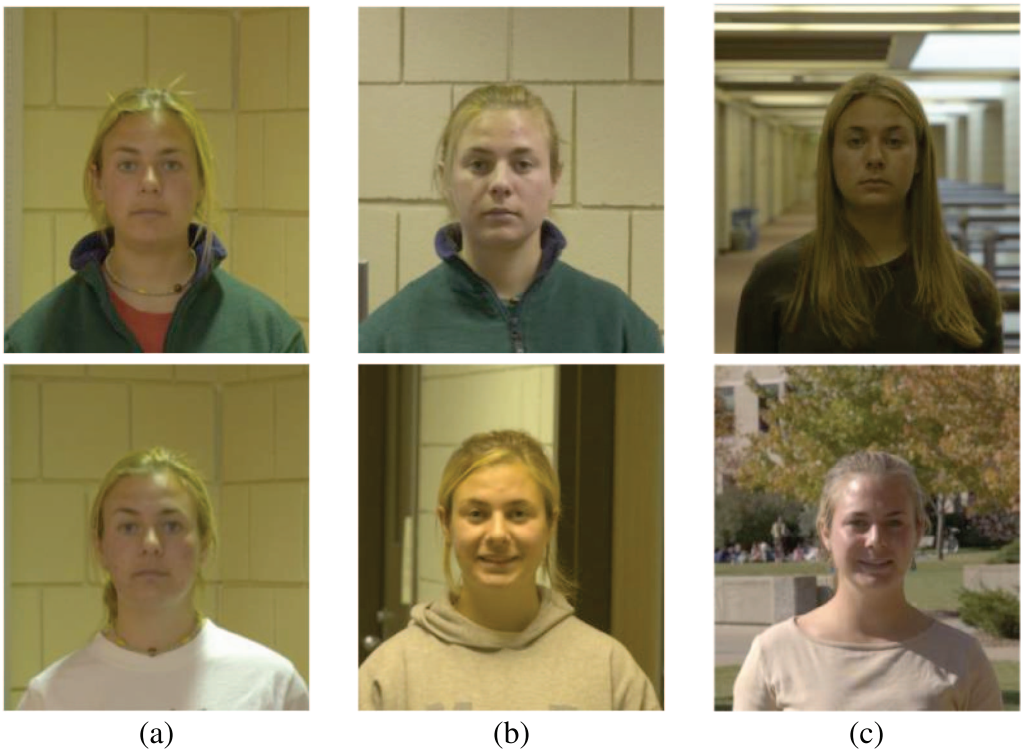


Fig. 1. Examples of face pairs of the same person from each of the partitions: (a) good, (b) challenging, and (c) very challenging.

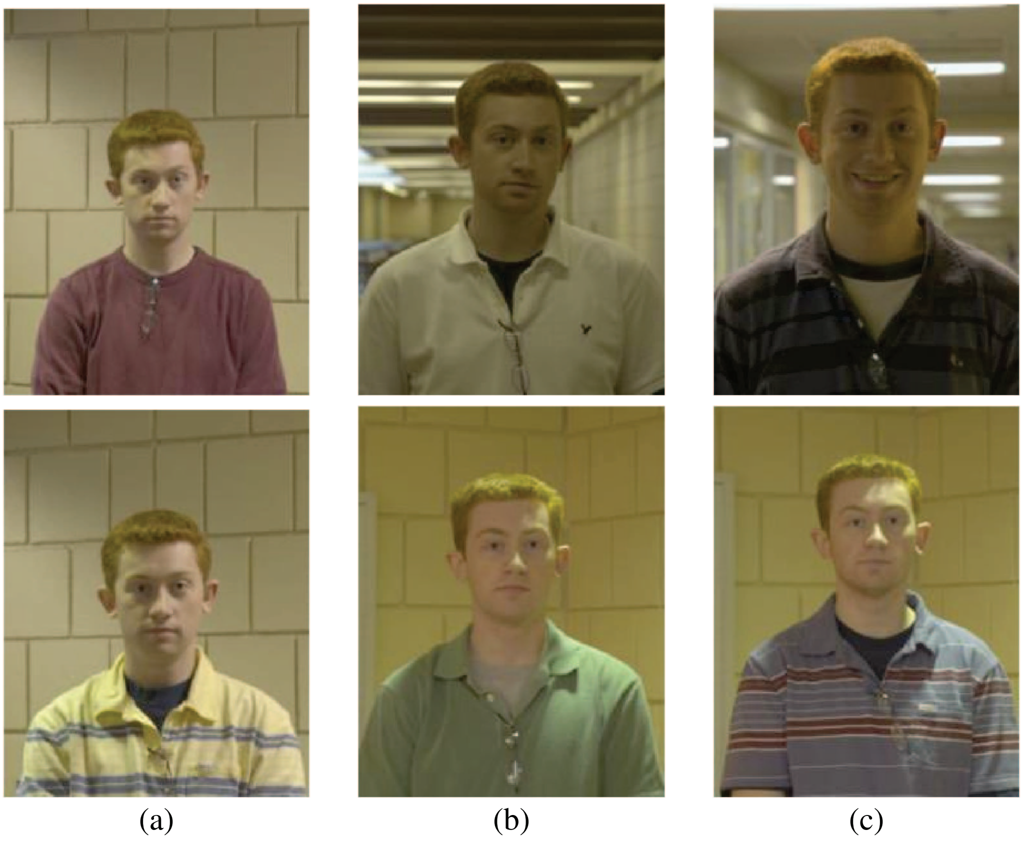


Fig. 2. Examples of face pairs of the same person from each of the partitions: (a) good, (b) challenging, and (c) very challenging.

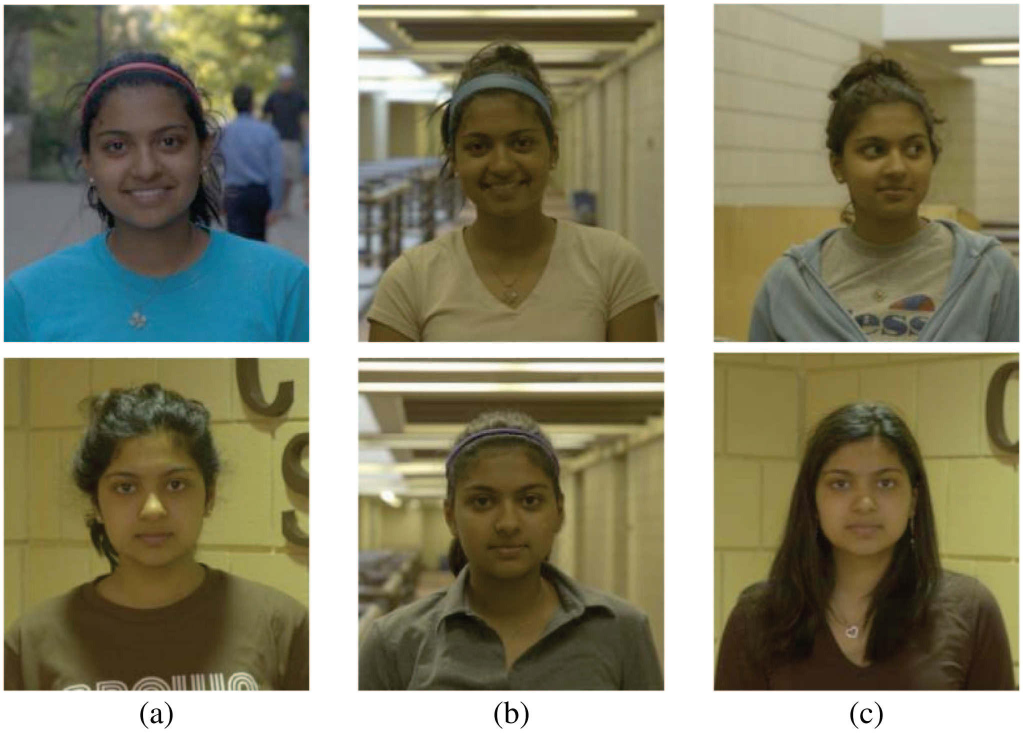


Fig. 3. Examples of face pairs of the same person from each of the partitions: (a) good, (b) challenging, and (c) very challenging.

the data. The similarity scores were selected to evenly sample the images in the experiment. The fused similarity scores are the sum of the individual algorithm similarity scores after the median has been subtracted and then divided by the MAD. If s_k is a similarity score for algorithm k and s_f is a fusion similarity score, then $s_f = \sum_k (s_k - \text{median}_k)/\text{MAD}_k$.

Figure 4 reports performance of the fusion algorithm on each of the partitions. Figure 5 shows the distribution of the match and non-matches for the fusion algorithm on all three partitions. The non-match distribution is stable across all three partitions. The match distribution shifts for each partition. The Good partition shows the greatest difference between the median of the match and non-match distributions and the least difference for the Ugly partition.

IV. Protocol

The protocol for the GBU Challenge Problem is one-to-one matching with training, model selection, and tuning completed prior to computing performance on the partitions. Consequently, under this protocol, the similarity score $s(t, q)$ between a target image t and a query image q does not in any way depend on the other images in the target and query sets. Avoiding hidden interactions between images, other than the two being compared at the moment, provides the clearest picture of how algorithms perform. More formally, any approach that redefines similarity $s(t, q; \mathcal{T})$ such that it depends upon the target (or query) image set $\mathcal{T}$ is NOT allowed in the GBU Challenge Problem.

To maintain separation of training and test sets, an algorithm cannot be trained on images of any of the subjects in the GBU Challenge Problem. It is important to note that there are images of the subjects in the GBU problem that are in the FRGC and the MBGC data sets. These images must be excluded from model selection, training, or tuning of an algorithm.

We illustrate acceptable and unacceptable training protocols with three examples. The first example is training of a principal components analysis (PCA) based face-recognition algorithm. In a PCA-based algorithm, PCA is performed on a training set to produce a set of Eigenfaces. A face is represented by projecting a face image on the set of Eigenfaces. To meet the training requirements of the protocol, images of subjects in the GBU must be excluded from the PCA decomposition that produces a set of Eigenfaces. The benchmark algorithm in Section V includes a training set that satisfies the training protocol.

A second example is taken from a common training procedure for linear discriminant analysis (LDA) in which the algorithm is trained on the images in a target set. Training an algorithm on a GBU target set the GBU protocol. Generally, it is well known that the performance of algorithms can improve with such training, but the resulting levels of performance typically do not generalize. For example, we've conducted experiments with an LDA algorithm trained on the GBU target images and performance improved over the baseline algorithm presented, see Section V. However, when we trained our LDA algorithm following the GBU protocol, performance did not match the LDA algorithm trained on a GBU target set.

The GBU protocol does permit image specific representations as long as the representation does not depend on other

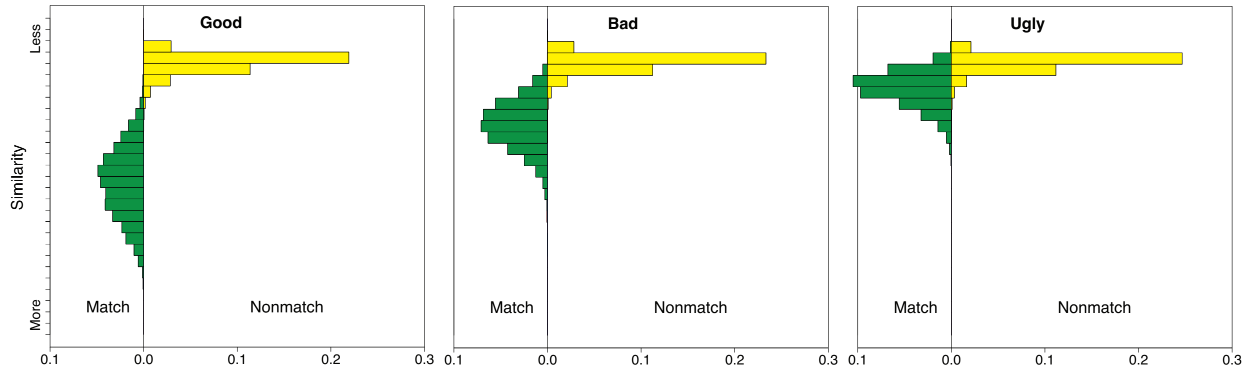


Fig. 5. Histogram of the match and non-match distributions for the Good, the Bad, & the Ugly partitions. The green bars represent the match distribution and the yellow bars represent the non-match distribution. The horizontal axes indicate relative frequency of similarity scores.

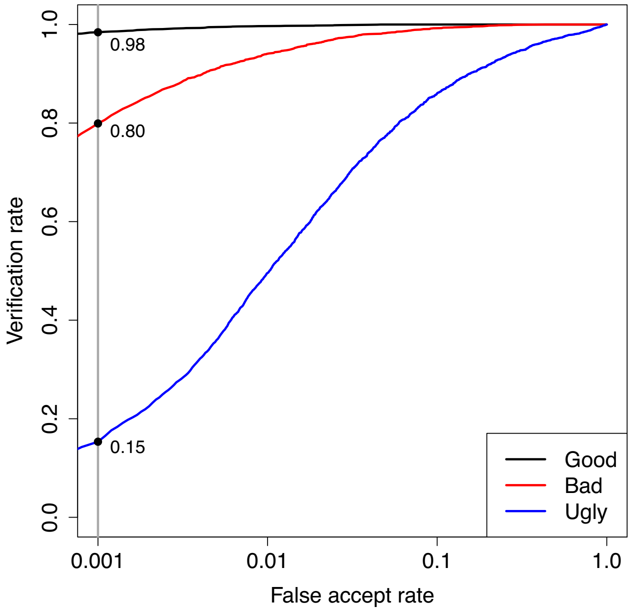


Fig. 4. ROC for the Fusion algorithm on the Good, the Bad, & the Ugly partitions. The verification rate for each partition at a FAR of 0.001 is highlighted by the vertical line at FAR=0.001.

images of other subjects in the GBU Challenge Problem. An example is an algorithm based on person-specific PCA representations. In this example, during the geometric normalization process, 20 slightly different normalized versions of the original face would be created. A person-specific PCA representation is generated from the set of 20 normalized face images. This method conforms with the GBU training protocol because the 20 face images and the person specific PCA representation are functions of the original single face image. When there are multiple images of a person in a target or query set, this approach will generate multiple image-specific representations. This training procedure does not introduce any dependence upon other images in the target set and consequently is permitted by the GBU protocol.

V. Baseline Algorithm

The GBU Challenge Problem includes a baseline face recognition algorithm as an entry point for researchers. The baseline serves two purposes. First, it provides a working example of how to carry out the GBU experiments following the protocol. This includes training, testing and evaluation using ROC analysis. Second, it provides a performance standard for algorithms applied to the GBU Challenge Problem.

The architecture of the baseline algorithm is a refined implementation of the standard PCA-based face recognition algorithm, also known as Eigenfaces [7][8]. These refinements considerably improve performance over a standard PCA-based implementation. The refinements include representing a face by local regions, a self quotient normalization step, and weighting eigenfeatures based on Fischer's criterion. We refer to the GBU baseline algorithm as local region PCA (LRPCA).

It may come as a surprise to many in the face recognition community that a PCA-based algorithm was selected for the GBU benchmark algorithm. However, when developing the LRPCA baseline algorithm, we explored numerous standard alternatives, including LDA-based algorithms and algorithms combining Gabor based features with kernel methods and support vector machines. For performance across the full range of the GBU Challenge Problem, our experiments with alternative architectures never resulted in overall performance better than the GBU baseline algorithm.

A. A Step-by-step Algorithm Description

The algorithm's first step is to extract a cropped and geometrically-normalized face region from an original face image. The original image is assumed to be a still image and the pose of the face is close to frontal. The face region in the original is scaled, rotated, and cropped to a specified size and the centers of the eyes are horizontally aligned and placed on standard pixel locations. In the baseline algorithm, the face chip is 128 by 128 pixels with the centers of the eyes spaced 64 pixels apart. The baseline algorithm runs in two modes: partially and fully automatic. In the partially automatic mode

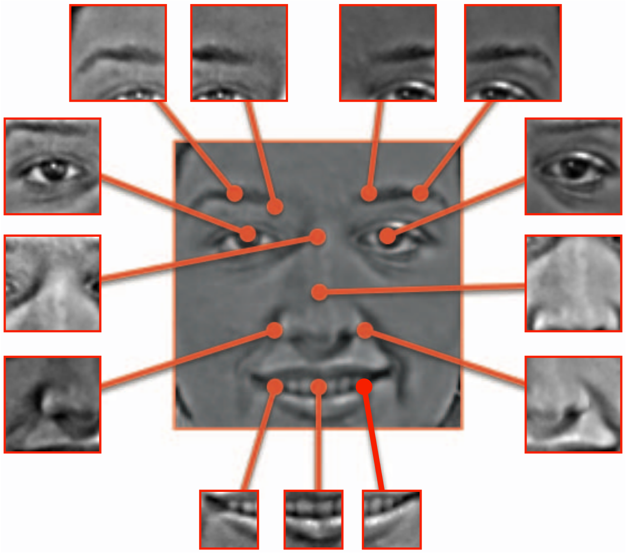


Fig. 6. This figure shows a cropped face and the thirteen local regions. The crop face has been geometrically normalized and the self quotient procedure performed.

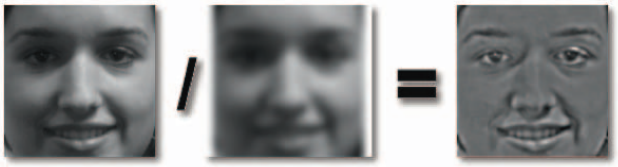


Fig. 7. This figure illustrates the computation of a self-quotient face image. The face image to the left is a cropped and geometrically normalized image. The image in the middle is the geometrically normalized image blurred by a Gaussian kernel. The image on the left is a self-quotient image. This image is obtained by pixel-wise division of the normalized image by the blurred image.

the coordinates of the centers of the eyes are provided; in the fully automatic mode, the centers of the eyes are located by the baseline algorithm.

In the LRPCA algorithm, the PCA representation is based on thirteen local regions and the complete face chip. The thirteen local regions are cropped out of a normalized face image. Some of the local regions overlap, see Figure 6. The local regions are centered relative to the average location of the eyes, eyebrows, nose and mouth.

The next step normalizes the 14 face regions to attenuate variation in illumination. First, self quotient normalization is independently applied to each of the 14 regions [9]. The self quotient normalization procedure first smoothes each region by convolving it with a two-dimensional Gaussian kernel and then divides the original region by the smoothed region, see Figure 7. In the final normalization step, the pixel values in each region are further adjusted to have a sample mean of zero and a sample standard deviation of one.

During training, 14 distinct PCA subspaces are constructed, one for each of the face regions. From each PCA decomposition, the 3*rd* through 252*th* eigenvectors are retained to represent the face. The decision to use these eigenvectors was based upon experiments on images similar to the images in the GBU Challenge Problem. A region in a face is encoded by the 250 coefficients computed by projecting the region onto the region's 250 eigenvectors. A face is encoded by concatenating the the 250 coefficients for each of the 14 regions into a new vector of length 3500.

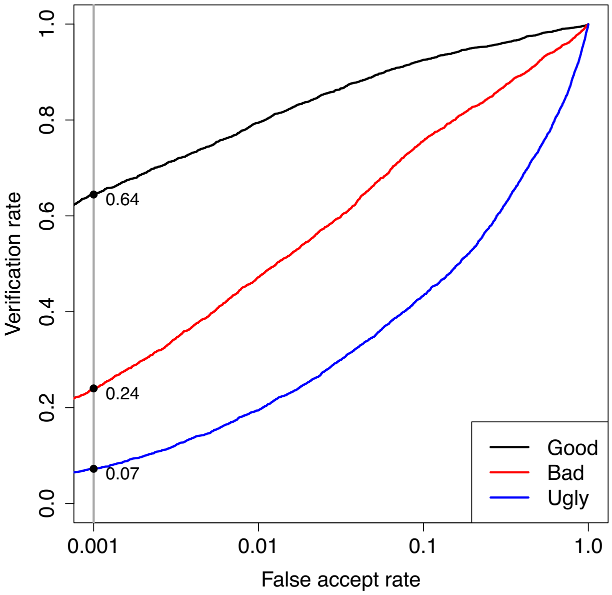


Fig. 8. ROC for the LRPCA baseline algorithm on the GBU partitions. The verification rate for each partition at a FAR of 0.001 is highlighted by the vertical line at FAR=0.001.

Each dimension in the PCA subspace is further scaled. First, the representation is whitened by scaling each dimension to have a sample standard deviation of one on the training set. Next, the weight on each dimension is further adjusted based on Fisher's criterion. This weight is computed based on the images in the training set. The Fisher's criterion weight emphasizes the dimensions along which images of different people are spread apart. The weight attenuates the dimensions along which the average distance between images of the same person and images of different people are roughly the same.

When used for recognition, i.e. during testing, images are first processed as described above and then projected into the 14 distinct PCA subspaces associated with each of the 14 regions. The coordinates of images projected into these spaces, 250 for each of the 14 regions, are then concatenated into a single feature vector representing the appearance of that face. This produces one vector per face image; each vector contains 3,500 values. The baseline algorithm measures similarity between pairs of faces by computing the Pearson's correlation coefficient between pairs of these vectors. The performance of the baseline algorithm on the GBU Challenge Problem is summarized in Figure 8. A comparison of performance of the Fusion and the LRPCA-baseline algorithm is given in Table I.

A recent area of interest in face recognition and biometrics is recognition from the ocular region of the face. There is interest in recognition from both near infrared and visible imagery. The region-based design of the LRPCA algorithm allows for baselining ocular performance on the GBU partitions. Baseline performance for the left ocular is computed from three of the 14 regions. The regions are the left eye and two left eye brow regions. For the right ocular region, performance is computed from the right eye and two right eye brow regions. The left eye (resp. right eye) are with respective to the subject; e.g., the left ocular region corresponds to a subject left eye. Performance for the LRPCA-ocular baseline for the left and right ocular regions is given in Figure 9.

A summary of performance of the Fusion, the LRPCA-face baseline and the LRPCA-ocular baseline algorithms are given in Table I.

TABLE I

PERFORMANCE OF THE FUSION, THE LRPCA-FACE BASELINE AND THE LRPCA-OCULAR BASELINE ALGORITHMS. FOR THE OCULAR BASELINE, PERFORMANCE IS GIVEN FOR BOTH THE LEFT AND THE RIGHT OCULAR REGIONS. THE VERIFICATION RATE AT A FAR = 0.001 IS GIVEN.

			LRPCA-ocular	
Partition	Fusion	LRPCA-face	left	right
Good	0.98	0.64	0.47	0.46
Bad	0.80	0.24	0.16	0.17
Ugly	0.15	0.07	0.05	0.05

VI. DISCUSSION AND CONCLUSION

This paper introduces the Good, the Bad, & the Ugly Challenge Problem. The main goal of the challenge is to encourage the development of algorithms that are robust to recognizing frontal faces taken outside of studio style image collections. The three partitions in the GBU Challenge Problem emphasize the range of performance that is possible when comparing faces photographed under these conditions. This structure allows for researchers to concentrate on the "hard" aspects of the problem while not compromising performance on the "easier" aspects.

Partitioning the challenge by levels of difficulty is the most prominent feature of the GBU Challenge Problem design. Another is controlling for the "recognizability" of people by selecting images of the same 437 people for inclusion in each of the GBU partitions. The data in the three partitions is further balanced so as to ensure that for each person the number of target and query images in each partition is the same. The design of the GBU Challenge Problem means that any difference in performance observed between partitions cannot be attributed to differences between people or numbers of images for individual people.

The unique design of the GBU Challenge Problem allows researchers to investigate factors that influence the performance of algorithms. O'Toole et al. [10] looks at the demographic effects on the nonmatch distribution. Beveridge et al. [11] shows that the quality of face images comes in pairs. Additional possible lines of investigation include understanding the factors that characterize the difference in match face pairs across the partitions. A second line of research is characterizing the recognizability of a face; e.g., the biometric zoo. A third line of research is developing methods for predicting performance of face recognition algorithms. The design of the GBU Challenge Problem encourages both the development of algorithms, and the investigation of methods for understanding algorithm performance.

ACKNOWLEDGMENTS

The authors wish to thank the Federal Bureau of Investigation (FBI) and the Technical Support Working Group (TSWG) for their support of this work. The identification of any commercial product or trade name does not imply endorsement or recommendation by the Colorado State University, the National Institute of Standards and Technology or the University of Texas at Dallas. The authors thank Jay Scallan for his assistance in preparing the GBU challenge problem.

REFERENCES

[1] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing Journal*, vol. 16, no. 5, pp. 295–306, 1998.

[2] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. PAMI*, vol. 22, pp. 1090–1104, October 2000.

[3] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 947–954.

[4] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale results," *IEEE Trans. PAMI*, vol. 32, no. 5, pp. 831–846, 2010.

[5] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan III, and S. Weimer, "Overview of the Multiple Biometrics Grand Challenge," in *Proceedings Third IAPR International Conference on Biometrics*, 2009.

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

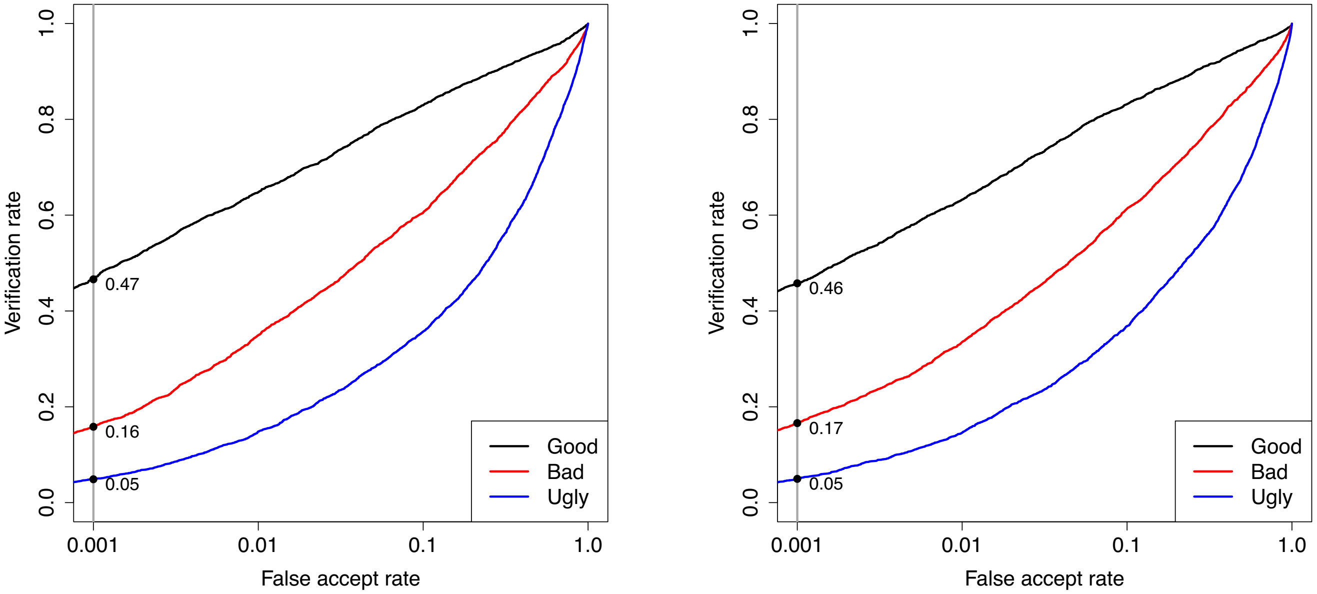
[7] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[8] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE Trans. PAMI*, vol. 12, no. 1, pp. 103–108, 1990.

[9] H. Wang, S. Li, Y. Wang, and J. Zhang, "Self quotient image for face recognition," in *Proceedings, International Conference on Image Processing*, vol. 2, 2004, pp. 1397–1400.

[10] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop, "Demographic effects on estimates of automatic face recognition performance," in *Proceedings, Ninth International Conference on Automatic Face and Gesture Recognition*, 2011.

[11] J. R. Beveridge, P. J. Phillips, G. H. Givens, B. A. Draper, M. N. Teli, and D. S. Bolme, "When high-quality face images match poorly," in *Proceedings, Ninth International Conference on Automatic Face and Gesture Recognition*, 2011.



(a)

(b)

Fig. 9. ROC for the LRPCA-ocular baseline algorithm on the Good, the Bad, & the Ugly partitions. In (a) performance is for the left ocular region that consists of the left eye and two left eye-brow regions; performance in (b) is for corresponding right ocular regions. The verification rate for each partition at a FAR of 0.001 is highlighted by the vertical lines at FAR=0.001.