On the Existence of Face Quality Measures

P. Jonathon Phillips* J. Ross Beveridge† David Bolme‡ Bruce A. Draper†,
Geof H. Givens§ Yui Man Lui† Su Cheng* Mohammad Nayeem Teli† Hao Zhang†

Abstract

We investigate the existence of quality measures for face recognition. First, we introduce the concept of an oracle for image quality in the context of face recognition. Next we introduce greedy pruned ordering (GPO) as an approximation to an image quality oracle. GPO analysis provides an estimated upper bound for quality measures, given a face recognition algorithm and data set. We then assess the performance of 12 commonly proposed face image quality measures against this standard. In addition, we investigate the potential for learning new quality measures via supervised learning. Finally, we show that GPO analysis is applicable to other biometrics.

1. Introduction

Most research into quality measures in biometrics implicitly assumes that image quality measures exist. The assumption is manifested in the structure of the papers, which evaluate the effectiveness of one or more quality measures. A search for effective quality measures assumes their existence.

We start from an existential perspective: do quality measures exist? This question immediately raises two additional questions. First, what is a quality measure? Second, what does it mean for a quality measure to exist? Our answer to the first question is that an image quality measure should be predictive of recognition performance [6],[11]. This may not be obvious, because quality measures describe images, while similarity is a function over pairs of images. A high quality face image, however, is similar to other images of the same person, but different from images of other people. As a result, a high quality image is one that causes the overall system performance to improve when it is added to the data set, or to deteriorate when it is removed. The opposite is true for low quality images.

Addressing the existence of quality measures is tricky. To lay the foundation for this discussion, we introduce an oracle for image quality. The oracle ranks the images in a data set according to their quality, given a recognition algorithm. Since it is generally not practical to compute an oracle's answer, we approximate this ranking by greedy pruned ordering (GPO). GPO analysis provides an estimated upper bound for quality measures given an algorithm and a data set. Quality measures better than the GPO probably do not exist.

The performance of twelve face quality measures from the literature are compared to the GPO derived upper bound. We also examine the ability to learn near optimal quality measures. Experiments are conducted on two data sets[2]. The first data set consists of the Good and Ugly partitions from the Good, Bad, and Ugly (GBU) face challenge problem [10]. All the faces are nominally frontal and the majority are visually high quality to humans. On the Good partition, the false reject rate (FRR) is 0.02 at a false accept rate (FAR) of 1 in a 1000; on the Ugly partition, the corresponding FRR is 0.85. The goal of the quality experiments on the Good-Ugly data set is to separate the images from the Good and the Ugly partitions.

The second data set consists of images taken with digital point and shoot cameras, the PaSC data set[3] [5]. The PaSC images were taken in ambient lighting with a set of five point and shoot cameras. The visual quality in the image set ranges from poor to high. The goal of experiments on the PaSC data is to measure the effectiveness of quality measures on data sets with a large range of visual quality.

*P. J. Phillips and S. Cheng are with the National Institute of Standards and Technology, 100 Bureau Dr., MS 8940 Gaithersburg MD 20899, USA.

†J. R. Beveridge, B. A. Draper, Y-M Lui, M. N. Teli, and H. Zhang are with the Department of Computer Science, Colorado State U., Fort Collins, CO 46556, USA.

‡D. S. Bolme is with the Oakridge National Laboratory, Knoxville TN.

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

[1]CSU was funded in part by the Department of Defense. PJP and SC were funded in part by the Federal Bureau of Investigation. DB was at CSU when this work was performed. The identification of any commercial product or trade name does not imply endorsement or recommendation by Colorado State U., Oakridge National Laboratory, or NIST.

[2]For information on obtaining the data sets used in this paper and additional supporting material, goto `http://face.nist.gov`.

[3]The collection of the PaSC data set was funded under IARPA's BEST program.

2. Review of Prior Work

A number of papers have looked at face quality measures. Hsu and Martin [7] correlate human subjective judgments of image quality to machine recognition performance (they match). Adler and Dembinsky [2] correlate a black-box vendor-supplied quality measure and to recognition performance. Abdel-Mottalib and Mahoor [1] study the effectiveness of a machine quality measure based on kurtosis, Weber [13] evaluates a quality measure based on sharpness and Beveridge et al. [4] measure quality via edge density.

Phillips and Beveridge [9] present a theoretical argument linking the concept of quality measures used to predict recognition performance back to the original problem of recognition. They conclude that, in the limit, a perfect assessment of quality pre-supposes understanding how to construct a perfect recognition algorithm. While the approach in this paper is very different, this work builds on [9] in so much as it establishes an empirical method, GPO, for establishing the best performance any quality measure could achieve on a particular combination of algorithm and data set .

3. Fundamentals and Oracle Analysis

The starting point for our discussion is an oracle model. In our model, the oracle is asked which images, if removed, would most improve the system's performance, and the oracle always provides the correct answer. For example, given a set of images and an algorithm, the oracle might be asked which 10 images, if removed, would produce the lowest FRR at a FAR of 1 in 1000. This question is roughly equivalent to asking what are the ten lowest quality images, therefore the oracle can be used to rank images by image quality. A second related question is, given a set of images and an algorithm, identify the smallest set of images which, if removed, drop the FRR to a desired goal. This second question is roughly equivalent to asking what are the low quality images that prevent system performance from reaching a performance goal.

On a dataset where potential quality measures are to be evaluated, the answers to questions asked of the quality oracle can be computed. In principal, an exhaustive search through all subsets of N images in the data set would lead to discovering which set of images is the answer to the question posed above. Such an oracle would consider the implications of removing all *combinations* of images. The choice of images to remove represents the best that could possibly be achieved and does so without regard to any underlying measured properties of the images themselves. This provides a gold standard for how an image quality measure ought to rank and select images for removal from a data set.

Given the exponential nature of the search and the requirement that the complete performance metric, for example FRR at a fixed FAR, be recomputed each time an image is removed, finding the true globally optimal solutions implied by the definition of the oracle is not feasible. However, a good approximation based upon a greedy heuristic is practical and of considerable value. Therefore, we introduce a technique called greedy pruned order (GPO) that iteratively prunes images from a data set. Greedy methods, such as the GPO, are standard approximations for combinatorial optimization problems.

GPO is an iterative algorithm. In each iteration, the image whose removal results in the greatest reduction in the error rate is pruned. The iterative procedure continues until the error rate is zero or all images have been pruned. Note the entire process never looks at image data itself; it is driven solely by the similarity scores already computed for the data set to which this analysis is being applied.

The ordering of images generated by the GPO procedure formally characterizes, for a given data set and algorithm, the very best any quality measure could do if asked to rank images for removal. It should not be confused with a quality measure derived from measurable properties of images. Instead, it should be thought of as a new yardstick against which to quantify the value of image quality measures. It represents a strong, well defined upper bound on what any image-derived quality measure could yield in terms of improved recognition.

4. Data sets

This section presents experiments designed to study the existence of face image quality measures from two perspectives. The first experiment considers the case where the data separate relatively well into two performance extremes. A quality measure in this context should identify the two extremes and hence separate the data. The second perspective is that of a more graduated situation involving a range of images varying between visually unrecognizable faces to visually high quality faces with everything in between. The goal on this data set is to find a quality measure that can handle this range of quality and rank choices of what images to retain accordingly.

One of the major factors effecting face recognition performance is pose. Nominal pose can be established by an acquisition policy or approximately estimated. In our quality studies, we assume an acquisition policy that all faces are nominally frontal. This allows our study to concentrate on face quality measures that are harder to identify.

Performance extremes are studied using the union of the Good and Ugly partitions from the GBU data set [10]. On the Good partition, performance on the FRVT 2006 fusion has a FRR of 0.02 at a FAR of 1 in 1000, and a corresponding FRR of 0.85 on the Ugly partition. The images were taken with a Nikon D70 and most are visually high quality.

The faces are nominally frontal and all images were taken within 10 months. The Good and Ugly partitions have the same number of images per person. These constraints mean that differences in performance between the partitions depends on variations of a subject's imaging conditions, not aging, different subjects, or pose. Experiments on this data set measure the ability of quality measures to divide the face images into their original partitions. Experiments are based on the FRVT 2006 fusion algorithm.

The second data set consists of 4688 still images taken with a digital point and shoot camera or cell phone camera: the PaSC data set [5]. All the faces in the images are nominally frontal and were taken between January and April 2011. The photos were taken with 5 point and shoot cameras at 8 locations. These images include complications that are rare in previous publicly released data sets, including poor imaging conditions, blur, over and under exposure, and compression artifacts. This is despite mega-pixel counts of 12 to 14 on these cameras. The majority of the PaSC images contain either the full body or body above the torso, and have sufficient pixels on the face to support recognition. Figure 1 shows a sampling of the visual quality of the faces. Experimental results for the PaSC images are based on face detection and recognition by the Pittsburgh Pattern (PittPatt) Recognition SDK 5.2.2.

Figure 1. Eight faces showing a sampling of the visual quality of still images in the PaSC.

5. Quality Assessment

It is well established in biometrics that performance of an algorithm needs to report both type I and type II errors, with FRR and FAR as the basis for analysis. In this section we extend the traditional FRR and FAR trade-off to include quality measures, which follows previous work [6][11].

Quality measures are included in the characterization of type I and type II errors in the following manner. A new threshold, a quality threshold, is introduced and then a data set is pruned to leave only those images above the quality threshold. Shifting the quality threshold and computing a performance metric, in particular the FRR at FAR of 1 in 100, for each threshold, relates pruning by the metric

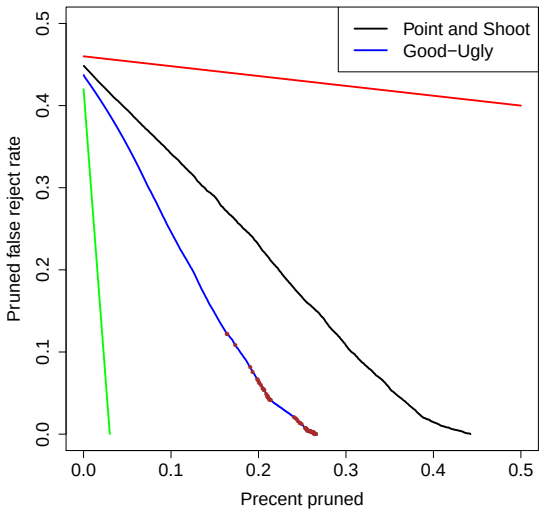


Figure 2. Greedy pruned order analysis on the PaSC and Good-Ugly data sets. For the Good-Ugly data, iterations where an image from the Good partition was removed is marked in brown. The red line is a nominal GPO curve when quality measure are not effective. The green line is a nominal GPO where there could exist an effective quality measures, see Section 9.

to recognition performance. This new curve, performance versus quality threshold, captures the trade-off between algorithm performance and the fraction of images pruned.

This analysis indirectly measures the effect of quality measures on FAR. We also present a method for directly measuring the effect of quality pruning on FAR. Based on the unpruned set of images, we compute a global threshold λ_g that yields a FAR = 0.01. To obtain a direct measure of the effect of pruning on FAR, we calculate the FAR at the global threshold λ_g on sets of quality pruned images.

6. Greedy Pruned Order Experiments

To provide insight into the potential for face image quality measures in face recognition, GPO order analysis was conduced on the both the PaSC and Good-Ugly data sets. This analysis is reported in Figure 2, where the vertical axis reports the FRR at a FAR of 1 in 100 and the horizontal axis is the fraction of images pruned. For the PaSC data set 44% of the images needed to be pruned to achieve FRR = 0 at a FAR of 1 in 100. On the Good-Ugly data set, 27% of the images needed to be pruned to achieve the same result. A total of 1154 images were removed, and 71 of these images were from the Good partition.

Both data sets had similar FRRs prior to pruning. The sharper drop in the Good-Ugly data set could be due to the error structure in this data set. The errors were concentrated

in the Ugly partitions and a large majority of the pruned images were from the Ugly partition. Remember, GPO analysis provides best case upper bounds on quality metric performance, given an algorithm and data set.

7. Computational Quality Measures

Many quality measures have been proposed in the literature for faces. In this section we evaluate the performance of twelve popularly proposed quality measures.

7.1. Quality Measures Investigated

The quality measures we investigate fall into three categories. The first consists of measures that can be computed from images. The second consists of camera settings that can be found in the exchangeable image file format (EXIF) header. The third category consists of our support vector machine (SVM) quality measures learned from the quality measures in the first and second categories. These measures explore the possibility that combinations of common quality measures may be effective.

The quality measures examined are (in alphabetical order):

Edge Density : A measure of the strength of edges in images that has been used to predict the performance of algorithms in prior work [4]. It was designed to measure focus but also responds to strong lighting [4].

Exposure time : The time the camera shutter was open as recorded in the image EXIF data. Longer shutter times indicate dimmer illumination and possibly increased motion blur if the scene is not static.

Eye distance : The distance between the eyes in pixels. Related to distance from the camera but a more direct measurement of what matters to algorithms. Typically more pixels on the face is better.

Face saturation : A measure of the number of face pixels that are saturated in the image, i.e., driven to the maximum intensity value. This is an indicator of over exposure.

Face standard deviation : The standard deviation of the face pixel values. This relates to the contrast of the face region in the image.

F stop : A measure of the amount of light that passes through the optics of the camera and is the ratio of the focus distance to the effective aperture. Recorded in the EXIF header.

Focus : The Strong Edge Motion Compensated (SEMC) focus measure from the face region of an image [3].

ISO speed rating : An EXIF camera setting that compensates for low light conditions. Typically, low ISO speeds near 100 indicate plenty of lighting and a good quality exposure. Higher ISO speeds indicate that the camera has used electronic gain to boost the signal from the sensor under poor illumination.

Left-right hist : A measure of the directionality of the illumination on the face. It is the intersection of the intensity value histograms taken from the right and left sides of the face. A value of zero indicates even illumination. Assumes frontal images.

Illumination direction : An estimate of lighting direction computed from the face image. Positive values indicate the face is illuminated from the front. Negative indicate strong side lighting that may be a problem for the algorithms.

Mean ratio : The ratio of the mean pixel value of the face region to the mean pixel value of the entire image. It is an indicator of whether the face is over- or under-exposed compared to the rest of the image.

Pose : An estimate of how close to frontal an image is where zero indicates the person is looking directly at the camera and larger values indicate increasing yaw angles.

SVM quality : A quality measure trained by an SVM to predict the quality of a face image. This features is a combination of the above twelve quality measures. The SVM was not trained on the GBU or PaSC data sets.

7.2. Experimental Results

For the PaSC and Good-Ugly data sets, performance of the quality measures is summarized in Figures 3 and 4. Both FRR and FAR are reported with 20% pruning. For the PaSC data set, quality results are reported for all measures described in Section 7.1. For the Good-Ugly data set, the quality measures focus, exposure time, and ISO rating were not available. To provide a base case upper bound for quality performance, the GPO analysis performance level is plotted.

For the PaSC data set, exposure time and face saturation were the most effective at reducing FRR at the 20% pruning level. In Figure 5 we examine their performance in greater detail. The change in FRR is plotted against the fraction of images pruned. The GPO best case upper bound is also plotted.

The results show that there is a substantial gap between the performance of the computational quality measures and the GPO upper bound. Also, in the FAR variation portions (green) of Figures 3 and 4, FAR increases in 13 out of 23 cases. This indicates more research is needed to better understand how pruning data sets based upon face image quality measures alters the non-match score population in general and FAR in particular.

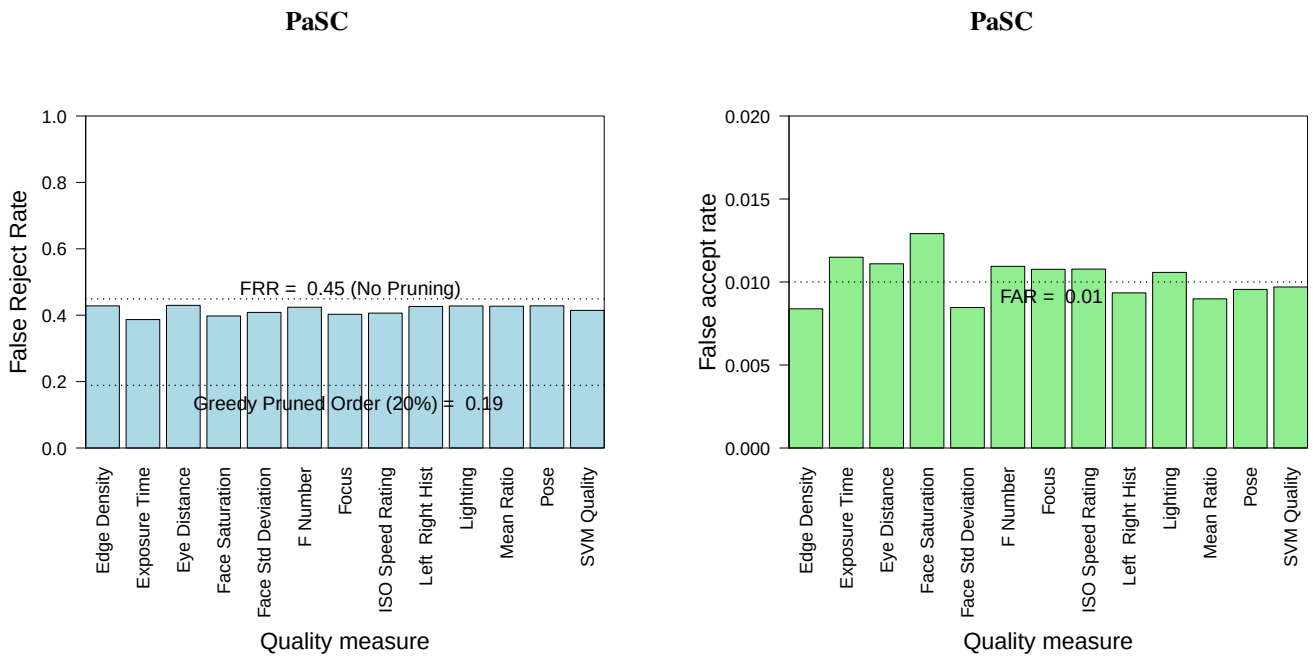


Figure 3. Performance of thirteen quality metrics on the PaSC images. In the left panel, FRR at a FAR of 0.01 is reported when the 20% lowest quality images are removed. The dashed line at FRR = 0.45 is performance with no pruning, and the line at FRR = 0.19 is the GPO for 20% pruning. In the right plot, FAR is reported when 20% of the lowest quality images are removed. The dashed line at FAR = 0.01 is performance with no pruning.

8. Learning Best Case Quality Measures

One approach to developing a quality measure is to learn it from the data. The goal is to learn a quality measure with performance close to the GPO upper bound. The learning task can be reduced to a two class problem. In the PaSC data set, the images are divided into two classes. The first class consists of images labeled as low quality. An image is labeled as low quality if it was pruned by the GPO analysis in Section 6. The second class consists of images labeled as high quality. An image is labeled as high quality if it was not pruned by the GPO analysis. If the performance of the classifier was 100%, then recognition performance on the high quality images would be FRR = 0 at a FAR = 0.01.

In this experiment, a classifier was trained to distinguish between low and high quality images as defined above. The classifier is principal component analysis (PCA) followed by linear discriminant analysis (LDA). The classifiers were trained on face regions after they had been normalized for size, rotation, and illumination; for details see Liu et al [8].

Separate training and test sets were used. For the low quality image class, images for the training and sets were selected from the first 1000 images pruned by the GPO method. For the high quality images, the training and testing sets consisted of images not pruned in the GPO analysis. Training and test sets consisted of 500 images randomly chosen (no overlap) from these sets. The experiment was repeated for different numbers of eigenvectors kept from the PCA components. The results are presented in Figure 6. Performance on both the training (blue) and test (green) sets is plotted. Classifier performance is characterized by the number of eigenvectors kept for the LDA classifier. The first n eigenvectors are kept in increments of 50. Classifier training and testing was replicated 100 times. In Figure 6, the results are of the 100 replicates are reported with a box-whisker plot. For example, performance on the test set for 50 eigenvectors is the left most green box-whisker plot, which a median classification rate of 0.63. The corresponding performance rate on the training data, left most blue box-whisker plot, is 0.69.

The results of the experiment show that it is possible to train a classifier to separate high quality from low quality images over the training set. Unfortunately, the results do not generalize to the test set. In fact, better performance on the training set comes at the cost of reduced performance on the test set. This is a classic case of over training a classifier. The best performance achieved on the test set was with 100 eigenvectors and median accuracy of 0.65. Note, 0.50 is random performance.

A similar experiment was performed training a PCA-LDA classifier to learn the difference between images in

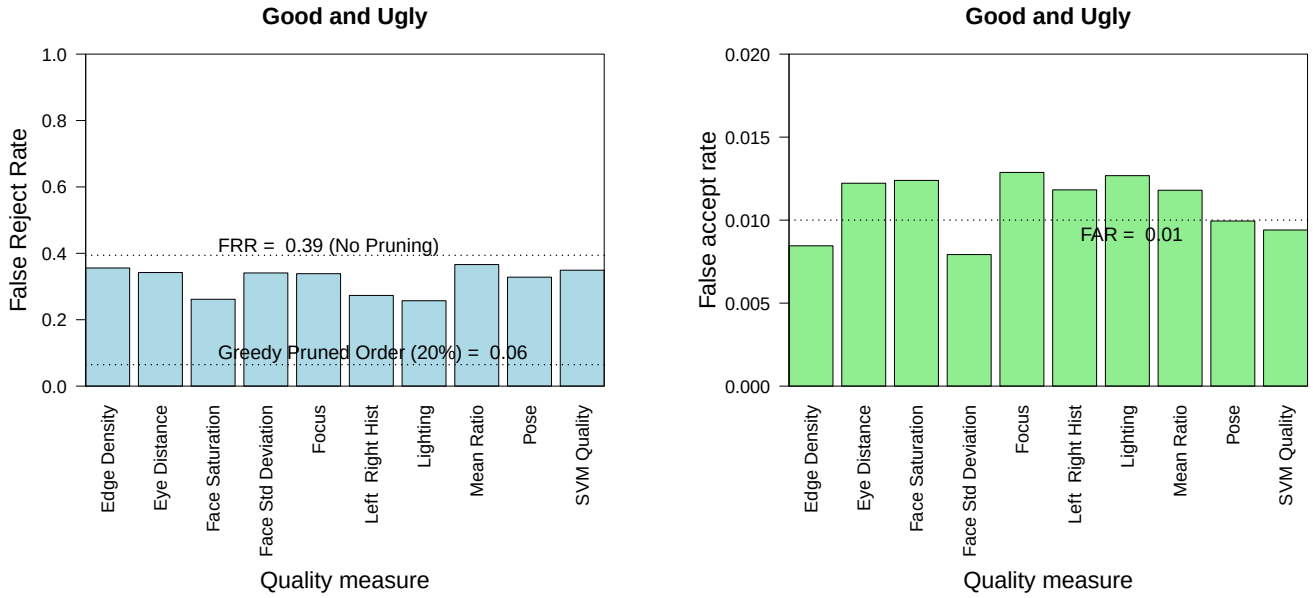


Figure 4. Performance of eleven quality metrics on the Good-Ugly images. In the left plot, FRR at a FAR of 0.01 is reported when the 20% lowest quality images are removed. The dashed line at FRR = 0.39 is performance with no pruning, and the line at FRR = 0.06 is the GPO for 20% pruning. In the right plot, FAR is reported when 20% of the lowest quality images are removed. The dashed line at FAR = 0.01 is performance with no pruning.

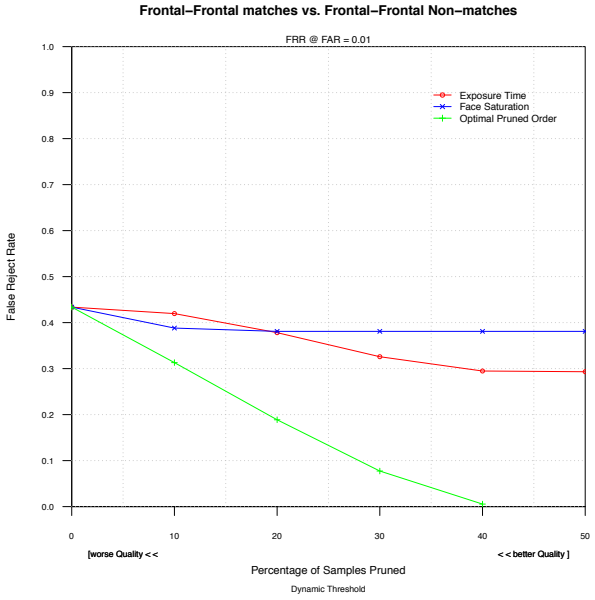


Figure 5. Quality performance trade-off plots for exposure time, face saturation, and GPO. The horizontal axis is fraction of images pruned by the three measures. The vertical axis is the corresponding FRR at a FAR of 0.01.

the Good and Ugly partitions. The results were similar and the conclusions the same.

9. Discussion

In the experiments conducted in this paper, the results of the GPO analysis provided best case performance upper bounds for computational quality measures. GPO analysis can provide insight into the overall effectiveness of any quality measure for a data set. This is highlighted in Figure 2, which has the GPO curves for both the PaSC and Good-Ugly data sets. Added to the figure are two hypothetical GPO curves. The green curve models a case were a quality measure is highly effective, and removal of a small fraction of the images greatly reduces the error rate. This situation would occur when the errors are concentrated in a small number of images. The red curve models the case where no quality measure would be effective. This situation occurs when errors are equally spread across all images. Thus, pruning a set of images would not substantially change the overall error rate. The GPO curves plotted for both face data sets are in between the two extremes. For the PaSC data set, 44% of the images need to be pruned to reduce the FRR to 0.0, and the FRR is 19% after 20% of the images are pruned. For the Good-Ugly data, the corresponding pruning percentages are 27% and 6%.

The formulation for GPO is general and is applicable to

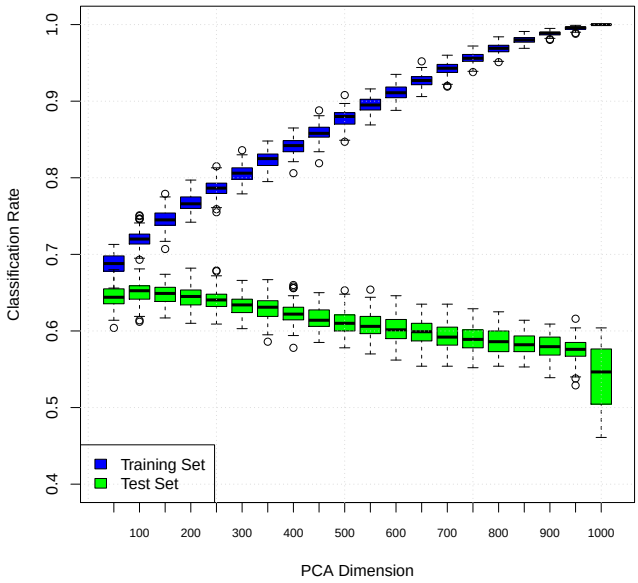


Figure 6. Accuracy of learning a quality measure on the PaSC images. Horizontal axis is number PCA dimensions kept for the LDA classifier. The vertical axis is classification rate reported with a box-whisker plot.

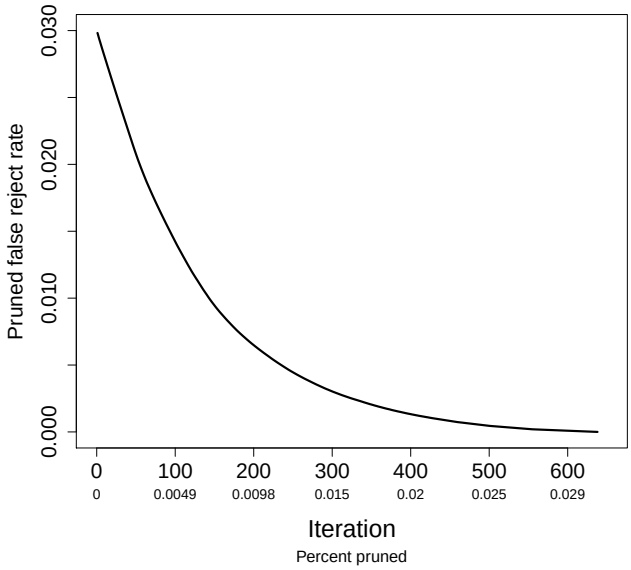


Figure 7. GPO analysis on the ICE 2006 data set.

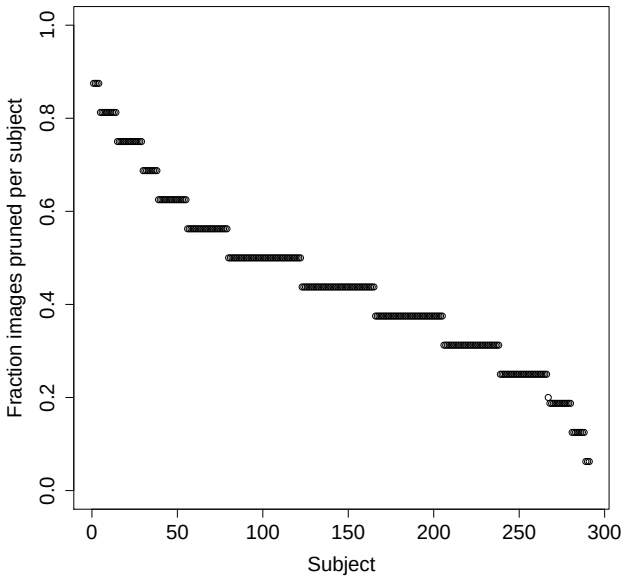
any biometric. To show the ability to generalize beyond face, Figure 7 shows a GPO analysis on iris recognition from the ICE 2006 [12]. The shape of the GPO curve for iris differs from that for face with the iris GPO curve resembling the function 1 over f.

This change for iris relative to face is noteworthy, and ties into a related concern. A quality measure might inadvertently prune a small set of subjects. To pursue this concern as well as explore the possible reason for the rapid drop in the GPO curve for iris, we looked at the fraction of images for each subject pruned by the GPO analysis. Figure 8 shows the results of subject analysis for the PaSC and ICE 2006 data sets. The horizontal axis is subject and the vertical axis is the fraction of images pruned in the GPO analysis. For example, in the ICE 2006, 34% of the iris images for subject 1 were pruned by the GPO. Because the minimum number of iris images per subject was 40, the high fraction of pruned images per subject was not caused by small sample sizes. Figure 8 suggests that for the PaSC data set the images pruned were relatively evenly distributed over the subjects. However, for the ICE 2006, there are a few subjects where a large portion of their images are pruned.

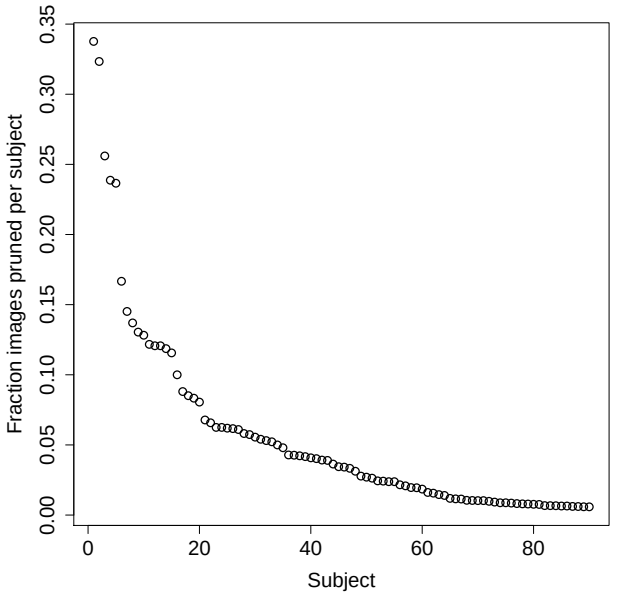
There are at least two sources of recognition errors. The first are imaging or environment covariates. In this case, errors are cause by poor imaging conditions or environmental considerations. One goal of quality measures is to identity the poor imaging conditions. The second are errors correlated with the subjects. An example of a source of errors correlated with a subject could be beards or contact lenses. In this case, the effect is that some subjects are harder to recognize and effective quality measures may prune subjects. The analysis of the errors from the ICE 2006 data suggests that this is a possibility. Also, our results suggest that subject based effects on quality measures may warrant further attention in the context of iris recognition. In general, these results suggest that subject based effects on quality measures may need to be investigated.

10. Conclusions

The Greedy Pruning Order analysis introduced in this paper represents a major step forward in formally bounding the improvement a face recognition system might achieve using an image quality metric to discard images prior to recognition. With GPO upper bounds established for two difficult face recognition data sets, 12 popular measures of face quality were compared to this upper bound. The results were not encouraging and gains using measured image properties showed room for improvement. Next, to test a strong machine learning alternative, an experiment with a trained quality classifier was carried out. The classifier could be made to perform well on the training data, but only at the expense of performance on the test data. Finally, both to underscore the generality of GPO analysis, and highlight



(a)



(b)

Figure 8. Subject impact on GPO analysis. Results on the (a) PaSC and (b) ICE 2006 data sets.

differences between face and iris recognition, GPO results on the ICE 2006 performance data were presented.

References

[1] M. Abdel-Mottaleb and M. Mahoor. Application notes-algorithms for assessing the quality of facial images. *IEEE Computational Intelligence Magazine*, 2:10–17, 2007.

[2] A. Adler and T. Dembinsky. Human vs. automatic measurement of biometric sample quality. In *CCECE*, 2006.

[3] J. 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 *Proceedings IEEE Computer Society and IEEE Biometrics Council Workshop on Biometrics*, 2010.

[4] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui. Focus on quality, predicting FRVT 2006 performance. In *Proceeding of the Eighth International Conference on Automatic Face and Gesture Recognition*, 2008.

[5] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE Conference on Biometrics: Theory, Applications and Systems*, 2013.

[6] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. PAMI*, 29:531–543, 2007.

[7] R. Hsu, J. Shah, and B. Martin. Quality assessment of facial images. In *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the*, pages 1–6. IEEE, 2007.

[8] Y. M. Lui, D. S. Bolme, P. J. Phillips, J. R. Beveridge, and B. A. Draper. Preliminary studies on the good, the bad, and the ugly face recognition challenge problem. In *CVPR Workshops*, pages 9–16, 2012.

[9] P. J. Phillips and J. R. Beveridge. An introduction to biometric-completeness: The equivalence of matching and quality. In *Third IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2009.

[10] P. J. Phillips, J. R. Beveridge, B. A. 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, and the ugly face recognition challenge problem. In *Proceedings Ninth IEEE International Conference on Automatic Face and Gesture Recognition*, 2011.

[11] P. J. Phillips and P. J. Flynn. Quality and demographic investigation of ICE 2006. In M. J. Burge and K. W. Bowyer, editors, *Handbook of Iris Recognition*. Springer, 2013.

[12] 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*, 32(5):831–846, 2010.

[13] F. Weber. Some quality measures for face images and their relationship to recognition performance. In *Biometric Quality Workshop. NIST, Maryland, USA*, 2006.