Predicting Good, Bad and Ugly Match Pairs

Gaurav Aggarwal, Soma Biswas, Patrick J. Flynn and Kevin W. Bowyer
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame
{gaggarwa, sbiswas, flynn, kwb}@nd.edu

Abstract

Several sources of variation in facial appearance have long been investigated that affect face matching performance. The recently introduced GBU challenge problem [1] indicates that even when the impact of most known factors is eliminated or significantly reduced by the data collection and experimentation protocol, there can be significant variation in performance across different partitions of the data. The GBU challenge problem consists of three partitions which are called the Good (easy to match), the Bad (average matching difficulty) and the Ugly (difficult to match). In this paper, we investigate various image and facial characteristics that can account for the observed significant difference in performance across these partitions. Given a match pair, we aim to predict the partition it belongs to. Partial Least Squares (PLS)-based regression is used to perform the prediction task. Our analysis indicates that the match pairs from the three partitions differ from each other in terms of simple but often ignored factors like image sharpness, hue, saturation and extent of facial expressions.

1. Introduction

Face recognition is one of the most active areas of research in the field of computer vision and pattern recognition and numerous algorithms have been proposed to address various aspects of the problem like variations in illumination, pose, expression, age, etc. This has given rise to a growing need for evaluating and understanding performance of different algorithms and commercial systems. There have been several recent efforts to perform large scale independent evaluations of face recognition systems in the form of Grand Challenges and Vendor Tests, e.g., Face Recognition Grand Challenge (FRGC) [2], Face Recognition Vendor Test (FRVT) [3], etc. These efforts have not only resulted in rank-ordering of different algorithms based on their performance on different datasets but also played an important role in development of improved algorithms based on their observations and analysis.

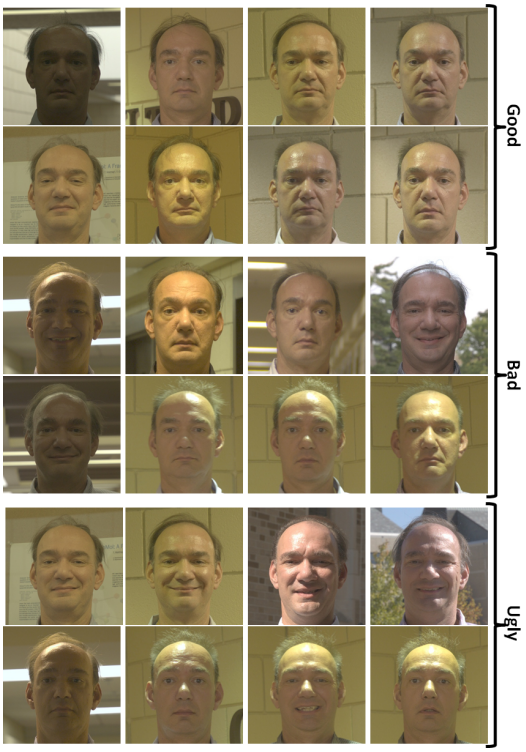

Figure 1. All query (top row) and target (bottom row) images of a subject from the three partitions in the GBU dataset.

The recently introduced GBU challenge problem [1] builds on the success of these evaluations to encourage development of algorithms that are robust to variations in frontal face images that are not acquired under studio-like controlled conditions. The GBU challenge presents three partitions of face images which are called the Good (easy to match), the Bad (average matching difficulty) and the Ugly (difficult to match). The partitions are created based on the performance of the top performers in the FRVT 2006 evaluation and signify the range of performance that is achieved with the state-of-the-art systems. What makes this problem interesting is that usual suspects like subject aging, sub-

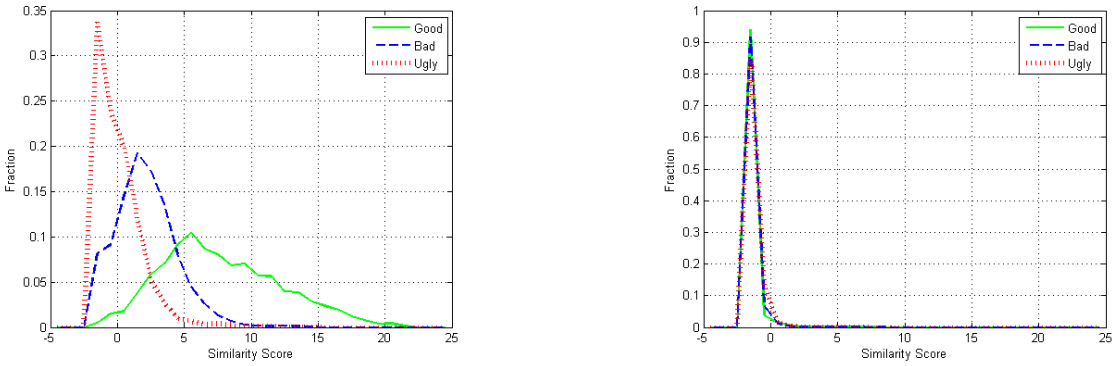

Figure 2. Match (left) and non-match (right) score distributions obtained using Pittpatt for the three partitions. Similar to the observation made in [1], the non-match score distribution is stable across the partitions while they differ in terms of match score distribution.

ject recognizability, pose variations and change in camera are not the factors that can lead to the performance difference across the three partitions. The potential effects due to these factors are virtually eliminated by following a strict data collection and partitioning protocol. All the images in the three partitions were acquired within the same academic year by the same model of camera in the frontal pose. In addition, all the three partitions consists of exactly same subjects with exactly same number of images. All target and query images of one subject from all the three partitions are shown in Figure 1. Though there are differences in illumination condition and facial expression across the images, these differences are present in all the three partitions as can be seen in Figure 1. The variation in performance across the three partitions is shown using Receiver Operator Characteristic (ROC) curves in Figure 3. In the absence of access to the top performing algorithms in the FRVT 2006, a well-known commercial face recognition system, Pittpatt [4] is used to compute the similarity scores. Although the performance obtained is slightly worse than what is shown in [1], the performance difference across the three partitions is equally prominent. The corresponding match (left) and non-match (right) score distributions are shown in Figure 2. The three partitions appear to have very similar non-match score distributions, therefore much of the variation in the performance can be attributed to the difference in match scores. Therefore, in this investigation much of the attention is devoted to factors that separate match pairs across the three partitions.

The unique design of the partitions in the GBU challenge problem leads to an interesting question. *What is it that is influencing the performance so drastically across the three partitions?* or in other words, *What are the factors that differ across the partitions even though such a strict protocol is used to both acquire images and to partition the dataset?* In this paper, we make an attempt to answer these questions. Since the apparent variations in illumination condition and facial expression are present in all the three partitions, a coarse classification of images as controlled/uncontrolled illumination or neutral/smiling expression as normally done [5], may not be very fruitful. To this end, we investigate characteristics like image sharpness, shadow pattern, hue and saturation content in addition to metrics quantifying extent of facial expression. The characteristics are chosen to capture observed variations in the images at a finer scale (compared to controlled/uncontrolled or neutral/smiling classification) to explain the performance difference.

Recent studies [6] have shown that quality of face images comes in pairs. It is the difference between the two images that affects similarity score instead of the individual quality of compared images considered in isolation. Along these lines, we characterize each match pair using the difference in investigated characteristics. We observe that the three partitions differ significantly based on the difference of these characteristics. Partial Least Squares (PLS)-based regression [7] is used to highlight the combined usefulness of these characteristics to distinguish between the three partitions.

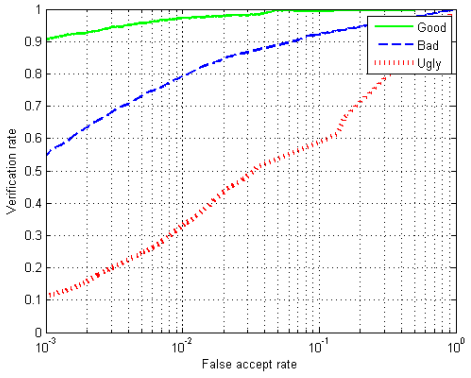

Figure 3. ROC curves for the three partitions obtained using Pittpatt as the matching engine.

The rest of the paper is organized as follows. The fol-

lowing section discusses a few significant related works from the literature. A brief description of the partitions is provided in Section 3. Description of various characteristics used in the investigation along with their effectiveness in separating the three partitions is provided in Section 4. A detailed analysis that includes PLS-based regression to combine various characteristics is provided in Section 5.

2. Related Work

The existing literature in this topic can broadly be divided into three main categories: a) works that define quality measures for individual images, b) works that define quality measures based upon properties of the compared image pair, and c) works that investigate effects of various covariates like age, gender, race, etc., on the matching performance. Grother and Tabassi [8] formalize the concept of sample quality as a scalar quantity that is monotonically related to the performance of the matcher. Their work asserts that quality measures should be developed to target matching performance and not just human perception of quality. Werner and Brauckmann [9] suggest incorporating digital characteristics like image resolution and compression measures, and facial characteristics like size and contrast, in addition to traditional quality metrics like brightness, contrast, etc., to measure goodness of a face image. Along similar lines, Weber [10] discuss two aspects of sample quality: character (features of sample source, e.g., pose, expression, etc.) and fidelity (accuracy with which the sample represents the source, e.g., sharpness, resolution, etc.) and investigate their impact on system performance. Hsu et al. [11] investigate various image-specific and face-specific quality metrics and infer that though overall quality scores are positively correlated with human ratings, humans appear to give importance to a few metrics that are not very critical in machine recognition process. A system for evaluating the quality of face images according to guidelines proposed by International Civil Aviation Organization (ICAO) is presented by Subasic et al. [12]. All these approaches define quality for individual face images.

Recent advances indicate importance of looking at facial quality in terms of pairs of images compared instead of assigning quality values to individual images. Phillips and Beveridge [13] characterize quality as an interaction between compared pairs of biometric samples, while showing the theoretical equivalence between perfect matching and perfect quality analysis. While highlighting this importance, Beveridge et al. [6] show the presence of a large number of contrary images as opposed to always-hard images in the GBU dataset. A contrary image is defined as one which has high match score with at least one other image but gives rise to poor similarity scores when compared to other good quality images.

Recently, researchers have also started to investigate various covariates that affect algorithm performance. Givens et al. [14] use a generalized linear model to analyze the effect of different covariate factors such as age, gender, race, facial hair, etc. on the performance of three algorithms. To estimate the effect of different covariates on performance, Beveridge et al. [5] fit a Generalized Linear Mixed Model (GLMM) with Bernoulli response and random effects for subjects. The analysis is performed on the FRGC data using both subject and image covariates.

3. GBU Dataset

Thorough details of the GBU dataset are provided in a recent publication [1]. Here we summarize a few relevant details for completeness. The GBU partitions are constructed from the FRVT 2006 [3] data acquired at University of Notre Dame. To control for the variation in "recognizability" of different subjects, each of the three partitions consists of same number of images of each person. To further control sources of variations, all images are acquired using the same camera model and in the frontal pose. All the images were acquired in the same academic year so that effect of aging on facial appearance can be reduced. The images were partitioned into the three subsets using a greedy strategy based on similarity scores obtained from the top three performing algorithms in the FRVT 2006 evaluation. Each of the three partitions in the GBU dataset consists of two sets of images, namely, a target set and a query set. All target and query sets contain 1085 images of 437 unique subjects. For each partition, algorithms are evaluated by computing similarity scores between all pairs of images across target and query sets, resulting in 3297 match pairs and $1,173,928$ non-match pairs.

4. Face Image Characterizations

The characterizations explored by us can be broadly divided into two categories: 1) image-specific characterizations, and 2) face-specific characterizations. Image-specific characterizations are generic image properties that have no relation whatsoever to the fact that images being analyzed are face images. On the other hand, face-specific characterizations are used to quantify facial properties which may not have any meaning for non-face images. The image-specific characterizations analyzed in this investigation include image sharpness, image hue content, image saturation content and image intensity content. Face-specific characterizations are included to account for variations in facial expression observed in the GBU dataset. We use several geometric measurements around the mouth region in an attempt to characterize these variations and to investigate if these play any role in the performance difference across the the three partitions in the GBU dataset. We now describe each of these characterizations in detail.

4.1. Image Sharpness-based Measures

The extent to which a face image is in focus has been shown to be a factor that influences matching performance [5]. Typically accurate knowledge about the degree of focus for an image is not available but reasonable estimates can be made from the image. We compute a sharpness metric that is shown to be sensitive to blur [15]. The value of this metric drops when the image becomes blurred. This makes it suitable to evaluate visual quality. The computation of this metric involves Singular Value Decomposition (SVD) of the local image gradient matrix. Given an image region of interest $g(x, y)$, its gradient matrix is defined as

$$\boldsymbol{G} = \begin{bmatrix} g_x(1) & g_y(1) \\ \vdots & \vdots \\ g_x(k) & g_y(k) \\ \vdots & \vdots \\ g_x(N) & g_y(N) \end{bmatrix} \tag{1}$$

where $g_x(k)$ and $g_y(k)$ denote x and y gradients at pixel location k, respectively. One can estimate local dominant direction and its strength for this region by computing SVD of the gradient matrix $\boldsymbol{G}$ as follows

$$\boldsymbol{G} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \boldsymbol{U} \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} \tag{2}$$

The singular values s_1 and s_2 represent the energy along directions $\boldsymbol{v}_1$ and $\boldsymbol{v}_1$, respectively. Xhu and Milanfar [15] show the behavior of dominant singular value s_1 on several different kinds of patches illustrating its usefulness as a sharpness metric.

Input face images are divided into regular rectangular regions and sharpness is computed for each of these regions resulting in a vector representing sharpness content of the entire face image. Given the sharpness vectors for each image in the dataset, we explore its usefulness to explain the performance difference across the three partitions. Given a match pair, Euclidean distance between the two sharpness vectors is computed. The underlying intuition is that a match pair should be relatively easy to verify if the images are similar as far as the sharpness measure is concerned. Figure 4 shows the distributions of difference in the sharpness measure obtained for match pairs for the three partitions. The three distributions differ significantly indicating that this sharpness-based measure is one of the factors that are responsible for the performance difference.

In addition to the described sharpness-based measure, we investigated an *edginess* based measure of image focus [5]. Given a face image, the approach involves convolving the image with Sobel edge operator followed by measurement of edge density. Edge density in a region is measured by collecting the gradient magnitudes in that region. Input face images are divided into regular rectangular regions and focus is computed for each of these regions resulting in a vector representing focus content of the entire face image. The distributions obtained for the three partitions using this measure appear quite similar to the ones shown in Figure 4 and are omitted due to space constraints.

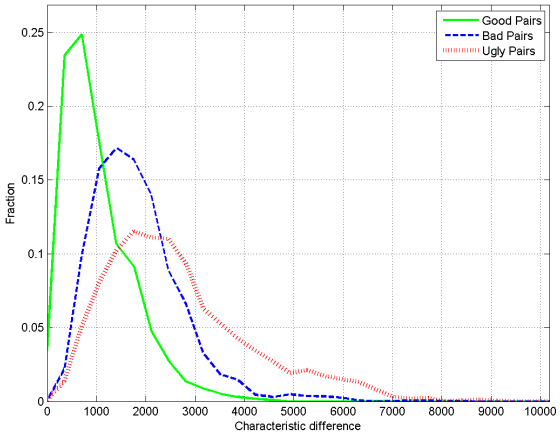

Figure 4. Distributions of the difference in the SVD-based sharpness measure in match pairs for the three partitions in the GBU dataset.

4.2. HSV space-based Measures

A quick glance at the images in the GBU dataset indicates a certain degree of variation in hue and saturation levels in the image. To investigate these variations for possible impact on the observed difference in matching performance, we compute simple hue and saturation-based measures. Input face images are transformed from RGB to HSV color space followed by computation of average hue and saturation values. The averages are computed for each of the small rectangular regions as done for the sharpness-based metrics.

Given a match pair, Euclidean distances between the two hue and two saturation vectors are computed. Figure 5 and Figure 6 show the distributions of difference in hue and saturation measures, respectively obtained for match pairs from the three partitions in the dataset. For both hue and saturation-based metrics, the three distributions differ significantly indicating that they play a role in the performance difference across the three partitions.

4.3. Measures to Characterize Shadow Patterns

A few images in the GBU dataset appear to have cast shadow patterns on the face regions that may be playing a role in the performance difference across the three partitions. Accurate detection and characterization of shadow patterns may require accurate face modeling which is out of the scope of this investigation. Instead, we divide the face

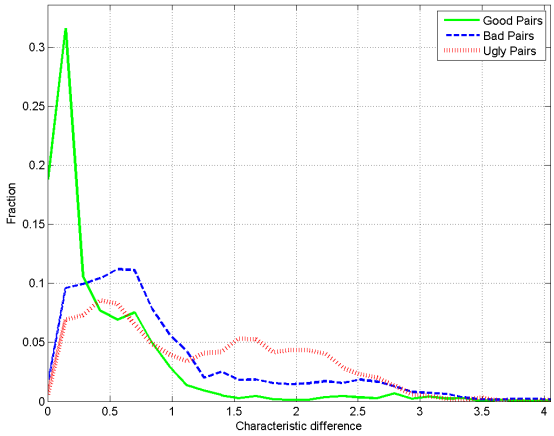

Figure 5. Distributions of the difference in the hue content-based measure in match pairs for the three partitions in the GBU dataset.

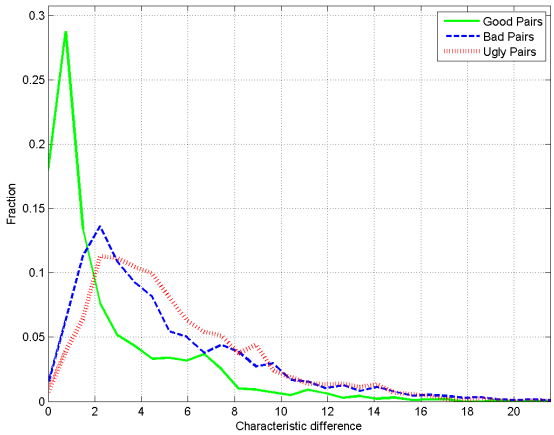

Figure 6. Distributions of the difference in the saturation content-based measure in match pairs for the three partitions in the GBU dataset.

image into regular rectangular regions and compute average intensities in each of these regions to capture intensity variations across the image due to any cast shadows. Raw image intensities are normalized by median intensity value for the entire image before computing the averages. This results in a shadow-indicator vector for each input image that consists of these average intensities.

Given a match pair, Euclidean distance between these shadow vectors is computed. Figure 7 shows the distributions of difference in shadow vectors obtained for match pairs from the three partitions in the dataset. The three distributions appear to differ significantly, indicating that they play a role in the performance difference across the three partitions. We also explored value from HSV-space in place of image intensity to capture shadow patterns. The resulting distributions appear very similar to the ones shown in Figure 7.

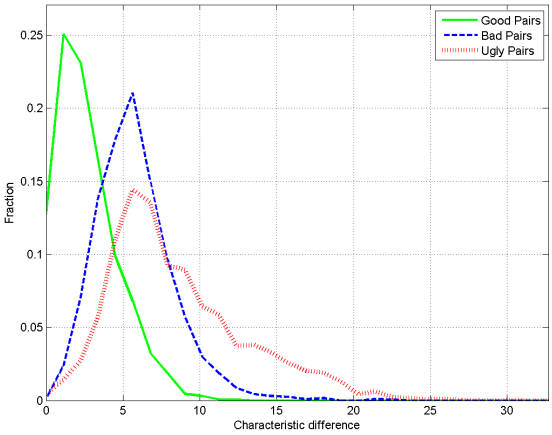

Figure 7. Distributions of the difference in the intensity-based shadow vectors in match pairs for the three partitions in the GBU dataset.

4.4. Measures to Characterize Extent of Facial Expression

The GBU dataset consists of face images with both neutral and smiling facial expressions. Therefore, as part of our investigation, we are interested in knowing if expression plays a role in the performance discrepancy across the partitions. Instead of classifying images coarsely as neutral or smiling as typically done [5], we make an attempt to quantify and estimate the extent of expression. For this, one needs to locate various primary facial landmarks. This is done automatically using publicly available Active Shape Model (ASM) library known as STASM [16]. STASM improves over the traditional ASM by incorporating a few simple but effective extensions that include a) fitting mode landmarks than are actually needed, b) selectively using two-dimensional templates in the ASM model instead of one-dimensional template, and c) relaxing the shape model when advantageous. Figure 8 shows an example face image with the facial landmarks as identified by STASM. Seven measurements around mouth region are included in the analysis that include outer width of the mouth, inner vertical gap and outer vertical gap (marked in the figure).

Given a match pair, differences between these various measurements are computed and used for analysis. Figure 9 shows the distributions of differences in length of one of the measurements for match pairs for the three partitions. The distributions indicate that extent of facial expression appears to be a factor that plays a role in the performance difference across the three partitions. The plots for the other

measurements show similar pattern and are omitted in the want of space.

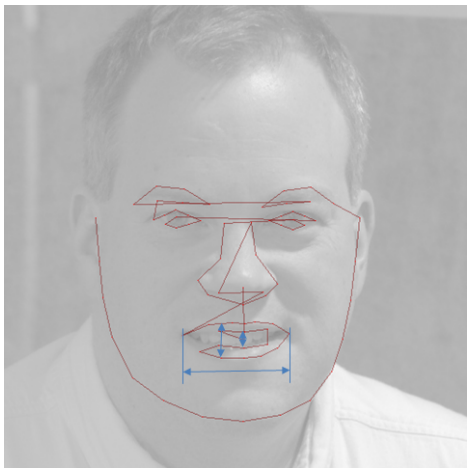

Figure 8. Facial landmarking using STASM [16] library to measure extent of facial expressions. Several measurements around mouth region are included in the analysis that include outer width of the mouth, inner vertical gap and outer vertical gap (marked in the figure).

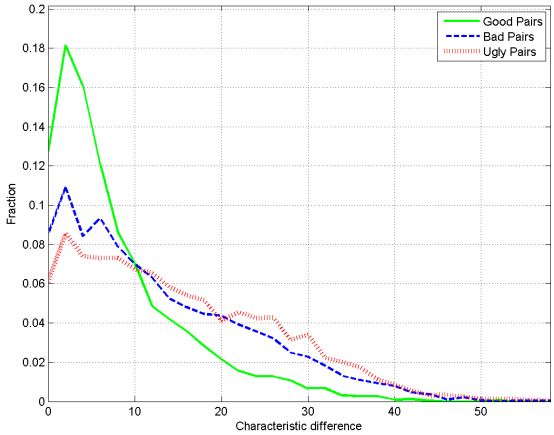

Figure 9. Distributions of the difference in the outer width of mouth in match pairs for the three partitions in the GBU dataset.

5. Analysis of Characterizations

On the whole, we investigated thirteen pair-wise measures that include two for image sharpness, two for HSV-based metrics, two to capture shadow patterns and seven to capture extent of facial expressions. All measures other than expression-based measures are computed locally by dividing the face region into regular 7×7 rectangular grid. As shown in the previous section, each of these measures appear to be correlated with the performance difference across the three partitions in the GBU dataset to a certain degree. The leads us to a few interesting issues that are discussed as follows:

5.1. How correlated are these measures?

Out of the thirteen measures analyzed in the previous section, a few of them try to capture very similar characteristics from the input face images. Therefore, we perform correlation analysis over them to understand how different are these measures from one another. Figure 10 shows a matrix consisting of Spearman's rank correlation coefficient obtained for each pair of the investigated measures. The coefficient indicates how well the relationship between two variables can be described using a monotonic function. As expected, the two image-sharpness based measures and the two shadow-characterizing measures appear to be very highly correlated. There appears to be very low correlation between image-specific characterizations and expression-based characterizations.

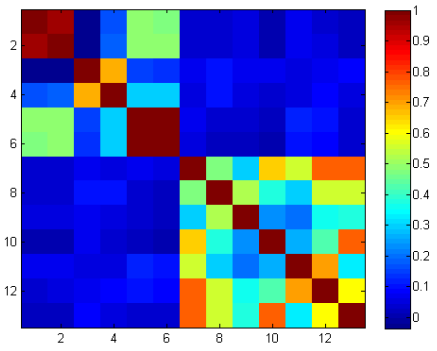

Figure 10. Spearman's rank correlation analysis over the thirteen characterization. 1 and 2 correspond to image-sharpness metrics, 3 and 4 correspond to hue and saturation content-based metrics, 5 and 6 correspond to metrics to characterize shadow patterns, and $7-13$ correspond to expression-based geometric measures (Best viewed in color).

5.2. Distributions for Non-match Pairs

So far, we have considered the effects of various characterizations only on match pair comparisons. Here, we illustrate the relation of a few of the investigated measures to the non-match comparisons for the three partitions. Figure 11 and Figure 12 show the comparison of distributions for match and non-match pairs for the focus-measure and an expression-measure, respectively. Interestingly, though the match and non-match distributions for the good partition differ significantly, this difference is relatively negligible for the bad and the ugly partitions. The same behavior is observed for the other metrics also, the plots for which are omitted due to space constraints. This observation further indicates that the investigated characterizations point

towards much better performance for the good partition as compared to the bad and the ugly partitions.

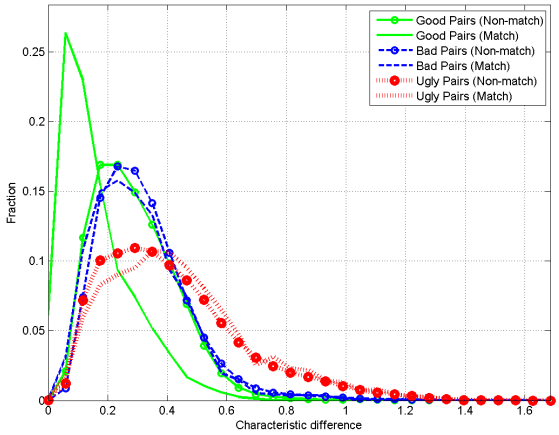

Figure 11. Distributions of the difference in the focus-measure in match and non-match pairs for the three partitions in the GBU dataset.

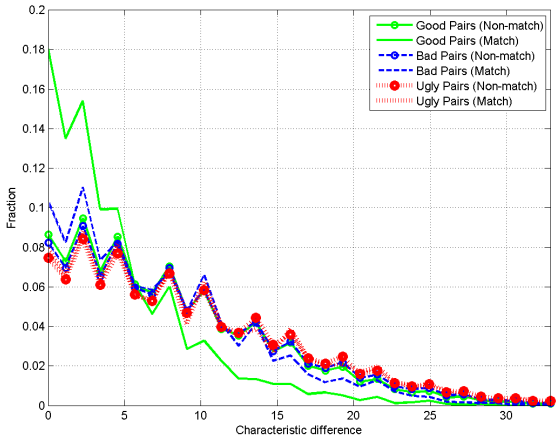

Figure 12. Distributions of the difference in the vertical outer mouth gap in match and non-match pairs for the three partitions.

5.3. Fusion of Characterizations

Given a pair of matching images, the underlying goal of this investigation is to predict either its similarity score or the partition it belongs to. To this end, we use Partial Least Squares (PLS)-based regression to fuse match-pair quality evidence obtained from the investigated characteristics. PLS regression generalizes and combines concepts from principal component analysis and multiple regression [7]. Like any other regression technique, the goal of PLS regression is to predict dependent variables from a set of independent variables. In this scenario, the dependent variable can either be the similarity score or the partition indicator variable that indicates whether a match pair belongs to the good, the bad or the ugly subset. Readers are referred to [7] for more details on PLS regression.

We perform PLS regression twice: once using the similarity scores for match pairs obtained from Pittpatt as the dependent variable, and second using partition indicator variable as the dependent variable. In both settings, the presented quality evidences are used as the independent variables. Figure 13 shows the predicted similarity scores obtained for match pairs from the three partitions. Ideally, these distributions should have been identical to the ones shown in Figure 2 (left). Though not perfectly similar to the desired distributions, the distributions indicate the usefulness of the proposed pair-wise measures to predict similarity score. PLS regression allows one to find out how much variation in dependent and independent variable is captured by the estimated latent vectors. We observe that while virtually all the variation in independent variables is accounted for, only about 64% of the variation in dependent variable (here similarity score) is captured, thereby indicating presence of other factors that have not been investigated in this paper. Figure 14 shows the predicted partition indicator variable obtained for match pairs from the three partitions. The three distributions are reasonably separated indicating the effectiveness of the proposed pair-wise measures in distinguishing between the three partitions.

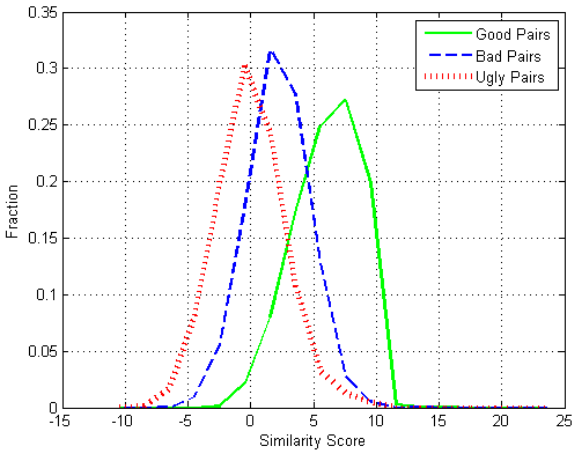

Figure 13. Predicted similarity scores obtained using PLS-based regression over match pairs.

6. Summary

The paper presented an investigation to explore the factors behind significant difference in performance across the three partitions in the GBU dataset. The strict acquisition and partitioning protocol of the dataset ensured that the usual suspects for performance degradation are not respon-

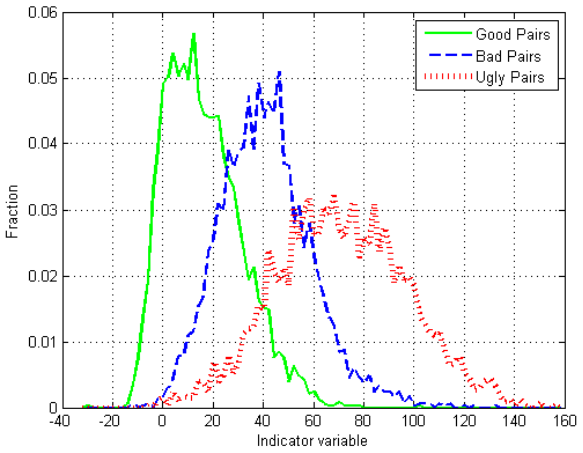

Figure 14. Predicted partition indicator variable obtained using PLS-based regression over match pairs.

sible for the performance degradation. We showed several image-specific and facial expression-specific metrics computed in a pair-wise fashion for match pairs that appear to play a role in the performance degradation across the partitions. PLS regression based fusion of the proposed metrics further indicated their effectiveness in being able to distinguish between the three partitions. We believe that reliably identifying these measures and understanding their behavior on performance can potentially serve three important goals - to predict the performance of algorithms on novel data; to design appropriate algorithms to account for variations in these characteristics; and to design appropriate acquisition environments at prospective sites to optimize performance.

References

[1] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahizada, and S. Wiemer, "An introduction to the good, the bad, and the ugly challenge problem," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2011, pp. 346–353. 1, 2, 3

[2] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 947–954. 1

[3] 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," in *NISTIR 7408*, 2007. 1, 3

[4] "Pittsburgh pattern recognition," *http://www.pittpatt.com/*. 2

[5] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper, "Factors that influence algorithm performance in the face recognition grand challenge," *Computer Vision and Image Understanding*, vol. 113, no. 6, pp. 750–762, 2009. 2, 3, 4, 5

[6] 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 *IEEE International Conference on Automatic Face and Gesture Recognition*, 2011, pp. 572–578. 2, 3

[7] H. Abdi, "Partial least squares regression (PLS-regression)," in *Encyclopedia for Research Methods for the Social Sciences*, M. L. Beck, A. Bryman, and T. Futing, Eds., 2003, pp. 792–795. 2, 7

[8] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 531–543, April 2007. 3

[9] M. Werner and M. Brauckmann, "Quality values for face recognition," in *NIST Biometric Quality Workshop*, 2006. 3

[10] F. Weber, "Some quality measures for face images and their relationship to recognition performance," in *NIST Biometric Quality Workshop*, 2006. 3

[11] R.-L. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research*, 2006. 3

[12] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec, "Face image validation system," in *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, 2005. 3

[13] P. J. Phillips and J. R. Beveridge, "An introduction to biometric-completeness : The equivalence of matching and quality," in *Proceedings of the IEEE International Conference on Biometrics: Theory, applications and systems*, 2005, pp. 414–418. 3

[14] G. H. Givens, J. R. Beveridge, P. Draper, B. A. Grother, and P. J. Phillips, "How features of the human face affect recognition: a statistical comparison of three face recognition algorithms," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2004, pp. 381–388. 3

[15] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *IEEE International Workshop on Quality of Multimedia Experience*, 2009, pp. 64–69. 4

[16] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Proceedings of European Conference on Computer Vision*, 2008. 5, 6