NISTIR 7759

When High-Quality Face Images Match Poorly

Beveridge, J. R.
Phillips, P. J.
Givens, G. H.
Draper, B. A.
Teli, M. N
Bolme, D.S.

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 7759

When High Quality Face Images Match Poorly

Beveridge, J. R.
Colorado State University

Phillips, P. J.
NIST
Information Access Division

Givens, G. H.
Draper, B. A.
Teli, M. N
Bolme, D.S.
Colorado State University

March 2011

U.S. Department of Commerce
Gary Locke, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Director

When High-Quality Face Images Match Poorly

J. Ross Beveridge, P. Jonathon Phillips, Geof H. Givens, Bruce A. Draper,
Mohammad Nayeem Teli and David S. Bolme

***Abstract*— In face recognition, quality is typically thought of as a property of individual images, not image pairs. The implicit assumption is that high-quality images should be easy to match to each other, while low quality images should be hard to match. This paper presents a relational graph-based evaluation technique that uses match scores produced by face recognition algorithms to determine the "quality" of images. The resulting analysis demonstrates that only a small fraction of the images in a well-studied data set (FRVT 2006) are low-quality images. It is much more common to find relationships in which two images that are hard to match to each other can be easily matched with other images of the same person. In other words, these images are simultaneously both high and low quality. The existence of such contrary images represents a fundamental challenge for approaches to biometric quality that cast quality as an intrinsic property of a single image. Instead it indicates that quality should be associated with pairs of images. In exploring these contrary images, we find a surprising dependence on whether elements of an image pair are acquired at the same location, even in circumstances where one would be tempted to think of the locations as interchangeable. The results presented have important implications for anyone designing face recognition evaluations as well as those developing new algorithms.**

I. Introduction

In the field of biometrics, there is considerable interest in identifying quality measures [6]. A quality measure can be defined as any measurable property of an image that is predictive of face recognition performance. An example of a quality measure is edge density in the facial region of an image, which was shown by Beveridge et al. to relate to face recognition performance in the Face Recognition Vendor Test (FRVT) 2006 evaluation [13]. The motivation for finding quality measures is to provide feedback to operators to help them collect good images, and to predict how well a face recognition algorithm will work on a new data set or in the context of a federated[1] system.

An open question is whether most face recognition failures are caused by low-quality images or by pair-wise inconsistencies between target and query images. This paper presents a novel analysis suggesting that, at least for the FRVT 2006 data set, low-quality images are relatively rare. More common are what we call "contrary images": images that have a contrary nature with respect to quality in so much as their quality is simultaneously high and low as defined by how they match to other images of the same person.

We compare the relative frequency of these two types of failures by analyzing a graph of pair-wise similarity scores for the FRVT 2006 data set. This graph allows us to single out low-quality images which match poorly against every other image of the same person. It also allows us to single out contrary image by defining an appropriate relationship between a 4-tuple of images. The resulting analysis shows that low-quality images are less common than contrary images, and this in turn suggests that two high-quality face images may, when compared to each other, match poorly.

This result leads to the question of what factors might cause two high-quality images of the same face not to match. In all likelihood there are many factors, including differences in lighting, pose and expression, but our analysis also strongly suggests another culprit: location. For all the contrary images we found, the match pair leading us to consider the contrary image to be of high quality was almost always taken at the same location, and the match pair leading us to consider the contrary image to be of low quality was always taken at a different location. This does not imply that traditional factors such as lighting are not important. Lighting angles are in part a function of location, so the two factors are inherently confounded. Nonetheless, it is striking that same/different location is such a strong predictor of whether two images will match well.

The findings presented here underscore the importance of characterizing facial biometric quality in terms of pairs of images as opposed to single images. This is important to the biometrics community as a whole, and particularly to those endeavoring to establish quality guidelines for facial biometrics. The findings with respect to the importance of location have significant implications for the design of future face recognition systems, particularly federated systems. This even holds for situations where the locations might be thought of as interchangeable, for example when both locations are indoors under fluorescent lighting. Finally, the importance of location suggests factors at work within state-of-the-art face recognition algorithms that make them unexpectedly sensitive to environmental factors associated with location. Better understanding why this sensitivity exists, and then developing algorithms robust with respect to such factors, is clearly an important challenge.

The next section provides background on prior work related to biometric quality in the context of face recognition.

This work was supported by the Technical Support Working Group (TSWG) under Task SC-AS-3181C. All authors except P. Jonathon Phillips are with Colorado State University. P. Jonathon Phillips is with NIST. P. Jonathon Phillips thanks the Federal Bureau of Investigation (FBI) for their support of this work. The identification of any commercial product or trade name does not imply endorsement or recommendation by the Colorado State University and the National Institute of Standards. Communication may be directed to J. Ross Beveridge at ross@cs.colostate.edu.

[1]A federated data set is one in which images collected by one agent are matched to images collected independently and under different conditions by another.

Section III introduces the data set and algorithm used in our analysis. Section IV presents the match and non-match score distributions used in our analysis and provides empirical justification for our decision to concentrate our analysis on match scores. Section V introduces our new methodology for framing questions about match quality in terms of patterns in a match graph, and presents the results that give rise the conclusion that quality is often a property of a match pair rather than a single image. Section VI presents our findings with respect to the critical role that location is playing in making a pair of images easier or harder to recognize.

II. Background

While this paper clearly demonstrates a fundamental limitation of approaches which define quality for individual images, as is so often the case, the fact that a task cannot be achieved in all cases does not mean it is not worth pursuing. Considerable work has already been done in the area of predicting face recognition performance based on measures associated with images, and here we review briefly some of this work.

To begin with definitions, Grother and Tabassi [6] define a quality measure as a number that relates an image's quality to a recognition system and is predictive of how well the system will perform recognition. This definition is consistent with most work on biometric quality, and a number of recent papers [8], [16], [5],[18], [17], [7], [4], [10] have looked at general image properties, such as contrast, sharpness, and illumination intensity. Luo [8] presents an instance of a general framework where quality is measured using Radial Basis Function (RBF) without relying on reference images for assessing quality. Subasic et al. [16] evaluate the quality of face images in travel documents according to the guidelines set by the International Civil Aviation Organization (ICAO). The quality of an image is represented by a fuzzy value. Fronthaler et al. [5] study an orientation tensor of an image with a set of symmetry descriptors which can be varied according to the application. They provide empirical results on fingerprints and show the applicability of the approach to assess face quality as well. Werner et al. [18] and Weber [17] recommend combining photographic (brightness, contrast, etc.,) and feature level (pose, expression, etc.,) scores to assess quality of images. Hsu et al. [7] showed the consistency and discrepancy between human quality ratings and machine quality scores using a classification-based score normalization process for various quality metrics. Fourney et al. [4] define an image's quality based on its potential to lead to a correct identification when used with existing face recognition software. Nasrollahi et al. [10] measure quality of faces in video sequences by combining features like out-of-plane rotation, sharpness, brightness and resolution.

Defining quality based upon properties of a pair of biometric signatures being matched is becoming more common. For example, in the context of iris and fingerprint recognition, Nandakumar et al. [9] define a single quality metric for each template-pair query based on the local image quality measures rather than estimating the quality of the template and the query images individually. Also, Phillips and Beveridge[11] in a theoretical exploration of the limits of quality measures define quality as a function of a pair of images, not a single image. Their key finding is theoretical, showing that the task of producing a perfect quality measure reduces to the problem of constructing a perfect recognition algorithm.

There is also a line of work [1], [2] that uses statistical models to relate covariates to recognition performance, and this work has resulted in a number of findings regarding how image properties influence the probability a person will be correctly verified given a query and target match pair. One finding in particular shows that edge density in the region of the face is a strong predictor of recognition performance.

Sheirer and Boult [14], [15] have pursued a different line of work closely related to biometric quality concerned with explicitly predicting when a recognition algorithm has failed. A notable aspect of this work is the formal characterization of the non-match distribution as a Weibull distribution and the subsequent ability to frame questions of when an algorithm has succeeded as a hypothesis test relative to the underlying distribution.

There is also a literature concerned with how people assess quality, and while this avenue of work does not meet the standard of relating a measurable property of an image to recognition performance, it is nonetheless interesting and of importance to people responsible for fielding systems and training individuals to use these systems. One recent example of such work [3] presents a study involving 87 people and their subjective assessment of a number possible face image quality factors.

III. The Good, the Bad & the Ugly Challenge

The face image data for the Good, the Bad & the Ugly Challenge Problem (GBU) comes from the Notre Dame multi-biometric data set collected as part of FRVT 2006 [13], [12]. The challenge problem is partitioned into three subsets: the Good, the Bad, and the Ugly. Each partition contains 2170 images of 437 people. These images are further split into two equal sized groups, a target set and a query set. Evaluation is carried out by measuring how well face recognition algorithms recognize images from the query set matched against the target set.

The three partitions were carefully constructed to introduce a wide range of difficulty, and this was done using similarity scores from a fusion algorithm created out of results from three distinct top-performing FRVT 2006 algorithms. The resulting verification rates for the fusion algorithm at a false accept rate (FAR) of 0.001 are 0.98, 0.80, and 0.15 for the good, the bad, and the ugly partitions respectively. The match and non-match distributions for this fusion algorithm across the three partitions are shown in Figure 1 and will be discussed further in the next section.

The images in the GBU were selected from a larger pool of 9307 images of 570 people. In addition to introducing a preference for easy matches in the good partition and very difficult matches in the ugly partition, the selection process

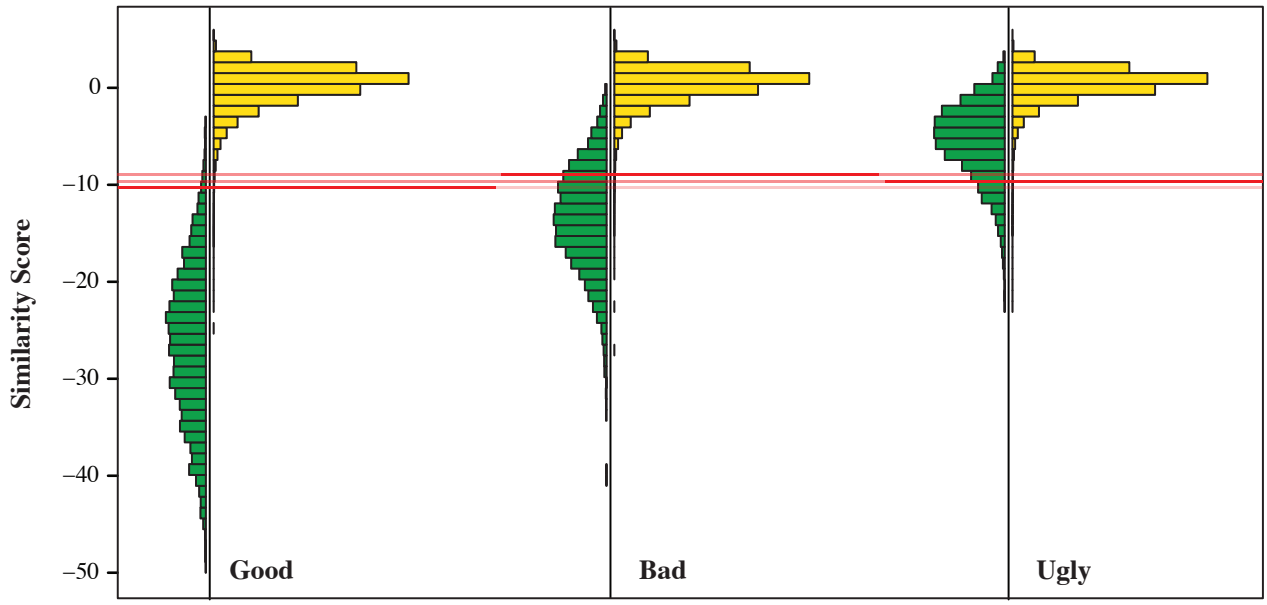

Fig. 1. Match score and non-match score distributions for the Good, the Bad & the Ugly partitions. More negative scores are better, match score distributions are green and non-match distributions are yellow. The FAR=0.001 thresholds for the three distributions are nearly the same and are indicated by red horizontal lines.

took account of several other constraints. First, no image could be a member of more than one partition. Second, images of the same 437 people are present in each partition and there are between 1 and 4 images per person in the target and query sets. Further, for each person, the same number of images are present in every partition. Finally, no pair of query and target images of the same person in the same partition are taken on the same day. We will return to the larger superset of 9307 images later in this paper, since we will use the superset of similarity scores in our relational match graph constructions.

In the Good, the Bad & the Ugly Challenge Problem, all images were acquired using a high quality digital camera, specifically a six megapixel Nikon D 70. All the photographs were posed with a person standing in front of a camera mounted at eye level on a tripod in a well lit setting. Indoors, the settings were typically hallways. Outdoors, the settings were either in an open area or against a building backdrop. In all of the images, the person being photographed was asked to look at the camera.

IV. True Matches Gone Bad

Our analysis focuses exclusively on match scores, and more generally on the properties of these pairs of images of the same person. In so doing, we are neglecting the role that non-match pairs might play in complicating the recognition task. In essence, we are studying true matches gone bad while ignoring the case where false matches turn good. The justification for this simplification is apparent in the non-match score distributions shown in Figure 1.

For the non-match scores, the distributions appears essentially equivalent for the three partitions. In contrast, the shift in the match score distributions moving between partitions is striking. Absent the evidence shown in these distributions, one might wonder if part of what makes the Bad and Ugly partitions more difficult is false matches being assigned uncharacteristically good similarity scores. However, given the stability of the non-match distribution between partitions, this does not seem to be the case.

Further, note the red lines in Figure 1. They represent the verification thresholds based upon FAR = 0.001 and each of the three non-match distributions. These thresholds are -10.21, -8.90 and -9.59 for the good, the bad, and the ugly partitions respectively.

As another measure of the stability of the tails of the non-match distributions, we calculated a verification threshold over all three partitions taken as a whole (-9.68). We then counted how many false accepts occur in each of the three partitions using this pooled threshold. For the good, the bad, and the ugly partitions there are 1683, 724 and 1133 false accepts, respectively. Considering that we are looking at only the tip of each distribution's tail, the fact that there are a significant number of false accepts in every partition provides additional evidence that the non-match distributions are behaving similarly to each other.

V. Quality Comes in Pairs

For face recognition, it has become clear to us that given a choice between discussing biometric quality in terms of single images, or instead in terms of pairs of images, it is more useful to think about biometric quality in terms of pairs of images. Here we lay out some of the strongest empirical evidence we have encountered so far in support of the maxim: "quality comes in pairs."

To develop the argument, consider for a moment the implications of presuming the opposite. In other words, consider what must logically follow from a presumption that the primary responsibility for failures in face recognition is explained by properties of individual images. The first implication is that it should be relatively straightforward to identify the best and worst quality images by the simple fact they are either always easy or always hard to recognize. The second implication is that once a match pair indicates an image is hard to recognize, that same image should never participate in a different easily recognized match pair. As we are about to illustrate, both situations above can be formally expressed in terms of subgraphs in a match graph.

A. Match Graphs

Consider a graph in which there is one vertex for each face image in an evaluation. Next, create annotated edges between vertices based upon similarity scores. The most elemental form of such a graph would include an edge annotated with a similarity score for every comparison carried out by a face recognition algorithm. For our purposes here, edges will only be defined between pairs of images of the same person. This simplification results in a clean partition of the full match graph into a series of connected subgraphs, one per person.

For simplicity, similarity scores are mapped to the categorical labels "Hard", "Easy" and "Medium", and in this analysis "Medium" edges will be ignored. One can imagine a multitude of ways of coming up with such a categorization; we define an easy match pair based upon the match scores from the good partition. Specifically, we determine the similarity score threshold that defines the best 40% of the matches in that partition and assign any match with a score better than this threshold the label "easy". To define a hard match pair, we turn to the ugly partition and find the similarity score threshold that defines the worst 40% of the matches. For the fusion algorithm these two threshold scores are -24.0 and -6.1; a match pair with a score less than -24.0 is easy and a match pair with a score greater than -6.1 is hard.

As commonly carried out, an evaluation over the good, the bad, and ugly data sets would not include similarity scores between pairs of images straddling two of the partitions. In other words, we would not have a similarity score relating an image of a person in the good partition to a different image of the same person in the bad or ugly partition. Such a limitation is unnecessary and undesirable for our match graphs. Therefore, we have chosen to expand our analysis to use all the available images and match scores from the original superset of data used to construct the GBU partitions. Specifically, 9307 images of 570 people.

Applying our definitions of easy and hard to the match pairs derived from these images, we obtain $14,517$ easy pairs and $10,868$ hard pairs. That leaves $35,564$ pairs as neither easy nor hard, for a total of $60,949$ match pairs. Recall from above that because it is well understood that images taken at or near the same time are more easily recognized, we removed from consideration match pairs taken on the same day.

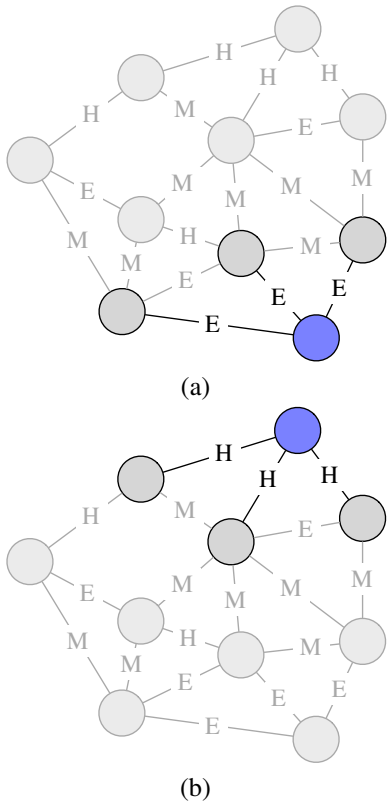

Fig. 2. Examples illustrating subgraph patterns associated with images that are: a) always easy to recognize and, b) always hard to recognize.

B. Once Easy Always Easy and Once Hard Always Hard

If one believes quality is intrinsic to single face images, then it follows that an image which is easy to recognize in one circumstance should always be easy to recognize. Similarly, an image which is intrinsically hard to recognize should always be hard to recognize. These ideas may be formalized as patterns in the match graph.

Specifically, the pattern illustrated in Figure 2a is one in which all edges leaving a specific vertex are labeled easy. Note that while in the illustration there are three edges, in general there will be more. The pattern illustrated in Figure 2b is one in which all edges leaving a specific vertex are labeled hard. Again, while the illustration shows three edges, in general the number of edges will vary and typically be greater than three.

When we search the match graph for instances of the once-easy-always-easy pattern, we find no instances. In other words, not one of the 9307 face images satisfies the constraint that it easily matches every other image of the same person. The total absence of such consistently easy to recognize images may in part be explained by an asymmetry in recognition. In other words, one would expect an overall low quality match from a comparison between a high-quality image and a low quality image. Still, it is striking that the pattern is never observed.

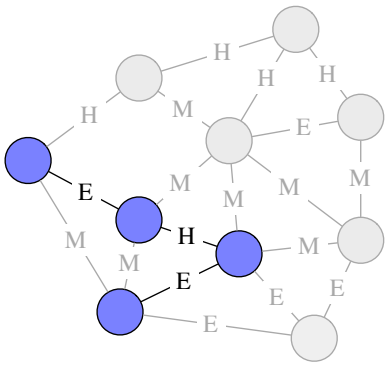

Fig. 3. An example illustrating a subgraph pattern in which images are simultaneously "hard" and "easy" to recognize.

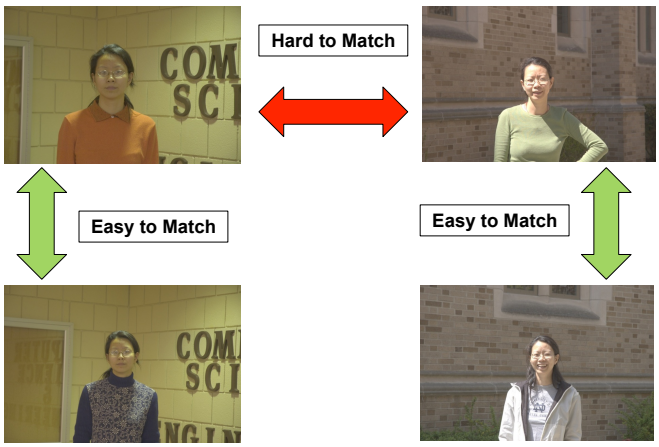

Fig. 4. The four images shown share an easy-hard-easy relationship. The two images across the top are hard to recognize. The pairs of images in the left and right columns are easy to recognize..

When we search the match graph for the once-hard-always-hard pattern, we find 86 examples, and the scarcity of such images is more difficult to explain if one truly believes quality resides intrinsically within individual images. If one believes the asymmetry argument used above, one would expect many images to satisfy the once-hard-always-hard pattern. Such a belief seems unsupported given that we see fewer than 100 images out of over 9000 that are consistently hard to recognize.

C. Contrary Images

We can go further with the analysis of the match graph, and define a four way interaction between images that indicates that an image is hard to recognize in one match pair and easy to recognize in another match pair, and further, that the other image for which matching is hard is itself easily recognized when compared to yet a different image. As stated above, we call these contrary images. While the English becomes a bit strained when describing contrary images, the concept is clearly illustrated in terms of the match graph as shown in Figure 3. Note while our previous relations were defined over all the edges leaving a single image, this pattern is defined over a 4-tuple of images related through three edges where the center edge is labeled hard and the two adjacent edges are labeled easy.

If we can find instances of such 4-tuples, then we will have found contrary images. When we search the match graph for this pattern, we not only find an instance, we found 221 such 4-tuples that include 214 distinct contrary images. Clearly, contrary images are not that unusual, and what is interesting is to begin to try to understand why these situations arise.

VI. Location, Location, Location

To explore what might be giving rise to images that are simultaneously easy and hard to recognize, we created quadrature graphics showing the four images involved in such a way that we could rapidly scan for the hard relationship across the top row and the two easy relationships down the two columns. An example is shown in Figure 4 where both images in the top row are contrary images.

It would be a mistake to dismiss too quickly the many factors that likely play some role in creating images that are both hard to recognize and easy to recognize depending upon the other image involved. Certainly lighting is important in this context, and of course lighting and location are related. Also, facial expression is certainly playing some role.

Those caveats aside, the most singularly striking aspect of the 221 image 4-tuples we inspected is that the columns are almost always taken at the same location. Indeed, when both images are taken indoors, they are always taken at the same location. Further, there is never a case where a hard pair, the top row, are taken at the same location.

A. Same vs. Different Location Performance across the GBU

The tuple analysis indicates images taken at the same location are more easily matched than those taken at different locations. However, as solid a finding as this is, it represents behavior in the extremes of the data set, specifically the easiest match pairs in the Good partition and the hardest match pairs in the Ugly partition. Therefore, how location effects verification rates over all the data in each of the GBU partitions still remains to be seen.

To address this question, we carried out a statistical analysis of the fusion algorithm performance data for the GBU partitions. While we could have simply reported single numbers, the verification rates for same and different locations across the GBU partitions, this would tell us nothing about variability in these performance numbers. So instead we've carried out a bootstrap analysis in order to estimate the distribution of verification rates over these cases. The details of this analysis follow. The punchline is evident directly from the results presented in Figure 5a: changing location significantly drops verification rates in all three partitions.

To explain our analysis more precisely, we bootstrapped verification rates for the good, bad, and ugly partitions split by whether the target and query image locations matched.

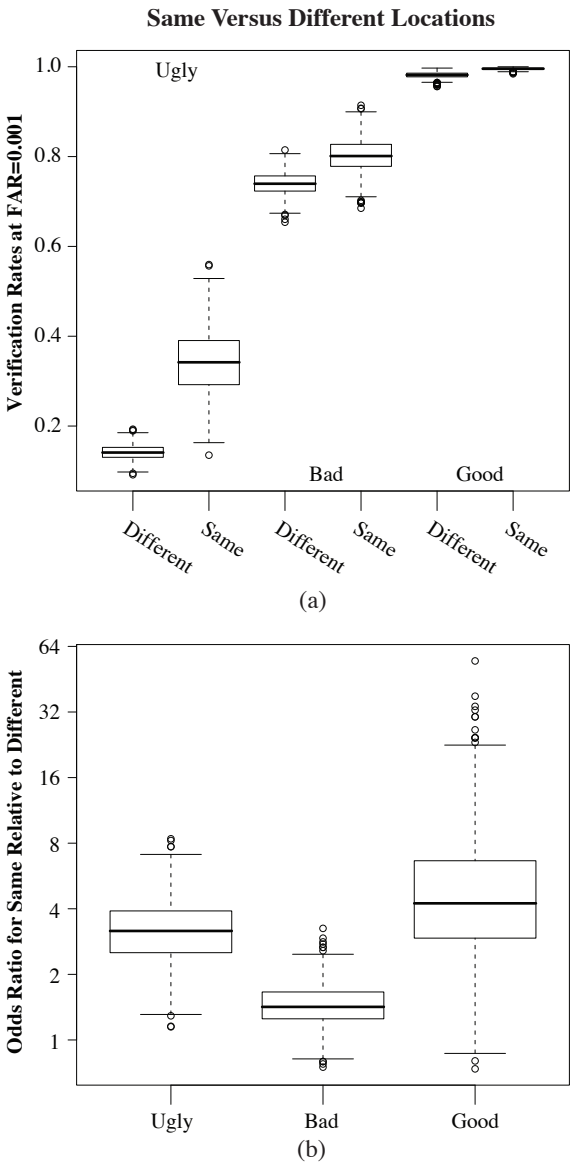

Fig. 5. Verification rates and odds ratios comparing images acquired at the same versus different locations. a) for each partition, the different and same location rates, b) for each partition the odds ratio for same relative to different.

For each of the 3×2 combinations of these variables, we calculated the empirical verification rate in our sample. This can also be viewed as the population-weighted predicted verification rates marginalized over the random effects.

Next, a bootstrap dataset was created by resampling subjects[2] with replacement. The new dataset was assembled using the accumulated data from this collection of pseudo-subjects. Each bootstrap dataset may, therefore, have a different number of trials than the original one. From this pseudo-dataset a new set of verification rates were calculated. We repeated this process 1000 times.

Figure 5a shows boxplots of the bootstrap distributions of verification rates at FAR = 0.001 split by partition and whether locations were the same or different. The difference in verification rates between the good, the bad, and the ugly partitions are self-explanatory. Next, note that for each partition verification rate tends to be higher when locations are the same. Notwithstanding these two observations, it is noteworthy that neither the incremental benefit of transition up from ugly to bad to good, nor the incremental benefit of same-location within each of these partitions, is constant. Instead, both are reduced as partition quality increases. For the good partition, this can be explained by the overall excellent verification performance: there is little room left for improvement.

As we delve deeper, however, more subtle differences are also apparent. We can see that for the ugly and bad partitions, the bootstrap variance of verification rates is considerably greater when location matches than when it doesn't. This suggests that matching location, while important, does not fully explain verification. Moreover, the effect is deceptively large. Remember that each of the thousand points in one boxplot is the *mean* of about 1000 individual verification outcomes over a pseudo-dataset representative of the sort one might obtain in the intended sampling population. Thus, to see the downward whisker of a same-location box overlap with the upward whisker of a different-location box is not to observe that some individuals are poorly verified even when the location is held constant, but instead to learn that the verification rate for *an entire population* is highly variable when same-location matches are attempted.

The magnitude of variation here is very large, considering that these large samples of subjects should be interchangeable. This suggests that there must be a strong subject-specific effect on verification of same-location images, and that there must likely be substantial effects of other unobserved variables as well.

Another way to examine these results is to consider odds ratios. Figure 5b shows boxplots of odds ratios for verification for same-location relative to different-location, split by the good/bad/ugly partitioning variable. These bootstrapped values were calculated in the same manner described above. Note that the vertical axis uses a $\log_2$ scale. Thus, for example, the median odds ratio for the ugly partition is about 3, meaning that if a *population* of subjects with ugly different-location images were somehow transformed into a *population of subjects* with ugly same-location images, the overall population-weighted verification odds would more than triple. The actual medians and 90% probability intervals are 3.2 (1.8, 5.5), 1.4 (1.0, 2.1), and 4.2 (1.7, 18.0) for the ugly, the bad, and the good partitions, respectively.

VII. Conclusion

Two aspects of the work just presented have broad implications for face recognition research. First, the discovery of more contrary images than always-hard images highlights the difficulty inherent in thinking of face quality as an intrinsic property of one image. Recall that a contrary image is of high-quality as implied by at least one match, but nonetheless

[2]We are not sampling images or image pairs because our analysis assumes people/subjects are interchangeable.

gives rise to a poor match when compared to an alternative high-quality image.

The second major finding is a surprising and important dependence upon the location where images are acquired. This dependence suggests a sensitivity to location in scenarios where one might expect one location to behave more or less like another. This location dependency has important implications for those designing algorithm evaluations and also for algorithm developers for whom the challenge will be to lessen such dependencies.

In terms of methodology, formulating questions about pairs of matching face images in terms of a relational match graph is, to our knowledge, new. We expect it is an approach to the study of how algorithms behave that will be open to expansion and elaboration in the future.

REFERENCES

[1] 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*, 113(6):750–762, June 2009.

[2] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, and Y. M. Lui. FRVT 2006: Quo vadis face quality. *Image and Vision Computing*, 28(5):732 – 743, 2009.

[3] O. Y. G. Castillo. Report: Survey about facial image quality. Technical report, Fraunhofer EPrints [http://publica.fraunhofer.de/eprints.har] (Germany), December 2006.

[4] A. Fourney and R. Laganiere. Constructing face image logs that are both complete and concise. In *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on*, pages 488–494. IEEE, 2007.

[5] H. Fronthaler, K. Kollreider, and J. Bigun. Automatic image quality assessment with application in biometrics. *Computer Vision and Pattern Recognition Workshop*, 0:30, 2006.

[6] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. Pattern Analysis Machine Intelligence*, 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] H. Luo. A training-based no-reference image quality assessment algorithm. In *International Conference on Image Processing*, 2004.

[9] K. Nandakumar, Y. Chen, A. Jain, and S. Dass. Quality-based score level fusion in multibiometric systems. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 473–476. IEEE, 2006.

[10] K. Nasrollahi and T. Moeslund. Face quality assessment system in video sequences. *Biometrics and Identity Management*, pages 10–18, 2008.

[11] P. J. Phillips and J. R. Beveridge. An introduction to biometric-completeness: the equivalence of matching and quality. In *BTAS'09: Proceedings of the 3rd IEEE international conference on Biometrics: Theory, applications and systems*, pages 414–418, Piscataway, NJ, USA, 2009. IEEE Press.

[12] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, H. S. Y. M. Lui, and S. Weimer. An Introduction to the Good, the Bad, & the Ugly Face Recognition Challenge Problem. In *Proceedings, Ninth International Conference on Automatic Face and Gesture Recognition*. IEEE Computer Society Press, March 2011.

[13] 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 Transactions on Pattern Analysis and Machine Intelligence*, 32(5):831 – 846, May 2010.

[14] W. Scheirer and T. Boult. A fusion-based approach to enhancing multi-modal biometric recognition system failure prediction and overall performance. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–7, 2008.

[15] W. J. Scheirer. *Improving the Privacy, Security, and Performance of Biometric Systems*. PhD thesis, University of Colorado, Colorado Springs, May 2009.

[16] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec. Face image validation system. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 30–33. IEEE, 2005.

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

[18] M. Werner and M. Brauckmann. Quality values for face recognition. In *NIST Biometric Quality Workshop*, 2006.