NISTIR 7674

Quantifying How Lighting and Focus Affect Face Recognition Performance

Phillips, P. J.
Beveridge, J. R.
Draper, B.
Bolme, D.
Givens, G. H.
Lui, Y. M.

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

NISTIR 7674

Quantifying How Lighting and Focus Affect Face Recognition Performance

Phillips, P. J.
NIST
Information Access Division

Beveridge, J. R.
Draper, B.
Bolme, D.
Givens, G. H.
Lui, Y. M.
Colorado State University

February 2010

U.S. Department of Commerce
Gary Locke, Secretary

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

Quantifying How Lighting and Focus Affect Face Recognition Performance*

J. Ross Beveridge David S. Bolme Bruce A. Draper Geof H. Givens Yui Man Lui
Colorado State University
Fort Collins, CO 80521, USA
beveridge@cs.colostate.edu

P. Jonathon Phillips
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

Recent studies show that face recognition in uncontrolled images remains a challenging problem, although the reasons why are less clear. Changes in illumination are one possible explanation, although algorithms developed since the advent of the PIE and Yale B data bases supposedly compensate for illumination variation. Edge density has also been shown to be a strong predictor of algorithm failure on the FRVT 2006 uncontrolled images: recognition is harder on images with higher edge density.

This paper presents a new study that explains the edge density effect in terms of illumination and shows that top performing algorithms in FRVT 2006 are still sensitive to lighting. This new study also shows that focus, originally suggested as an explanation for the edge density effect, is not a significant factor. The new lighting model developed in this study can be used as a measure of face image quality.

1. Introduction

Face recognition technology has made significant progress since the early 1990's. Current face recognition systems are highly accurate for face images collected in studios with consistent pose, focus and lighting. The Face Recognition Vendor Test (FRVT) 2006 showed that it is possible to achieve a false reject rate (FRR) of 0.008 at a false accept rate (FAR) of 0.001 for well controlled images [16]. For many in the research community, this is now considered to be a "solved problem".

*The work was funded in part by the Technical Support Working Group (TSWG) under Task T-1840C. PJP was supported by the Department of Homeland Security, Director of National Intelligence, Federal Bureau of Investigation and National Institute of Justice. The identification of any commercial product or trade name does not imply endorsement or recommendation by Colorado State University or the National Institute of Standards and Technology.

The FRVT 2006 also included a set of uncontrolled images taken in hallways or outside with variations in lighting, focus, near fontal pose, and expression, among other factors. For these images, the best reported performance was a FRR of 0.12 at an FAR of 0.001; among the better algorithms, the FRR ranged from 0.12 to 0.38 at an FAR of 0.001 [16, 3]. Two other useful uncontrolled data sets are Labeled Faces in the Wild [8] and the PubFig [11]. Both include images collected from the World Wide Web without explicitly controlling for factors such as lighting, pose, focus or expression. Error rates for recent algorithms on Labeled Faces in the Wild range around 0.20 [17] to 0.15 [11]. For the PubFig data set the best total error rate is 0.22 [11].

The face recognition research community has considered a number of factors to explain what makes recognition harder on uncontrolled data. The most common explanations involve illumination, pose and expression [14, 19, 12]. Other factors studied include age of the person, gender, expression, image resolution, and time between images [13].

Of these, changes in illumination ranks high on most researcher's list of factors making face recognition hard. To study the effects of changes in pose and illumination, the Carnegie Mellon University Pose, Illumination, and Expression (PIE) and Yale B data sets [19, 12] were created. In these data sets, changes in pose and illumination are systematically varied. Since the release of these data sets, there has been a dramatic decrease in the error rate associated with change in illumination direction. However, up to now it has not been possible to assess quantitatively the impact of changing illumination on data sets such as FRVT 2006 that do not explicitly control for illumination direction.

Other directly measurable quality measures of face images have been proposed that predict when algorithms will fail. In particular, Beveridge et al. [3] show that higher edge density is strongly associated with recognition failure on the FRVT 2006 uncontrolled images. What they did not provide was a physical explanation. Is edge density reflecting

changes in focus? Is it responding to changes in illumination? Perhaps neither; edge density could be responding to some other unidentified aspect of the images.

To answer these questions, this paper presents a model for estimating illumination direction based on prior work by Sim and Kanade [20]. It also introduces a new measure of image focus. Finally, it presents a series of of four experiments on the same FRVT 2006 data studied in [3] that unravel the relative importance of lighting and focus, and relate both back to the original finding that higher edge density predicts recognition failure.

The most significant finding is that lighting accounts for the edge density effect. In other words, the best explanation for why edge density was previously found to be such a strong predictor of recognition performance is that it was indirectly capturing information about lighting. The second most significant finding is that focus, when quantified using a better measure, plays little or no role in making recognition easier or harder. This study represents the first time, to our knowledge, that an illumination measure has been directly incorporated into a multi-factor study of face recognition performance on a large uncontrolled data set.

2. Motivation

Beveridge et al noted that low edge densities predict successful recognition in FRVT2006. Looking more closely at the data, we noticed that high edge density was often associated with outdoor images taken under harsh lighting, while low edge densities were associated with more uniform lighting. Furthermore, it appeared that recognition often failed when the face was illuminated from the side by direct sunlight. This makes sense to anyone familiar with the algorithms, since such lighting introduces both strong shadows and strong asymmetry.

The apparent association between strong side lighting and recognition failure raises the question of whether lighting conditions might explain recognition performance as well or better than edge density, while supplying a physical explanation. Telling the face recognition community that algorithms fail for images with high edge density raises the question of how one operationally 'controls' edge density. An explanation in terms of how the face is illuminated is qualitatively different and operationally more useful.

A separate lingering question concerns image focus. Reviewing the images by hand reinforced our belief that edge density was responding to many aspects of the scene and images. Poorly focused images did have low edge density, but edge density clearly responds to other factors, too. From an operational standpoint, knowing if poor focus contributes to recognition failure is important, so the need for an effective quantitative measure of image focus persists.

In response to the questions raised by our visual inspection of the data, we have developed a lighting model and a focus measure. These measures are described in Section 4. Armed with these quantitative means of assessing lighting and focus, we investigated four conjectures:

Conjecture 1 Lighting subsumes setting (i.e., indoors versus outdoors) as a predictor of performance.

Conjecture 2 Lighting subsumes edge density as a predictor of performance.

Conjecture 3 Focus explains recognition failure better than edge density.

Conjecture 4 Since focus is distinct from lighting, predictions about recognition failure using both will be better than either alone.

The basis for these conjectures is our hypothesis that edge density and setting are poor quality measures that only indirectly relate to performance. Edge density and setting confound different aspects in which images (and hence recognition performance) may vary. In comparison, lighting and focus isolate two physical components of image quality. Ideally they should be nearly orthogonal, and they may capture nearly all of the signal inherent in the inferior measures and perhaps additional information as well. Experiments designed to test each of these conjectures and the interpretation of their results are presented in Section 5.

3. Background

The work presented here draws upon prior work in lighting estimation, focus estimation and performance evaluation. Prior work in each of these areas is reviewed here.

3.1. Lighting Estimation

There are many sources of variations that can confuse a person's identity. Illumination is one of those variations in imagery. In fact, it has been argued that changes in illumination can make two images of the same person less similar than two images of different people [14]. A considerable literature has been developed that addresses a variety of techniques for estimating illumination, and in particular, Sim and Kanade [20] were the first to employ kernel regression for illumination estimation.

The model presented in Section 4.1 also uses kernel regression and is similar in general approach to that of Sim and Kanade. It should be noted, however, that the primary goal of Sim and Kanade was to remove lighting artifacts. Our goal is to estimate the lighting angle, from frontal to side. Hence, the approaches are related, but not identical.

3.2. Focus Estimation

Krotkov proposed edge density as a measure of focus in 1989 [10]. It was compared to a number of alternatives,

and found superior. In particular, it was found to do a better job than alternative approaches that were based upon the spectral energy signature of an image. Beveridge et al. [3] adapted the measure slightly by restricting it to only the region of image containing the face.

3.3. Performance Evaluation

Empirical evaluation and performance evaluation have a long and rich history in computer vision in general and face recognition in particular [7, 9, 21]. For example, over the past decade there have been a series of empirical evaluation workshops associated with CVPR. In face recognition, there have been a string of evaluations starting with FERET[15] and moving through FRVT 2006 [16]. There is also increasing interest in scenarios more closely associated with the web: see for example the Labeled Faces in the Wild results page [1].

Two lines of work are particularly relevant to the results presented here. The first is that of Beveridge et al. [4, 3] that has proposed the use of a Generalized Linear Mixed Effect Model (GLMM) to relate a set of factors, i.e., covariates, to the probability that a face recognition algorithm will succeed. As already discussed above, that work established that a particularly simple image measure, edge density, is strongly related to face recognition performance in the FRVT 2006. What that work did not do was to provide, in any substantive manner, a practical explanation of why.

The second line of closely related research includes efforts to predict the success or failure of biometric algorithms, in our case, face recognition algorithms, based on objective quantitative information available at the time of recognition. If one associates these measures with biometric quality, the task is equivalent to that of predicting recognition success based upon biometric quality. One recent good example of such work is that of Grother and Tabassi [6]. It should also be noted that prediction of recognition success may also be based upon sets of related match scores rather than on a directly measurable property of the biometric; an intriguing example of this approach is the recent work by Scheirer and Boult [18].

Our work below will bring together some aspects of each approach. The starting point in the analysis will be a GLMM comparable to that developed by Beveridge et al. From the second line of work, we will use the model's ability to predict success or failure as a way to assess the value of an associated factor. The factors examined here are edge density, focus, lighting and setting.

4. Approach

This section sets forth our new lighting estimation model, focus measure, and the approach that will be used to test when one covariate subsumes another.

[1] http://vis-www.cs.umass.edu/lfw/results.html

4.1. Lighting Estimation

The illumination model is trained on the CMU-PIE [19] imagery and is constructed to return a single number indicating the extent to which the face is being lit from the side. Recall that CMU-PIE has images for 68 subjects, and each subject has 24 illumination variants. There are also images with the room lights on and the room lights off. Here we use only the images with room lights on.

To eliminate the identity effect, we average all 68 subjects along each illumination variant. Hence, 24 illumination variant images are used to train the lighting model.

The lighting model is defined as follows. Let μ_k be the average image in illumination variant k. We apply kernel regression [1] to estimate the lighting coefficients described as follows:

$$\begin{aligned} \hat{x} &= \frac{\sum_{k=1}^{n} \alpha_k \mu_k}{\sum_{k=1}^{n} \alpha_k} \\ \alpha_k &= \exp(-\frac{\parallel x - \mu_k \parallel^2}{\sigma^2}) \end{aligned} \tag{1}$$

where x is the estimated image, n is 24 (the number of variants), and σ is 10. As Equation (1) shows, $\hat{x}$ is the reconstructed lighting images from the training set.

According to the illumination cone principle [2], any image in the illumination cone can be reconstructed by a linear combination of extreme rays given as:

$$\hat{x} = \max(Bs, 0) \tag{2}$$

where B is an illumination basis, s is the lighting coefficients, and 0 is used to remove the negative values corresponding to the shadowed surface. Therefore, we obtain the lighting coefficients as:

$$s = B^{-1}\hat{x} \tag{3}$$

where B is the average illumination μ in our case. The dominant illumination direction is determined as:

$$d^* = \operatorname{argmax} s \tag{4}$$

and the lighting estimation γ is computed as:

$$\gamma = s(d^*)\mathrm{W}(d^*) \tag{5}$$

where W is a predefined weighted vector which is [0 0 -3 -3 -2 -2 -1 -1 1 1 1 1 1 1 1 -1 -2 -3 -2 -1 1 1 1 0]. An example of lighting estimation is given in Figure 1. The predefined weights correspond to the lighting directions from the CMU-PIE dataset where positive indicates frontal illumination, negative denotes side illuminations, and zero indicates that no flashes were applied.

Figure 1. Example of lighting estimation: left (-0.240), middle (-0.237), right (0.068)

4.2. A Strong Edge Motion Compensated Focus Measure

The problem with edge density as a measure of focus across a range of images is that edge density responds to many different aspects of the person and scene. These include lighting, glasses and hair across the forehead. To correct for this problem, a new motion blur compensated focus measure is introduced. The measure also starts by using Sobel operators to compute image gradients. Instead of measuring the average gradient magnitude it measures the gradient of the strongest edges, since the gradient of the strongest edge is a much better predictor of image focus than the average gradient. It also measures motion blur, and equates motion blur with poor focus.

An image suffering from motion blur is smooth in the direction of motion and sharp in the orthogonal direction. To measure the strongest edges in the direction of motion, PCA is applied to the horizontal and vertical components of the image gradient vectors. If motion blur is present, the PCA component corresponding to the smallest principal component indicates the direction of motion. By projecting the gradient vectors onto the smallest principal component, the new measure estimates the gradient in the blurriest direction, and therefore measures blur caused by both focus and motion.

In our earliest version of this new algorithm, the largest edge in the direction of motion was used to estimate focus. However, as one might expect, this approach is unstable because the maximum value is often an outlier and not representative of the focus in the image. For this reason we select the edge strength associated with the 97.5% quantile. This is close enough to the maximum to measure the gradient of a strong edge, but also reduces the chance of selecting an outlier.

To compare the new focus measure to edge density, 54 images were hand selected to span a range from poorly focused to well focused. This set also included images with blur due to motion. The 54 images were then ranked by alternative measures of focus in order to assess how well these measures performed. As expected, the ranking imposed by edge density did not align well with our judgment about focus. In contrast, the Strong Edge Motion Compensated (SEMC) focus measure just described ranked images in a manner consistent with our judgement. Images with motion blur were included in this test, and the steps described above for handling motion blur resulted in those images being scored as out of focus.

Figure 2 compares some of the highest and lowest edge density images to some of the highest and lowest SEMC focus images. The highest edge density images are always outside with strong lighting on the face and shadows, while the lowest edge density images are always indoors with neutral lighting. The SEMC focus measure seems to break the dependency on lighting. Outdoor images with strong shadows are often included with the low group because of poor focus, and in-focus images with neutral lighting are often found in the high focus group.

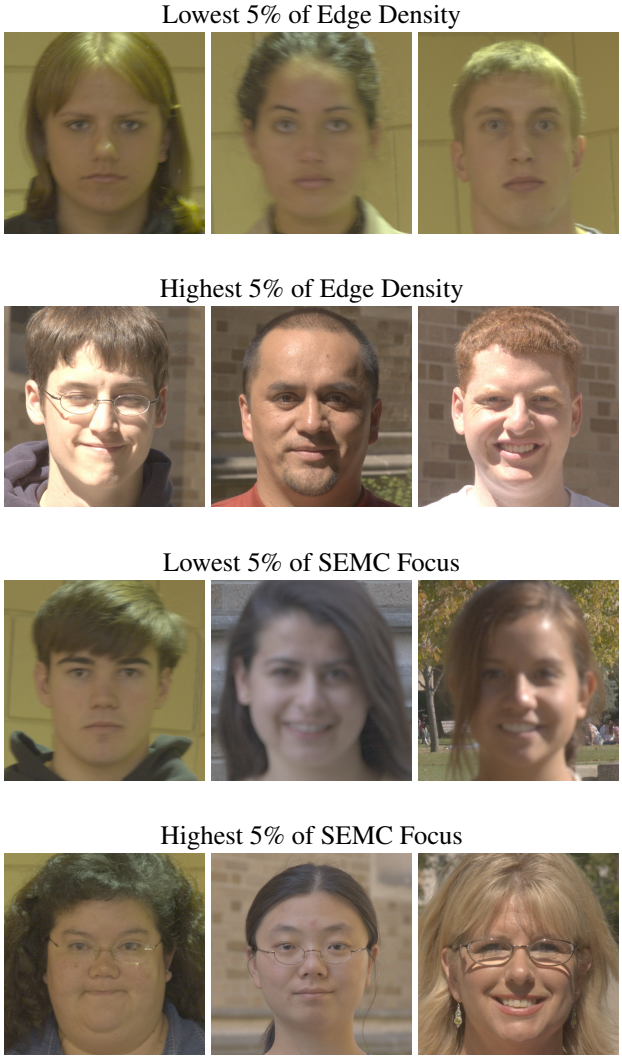

Figure 2. This figure shows some selected images from the highest 5% and lowest 5% of the edge density and SEMC focus measures.

4.3. Comparing GLMMs to Assess Importance

The conjectures posed in Section 2 mostly involve statements about the relative value of one factor (covariate) versus another. It is possible to approach such questions with a single statistical model including all significant covariates and their significant interactions. Indeed, the statements about the importance of edge density by Beveridge et al. [3, 5] does just that. However, one drawback of this approach is that for the sample sizes relevant here, most if not all of the proposed covariates and their interactions may turn out to be significant and hence warrant inclusion in the GLMM. The combinatorics of model selection quickly surpass one's ability to make sense of the results.

Another problem with that approach is its intent. With the GLMM approach, one obtains (for a given set of covariates) estimated probabilities of verification. The relative performance of models (and hence the comparative value of predictors in two competing models) is not necessarily directly related to the estimates of verification probabilities or model parameters. Instead, it is related to the proportion of the total variability in the data explained by each model.

A natural way to quantify that criterion is to ask how many cases each model predicts correctly. Note that each GLMM may be viewed as a mapping from the space of covariates to a predicted probability of successful verification at a specific false accept rate. Moreover, the prediction is relative to specific match pairs. For the FRVT 2006 analysis, the match pairs consist of a highly controlled target image and a less controlled query image taken either in a hallway or outdoors.

Thus, it is possible to test how well each GLMM predicts the actual outcomes, i.e. success or failure for each match pair. Specifically, the probability of verification may be thresholded in order to create a binary classifier: predict successful verification if the probability of success is above a threshold τ. In one sense, this does an injustice to the model in terms of absolute quality, because it throws away the fine gradations in predicted probabilities of verification. However, as a means of making relative comparisons between the predictive power of alternative covariates in the context of a larger statistical model, it works well. Indeed, as the threshold τ is varied it is possible to create a standard ROC plot with one curve per GLMM.

Each of the four conjectures in Section 2 will be tested by assessing the relative predictive power of a carefully chosen set of candidate models. Let us illustrate with the first Conjecture in Section 2, namely that lighting subsumes setting. To test this conjecture, four distinct GLMMs may be created. They are:

1 A baseline GLMM without either lighting or setting.

2 Baseline plus an effect for lighting.

3 Baseline plus an effect for setting.

4 Baseline plus effects for lighting and setting.

By comparing the performance of these four models, we can assess the relative importance of lighting and setting. A set of four ROC curves compares the models. When interpreting these ROC plots, it is important to remember that they are measuring the predictive power of alternative models for predicting whether verification succeeds or fails, the ROC plots are NOT reporting the False Accept Rate and False Failure Rate of the actual face recognition algorithms.

5. Findings

Beveridge et al. [3] reported results for a GLMM that used fused similarity scores from three top performing FRVT 2006 algorithms. Our first step was to replicate that GLMM and to then remove from it all effects associated with edge density[2] and setting. Note the resulting model still includes other covariates including gender, race, size of the face and whether the person was wearing glasses in the uncontrolled image.

The next step was to add back into this baseline GLMM main effects for covariates in the pattern described above in order to test each of the four conjectures presented in Section 2. The ROC plots for each associated experiment are presented in Figure 3. The findings are as follows.

5.1. Experiment 1: Lighting and Setting

Our conjecture, prior to running this experiment, was that lighting would subsume setting in the sense that whatever makes recognition more difficult outdoors would be entirely captured by our newly developed measure of lighting. The ROC presented in Figure 3 supports this conjecture. Note the order, from lowest to highest equal error rate is: Lighting+Setting, Lighting Only, Setting Only and Baseline. The primary result seen here is that models that use the lighting variable are superior to those that do not. There is a small but notable distinction between the Lighting+Setting model and the Lighting Only Model, indicating that lighting may not capture 100% of the complex amalgam of factors that inherently differ between settings. Nevertheless, it is remarkable how fully the relationship between setting and performance can be explained solely by our lighting measure.

5.2. Experiment 2: Lighting and Edge Density

Our conjecture was that lighting would subsume edge density. The ROC presented in Figure 3 fully supports this conjecture. Note the ROC curves for the Lighting Only and Lighting+Edge Density GLMMs are nearly identical,

[2] In [3] the edge density covariate was often described as FRIFM.

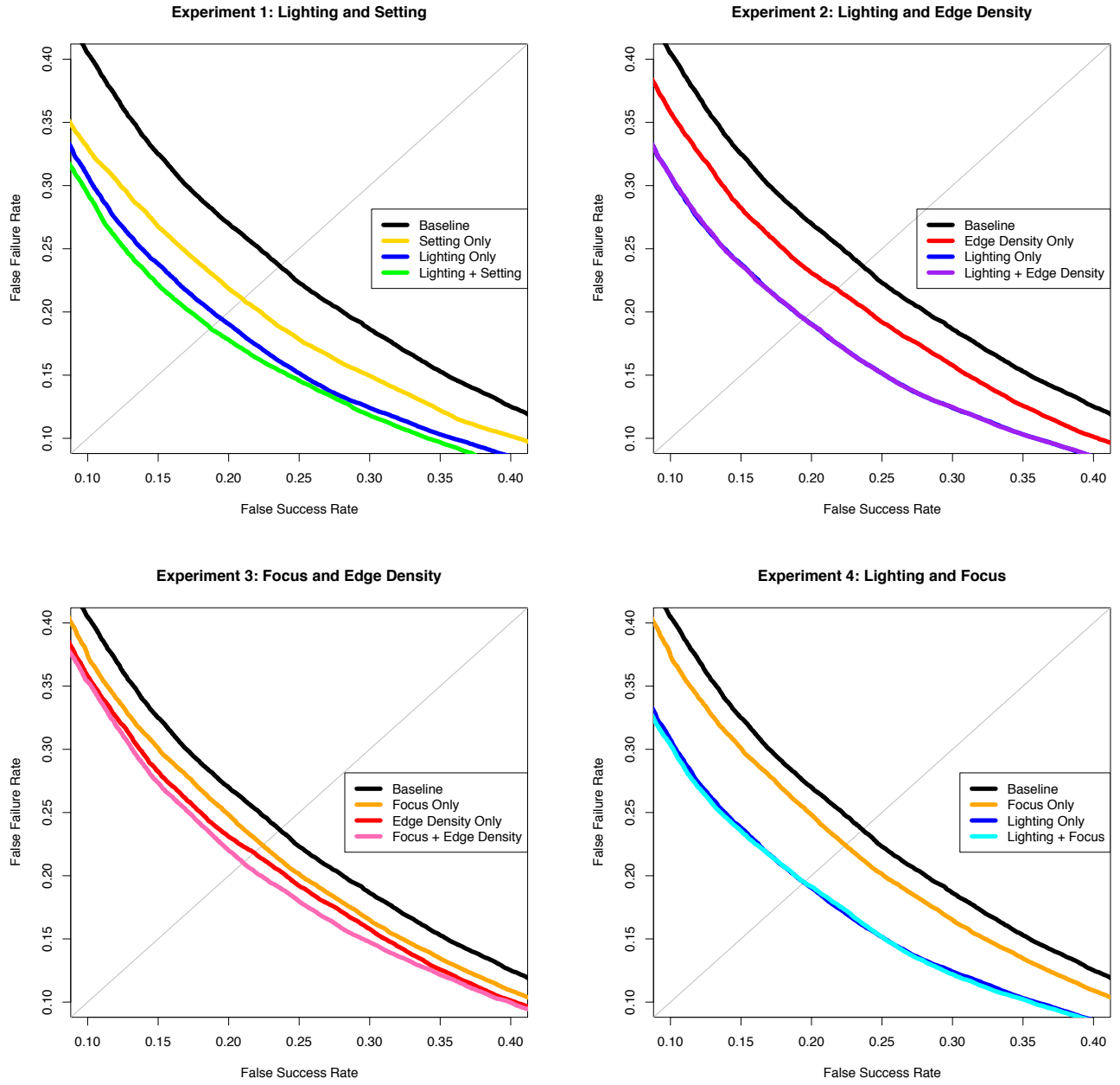

Figure 3. This figure shows ROC plots comparing the relative predictive power of alternative GLMMs which include combinations of covariates. Experiments 1 through 4 correspond to the conjectures 1 through 4 in 2. In the Experiment 2 plot the Lighting Only and Lighting+Edge Density lie on top of one another.

and in the plot one sees the Lighting+Edge Density curve nearly completely covering the Lighting Only curve. Also note that both curves have a lower equal error rate than the Edge Density Only curve. At least in this experiment, there appears to be no additional predictive value associated with knowing edge density once a good lighting estimate is available.

5.3. Experiment 3: Focus and Edge Density

Our conjecture was that a good measure of whether the face is in focus would better predict failure than edge density. The ROC presented in Figure 3 suggests this is not the case. Note the order, from lowest to highest equal error rate is: Focus+Edge Density, Edge Density Only, Focus Only and Baseline. In short, edge density as a measurable prop-

erty of the query face images is doing a slightly better job of predicting recognition failure then is the newly proposed focus measure. Also, since performance is slightly better when both variables are used, it appears that each variable is providing some independent information. Our results are consistent with the possibility that focus itself is not a major contributing factor to verification failure.

5.4. Experiment 4: Lighting and Focus

Our initial conjecture, based upon the fact that focus and lighting are operationally distinct, was that predictions based upon the two together would be better then either alone. As shown in Figure 3, the result does not support this conjecture. Instead, the two curves with lowest equal error rate are for Lighting Only and Lighting+Focus; the two curves lie nearly on top of each other. Further, the error rate for the Focus Only GLMM is much higher. Two things are apparent from this result. First, the newly proposed lighting model is doing a good job of predicting when verification will fail. In fact, the information is good enough that the introduction of the focus measure adds nothing. Second, in order to make sense of the lack of improvement when focus is added, we must again entertain the notion that focus is unrelated to verification failure.

In this experiment one might ask whether a focus effect is being masked by lighting. Theoretically, the effect of focus on performance might depend on the nature of the lighting, with the average effect across the lighting range being close to zero. To test this, we fit a fifth model that included both predictors plus their interaction. The result still showed no evidence that focus mattered. The same strategy was also used in Experiment 2, again yielding a negative finding.

5.5. Summarizing Effects

Beveridge et al. [3] reported effects in terms of estimated probability of verification associated with individual covariates and combinations of covariates. Here we do the same for the main effects of Lighting, SEMC Focus and Edge Density. These effects are reported for the models developed for Experiments 2, 3 and 4. They are summarized in Figure 4.

The vertical axis indicates the estimated probability that verification will succeed. For SEMC Focus and Edge Density, the horizontal axis represents standardized covariate values, where standardization in this context means the original values have been scaled and shifted to have a sample mean of zero and a sample standard deviation of one. The horizontal axis for Lighting was left in the original raw units of the illumination model, and the four values shown may be interpreted as follows: 0.05 indicates frontal lighting, -0.05 indicates modest side lighting, -0.10 indicates notable side lighting, and -0.18 indicates very strong side lighting. These three covariates were included in GLMMs for multiple experiments, and so the main effects for distinct experiments are shown. When reporting probability of verification estimates, all other covariates in the GLMM are set to baseline (default) values.

Lighting has a very strong effect. Estimated probability of verification for frontal lighted images is around 0.95 and drops down to roughly 0.65 for the images where side lighting is most pronounced. The effect is nearly identical for the Lighting+Edge Density (Exp2) and Lighting+Focus (Exp3) GLMMs. By any standard, this is a large main effect.

The SEMC Focus and Edge Density main effects viewed together tell an interesting story. First recall that Experiment 3 uses a GLMM including both SEMC Focus and Edge Density, and does not include Lighting. Consequently, a main effect associated with Experiment 3 is shown for both SEMC Focus and Edge Density. Moreover, larger values of SEMC Focus are associated with increased probability of verification. This is consistent with common expectation; recognition is easier when query images have a higher focus score. In contrast, increased Edge Density in Experiment 3 is associated with decreased probability of verification.

Note, however, that the main effect for both SEMC Focus and Edge Density essentially disappears when each is paired with Lighting, as in Experiments 4 and 2 respectively. This is further evidence that Lighting is a highly predictive covariate, and that when added to a GLMM, it subsumes much of what is otherwise attributed to SEMC Focus and Edge Density.

6. Conclusions

Prior work showed that face recognition on uncontrolled images remains challenging and that high edge density in images predicts failure. The study presented here introduces a new model for measuring illumination direction in images and shows that side lighting strongly degrades algorithm performance. In fact, the new lighting measure subsumes and explains in physical terms the previously reported edge density effect. More generally, we show that lighting remains a problem for state-of-the-art algorithms, circa FRVT 2006. Consequently, lighting direction can be viewed as an important quality measure that predicts face recognition performance.

References

[1] C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

[2] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions. *IJCV*, 28(3):245–260, July 1998.

[3] J. R. Beveridge, G. H. Given, P. J. Phillips, B. A. Draper, and Y. M. Lui. Focus on quality, predicting FRVT 2006 perfor-

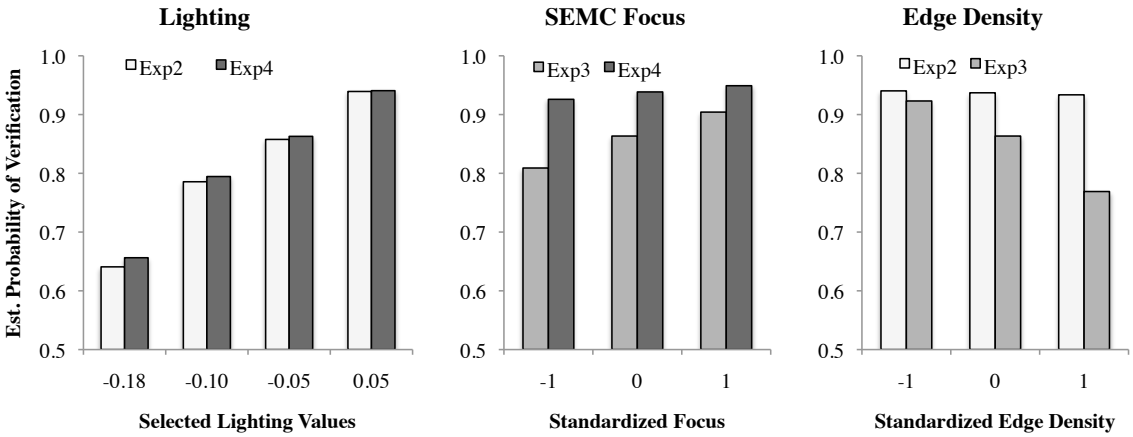

Figure 4. Main effects for lighting, focus and edge density for experiments where GLMM includes each covariate.

mance. In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, September 2008.

[4] 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.

[5] 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*, In Press:–, 2009.

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

[7] R. Haralick. Performance Characterization in Computer Vision. *CVGIP*, 60(2):245–249, September 1994.

[8] G. Huang, M. Ramesh, T. Berg, and E. Learned Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report Tech. Rep. 07-49, University of Massachusetts at Amherst, October 2007.

[9] K. W. Bowyer and P. J. Phillips (editors). *Empirical evaluation techniques in computer vision*. IEEE Computer Society Press, 1998.

[10] E. P. Krotkov. *Active Computer Vision by Cooperative Focus and Stereo*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1989.

[11] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV)*, October 2009.

[12] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[13] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips. A meta-analysis of face recognition covariates. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, September 2009.

[14] Y. Moses, Y. Adini, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. In *European Conference on Computer Vision, Marseille*, pages 286–296, 1994.

[15] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, October 2000.

[16] 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*, (in press)(DOI 10.1109/TPAMI.2009.59):1–1, 2009.

[17] N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2591–2598, June 2009.

[18] 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.

[19] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination and Expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003.

[20] T. Sim and T. Kanade. Combining models and exemplars for face recognition: An illuminating example. In *CVPR Workshop on Models versus Exemplars in Computer Vision, Hawaii*, 2001.

[21] N. A. Thacker, A. F. Clark, J. L. Barron, J. R. Beveridge, P. Courtney, W. R. Crum, V. Ramesh, and C. Clark. Performance characterization in computer vision: A guide to best practices. *Computer Vision and Image Understanding*, 109(3):305 – 334, 2008.