21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Quality Metrics for Practical Face Recognition

Ayman Abaza, Mary Ann Harrison
West Virginia High Technology Consortium
aabaza@wvhtf.org

Thirimachos Bourlai
West Virginia University
Thirimachos.Bourlai@mail.wvu.edu

Abstract

In biometric studies, quality evaluation of input data is very important, and has proven to have a direct relation with system performance. Quality measures can provide real-time feedback to reduce the number of poor quality submissions to the system. Another benefit is that they can predict and improve the authentication performance (e.g., by using quality-dependent thresholds). This paper main focus is image quality assessment for face recognition. First, we evaluate a number of techniques that measure image quality factors namely, contrast, brightness, focus, sharpness, and illumination. Second, via a set of experiments measuring the sensitivity of each matric to quality change, we select the most practical measure(s) for each quality factor. Finally, we propose a novel face image quality index (FQI) that combines the five aforementioned quality factors. Via a set of statistical significance tests, we illustrate and support that FQI is a promising quality measure that can be used as an alternative to some benchmark face image quality measures.

1 Introduction

Image quality measures for biometric samples, in particular, is a research field on its own and, in most cases, the evaluated quality factors are biometric modality specific. Two categories of quality measures can be distinguished: generic (can be used for any biometric modality) or biometric modality specific (viz., designed to address issues related to a specific modality such as iris, fingerprints, or faces). Some examples of generic quality measures that are used to qualify the perceived image degradation (typically, compared to a reference good quality image) are image contrast, brightness, sharpness and illumination. Face-based quality measures, ISO/IEC-19794-5, include but are not limited to: (a) Digital formatting of face images due to resolution, and grayscale contrast; (b) Scene deviation due to head rotation, illumination, eyes open versus close, glasses versus no glasses, and mouth open versus close; (c) Position of face as well as camera, exposure, image brightness, focus and sharpness.

Image quality metrics can be classified according to the availability of a reference image [15]: (i) Full-reference, which is the case with most existing approaches; (ii) "Blind", or no reference-based quality assessment, which is a more desirable approach (reference images are not always available). There are several image quality measures proposed in the literature, such as (a) Universal Quality Index (UQI) [14], (b) Average Image (AVI) [8], and (c) Wavelet-based (WB) face quality measure [13].

The main contributions of this work are: (a) evaluation study of various image quality measures for face images acquired from well-known face databases, (b) selection of the most practical measures, which do not need references and require less computational time, and (c) proposal of a new face quality index that is designed to model the quality of a detected face [1], by integrating several standard image quality measures.

The rest of the paper is organized as follows: Section 2 presents a number of image quality measures, followed by several experiments to (i) select the most practical matrices (section 3), (ii) combining these measurements (section 3.1), and (iii) applying the resulted quality index to real databases in section 4. Conclusions and future work are discussed in section 5.

2 Quality Measures for Face Images

Various image quality measures Q_m have been reported in the literature. Here, we summarize the main factors for face image quality measures (details in table 1):

Contrast: The image contrast can be defined as the root mean square of the image intensity C_{RMS} [5]. An-

[1] Face detection was performed using a commercial software developed by Pittsburgh Pattern Recognition (PittPatt) - http://www.pittpatt.com/

other definition for image contrast is the Michelson contrast reported in C_{Mic} [10].

Brightness: B_1 can be calculated as an average of the brightness component in the HSB (or HSV) color space, stands for (Hue, Saturation, Brightness). Bezryadin et al. [3] suggested another image brightness measure B_2, based on X, Y, and Z are Tristimulus values.

Focus: Yap and Raveendran [17] presented several definitions of image focus measures such as the L_1-norm F_1, and the energy of the Laplacian F_2.

Sharpness: Several image sharpness measurement techniques have been proposed in the literature; namely the average gradient (S_1 [8], and S_2 [5]), the Tenengrad S_3, and the adaptive Tenengrad S_4 [16].

Illumination: The illumination of an image can be the estimation of its luminance distortion I_1 [14]. Abdel-Mottaleb and Mahoor[1] calculated the illumination as the weighted sum of the mean intensity values of several regions of the image I_2.

3 Selection of Image-based Quality Measures

To evaluate the performance of the proposed face quality measures, 1040 face images from CASPEAL [4] database were used. Synthesized effect were added to change the image contrast, brightness and blurriness (details in section 4):

Contrast: The image was saturated at low and high intensities, in a step of 10%. $C_{RMS}, (corr = 0.996)$, represents the image contrast better than the Michelson contrast measure $C_{Mic}, (corr = 0.684)$, which depends only on image max and min intensity values. We denote the selected contrast measure by $C = C_{RMS}$.

Brightness: The image brightness was artificially adjusted via $gamma$ parameter, in steps of 10%. Both image brightness measurements (B_1 and B_2) achieve close performance ($corr_{B1} = 0.974, corr_{B2} = 0.993$). However the B_2 is very time consuming, 22 times compared to B_1 (based on our study); Hence, we decide to perform the brightness measure by $B = B_1$.

Focus / Sharpness: The used blurring factor was a circular averaging filter over a region of diameter that equals to 3-17 pixels in an increment of 2 pixels. Both image focus measures achieve close performance ($corr_{F1} = 0.752, and corr_{F2} = 0.608$). The computational complexity of both factors was also very close; hence, we decide to use an average $F = \frac{F_{L1}+F_{EL}}{2}$. For the four proposed image sharpness measures, the performance were ($corr_{S1} = 0.918, corr_{S2} = 0.917, corr_{S3} = 0.592, and corr_{S4} = 0.560$); and the computational complexity was very

Table 1. Face image quality measures (Q). Test image I(x,y) is of size $N \times M$, where μ, I_{min}, and I_{max} are the mean, minimum and maximum intensity values respectively. X, Y, and Z are Tristimulus values. G(x,y), G_{xx} and G_{yy} are the image gradient, and horizontal / vertical second derivatives. $\overline{r}$ and $\overline{t}$ are the variances of the reference image (r) and the test image (I), while σ_{rI} is the covariance of (r) and (I). w_i is the Gaussian weight of the i^{th} region.

Q	Method
C	$C_{RMS} = \sqrt{\frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[I(x,y)-\mu]^2}{MN}}$ $C_{Mic} = \frac{I_{max}-I_{min}}{I_{max}+I_{min}}$
B	$B_2 = \sqrt{D^2+E^2+F^2}$ $\begin{bmatrix} D \\ E \\ F \end{bmatrix} = \begin{bmatrix} 0.21 & 0.71 & 0.47 \\ 1.85 & -1.28 & -0.44 \\ -0.37 & 1.01 & -0.61 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$
F	$F_1 = \sum_{x=1}^{M}\sum_{y=1}^{N}\lvert G_{xx}(x,y)\rvert + \lvert G_{yy}(x,y)\rvert$ $F_2 = \sum_{x=1}^{M}\sum_{y=1}^{N}[G_{xx}(x,y)+G_{yy}(x,y)]^2$
S	$S_1 = \frac{1}{2}[\frac{1}{(N-1)M}\sum_{x=1}^{M}\sum_{y=1}^{N-1}\lvert I_{x,y}-I_{x,y+1}\rvert + \frac{1}{(M-1)N}\sum_{x=1}^{M-1}\sum_{y=1}^{N}\lvert I_{x,y}-I_{x+1,y}\rvert]$ $S_2 = \sum_{x=1}^{M-2}\sum_{y=1}^{N-2} G(x,y)$ $S_3 = \sum_{x=1}^{M}\sum_{y=1}^{N}(L_x.I_x^2 + L_y.I_y^2),\ where$ $L(x,y) = [I(x-1,y)+I(x+1,y) - I(x,y-1)-I(x,y+1)]^P$ $S_4 = \sum_{x=1}^{M}\sum_{y=1}^{N} L(x,y)[I_x^2+I_y^2],\ where$ $L(x,y) = [I(x-1,y)+I(x+1,y) - I(x,y-1)-I(x,y+1)]^P$
I	$I_1 = \frac{2\sigma_{rI}\overline{r}\overline{I}}{[\overline{r}^2+\overline{I}^2]}$ $I_2 = \sum_{i=1}^{16} w_i.\bar{I}_i,\ where$ $\bar{I}_i = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N} I(x,y).$

close; hence, we decide to use an average of the first two $S = \frac{S_1+S_2}{2}$.

Illumination: To evaluate the performance of the proposed illumination measures, seven sets from Yale [6] database were used. Each set contains 38 face images collected using a specific illumination setup. This empirical evaluation showed that the two image illumination measurements performed almost the same $(corr_{I1} = 0.938, and corr_{I2} = 0.881)$. We define the illumination measure by $I = I_1$ because I_2 needs to have a reference image.

3.1 Proposed Face Quality Index

Each of the above mentioned quality measures can only provide an estimate of a single quality factor in an image. However, several biometric application may require to have a more generic quality measure. Face quality index is obtained as follows:

- Each quality measure (defined as Q_m) needs to be normalized (between 0 and 1) that conveys meaningful interpretations for either poor and good qualities $f(Q_m) \rightarrow [0, ..., 1]$.
- Figure 1 shows the distributions of the quality measures, generated using the good set of FOCS database [2]. By studying these distributions, we decided to use Gaussian models $f(Q_m) = G(Q_m)$ as a mapping to the normalized range. This model maps quality measures of values closer to the distribution mean to values closer to 1, while maps values away from the mean to values closer to 0. For example, fig. 1- b) the brightness mean is 0.5, so $Q_m = 0.5$ will be mapped to $f(Q_m) = 1$; while $Q_m = 0 and 1$ will be mapped to $f(Q_m) = 0$.
- The proposed quality measures were integrate into a generic face quality index (FQI), where the geometric mean was used [9].

4 Experimental Results

In this section, we present various experiments to evaluate: (i) the performance of each quality factor, using simulated quality changing effects, and (ii) the performance of the proposed FQI indexes. We evaluated the effect of these factors to face recognition performance using images from the Yale [6], FERET [11] and MBGC[3] databases. We prepare: (i) **FTMC set:** A set of 238 subjects from FERET and 107 MBGC databases forming 345 gallery images and 345 corresponding probes. For training, 1530 images from subjects that are not included in galleries/probes set were used. A commercial face recognition system (PittPatt) was used to segment the face region and locate the eyes' centers. Each image was initially normalized by fixing the inter-pupillary pixel distance (IPD) to 75 pixels, performing an in plane rotation to set the line between the eyes to horizontal, and cropping the image to 250×200 pixels [7], such that eye-level is at 115, and left eye is at 62.5 pixel-level. (ii) **Yale set:** This database represents a real face database with various changes in illumination. For example, "$yaleB01\_P00A+000E+00$" belongs to subject no. 1, seen in pose ♯0, and the light source direction with respect to the camera axis is at 0 degrees azimuth ("A+000") and 0 degrees elevation ("E+00").

We conduct an experiment to compare the base performance of various Face Recognition (FR) algorithms using the FTMC set. In an identification experiment, rank-1 was used to represent the FR performance: Intensity-based techniques as Principal Component Analysis (PCA)[4] - 87.8% , and Independent Component Analysis (ICA) [2] - 85.8%, as well as distribution-based as Local Binary Pattern (LBP) [12] - 91.3%, and Local Ternary Patterns (LTP) [12] - 90.7%, and finally PittPatt, commercial software, 99.42%. In our study we decided to use LBP (PittPatt has an integrated quality evaluation/correction that we can not control).

To evaluate the performance of the used quality measurements and vs. the proposed FQI, we used the FTMC set for all the experiments except for the illumination assessment Yale set. We also perform a set of experiments to compare the proposed FQI to some benchmark measurements: (UQI) [14], (AVI) [8], and (WB) [13].

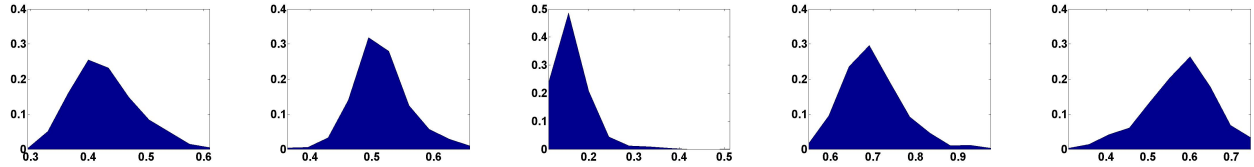
First, to evaluate how the contrast measure reflects the change in the image contrast, artificial contrast variation of the input face images are induced. For example, "$0.05 - 0.95$" maps the intensity values such that 10% of data is saturated at min and max intensities. Table 2-(Contrast) shows that FQI has a high correlation with recognition performance compared Wavelet based, which is comparable to average correlation and UQI.

Second, to evaluate how the proposed brightness measure reflects changes in image brightness, deviations in brightness intensity are induced. Brightness is artificially adjusted via $gamma$ parameter. In case $\gamma < 1$, the mapping is weighted towards higher (brighter) output values, and vice versa. Table 2-(Bright) illus-

[2] http://www.cse.nd.edu/cvrl/CVRL/Data_Sets.html

[3] http://www.cse.nd.edu/cvrl/CVRL/Data_Sets.html

[4] Linear Discriminant Analysis (LDA) did not perform good, as only 2 samples per subjects are available



(a)Contrast (b)Brightness (c)Focus (d)Sharpness (e)Illumination

Figure 1. Distribution of face image quality factors

Table 2. Face identification rank 1 "R1" using LBP and PittPatt. The probe face images qualities were artificially degraded: (a) contrast saturation, (b) brightness variation, and (c) blurring using circular averaging filter with variable diameter "D".

Contrast	*LBP*	*Pitt*	*FQI*	*AVI*	*UQI*	*WB*
Normal	91.3	99.4	0.37	0.89	0.13	0.47
10%	91.3	98.8	0.37	0.87	0.13	0.46
20%	90.4	99.1	0.32	0.84	0.13	0.45
30%	89.0	98.8	0.26	0.80	0.12	0.43
40%	82.6	97.4	0.21	0.78	0.12	0.42
50%	71.0	96.2	0.15	0.72	0.11	0.41
60%	52.8	74.2	0.12	0.71	0.09	0.40
70%	25.5	89.3	0.10	0.67	0.08	0.40
80%	4.64	83.8	0.08	0.63	0.06	0.42
CorLBP			0.87	0.94	0.99	0.33
Bright	*LBP*	*Pitt*	*FQI*	*AVI*	*UQI*	*WB*
$\gamma = 0.6$	91.0	99.1	0.08	0.91	0.14	0.68
$\gamma = 0.7$	91.0	99.1	0.18	0.91	0.14	0.62
$\gamma = 0.8$	90.7	99.1	0.31	0.90	0.14	0.56
$\gamma = 0.9$	91.0	99.1	0.40	0.90	0.14	0.52
Normal	91.3	99.4	0.37	0.89	0.13	0.47
$\gamma = 1.1$	91.3	99.4	0.32	0.87	0.13	0.44
$\gamma = 1.2$	91.3	98.3	0.25	0.85	0.12	0.41
$\gamma = 1.3$	91.0	99.4	0.18	0.83	0.12	0.38
$\gamma = 1.4$	90.4	99.4	0.14	0.82	0.11	0.36
CorLBP			0.33	0.15	0.19	0.03
Focus/ Sharp	*LBP*	*Pitt*	*FQI*	*AVI*	*UQI*	*WB*
Normal	91.3	99.4	0.37	0.89	0.13	0.47
$D = 3$	50.7	99.4	0.09	0.89	0.15	0.11
$D = 5$	36.8	99.1	0.06	0.89	0.15	0.11
$D = 7$	32.2	98.6	0.05	0.89	0.15	0.11
$D = 9$	30.1	98.6	0.05	0.90	0.15	0.12
$D = 11$	27.5	96.2	0.04	0.90	0.14	0.12
$D = 13$	25.2	91.6	0.04	0.90	0.13	0.13
$D = 15$	21.2	87.8	0.04	0.91	0.13	0.14
CorLBP			0.95	−.85	0.28	0.82

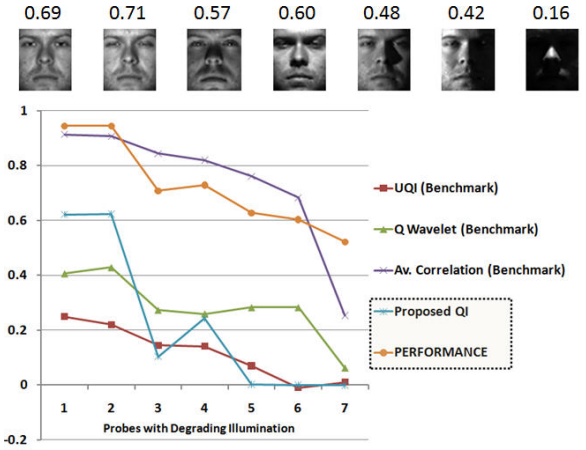


Figure 2. Evaluation of the proposed FQI vs. benchmark techniques designed to measure deviations in illumination.

trates how FQI is showing the best correlation; however it is very low. We attribute this to LBP being slightly affected by brightness variation.

Third, to evaluate how the proposed focus and sharpness measures reflects deviations in the image blurriness, focus and sharpness were changed by smoothing the input face images at various levels. Table 2-(Focus/Sharp) shows FQI can pick the deviation in correlation with recognition very efficiently.

Fourth, to evaluate the proposed illumination measure deviations, real data of various illumination changes from Yale set were used. In each experiment, different probes with various angles of the light source direction with respect to the camera, were used. Figure 2-(d) illustrates that FQI is selecting the deviation in intensity fairly good when compared to other benchmark techniques. When benchmark techniques are used we can see that: (a) WB quality measure is having a flat response across all illumination changes; (b) FQI index has better response compared to UQI and AVI.

In order to show that the proposed FQI is properly selecting the deviation in intensities, we also ran a non-parametric statistical hypothesis test, i.e., the Wilcoxon

signed-rank test. Wilcoxon test is as an alternative to the paired Students t-test since the distributions of the FQI and image quality measures cannot be assumed to be normally distributed. This test was used to compare the intensity changes of each quality factor to benchmark techniques as well as the proposed FQI. The best results were acquired in the brightness experiment. At the default 5% significance level, the Wilcoxon test fails to reject the hypothesis of zero median for the difference between the sample where the brightness of an input image varies, and the sample where FQI is computed for each intensity. In this case p-value=0.47, and h=0, and thus, we conclude that there is no difference between the samples. When running the same test using average correlation (which visually can be considered as the second best choice in selecting the deviation in intensity) instead of FQI, the Wilcoxon test rejects the hypothesis of zero median. In this case the p-value=0.0009 and h=1. When the other quality factors were investigated, FQI in most cases, proved to be a good alternative choice when compared to the best benchmark techniques.

5 Conclusions and Future Work

This paper presents several methods to evaluate face image quality measures. Gaussian models were used to normalize these measures, and a novel face quality index was used to integrate them. Experimental results indicated that: (i) the proposed FQI is capable of selecting deviations of quality factors; (ii) the benchmark techniques employed correctly respond to some of the changes in intensities, but not to all; and (iii) the proposed face quality index reflects the changes of input quality factors in correlation with recognition.

Our plan for future work includes: (a) enhancing face image quality index FQI by tuning the Gaussian models parameters, or by testing other models; (b) use real-world databases that represent changes of quality; and (c) investigating the image quality for other biometric modalities, namely the ear.

References

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

[2] M. Bartlett, H. Lades, and T. Sejnowski. Independent component representations for face recognition. In *proc. of the SPIE*, CA, USA, 1998.

[3] S. Bezryadin, P. Bourov, and D. Ilinih. Brightness calculation in digital image processing. In *proc. of the Int. Symposium on Technologies for Digital Fulfillment*, NV, USA, 2007.

[4] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, 38:149–161, 2008.

[5] X. Gao, S. Li, R. Liu, and P. Zhang. Standardization of face image sample quality. In *proc. of ICB*, Korea, 2007.

[6] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell. PAMI*, 23:643–660, 2001.

[7] B. Klare and A. Jain. On a taxonomy of facial features. In *Proc. of BTAS*, DC, USA, 2010.

[8] K. Kryszczuk and A. Drygajlo. On combining evidence for reliability estimation in face verification. In *proc. of EUSIPCO*, Italy, 2006.

[9] K. Kryszczuk, J. Richiardi, and A. Drygajlo. Impact of combining quality measures on biometric sample matching. In *proc. of the IEEE BTAS*, DC, USA, 2009.

[10] A. Michelson. *Studies in optics*. University of Chicago Press, 1927.

[11] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The feret evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell. PAMI*, 22:1090–1104, 2000.

[12] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Pattern Anal. Mach. Intell. PAMI*, 19:1635 – 1650, 2010.

[13] M. Vatsa, R. Singh, and A. Noore. SVM-based adaptive biometric image enhancement using quality assessment. In *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, pages 351–367. 2008.

[14] Z. Wang and A. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9:81–84, 2002.

[15] Z. Wang, A. Bovik, H. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13:600–612, 2004.

[16] Y. Yao, B. Abidi, N. Kalka, N. Schmid, and M. Abidi. Improving long range and high magnification face recognition: database acquisition, evaluation, and enhancement. *Computer Vision Image Understanding*, 111:111–125, 2008.

[17] P. Yap and P. Raveendran. Image focus measure based on Chebyshev moments. *IEE Proceedings on Vision, Image and Signal Processing*, 151:128–136, 2004.