Design and evaluation of photometric image quality measures for effective face recognition

Ayman Abaza¹², Mary Ann Harrison¹, Thirimachos Bourla², Arun Ross⁴

¹Advanced Technologies Group, West Virginia High Technology Consortium Foundation, Fairmont, West Virginia, USA
²System and Biomedical Engineering, Cairo University, Giza, Giza 12613, Egypt
³Lane Department of Computer Science and Electrical Engineering, Statler College of Engineering and Mineral Resources, Morgantown, West Virginia, USA
⁴Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, USA
E-mail: aabaza@wvhtf.org

Abstract: The performance of an automated face recognition system can be significantly influenced by face image quality. Designing effective image quality index is necessary in order to provide real-time feedback for reducing the number of poor quality face images acquired during enrollment and authentication, thereby improving matching performance. In this study, the authors first evaluate techniques that can measure image quality factors such as contrast, brightness, sharpness, focus and illumination in the context of face recognition. Second, they determine whether using a combination of techniques for measuring each quality factor is more beneficial, in terms of face recognition performance, than using a single independent technique. Third, they propose a new face image quality index (FQI) that combines multiple quality measures, and classifies a face image based on this index. In the author’s studies, they evaluate the benefit of using FQI as an alternative index to independent measures. Finally, they conduct statistical significance Z-tests that demonstrate the advantages of the proposed FQI in face recognition applications.

1 Introduction

The performance of biometric systems in operational environments can be impacted by several factors [1], including the quality of the input biometric data (e.g. face image). Poor quality data can cause efficiency loss of the biometric system. Thus, assessing the quality of the input biometric data prior to processing, can be beneficial in terms of improving matching performance.

Before we discuss our proposed approach let us first introduce some image-related notations and terminology that will be used through the remainder of this paper:

- **Quality factors** are image quality attributes such as contrast, brightness, sharpness, focus and illumination.
- **Quality measures** are techniques that are used to quantify quality factors. These measures can be arranged in an array called ‘quality matrix’.
- **Quality index** is a single number that represents the overall image quality of a biometric modality (e.g. face image).

Image quality measures (IQMs) are typically modality specific. Two categories of quality measures can be distinguished: generic (can be used for any biometric modality) or specific (viz. designed to address issues related to a specific biometric modality such as iris [2], fingerprints [3] or faces [4–6]). For the face modality, based on two-dimensional (2D) visible images, generic IQMs such as average image (AVI) [7], universal quality index (UQI) [8] and IQM [9] can be used. Biometric researchers have also developed modality-specific image quality assessment measures such as those based on redundant wavelets [10].

Face-based quality factors can be categorised in many different ways. One such categorisation, based on ISO/IEC standards, is described as follows (see Table 1):

- Factors related to the digital formatting of face images such as spatial resolution and contrast of grey-scale images.
- Factors related to scenes where faces are present such as head rotation, illumination, eyes, glasses and mouth.
- Factors related to photographic clauses such as head position in the image, exposure, brightness, focus and sharpness.

Several techniques have been proposed in the literature that discuss the benefits of using image quality factors for solving various face recognition related problems. However, biometric systems are expected to determine which technique to use to compute a specific quality factor. For example, the sharpness factor can be assessed using several techniques [12, 13]. The decision to select one technique over another is problem/application specific and often is made based on experience. However, such a heuristic decision making process becomes even more complicated when multiple image quality factors are considered (sharpness, illumination, focus etc.). Processing time can be
Section 4 presents a study on the evaluation of some of the evaluation criteria for image quality assessment databases used throughout this paper. Section 3 summarises some of the evaluation criteria for image quality assessment measures. To the best of our knowledge, this objective is consistent with matching or recognition accuracy. Table 2 summarises some of the evaluation criteria for image quality assessment measures. To the best of our knowledge, Hsu et al. [4] presented the only face quality assessment measure that is driven by face-based matching scores.

Other similar case studies for image quality assessment include:

3 Quality measure evaluation criteria

Sheikh et al. [20] defined the goal of quality assessment measures to be: ‘objective evaluation of quality in a way that is consistent with subjective human evaluation’. However, in the field of biometrics, this objective is consistent with matching or recognition accuracy. Table 2 summarises some of the evaluation criteria for image quality assessment measures. To the best of our knowledge, Hsu et al. [4] presented the only face quality assessment measure that is driven by face-based matching scores.

Table 2 Classification of image quality assessment measures based on evaluation criteria

<table>
<thead>
<tr>
<th>Method</th>
<th>Evaluation criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQM algorithm [9]</td>
<td>quality evaluation by human participants</td>
</tr>
<tr>
<td>pose estimation, sharpness, brightness and spatial resolution [21]</td>
<td>quality-based rankings compared with human rankings</td>
</tr>
<tr>
<td>focus [22]</td>
<td>synthesised degradation using Gaussian noise</td>
</tr>
<tr>
<td>illumination [12]</td>
<td>using various illuminated data from ‘Yale database’</td>
</tr>
</tbody>
</table>

2 Databases used

We evaluated a set of established IQMs using the following face databases: CASPL [14], Yale [15], FERET [16], MBGC [17], good–bad–ugly [18] and QFIRE [19]. Several face data sets, generated from the aforementioned databases, were evaluated:

- **Yale**: 38 gallery images from Yale [15] database where the light source direction with respect to the camera axis is at 0° azimuth and 0° elevation; and several probes data sets with illumination changes azimuth or elevation with respect to the camera axis.
- **FTMC**: A set of 238 subjects from FERET [16] and 107 subjects from MBGC [17] databases forming 345 gallery images and 345 corresponding probes.
- **Good–ugly**: 1085 image pairs from the good set and 1085 image pairs from the face and ocular challenge series (FOCS) database [18] were used.
- **QFIRE**: A set of 90 subjects from QFIRE [19] database were used: (i) 1800 face images from the normal setting images; (ii) 1080 face images extracted randomly from videos captured at 5, 15 and 25 feet while adjusting the focal plane of the camcorder across the full range; and (iii) 3240 face images extracted randomly from videos captured at three different illumination settings, that is, low, medium and high.

1. Perform a comparative study of various techniques that have been used to measure quality factors such as contrast, brightness, focus, sharpness and illumination. In particular, we evaluate the correlation between the measure used and a known (manually adjusted or computed) degradation in image quality (see Section 4).
2. Determine the most practical set of quality measures based on their correlation with systematic image degradation and computational speed.
3. Propose an alternative face image quality index (FQI) to predict face matching performance (see Fig. 1).

For the purpose of this study, we utilised several face databases, viz. CASPL [14], YALE [15], FERET [16], MBGC [17], FOCS [18] and QFIRE [19]. The rest of this paper is organised as follows. Section 2 describes the databases used throughout this paper. Section 3 summarises some of the evaluation criteria for image quality assessment measures. Section 4 presents a study on the evaluation of various IQMs, followed by selection of the practical measures Section 5, and integration of the selected measures into a face quality index Section 6. Section 7 presents the results of applying our proposed image quality index to both real data as well as data in which different image quality factors were manually adjusted, that is, synthetically changed (in the rest of the paper such data will be called ‘simulated’ data), followed by a case study to show beneficial usage of the proposed quality index. Conclusions and future work are discussed in Section 8.
• Yao et al. [13], measured the sharpness of face images, where the authors first, enhanced these images and, then, they performed face recognition before, finally comparing it to using un-enhanced images.
• Poh et al. [23], used a fusion algorithm that attempts to select a subset of biometric modalities/systems in order to achieve the maximal generalisation performance. In [24], Poh et al. normalised the quality-based score, and, then, face recognition performance was compared with linear normalisation.
• Vatsa et al. [10] fused the quality-based score, and then face recognition performance was compared with linear normalisation.
• Bhatt et al. [25] proposed a framework for quality-based classifier selection, and, then, recognition performance was performed to regular fusion cases.
• Kryszczuk and Drygajlo [26] used signal quality measures and classifier scores to improve performance in uni- and multi-modal scenarios.

4 Quality factors and measures for face images
Various IQMs have been reported in the literature that were used for face recognition. The most frequently used ones are those measuring the following quality factors [11]: (a) brightness, (b) contrast, (c) focus, (d) sharpness and (e) illumination. As discussed in the introduction, for each of the aforementioned factor, multiple measures are available in the literature. In this paper, our goal is to design, develop and evaluate a unified technique that combines various IQMs and generates a single value that can be used to represent the level of overall quality of query face images when used in practical face recognition scenarios.

4.1 Contrast
Image contrast is the difference in colour intensities that makes an object (face) distinguishable. The face image contrast [12] can be measured using the following equation
\[
C_{\text{RMS}} = \sqrt{\frac{\sum_{x=1}^{M} \sum_{y=1}^{N} [I(x, y) - \mu]^2}{MN}}
\] (1)
where \(\mu\) is the mean intensity value of the test face image \(I(x, y)\) of size \(N \times M\).

Another technique for image contrast is the Michelson contrast measure [27]
\[
C_{\text{Mic}} = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} \] (2)
where \(I_{\text{min}}\) and \(I_{\text{max}}\) are the minimum and maximum intensity values of the test face image \(I\).

4.2 Brightness
Wyszecki and Stiles [28] define brightness as an attribute of a visual sensation according to which a given visual stimulus appears to be more or less intense; or, according to which the area in which the visual stimulus is presented appears to emit more or less light, and range variation in brightness from 'bright' to 'dim' [29].

The face image brightness measure (let us denote it by \(B_1\)) can be calculated as the average of the brightness component after converting it into the HSB (hue, saturation and brightness) domain [29]. To convert from RGB (red, green and blue) colours to HSB range, each component is first normalised to the \([0, 1]\) range as follows
\[
\begin{bmatrix}
  r \\
  g \\
  b
\end{bmatrix} = \frac{1}{255} \begin{bmatrix}
  R \\
  G \\
  B
\end{bmatrix}
\]
\[
B_1 = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left( \max(r, g, b) \right)
\] (3)

Bezryadin et al. [29] suggested another image brightness measure
\[
B_2 = \sqrt{D^2 + E^2 + F^2}
\] (4)
\[
\begin{bmatrix}
  D \\
  E \\
  F
\end{bmatrix} = \begin{bmatrix}
  0.2053 & 0.7125 & 0.4670 \\
  1.8537 & -1.2797 & -0.4429 \\
-0.3655 & 1.0120 & -0.6104
\end{bmatrix} \begin{bmatrix}
  X \\
  Y \\
  Z
\end{bmatrix}
\]
where \(X, Y\) and \(Z\) are the tristimulus values. To convert from RGB colours to \(XYZ\) range, each component is first normalised to the \([0-1]\) range
\[
\begin{bmatrix}
  X \\
  Y \\
  Z
\end{bmatrix} = \begin{bmatrix}
  0.4124 & 0.3576 & 0.1805 \\
  0.2126 & 0.7152 & 0.0722 \\
  0.0193 & 0.1192 & 0.9505
\end{bmatrix} \begin{bmatrix}
  r \\
  g \\
  b
\end{bmatrix}
\]

4.3 Focus and sharpness
Image focus refers to the degree of blurring of face images. For a thin lens, given an object (face) at distance \(O_d\), the image is formed at distance \(I_d\), the focal distance of the lens \(f\) is given by: \(1/f = 1/O_d + 1/I_d\). If the face is displaced from \(O_d\), the energy from the face through the camera lens is distributed over a circular patch on the image plane, thus will form a blurred face image [22].

Yap and Ravendran [22] presented several image focus measures such as the \(L_1\)-norm of the image gradient, and the energy of the Laplacian. The \(L_1\)-norm of the image is defined as
\[
F_{L_1} = \sum_{x=1}^{M} \sum_{y=1}^{N} |G_{xx}(x, y)| + |G_{yy}(x, y)|
\] (5)
and the energy of the Laplacian of the image as
\[
F_{EL} = \sum_{x=1}^{M} \sum_{y=1}^{N} (G_{xx}(x, y) + G_{yy}(x, y))^2
\] (6)
where \(G_{xx}\) and \(G_{yy}\) are the second derivatives in the horizontal and vertical directions, respectively.

Image sharpness describes the clarity of detail in a face image, and it refers to the degree of clarity in both coarse and fine details [12]. Several image sharpness measures have been proposed in the literature. Kryszczuk
Drygajlo [7] defined image sharpness measure as

\[
S_1 = \frac{1}{2} \left[ \frac{1}{(N-1)M} \sum_{x=1}^{M} \sum_{y=1}^{N-1} |I_{x,y} - I_{x+1,y}| \right. \\
+ \left. \frac{1}{(M-1)N} \sum_{x=1}^{M-1} \sum_{y=1}^{N} |I_{x,y} - I_{x+1,y}| \right] 
\]

where \( G(x,y) \) is the gradient value at \((x,y)\).

The Tenengrad sharpness measure is defined as

\[
S_2 = \sum_{x=1}^{M-2} \sum_{y=1}^{N-2} G(x,y) 
\]

where \( G(x,y) \) is the gradient value at \((x,y)\).

The adaptive Tenengrad sharpness measure [13] is defined as

\[
S_3 = \sum_{x=1}^{M} \sum_{y=1}^{N} (L_x \cdot L_x^2 + L_y \cdot L_y^2) 
\]

\[
L_x(x,y) = [I(x + 1, y) - I(x - 1, y)]^p \\
L_y(x,y) = [I(x, y + 1) - I(x, y - 1)]^p 
\]

where \( L_x, L_y \) are the weights in the horizontal and vertical directions, and \( L_x, L_y \) are the horizontal and vertical gradients obtained by applying the Sobel filter.

The contrast and the brightness measures, where the face image intensity values were mapped to new values in the output image, and time in milliseconds.

Table 3: Contrast and the brightness measures, where the face image intensity values were mapped to new values in the output image, and time in milliseconds

<table>
<thead>
<tr>
<th>CASPL</th>
<th>( C_{\text{RMS}} )</th>
<th>( C_{\text{Mic}} )</th>
<th>CASPEAL</th>
<th>( B_1 )</th>
<th>( B_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.4284</td>
<td>0.9277</td>
<td>( y = 0.5 )</td>
<td>0.6135</td>
<td>0.6824</td>
</tr>
<tr>
<td>10%</td>
<td>0.4736</td>
<td>0.9763</td>
<td>( y = 0.6 )</td>
<td>0.5382</td>
<td>0.6417</td>
</tr>
<tr>
<td>20%</td>
<td>0.5151</td>
<td>0.9931</td>
<td>( y = 0.7 )</td>
<td>0.4756</td>
<td>0.6054</td>
</tr>
<tr>
<td>30%</td>
<td>0.5578</td>
<td>0.9968</td>
<td>( y = 0.8 )</td>
<td>0.4226</td>
<td>0.5724</td>
</tr>
<tr>
<td>40%</td>
<td>0.6047</td>
<td>0.9999</td>
<td>( y = 0.9 )</td>
<td>0.3774</td>
<td>0.5424</td>
</tr>
<tr>
<td>50%</td>
<td>0.6564</td>
<td>1.0000</td>
<td>normal</td>
<td>0.3133</td>
<td>0.5148</td>
</tr>
<tr>
<td>60%</td>
<td>0.7119</td>
<td>1.0000</td>
<td>( y = 1.1 )</td>
<td>0.3045</td>
<td>0.4893</td>
</tr>
<tr>
<td>70%</td>
<td>0.7731</td>
<td>1.0000</td>
<td>( y = 1.2 )</td>
<td>0.2753</td>
<td>0.466</td>
</tr>
<tr>
<td>80%</td>
<td>0.8393</td>
<td>1.0000</td>
<td>( y = 1.3 )</td>
<td>0.2497</td>
<td>0.4444</td>
</tr>
<tr>
<td>90%</td>
<td>0.9086</td>
<td>1.0000</td>
<td>( y = 1.4 )</td>
<td>0.2271</td>
<td>0.4245</td>
</tr>
<tr>
<td>corr</td>
<td>0.9960</td>
<td>0.6840</td>
<td>corr</td>
<td>0.974</td>
<td>0.993</td>
</tr>
<tr>
<td>time, ms</td>
<td>4.1</td>
<td>11.3</td>
<td>time, ms</td>
<td>230</td>
<td>11</td>
</tr>
</tbody>
</table>

4.4 Illumination

Luminance distortion is one of the measures of the image factor related to illumination. The term ‘luminance’ is used to describe the amount of light that passes through or is emitted from a particular area of the image. The UQI is a combination of three main factors: loss of correlation, luminance distortion and contrast distortion. The luminance distortion is defined as

\[
I_1 = \frac{2 \sigma_r \overline{\rho}}{\overline{\rho}^2 + \overline{\tau}^2} 
\]

where \( \overline{\rho} \) and \( \overline{\tau} \) are the variances of the reference \( (r) \) and test image \( (t) \), respectively, and \( \sigma_{\rho \tau} \) is the covariance of \( (r) \) and \( (t) \).

Another image illumination measure [30] is calculated as the weighted sum of the mean intensity values of the image divided into \((4 \times 4)\) blocks.

\[
I_2 = \sum_{i=1}^{4} \sum_{j=1}^{4} w_{ij} \overline{I}_{ij} \\
\overline{I}_{ij} = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} I(x,y) 
\]

where \( w_{ij} \) is the weight factor of each block. Abdel-Mottaleb and Mahoor [30] defined a Gaussian mask to add weights to various blocks of the face. This has the effect of assigning large weights to the blocks in the middle of the image and small weights to image boarders.

5 Selection of quality measures

In this work, in order to evaluate the performance of various face quality measures, face images from CASPL [14] and Yale [15] databases were used. We synthetically change CASPL images by adjusting the contrast, brightness and blurriness. What follows is a description of the process we used to assess the performance of each quality measure, by calculating the correlation coefficient:

Contrast: The CASPL face images were saturated at low and high intensities (see Table 3, and Fig. 2), in a step of 10%. For example, for the 10%, the values in intensity image I, range \([LOWIN = 0, HIGHIN = 1]\), are mapped to new values \([LOWOUT = 0.05, HIGHOUT = 0.95]\). Values below LOWIN and above HIGHIN are clipped; that is, values below LOWIN map to LOWOUT, and those above HIGHIN map to HIGHOUT. \( C_{\text{RMS}} \). corr = 0.996, represents the face image contrast factor better than the Michelson contrast measure \( C_{\text{Mic}} \). corr = 0.684. Michelson contrast measure does not represent the image contrast factor well, we return this to the fact that Michelson contrast measure depends only on the maximum and minimum values of the face. Hence, we denote the selected contrast measure by \( C = C_{\text{RMS}} \).

Brightness: The image brightness was artificially adjusted via the ‘gamma’ parameter, shown in Table 3 and Fig. 3, in 10% steps. Gamma specifies the shape of the curve describing the relationship between the values in input and output images. Both face image brightness measures \((B_1 \text{ and } B_2)\) achieve...
the same performance \((\text{corr}_{B_1} = 0.974, \text{corr}_{B_2} = 0.993)\).

However, the computation of \(B_2\) is very time consuming, about 22 times compared to that required by \(B_1\). Hence, the brightness measure is \(B_1\).

**Focus and sharpness:** Each image, under study, was blurred using a circular averaging filter over a region of diameter equals to 3–19 pixels, in increments of 2 pixels (see Table 4 and Fig. 4). Empirical evaluation suggested that: (i) the two measures of the face image focus factor achieve almost the same performance \((\text{corr}_{FL_1} = 0.752, \text{corr}_{FEL} = 0.608)\) and they require almost the same computational time. Hence, we decide to use an average of the two measures \(F = (F_{L_1} + F_{EL})/2\) and (ii) similarly, in terms of sharpness, we use an average of the first two sharpness measures, that is, \(S = (S_1 + S_2)/2\).

**Illumination:** Seven sets from the Yale [15] database, each captured under different illumination conditions were used, as shown in Table 5. From the empirical evaluation performed, we show that the two image illumination measures \(I_1\) and \(I_2\) achieve almost the same performance \((\text{corr}_{I_1} = 0.938, \text{corr}_{I_2} = 0.881)\). However, \(I_2\) does not need a reference image, and requires less processing time to be computed. Hence, the illumination measure is \(I_2\).

## 6 Proposed FQI

Each of the aforementioned IQMs can only provide an estimate of a single image quality factor. However, several biometric applications would require to have a general quality index used to indicate the overall quality of input data (e.g., face images, iris images etc.). The general quality can be used to: (i) reduce the number of poor quality face images acquired during enrollment thereby improving matching performance; and (ii) add weights in case of integrating the matching scores of several probes. Integrating these quality measures into a FQI ensures the collection of good quality images during enrollment. In this paper, we examine several criteria to calculate our proposed FQI and, via a set of experiments, illustrate that this FQI can complement the usage of conventional IQMs.

### 6.1 Quality measures fusion schemes

Grother and Tabassi [31] combined ‘normalised’ quality measures, that is, in the range \([0, 1]\), where ‘0’ corresponds to the same performance \((\text{corr}_{B_1} = 0.974, \text{corr}_{B_2} = 0.993)\).

### Table 4 Focus and the sharpness measures, where smoothing of the input face images was introduced by using a circular averaging filter (denoted as ‘d’) with various diameters, and time in milliseconds

<table>
<thead>
<tr>
<th>CASPL Diameter</th>
<th>Focus (F_1)</th>
<th>Focus (F_2)</th>
<th>Sharpness (S_1)</th>
<th>Sharpness (S_2)</th>
<th>Sharpness (S_3)</th>
<th>Sharpness (S_4)</th>
<th>Time, ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.0374</td>
<td>0.2879</td>
<td>0.7161</td>
<td>0.6819</td>
<td>0.4748</td>
<td>0.4354</td>
<td>10.2</td>
</tr>
<tr>
<td>(D = 3)</td>
<td>0.0125</td>
<td>0.0259</td>
<td>0.4573</td>
<td>0.4363</td>
<td>0.0296</td>
<td>0.0014</td>
<td>7.1</td>
</tr>
<tr>
<td>(D = 5)</td>
<td>0.0072</td>
<td>0.0074</td>
<td>0.3641</td>
<td>0.3460</td>
<td>0.0071</td>
<td>0.0002</td>
<td>4.2</td>
</tr>
<tr>
<td>(D = 7)</td>
<td>0.0049</td>
<td>0.0032</td>
<td>0.3035</td>
<td>0.2876</td>
<td>0.0024</td>
<td>4 \times 10^{-5}</td>
<td>4.3</td>
</tr>
<tr>
<td>(D = 9)</td>
<td>0.0038</td>
<td>0.0016</td>
<td>0.2566</td>
<td>0.2427</td>
<td>0.0010</td>
<td>1 \times 10^{-5}</td>
<td>4.3</td>
</tr>
<tr>
<td>(D = 11)</td>
<td>0.0028</td>
<td>0.0009</td>
<td>0.2174</td>
<td>0.2057</td>
<td>0.0004</td>
<td>7 \times 10^{-6}</td>
<td>6.6</td>
</tr>
<tr>
<td>(D = 13)</td>
<td>0.0022</td>
<td>0.0006</td>
<td>0.1859</td>
<td>0.1761</td>
<td>0.0002</td>
<td>4 \times 10^{-6}</td>
<td>3.8</td>
</tr>
<tr>
<td>(D = 15)</td>
<td>0.0019</td>
<td>0.0004</td>
<td>0.1610</td>
<td>0.1527</td>
<td>0.0001</td>
<td>3 \times 10^{-6}</td>
<td>3.8</td>
</tr>
<tr>
<td>(D = 17)</td>
<td>0.0016</td>
<td>0.0003</td>
<td>0.1413</td>
<td>0.1341</td>
<td>7 \times 10^{-5}</td>
<td>2 \times 10^{-6}</td>
<td>3.8</td>
</tr>
<tr>
<td>(D = 19)</td>
<td>0.0015</td>
<td>0.0003</td>
<td>0.1250</td>
<td>0.1187</td>
<td>4 \times 10^{-5}</td>
<td>1 \times 10^{-6}</td>
<td>3.8</td>
</tr>
<tr>
<td>corr</td>
<td>0.752</td>
<td>0.608</td>
<td>0.918</td>
<td>0.917</td>
<td>0.592</td>
<td>0.560</td>
<td>3.8</td>
</tr>
</tbody>
</table>
to bad quality and ‘1’ corresponds to good quality, for two compared fingerprints using several methods [32]:

- Minimum \( \bar{q} = \min(q_1, q_2) \)
- Geometric mean \( \bar{q} = \sqrt{q_1 \cdot q_2} \)
- Difference \( \bar{q} = |q_1 - q_2| \)
- Mean \( \bar{q} = (q_1 + q_2)/2 \)

Kalka et al. [33] used Dempster-Shafer theory approach as a combination scheme for ‘normalised’ quality measures. Given the belief that quality is bad (value = A), and the belief that quality is good (value = B), Kalka et al. [33] adopt Murphy’s combination rule to integrate beliefs. The following equation is a generalised expression for combining beliefs from \((k = 5)\) quality factors \(m_1\) to \(m_k\):

\[
\tilde{m}(A) = \frac{(m_{\sim 1}(A) \cdot m(A))^n}{(m_{\sim 1}(A) \cdot m(A))^n + (m_{\sim 1}(B) \cdot m(B))^n}
\]

where \(m(B) = 1 - m(A)\) since these propositions are complements of each other, and with equal probabilities, \((n = 0.5)\) for all evidence.

As each individual quality measure (e.g. the one estimating the contrast factor) yields a raw number in the range \([-\infty, \infty]\). This raw number needs to be mapped to a specific score range \([0, 1]\) that conveys meaningful interpretations from poor to good quality [4]. As it will be seen in Section 8, linear normalisation schemes are shown to be inefficient for combining quality measures. This is because, in practical scenarios, there are ranges corresponding to good and bad quality.

To find a better normalisation scheme, we studied the distributions of the quality measures. We used a subset from the ‘good’ set (as will be discussed in Section 8). Based on these distributions, we found the Gaussian models \(f(Q_m) = G(Q_m)\) to be a closer approximation for non-linear normalisation. This non-linear normalisation method was shown to be more efficient than the linear normalisation one.

The geometric mean was found to be the best fusing rule to integrate the above mentioned quality measures [34]. Thus, the Gaussian-based face quality index is defined as follows

\[
FQI = \sqrt[n]{G_c(C) \cdot G_b(B) \cdot G_f(F) \cdot G_s(S) \cdot G_i(I)}
\]

6.2 Neural network (NN)

In this paper, we proposed to use NN scheme to integrate the quality measures. One of the main advantages of NNs is the ability to use raw quality measures; in other words they do
We designed several NNs to classify the input images as either ‘good’ or ‘ugly’, and hence the expected matching performance using these images. Results from the FRVT 2006, showed the performance rates for the verification rate at false accept rate = 0.001 are [18]: GOOD = 0.98, and UGLY = 0.15.

We used the following scheme to combine the quality vectors of the probe and gallery by taking the minimum of each quality measure

$$Q(P, G) = \min(Q(P), Q(G))$$

$$= \min([C_p, B_p, F_p, S_p, I_p], [C_g, B_g, F_g, S_g, I_g])$$

$$= \min(C_p, C_g), \min(B_p, B_g), \min(F_p, F_g), \min(S_p, S_g), \min(I_p, I_g)$$  \hspace{0.5cm} (15)

where \(Q(P)\), and \(Q(G)\) are the quality vectors \(Q(P) = [C_p, B_p, F_p, S_p, I_p]\), and \(Q(G) = [C_g, B_g, F_g, S_g, I_g]\) for probe and gallery face images, respectively.

To classify the input, we applied the following hypothesis of good and ugly. To mark a face image as good, or low-quality, we define the following hypothesis: ‘high-quality face image persists yielding high-matching score regardless of the used matching technique or the matching image (i.e. various good probes of the same person), and vice versa’.

The NN output can be defined as

$$y^n = [C_t, B_t, F_t, S_t, I_t]$$

where \([C, B, F, S, I]\) are the quality measures, and the output is either good ‘1’, or bad ‘0’.

Figs. 5a and b show the performance of two face recognition techniques [a research technique namely local binary pattern (LBP) and a commercial software (PittPatt)], using the ‘good subset’ from the FOCS database. We kept the higher 50% for LBP and PittPatt, respectively, then we switched the gallery images for the same probe and apply the same rule

$$\min(M'(P, G_{i,1}), M'(P, G_{i,2})) > \text{Threshold}'$$  \hspace{0.5cm} (17)

where \(M\) is the matching score, \(P_i\), \(G\), are the probe and gallery images for sample \(i\), using recognition technique \(t\).

Figs. 5c and d show the performance of the two face recognition techniques using the ‘ugly subset’ from the FOCS database. We kept the lower 50% for LBP and PittPatt, respectively, then we switched the gallery images for the same probe and apply the same rule

$$\min(M'(P, G_{i,1}), M'(P, G_{i,2})) < \text{Threshold}'$$  \hspace{0.5cm} (18)

7 Experimental results

In this section, first, we evaluate several face recognition algorithms. Second, we present a set of experiments to evaluate the performance of independent quality measures against the proposed FQI index when using both simulated (image quality was synthetically changed) as well as real data. Finally, we present a case study on the beneficial usage of the proposed face quality index.

7.1 Evaluation of face recognition algorithms

For face detection we used a commercial software developed by the Pittsburgh Pattern Recognition (PittPatt) [http://www.pittpatt.com/]. For the selected FTMC (as well as the good–ugly, and the QFIRE data sets), PittPatt was used to segment the face region and locate eyes-centres. Each image was initially normalised by fixing the inter-pupillary pixel distance to 75 pixels. Then the face image is rotated to set the line between the eyes horizontally. Finally, the face image, 250 \(\times\) 200 pixels, [35] is segmented such that the eye-level is at 115 pixel-level, left eye is at 62.5 pixel-level.

Various face recognition algorithms that can be classified as, intensity-based like principal component analysis (PCA), and independent component analysis (ICA) [36]; distribution-based like LBP [37], and local ternary patterns

| Table 6 Face recognition performance using various techniques; Rank1 score represents the performance of identification experiments |
|-----------------|------------------|
| FTMC set       | Rank 1, %        |
| PCA            | 87.8             |
| ICA            | 85.8             |
| LBP            | 91.3             |
| LTP            | 90.7             |
| PittPatt       | 99.4             |
(LTP) [38] were used in an evaluation experiment. We conduct this comparison experiment (as shown in Table 6) using the FTMC data set. The commercial software (PittPatt) achieved the best performance followed by the LBP technique. We used LBP (a texture-based technique) in addition to the commercial face recognition system since we do not have control of the pre-processing step of the commercial system.

In this experiment, we used FTMC data set, which carries one training face image. We could not apply face recognition techniques which requires multiple training samples per subject [39–42].

7.2 Performance of various quality measures

In a first set of experiments to evaluate the performance of various quality measures, we used the Yale data set (real database that has various illumination setups), the QFIRE (real database that has various focus and illumination setups) and the FTMC data set by adding synthesised changes.

To evaluate how the contrast measure reflects the image contrast factor, artificial contrast variation of the input face images are induced. For example, '0.05–0.95' maps the intensity values in face image I to new values in the output image such that 10% of data is saturated at low and high intensities of I. This increases the contrast of the output image. Table 7 shows that: (i) face recognition performance degrades while contrast increases, (ii) LBP performance degraded dramatically with contrast and (iii) the proposed contrast measure is highly correlated with the image contrast change. Fig. 6a illustrates proper response to contrast changes.

To evaluate how the brightness measure reflects deviations of the brightness intensity factor, brightness is artificially adjusted via the \( \gamma \) parameter. This parameter specifies the shape of the curve describing the relationship between the values of the input and output images, after the brightness level is manually adjusted. In case \( \gamma \) is less than 1, the mapping is weighted towards higher (brighter) output values, and vice versa. Table 7 shows that: (i) face recognition slightly changes for the various brightness levels, and (ii) the proposed brightness measure picks the change in the image brightness. Fig. 6b illustrates proper response to brightness changes.

To evaluate how the focus and sharpness measures reflect deviations in the image blurriness. Focus and sharpness were changed by smoothing the input face images at various levels. The used smoothing factor is a circular averaging filter

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Face recognition performance (r1: rank1) using images where contrast and brightness intensities were changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>FTMC PittPatt R1, %</td>
</tr>
<tr>
<td></td>
<td>99.42</td>
</tr>
<tr>
<td>normal</td>
<td>98.84</td>
</tr>
<tr>
<td>0.05–0.95</td>
<td>99.13</td>
</tr>
<tr>
<td>0.1–0.9</td>
<td>98.84</td>
</tr>
<tr>
<td>0.2–0.8</td>
<td>97.39</td>
</tr>
<tr>
<td>0.25–0.75</td>
<td>96.23</td>
</tr>
<tr>
<td>0.3–0.7</td>
<td>94.20</td>
</tr>
<tr>
<td>0.35–0.65</td>
<td>89.57</td>
</tr>
<tr>
<td>0.4–0.6</td>
<td>84.35</td>
</tr>
<tr>
<td>0.45–0.55</td>
<td>74.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DATA</th>
<th>FTMC PittPatt R1, %</th>
<th>LB0050 R1, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.42</td>
<td>91.30</td>
</tr>
<tr>
<td>normal</td>
<td>98.84</td>
<td>91.30</td>
</tr>
<tr>
<td>( \gamma ) = 0.5</td>
<td>99.13</td>
<td>90.44</td>
</tr>
<tr>
<td>( \gamma ) = 0.6</td>
<td>98.84</td>
<td>88.99</td>
</tr>
<tr>
<td>( \gamma ) = 0.7</td>
<td>97.39</td>
<td>82.61</td>
</tr>
<tr>
<td>( \gamma ) = 0.8</td>
<td>96.23</td>
<td>71.01</td>
</tr>
<tr>
<td>( \gamma ) = 0.9</td>
<td>94.20</td>
<td>52.75</td>
</tr>
<tr>
<td>( \gamma ) = 1.0</td>
<td>89.57</td>
<td>25.51</td>
</tr>
<tr>
<td>( \gamma ) = 1.1</td>
<td>84.35</td>
<td>4.64</td>
</tr>
<tr>
<td>( \gamma ) = 1.2</td>
<td>74.20</td>
<td>0.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Face recognition performance using images where blurriness intensity was artificially changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>PittPatt %</td>
</tr>
<tr>
<td>FTMC</td>
<td>normal</td>
</tr>
<tr>
<td>disk = 3</td>
<td>99.42</td>
</tr>
<tr>
<td>disk = 5</td>
<td>99.13</td>
</tr>
<tr>
<td>disk = 7</td>
<td>98.55</td>
</tr>
<tr>
<td>disk = 9</td>
<td>98.55</td>
</tr>
<tr>
<td>disk = 11</td>
<td>96.23</td>
</tr>
<tr>
<td>disk = 13</td>
<td>91.59</td>
</tr>
<tr>
<td>disk = 15</td>
<td>87.83</td>
</tr>
<tr>
<td>disk = 17</td>
<td>82.03</td>
</tr>
<tr>
<td>disk = 19</td>
<td>71.88</td>
</tr>
</tbody>
</table>

Fig. 7 Examples for blurring face images: (upper) synthesised blurring using circular average filter (images from Yale [15]); (lower) real data from QFIRE database

---

© The Institution of Engineering and Technology 2014

IET Biom., pp. 1–11
doi: 10.1049/iet-bmt.2014.0022
denoted as ‘disk’) over a region of diameter equals to 3–19 pixels in an increment 2 pixels. Table 8 shows that: (i) PittPatt face recognition performance degrades when smoothing increases, and (ii) LBP, as a distribution-based, performance degraded dramatically with image blurring. Fig. 7 illustrates proper response to blurring variation using: (i) synthesised blurring by circular average filter with various diameters, and (ii) real data from QFIRE varying out-of-focus blur, where videos were captured at 5, 15 and 25 feet while adjusting the focal plane of the camcorder.

To evaluate the illumination measure deviations in the input illumination intensity, real data of various illumination changes from Yale set and QFIRE set were used. Fig. 8 illustrates proper response to illumination change using real data from: (i) YALE database, where the light source direction with respect to the camera axis, and (ii) QFIRE database, where three different levels of face contrast are achieved by three different illumination settings. Table 9 shows LBP performance was reasonable for minor illumination change, then degraded dramatically with major change.

### 7.3 Performance of various quality measures

In a second set of experiments, we trained several NNs to differentiate between ‘good’ and ‘ugly’. Table 10 shows 1-layer (six neurons) is yielding the best classification performance (81.02%).

Using the same data set, we compare the performance of NN combination of quality measures to other methods which does not need normalisation step, like logistic-regression, and support-vector-regression [43]; as well as other methods which need normalisation step like minimum, maximum, mean, geometric mean and Dempster-Shafer [33]. Also we recorded the performance using linear normalisation and Gaussian-models (as shown in Table 11).

| Table 10 Neural network |
|-------------------------|-----------------|-----------------|
| Layers | Nodes | Train perf, % | Test perf, % |
| 2 | 20-5 | 94.78 | 73.15 |
| 2 | 15-5 | 87.39 | 76.85 |
| 2 | 10-5 | 94.35 | 70.37 |
| 2 | 9-5 | 88.26 | 78.24 |
| 2 | 9-3 | 86.52 | 78.70 |
| 2 | 7-3 | 90.00 | 75.93 |
| 1 | 7 | 88.70 | 80.09 |
| 1 | 6 | 85.65 | 81.02 |
| 1 | 5 | 84.78 | 78.24 |
| 1 | 4 | 83.91 | 77.78 |

(i) Simulate the matching score using minimum of probe and gallery qualities. (ii) Classify the image as good or ugly.

| Table 11 Comparison of several quality measures fusion schemes, to classify the input image as ‘good’ or ‘ugly’ |
|-----------------|----------|----------|
| Fusion Rule | Linear, % Normalisation, % | Gaussian, % Models, % |
| minimum | 50.00 | 70.37 |
| maximum | 48.61 | 51.39 |
| mean | 48.15 | 72.69 |
| geometric-mean | 50.00 | 75.46 |
| dempster-shafer | 48.61 | 65.28 |
| logistic-regression | 71.30% | |
| support-vector-regression | 76.39% | |
| neural-network | 81.02% | |

![Fig. 9 Examples](www.ietdl.org)

a High quality (gallery)  
b High quality (probe)  
c Low-quality (probe) images (images from FOCS [18])

![Table 9 Face recognition performance using images where illumination intensity was changed](www.ietdl.org)

<table>
<thead>
<tr>
<th>Yale</th>
<th>LBP, %</th>
<th>(I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G = A + 000E + 00</td>
<td>100</td>
<td>0.694</td>
</tr>
<tr>
<td>P = A + 010E + 00</td>
<td>97.37</td>
<td>0.706</td>
</tr>
<tr>
<td>P = A + 000E - 35</td>
<td>92.11</td>
<td>0.570</td>
</tr>
<tr>
<td>P = A + 000E + 45</td>
<td>89.47</td>
<td>0.601</td>
</tr>
<tr>
<td>P = A + 050E + 00</td>
<td>26.32</td>
<td>0.479</td>
</tr>
<tr>
<td>P = A + 070E + 00</td>
<td>5.26</td>
<td>0.416</td>
</tr>
<tr>
<td>P = A + 000E + 90</td>
<td>0.157</td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 8 Examples deviations in illumination intensity from: (upper) Yale database; (lower) QFIRE database](www.ietdl.org)
Blind fusion (mean fusion), ugly probe and good probe, the enhanced to 69.00% using LBP, and 94.67%.

The propped face quality index, the system performance was PittPatt.

enhance the overall face recognition performance using identifiers, generated from two different matching scenarios: both the used for generating the face match scores. These results are instances of faces: two high-quality images (from the good indoors and outdoors. The partitions of interest are referred to as ‘good’ and ‘ugly’, that have an average identification accuracy of 0.98 and 0.15, respectively [18], Fig. 9 shows examples of high- and low-quality face images.

The used dataset composed of 300 subjects, three frontal instances of faces: two high-quality images (from the good dataset), and one low-quality image (from the ugly dataset). PittPatt [http://www.pittpatt.com/] software, and LBP were used for generating the face match scores. These results are generated from two different matching scenarios: both the gallery and probe are of high-quality, referred to as ‘good–good’ and the gallery is high-quality, but the probe is low-quality, referred to as ‘good–ugly’. Table 12 shows the following:

- Blind fusion (mean fusion), ugly probe and good probe, the identification rank 1 is 60.67% using LBP, and 92.33% using PittPatt.
- Selective fusion (basically to reject ‘ugly’) probes, using the propped face quality index, the system performance was enhanced to 69.00% using LBP, and 94.67%.

This case study shows that the proposed face quality index can be used to filter low-face quality image and hence to enhance the overall face recognition performance using PittPatt.

8 Conclusions and future works

In this paper, we first, evaluated a variety of face IQMs related to the following quality factors: contrast, sharpness, focus, brightness and illumination. We used both synthetic as well as real-world data. Then, we illustrated that the usage of supervised learning methods (e.g. NNs) is very important to understand the relationship between these measures and matching score prediction when used in practical face recognition scenarios. Our study resulted in the development of a more efficient FQI that manages to reflect the changes of input quality factors in correlation with face recognition performance. Experimental results indicate that certain image quality factors, namely contrast, sharpness and focus, highly affect the performance of texture-based face matching schemes such as LBP.

Our plan for future works includes: (i) studying other quality factors that affects face recognition performance (e.g. pose, image compression [44], reducing the spatial and grey-level resolution of the normalised images [45], or noisy night time images [46]); (ii) developing more sophisticated techniques that target to enhance the efficiency of the proposed face quality index; and (iii) investigating the image effect of various IQMs on other important biometric modalities, namely the human ear image.

9 References

11 Sang, J., Lei, Z., Li, S.Z.: ‘Face image quality evaluation for ISO/IEC standards 19794–5 and 29794–5’. Int. Conf. on Biometrics (ICB), Sassari, Italy, 2009
12 Gao, X., Li, S.Z., Liu, R., Zhang, P.: ‘Standardization of face image sample quality’. Int. Conf. on Biometrics (ICB), Seoul, Korea, 2007

| Table 12 Comparison of several identification performance (rank1) |
|---------------------|---------------------|
|                     | LBP, % | PittPatt, % |
| good–good           | 77.00  | 98.33       |
| good–ugly           | 2.00   | 1.33        |
| mean fusion         | 68.67  | 92.33       |
| selective           | 69.00  | 94.67       |


Phillips, P.J., Beieridge, J.R., Draper, B.A., et al.: ‘An introduction to the good, the bad, and the ugly face recognition challenge problem’. IEEE Int. Conf. on Automatic Face and Gesture Recognition and Workshops (FG), Santa Barbara, CA, USA, 2011


Nasrollahi, K., Moeslund, T.B.: ‘Face quality assessment system in video sequences’, European Workshop on Biometrics and Identity Management (BIOID), Roskilde, Denmark, 2008


Abaza, A., Harrison, M.A., Bourlai, T.: ‘Quality metrics for practical face recognition’. 21th Int. Conf. on Pattern Recognition (ICPR), Tsukuba, Japan, 2012


