

Comparison of Face Image Quality Metrics

Kevin O'Connor, Gregory Hales,
Jonathon Hight
Biometric Standards, Performance and
Assurance Laboratory
Purdue University, Department of
Technology, Leadership and Innovation
West Lafayette, IN

Shimon Modi, Ph.D.
Center for Development of Advanced
Computing,
Mumbai, India

Stephen Elliott, Ph.D.
Biometric Standards, Performance and
Assurance Laboratory
Purdue University, Department of
Technology, Leadership and Innovation
West Lafayette, IN

Abstract—Automated face recognition offers an effective method for identifying individuals. Face images have been used in a number of different applications, including driver’s licenses, passports and identification cards. To provide some form of standardization for photographs in these applications, ISO / IEC JTC 1 SC 37 have developed standardized data interchange formats to promote interoperability. There are many different publically available face databases available to the research community that are used to advance the field of face recognition algorithms, amongst other uses. In this paper, we examine how an existing database that has been used extensively in research (FERET) compares with two operational data sets with respect to some of the metrics outlined in the standard ISO / IEC 19794-5. The goals of this research are to provide the community with a comparison of a baseline data set and to compare this baseline to a photographic data set that has been scanned in from mug-shot photographs, as well as a data set of digitally captured photographs. It is hoped that this information will provide Face Recognition System (FRS) developers some guidance on the characteristics of operationally collected data sets versus a controlled-collection database.

Keywords—*face recognition; image quality; law enforcement; biometrics*

I. INTRODUCTION

Face recognition has been used extensively to verify the identity of an individual. However, the performance of such systems is constrained by the quality of the images in the data set. Publicly available face data sets have contributed to the development of many face recognition algorithms. In fact, evaluations such as face-recognition vendor test (FERET) [1] and academic data sets such as the Pose, Illumination, and Expression database [2] have served this purpose [3]. Many times, operational data sets are not available to the academic community for a number of reasons, such as privacy rules and existing regulations. This limitation is unfortunate because in a recent report, it was noted that face images captured at U.S. ports of entry “do not conform either to the national FR (face recognition) standard adopted by DHS or to FACESTD, the international standard specified in the Registry of USG Recommended Biometric Standards” [4] (page 12). Moreover,

such operational environments introduce challenges such as pose, head size, non-uniformity of lighting and general illumination issues. The National Institute of Standards and Technology (NIST) also conducted a study of face images in 2004 and compared operational data from these ports of entry to the FERET data set. They concluded that the operational images present challenges, which include faces that are not centered, non-frontal head poses and poor illumination / backgrounds [5]. Therefore, examining data sets against a common list of standard metrics of image quality may provide valuable information to facial recognition algorithm developers. In 2006, L-1, a commercial identity solutions provider, conducted a quality assessment of facial images. One of the goals of their research was to contribute to the quality assessment framework of the ISO/IEC 19794-5 standard. A summary of their report can be found here [6]. The metrics used in that particular study map to the same metrics used in this paper.

It is not only important to compare one data set against another data set by using the same metrics, but it is of interest to the community that older images that have been collected historically be included. For example, many mug-shot photographs were collected using film and subsequently printed out and stored. Over time, these photographs have aged. Also, these photographs were captured when acquisition best practices were not in place (or adopted by the organization). Large amounts of legacy data of this type are stored by organizations, and these data may be the only photographic record available of a given individual. Understanding the quality of such legacy data and how it will be processed by a face recognition system are therefore important.

II. RELATED WORK

The motivation behind this work stemmed from an initial request to analyze mug-shot photographs collected in a correctional facility, to identify problematic standardized image quality metrics, and to optimize the capture process. The goal was to improve the image quality of photographs taken in an operational environment [7]. That study only examined images captured by a digital camera. However, paper-based

photographs were also made available for analysis. To provide more insight into the results, the comparison with the FERET data set was used. The impact of poor-quality data has been studied at length [8]. In that report, the authors make a number of observations, including that the detection and measurement of quality is necessary for dealing with such poor-quality data. Therefore, a methodology for accurately quantifying quality is required if “prevention and mediation” are to occur (page 4). Prior to the publication of that report, NIST published a best-practice document for capturing mug shots in the law enforcement environment [5]. That document has been used as a reference tool for law enforcement agencies worldwide.

III. METHODOLOGY

Three different data sets were used in the analysis. Two data sets were collected in an operational environment. The first data set, called *Legacy* in this paper, consisted of photographs that were captured over a number of years. Some of these photographs were of very poor visual quality and had not been stored in optimal conditions, nor were they printed on archival-quality paper. Some of the photographs had holes in them where they had been filed in a ring binder, or they had started to change color because of their age. A Kodak i1220 scanner was used to convert these images to digital format for subsequent processing in the image quality tool. The scanner had the following resolution: 300 dpi, 24-bit color, media type of card stock, document type of photograph, with no compression. A total of 9,233 photographs were scanned and then subsequently cropped to remove any identifying artifacts, such as their correction number. This was a requirement of the University’s Institutional Review Board (IRB). A second data set, called *Electronic* data set, consisting of 49,694 images was collected, of which 48,786 were used for this study. All of the metadata for these images also had to be removed because of the IRB restrictions on this data set; however, these images would have been collected around 2009 / 2010. The third data set was the *FERET* data set, which consisted of 4,063 images. All three of these data sets were processed by a commercially available image quality algorithm. There were several reasons for choosing a commercial image quality algorithm, as opposed to creating one for the purposes of this study. First, the experiment is agnostic to the type of image quality tool that is being used because the output of the tool is a standardized set of metrics. Therefore, the motivation was to examine the differences in the standardized metrics, as opposed to creating a novel approach to extracting such information. The tool was not the focus; rather, the image quality results were the focus. Second, the tool was well understood within the biometric community and had been used to perform reliable and repeatable studies. The software provided similar metrics as shown in [6]. Each of the image quality metrics were clustered together: format, digital, photographic, scene, and algorithmic. In total, 36 image quality metrics were evaluated. These five clusters are described in more detail in ISO/IEC 19794-5 [9]. Out of these 36 image quality metrics, 28 provided a range score between zero and 10, which was used for subsequent statistical analysis. These 28 image quality metrics were banded in three ranges: 0–3.9 indicates a poor score, 4–6.9

indicates a medium score, and 7–10 indicates a good score. An analysis of variance (ANOVA) method of analysis was used to determine whether the means of the different groups were equal. Parametric tests involve hypothesis testing that have a strict set of assumptions that have to be met [10]. The ANOVA results can be divided into two segments: the variation that is explained by the model (1) and the variation that is not, which is called the error (2). Both are used to calculate the *F*-statistic (3) testing the hypothesis $H_0 = \mu_1 = \mu_2 = \dots = \mu_1$. The results are described as p-values. When the H_0 is rejected, the variation of the model (SSM) tends to be larger than the error (SSE), which corresponds to a larger *F* value. This is represented by the equation below:

$$SSM = \sum (\hat{Y}_i - \bar{Y})^2, dfM = 1, MSM = SSM / dfM \quad (1)$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2, dfE = n - 2, MSE = SSE / dfE \quad (2)$$

$$F = MSM / MSE \sim F(dfM, dfE) \quad (3)$$

$$F \geq F(1 - \alpha, dfM, dfE) \quad (4)$$

We have used this methodology in other experiments. Following this initial analysis, a Tukey’s test was conducted to determine which means were significantly different from one another. The test compares the means of every treatment with every other treatment. This method is more suitable for multiple comparisons [11].

The hypothesis for this experiment was as follows:

$$H_0: \mu_{iqcL} = \mu_{iqcE} = \mu_{iqcF} \quad (5)$$

$$H_a: \mu_{iqcL} \neq \mu_{iqcE} \neq \mu_{iqcF} \quad (6)$$

where *iqc* is the individual image quality metric, *L* is the Legacy data set, *E* is Electronic, and *F* is the FERET. The alpha (α) for each of the tests was set at 0.05.

IV. RESULTS

The first set of analyses was to examine the image quality scores of the three data sets. The ANOVA revealed that for the 28 variables that were used in the statistical analysis, the null hypothesis was rejected at $\alpha = 0.05$. The results are presented in Tables I–VI, showing overall image quality as well as the five clusters (format, digital, photographic, scene, and algorithmic) from ISO / IEC 19794-5.

TABLE I. OVERALL

Quality Metric	Legacy	Electronic	FERET
Overall	7.14	6.24	7.28

Table I shows the *overall* quality metric revealed interesting results. We would expect that the FERET data set would have the best image quality, and the Legacy data set the worst. This was not the case. The more current, digitally captured photographs (the Electronic data set) had the worst

image quality of the three data sets. This result indicates that artifacts other than the acquisition technology might affect the image quality. A graphical representation of the overall quality score is shown below in Figure 1.

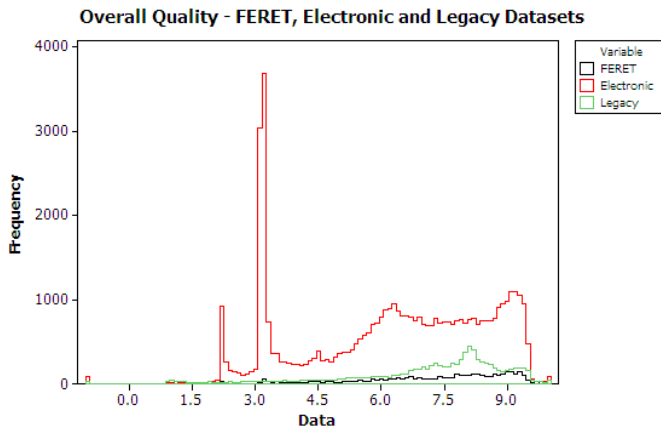


Figure 1. Overall quality distribution of the three data sets

The distribution of the *overall* quality scores for the Legacy, Electronic and FERET data sets showed that the Legacy and FERET data sets had similar distributions, which were slightly skewed toward the “good-quality” metric. However, the bulk of the Electronic photographs have a spike between 3.0 and 3.5, which is in the “poor-quality” metric band. Although there are significantly more photographs in the Electronic data set, the distribution is clearly different than the other two data sets.

TABLE II. FORMAT

Quality Metric	Legacy	Electronic	FERET
Compression Artifacts	6.74	6.79	5.08

As can be seen in Table II, the compression artifact is of medium quality for all three of the data sets.

TABLE III. DIGITAL

Quality Metric	Legacy	Electronic	FERET
Contrast	6.39	6.90	6.48
Scanning Artifacts	6.68	7.12	7.79
Interlaced	9.77	7.28	8.78
Sensor Noise	7.35	6.80	5.64

Table III shows that no one data set provided a clear advantage over the other with respect to the image quality metrics.

TABLE IV. PHOTOGRAPHIC

Quality Metric	Legacy	Electronic	FERET
Centered	5.89	5.28	6.60
Cropping	9.95	9.95	9.99
Focus	7.63	4.24	5.07
Motion Blur	8.01	7.96	8.41
Exposure	7.21	6.96	7.74
Unnatural Color	6.93	7.38	6.62

It should be noted that the Legacy data set was manually cropped to remove any identifiers, and therefore, this is probably an artificially high metric for this type of photograph. The other two databases were not cropped. It is interesting to note that the Electronic data set has the lowest score with respect to centering, focus, and motion blur. This result correlates with the related results of the point of entry, which included non-centered faces and blurry or poorly illuminated faces [4]. These results are similar to those in [7]. Clearly, the operational images are not conforming to the Standard Geometric Characteristics shown in Figure 2.

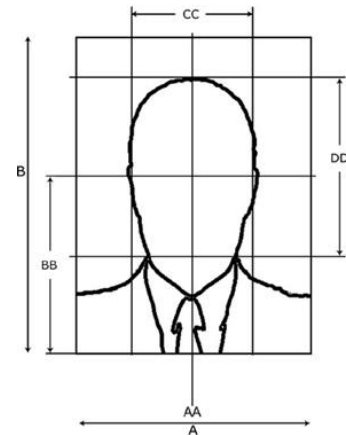


Figure 2. Standard Geometric Characteristics [HYPERLINK "" | "ISO5" 1].

Tables V and VI show the scene and algorithmic clusters of image quality variables.

TABLE V. SCENE

Quality Metric	Legacy	Electronic	FERET
Eyes Clear	9.53	9.45	8.77
Glare Free	6.58	6.63	6.56
Sunglasses	6.81	5.28	6.01
Eyes Open	8.38	7.94	7.77
Shadows in the Eye Sockets	8.64	8.08	8.13
Uniform Lighting	5.39	4.82	4.61

<i>Quality Metric</i>	<i>Legacy</i>	<i>Electronic</i>	<i>FERET</i>
Hot Spots	5.36	4.92	6.09
Facial Shadows	7.02	7.56	8.12
Background Uniformity	3.85	6.86	5.94
Background Brightness	4.00	3.38	5.31
Background Shadows	6.27	5.10	7.94
Frontal Pose	6.90	7.37	8.40

TABLE VI. ALGORITHMIC

<i>Quality Metric</i>	<i>Legacy</i>	<i>Electronic</i>	<i>FERET</i>
Faceness	8.94	9.37	9.43
Texture	7.54	3.47	4.18

The second analysis was Tukey's test, the results of which can be deduced from the results in Tables I–VI. Because of the IRB constraints, we were unable to process these through a face recognition algorithm and retrieve performance scores. However, we did complete an analysis on the metric Failure to Extract. A Failure to Extract is concerned with samples that are unable to be processed completely. A Failure to Extract may occur for a number of reasons, such as feature segmentation, extraction, or quality control. A failure to extract could be a failure to feature extract or failure to determine quality [11]. The results of the FTX rates for each database are given in Table VII.

TABLE VII. FAILURE TO EXTRACT

<i>Database</i>	<i>Legacy</i>	<i>Electronic</i>	<i>FERET</i>
FTX	810	1791	8
Total Images	9232	49692	4063
FTX rate	8.77%	3.60%	0.19%
Overall quality score (Table I)	7.14	6.24	7.28

The results show that the FERET data set had the lowest FTX at 0.19%, and the Legacy data set was the worst at 8.77%. This finding does not correlate with the overall quality score, repeated for clarity from Table I.

V. CONCLUSIONS AND RECOMMENDATIONS

The results of this analysis indicate that there is still much to be done to improve the image quality of operational data sets. Additionally, this paper reports the image quality results of operationally collected data sets and compares them to a publically available data set. The FERET data set, which has been used to train face recognition algorithms, had a better

overall image quality and better results for many of the image quality metrics. Those agencies that collect images in the field need to be aware of image quality deficiencies and should strive, where appropriate, to collect images that have better image quality. However, because of operational constraints, such a desire might be unobtainable. Therefore, algorithm developers should be able to work with operationally gathered data sets to realistically model such environments, and organizations should try and make such data sets available in accordance with the appropriate rules and regulations. While this research was unable to generate performance results for the data sets, it is evident from the image quality scores and extraction rates that developers may want to rethink the training process to include operational data or data that are captured under similar constraints if actual operational data are not available.

VI. REFERENCES

- [1] P. J. Phillips, J. Wechsler, and P. J. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image Vis. Comput.*, vol. 16, pp. 295-306, 1998.
- [2] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615-1618, Dec. 2003.
- [3] P. Mohanty, S. Sarkar, R. Kasturi, and P. J. Phillips, "Subspace Approximation of Face Recognition Algorithms: An Empirical Study," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 734-748, Dec. 2008.
- [4] DHS, "Facial Image Quality Improvement and Face Recognition Study Final Report," United States Visitor and Immigrant Status Indicator Technology (US-VISIT) Program, 2007.
- [5] NIST, "Best Practices Recommendation for Capturing Mug-shots and Facial Images, Version 2," National Institute of Standards and Technology, 1997.
- [6] R.-L. V. Hsu, J. Shah, and B. Martin, "Quality Assessment of Facial Images," in *Biometrics Consortium*, Baltimore, MD, 2006, pp. 1-18.
- [7] G. T. Hales, "Evaluation of the Indiana Department of Corrections Mug Shot Capture Process," MS Thesis, Purdue University, West Lafayette, IN, 2010.
- [8] A. Hicklin and R. Khanna, "The Role of Data Quality in Biometric Systems," Mitretek Systems, 2006.
- [9] ISO/IEC: 19794-5 Information Technology – Biometrics – Biometric data interchange formats – Part 5: Face image data. 2005.
- [10] NIST / SEMATECH. (2006) e-Handbook of Statistical Methods.
- [11] E. P. Kukula and S. J. Elliott, "Beyond current testing standards: A framework for evaluating human-sensor interaction," in *International Biometric Performance Testing Conference*, Gaithersburg, MD, 2010.

