On the assessment of face image quality based on handcrafted features

Olaf Henniger[1], Biying Fu[1], Cong Chen[2]

Abstract: This paper studies the assessment of the quality of face images, predicting the utility of face images for automated recognition. The utility of frontal face images from a publicly available dataset was assessed by comparing them with each other using commercial off-the-shelf face recognition systems. Multiple face image features delineating face symmetry and characteristics of the capture process were analysed to find features predictive of utility. The selected features were used to build system-specific and generic random forest classifiers.

Keywords: Biometrics, face recognition, face image quality.

1 Introduction

Not all biometric samples are equally well suited for the automated recognition of individuals. The utility of a biometric sample, i.e. its usefulness for telling mated and non-mated samples apart, can be expressed by a quality score [ISO16]. The quality score can be used, e.g., for deciding whether the re-acquisition of data is deemed necessary, for weighting partial results in multi-biometric systems, or for selecting the best (in some sense) from a series of captured biometric samples. The utility of a biometric sample depends on its faithfulness to its source (i.e. fidelity) and the distinctiveness of the biometric features (i.e. character) [ISO16]. The utility of a biometric sample can be quantified after comparing it with mated and non-mated samples from a dataset. Hence, utility depends on the underlying dataset and on the feature extraction and comparison algorithms used.

Fields holding biometric sample quality scores were introduced into several standardized biometric data interchange formats [ISO19]. In these data formats, if a quality score is reported, valid values are integers between 0 and 100. According to [ISO18], quality scores from 0 to 25 should indicate unacceptable quality, from 26 to 50 marginal quality, from 51 to 75 adequate quality, and from 76 to 100 excellent quality. The calibration of the boundaries between the levels of quality is a considerable challenge.

Related work on predicting the utility of biometric samples concentrated on finger images [TWW04, T^+16, ISO17] and iris images [TGS11, ISO15]. There is no standard yet for how to assess face image quality. To better understand face image quality assessment, algorithms can currently be submitted to NIST for evaluation [G^+20]. A Technical Report

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany,
{ olaf.henniger | biying.fu }@igd.fraunhofer.de

[2] Technical University of Darmstadt, Department of Computer Science, Darmstadt, Germany

[ISO10] outlines a set of face image features that could be useful in calculating quality scores, but does not specify how to aggregate the individual feature values. It is currently under revision and amenable to contributions. A Technical Report on portrait quality [ICA18] includes requirements and recommendations on how to capture suitable reference face images, but does not specify how to assess the quality of face images captured in an arbitrary environment. Tools for automatically checking compliance to the ICAO requirements compute a number of individual scores, which however are not aggregated into an overall quality score [F^+12].

For assessing face image quality, proprietary algorithms trained for particular face recognition algorithms are in use. This paper investigates how to predict the utility of face images across multiple state-of-the-art face recognition algorithms. We followed a supervised machine-learning approach similar to the one applied in finger image quality assessment [T^+16, ISO17]. The goal was to learn a mapping from a face image feature vector to a scalar quality score. In contrast to [HO^+19], which is about face image quality assessment using automatically learned features, we took “handcrafted” features into consideration, which were drawn from [ISO10]. The strength of handcrafted features is their explainability, which helps avoid using features with a potential demographic bias.

The rest of this paper is organized as follows: Section 2 describes the data available for this study. Section 3 deals with the a-posteriori assessment of the quality of a biometric sample by comparing it with other samples. Section 4 considers the a-priori quality assessment, i.e. predicting the utility of a sample without comparing it first with other samples. Section 5 assesses the accuracy of the utility-prediction model using a testing dataset. Finally, Section 6 summarizes the results and gives an outlook on future research work.

2 Underlying data

2.1 Face image dataset

The publicly available NIST Special Database 32 [C^+09] was used in the analysis. It consists of 712 face images of 380 different test subjects. 686 images are frontal or nearly frontal face images. 26 images are full-profile or nearly full-profile images. For 145 test subjects, the dataset contains more than one different images, and for 69 test subjects even more than two different images. For one test subject, the dataset contains three identical images, only one of which was used in this study to avoid bias. For the other test subjects, there is only one image. The sizes of the uncropped frontal images range from 240×240 pixels to 1824×1170 pixels. The majority of images is of size 480×600 pixels.

2.2 Similarity scores

Two commercial off-the-shelf face identification systems were used to calculate similarity scores. In the following, they are referred to as System 1 and System 2. The systems are treated as “black boxes” as we do not target at a comparative technology evaluation. The systems underwent the deep-learning revolution and successfully participated in NIST’s recent Face Recognition Vendor Tests [GNH18a, GNH18b]. Both systems were configured so that each search returned 100 candidates, for whom the similarity scores were logged.

For both systems, all and only frontal face images were attempted to enrol into the reference database. Only frontal face images were enrolled because many face recognition systems do store frontal face images as reference images (e.g. in ePassports, forensic databases, entry/exit systems). System 1 encountered three failures to enrol. System 2 encountered no failures to enrol.

After enrolment, for each system, the reference database was searched against all (frontal and profile) face images. For utility assessment, comparisons of images with themselves were not taken into consideration. System 1 encountered three failures to extract, for the same images for which it encountered failures to enrol. System 2 encountered no failures to extract. For all frontal probe images without failure to extract, both systems returned all mated reference identifiers in the candidate lists. For profile probe images, neither system returned all mated reference images in the candidate lists. For this lack of mated similarity scores for profile images, we limited the study exclusively to the (nearly) frontal face images. Taking into account only frontal images, both systems returned all mated reference identifiers at the head of their candidate lists, i.e. all mated similarity scores were greater than any non-mated similarity score. Hence, despite their diversity, all frontal images in the dataset, except for the ones with a failure to extract, could be regarded as excellent quality with respect to state-of-the-art face recognition systems.

2.3 Training subset and testing subset

We partitioned the face image dataset randomly into nearly equally large disjoint training and testing subsets, leaving the subsets of face images for the same test subject undivided. The training dataset consisted of 345 face images of 190 test subjects. The testing dataset consisted of 339 face images of 190 test subjects.

3 A-posteriori assessment of utility

The utility of a biometric sample can be predicted in several ways, e.g.:

- For NIST’s fingerprint image quality (NFIQ) assessment algorithm, version 1, the utility of a biometric sample was defined as the normalized difference between the mean of the sample’s mated non-self similarity scores and the mean of the sample’s non-mated similarity scores. Predicting a real-valued scalar is a regression problem. However, as regression methods failed to give adequate predictions, for NFIQ 1.0, the machine-learning problem was restated in terms of classification into utility classes (excellent, very good, good, fair, or poor utility) [TWW04].

- For the new and improved version NFIQ 2.0, a random decision forest was trained for binary classification into two utility classes (high or low utility). The trained random decision forest outputs class membership along with its probability. The quality score is the probability that an image belongs to the high-utility class multiplied by 100 and rounded to the nearest integer [T^{+}16, ISO17].

The normalized difference between the mean of a sample's mated non-self similarity scores and the mean of its non-mated similarity scores can be calculated only if a sufficient number of randomly distributed mated and non-mated similarity scores were available. However, because similarity scores were available only for the most similar candidates, we chose another measure of separation of mated and non-mated similarity score distributions. For each frontal face image with more than one mated non-self similarity score, we computed a utility score as the normalized difference between the arithmetic mean of mated non-self similarity scores and the maximum non-mated similarity score. Images with only one mated non-self similarity score were not considered because the arithmetic mean of mated non-self similarity scores would be the same for both compared images, independent of their quality.

Using the training dataset, we built regression models between the utility score and multiple features specified in Section 4.1. However, these models failed to give adequate predictions of utility in the testing dataset. Therefore, like NFIQ 2.0, we tried binary classification into two utility classes. We selected face images of high and low quality as follows:

1. Class 1: All images whose minimum mated score was greater than the threshold value that corresponds to FNMR = 60% were labelled as high quality.
2. Class 0: All images whose maximum mated score was less than the threshold value that corresponds to FNMR = 30% were labelled as low quality.

The boundaries are arbitrary. They were chosen so that in the given training dataset about 40 images were of Class 1 and about 40 images were of Class 0 for either system. The remaining images neither belong to Class 1 nor to Class 0.

For each face recognition system, a specialized quality prediction model can be constructed. However, it would be useful to build a generic face image quality assessment model independent of particular face recognition systems. For this purpose, we formed unions and intersections of the Classes 1 and 0 of System 1 and System 2, respectively:

- The union of the Class 1 training sets consisted of 47 images that were of high quality for either System 1 or System 2. The union of the Class 0 training sets consisted of 49 images that were of low quality for either System 1 or System 2.
- The intersection of the Class 1 training sets consisted of 25 images that were of high quality for both System 1 and System 2. The intersection of the Class 0 training sets consisted of 36 images that were of low quality for both System 1 and System 2.

4 A-priori assessment of sample quality

4.1 Selection of face image features

Several face image features that could be suitable for predicting utility were proposed in [ISO10]. We coded the feature extraction in Python and extracted a feature vector consisting of the following elements from each face image:

- left-right (lighting and pose) symmetry [GLZ07, ISO10] calculated as sum of differences of normalized pixel luminance values of the left and right halves of the face and as cross-entropy (CE), Kullback-Leibler (KL) divergence, and intersection of histograms of
  - normalized pixel luminance values of the left and right halves of the face and
  - LBP (local binary pattern) filtered pixel luminance values of the left and right halves of the face, respectively,
- characteristics of the capture process: contrast, global contrast factor, measures of image brightness (mean, variance, skewness, and kurtosis of pixel luminance values), exposure, sharpness, inter-eye distance, and blur [ISO10].

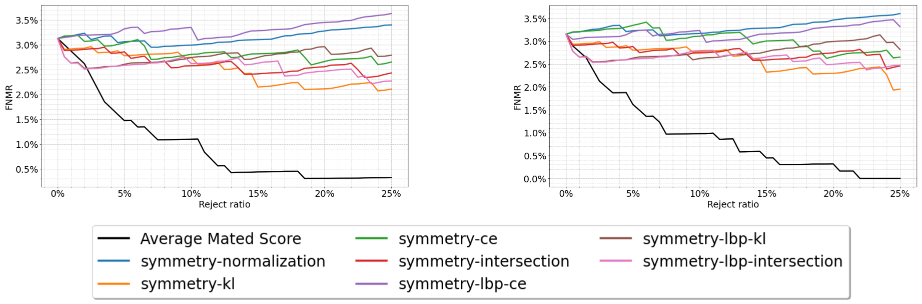
Tab. 1 shows the coefficients of the Spearman's rank correlations between the face image features and utility scores for System 1 and System 2.

Tab. 1: Spearman's rank correlation coefficients for the face image features under consideration

	symmetry-normalization	symmetry-KL	symmetry-CE	symmetry-intersection	symmetry-LBP-CE	symmetry-LBP-KL	symmetry-LBP-intersection	contrast	global contrast factor	mean of luminance	variance of luminance	skewness of luminance	kurtosis of luminance	exposure	sharpness	inter-eye distance	blur
System 1 utility score	4	-21	24	14	-8	-5	7	24	-24	27	19	-25	-26	32	27	-2	25
System 2 utility score	5	-14	27	7	-2	-10	12	21	-8	20	17	-12	-14	28	24	11	6

Within the training dataset, higher correlations with the utility scores were observed for exposure, mean, variance, skewness, and kurtosis of pixel luminance, left-right symmetry calculated as cross-entropy of histograms of normalized pixel values, sharpness, blur, global contrast factor, and contrast. Variance of pixel luminance was strongly correlated with contrast. Skewness of pixel luminance was strongly correlated with kurtosis. Therefore, the variance and skewness of pixel luminance need not be used in the training process.

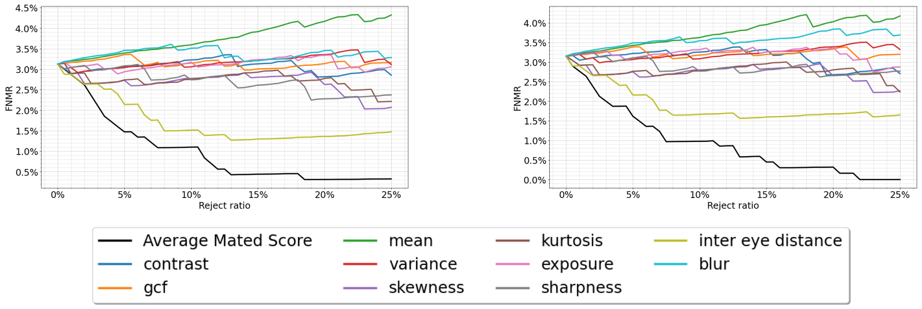
Fig. 1 shows error vs. reject curves (ERCs) for the symmetry features, Fig. 2 for the other features. An ERC shows the dependence of the FNMR at a fixed decision threshold on



(a) System 1

(b) System 2

Fig. 1: Error vs. reject curves for symmetry features



(a) System 1

(b) System 2

Fig. 2: Error vs. reject curves for other face image features

the percentage of reference and probe images excluded based on unfavourable feature values [GT07]. The ERCs vary for different decision threshold values. The thresholds were set to give an initial FNMR value of about 3%. The ERCs show that exclusion based on the values of the features with higher correlation with utility led to reduced FNMR values within the training dataset. In addition, exclusion of images based on inter-eye distance, left-right symmetry calculated as Kullback-Leibler divergence of histograms of normalized pixel values, histogram intersection of normalized pixel values, and histogram intersection of LBP filtered pixel values led to reduced FNMR values within the training dataset. Left-right symmetry calculated as histogram intersection of normalized pixel values was strongly correlated with that calculated as Kullback-Leibler divergence of histograms of normalized pixel values and, therefore, need not be used in the training process.

From the above evaluations, we selected the following features for use in the training:

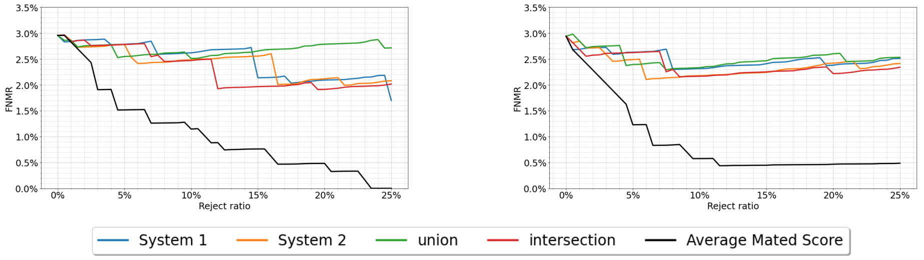
- left-right symmetry calculated as cross-entropy of histograms of normalized pixel values, as Kullback-Leibler divergence of histograms of normalized pixel values, and as histogram intersection of LBP filtered pixel values,
- from the characteristics of the capture process: contrast, global contrast factor, mean and kurtosis of pixel luminance, exposure, sharpness, inter-eye distance, and blur (i.e. all except of variance and skewness of pixel luminance).

4.2 Building utility-prediction models

For predicting the utility of face images within System 1 and System 2 and in general, random decision forests were built in Python using the training dataset. To find optimal parameter settings for the random forests, a grid search was applied, and all possible parameter combinations within the search space were verified with 3-fold cross-validation.

5 Evaluation of the accuracy of the utility-prediction model

To evaluate the accuracy of the utility-prediction models, the models trained for System 1 and System 2 and the models built using the union and the intersection of the images selected for System 1 and System 2 were used to generate quality scores for the testing data. Fig. 3 shows the ERCs with respect to these quality scores, starting at an FNMR value of about 3%. The ERCs show that exclusion of images with low quality scores lead to a reduced FNMR within the testing dataset. The model built using the intersections of images provided better results than the model built using their union did.



(a) System 1

(b) System 2

Fig. 3: Error vs. reject curves for quality scores

6 Summary and outlook

This study provided preliminary insights into features usable to predict the utility of face images within face recognition systems. Future work will expand the range of features and face recognition systems and the size and diversity of datasets explored. A next step is the evaluation of features expressing the degree of ICAO compliance. Another step is the consideration of a face image dataset containing images of all quality levels, including marginal and unacceptable quality.

Acknowledgments

This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[C$^+$09] S. Curry et al. NIST Special Database 32 – Multiple Encounter Dataset I (MEDS-I) – Data Description Document. NIST Interagency Report 7679, NIST, 2009.

[F$^+$12] M. Ferrara et al. Face Image Conformance to ISO / ICAO Standards in Machine Readable Travel Documents. *IEEE Trans. Inf. Forensics Secur.*, 7(4):1204–1213, 2012.

[G$^+$20] P. Grother et al. Ongoing Face Recognition Vendor Test (FRVT) – Part 5: Face Image Quality Assessment. Draft NIST Interagency Report, NIST, 2020. Retrieved from https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.

[GLZ07] X. Gao, S.Z. Li, and P. Zhang. Standardization of face image sample quality. In S.-W. Lee and S.Z. Li, editors, *Proc. of ICB*, 2007.

[GNH18a] P. Grother, M. Ngan, and K. Hanaoka. Ongoing Face Recognition Vendor Test (FRVT) – Part 1: Verification. Draft NIST Interagency Report, NIST, 2018. Retrieved from https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.

[GNH18b] P. Grother, M. Ngan, and K. Hanaoka. Ongoing Face Recognition Vendor Test (FRVT) – Part 2: Identification. NIST Interagency Report 8238, NIST, 2018.

[GT07] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):531–543, April 2007.

[HO$^+$19] J. Hernandez-Ortega et al. FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In *Int. Conf. on Biometrics (ICB)*, 2019.

[ICA18] Portrait quality (reference facial images for MRTD). ICAO Technical Report, 2018.

[ISO10] Information technology – Biometric sample quality – Part 5: Face image data. Technical Report ISO/IEC TR 29794-5, 2010.

[ISO15] Information technology – Biometric sample quality – Part 6: Iris image data. International Standard ISO/IEC 29794-6, 2015.

[ISO16] Information technology – Biometric sample quality – Part 1: Framework. International Standard ISO/IEC 29794-1, 2016.

[ISO17] Information technology – Biometric sample quality – Part 4: Finger image data. International Standard ISO/IEC 29794-4, 2017.

[ISO18] Information technology – Biometric application programming interface – Part 1: BioAPI specification. International Standard ISO/IEC 19784-1, 2018.

[ISO19] Information technology – Extensible biometric data interchange formats – Part 1: Framework. International Standard ISO/IEC 39794-1, 2019.

[T$^+$16] E. Tabassi et al. NFIQ 2.0 – NIST Fingerprint Image Quality. Draft NIST Interagency Report, NIST, 2016. Retrieved from https://www.nist.gov/document/nfiq2reportpdf.

[TGS11] E. Tabassi, P. Grother, and W. Salamon. Performance of iris image quality assessment algorithms. NIST Interagency Report 7820, NIST, 2011.

[TWW04] E. Tabassi, C.L. Wilson, and C.I. Watson. Fingerprint image quality. NIST Interagency Report 7151, NIST, 2004.