A. Brömme, N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb,
A. Sequeira, M. Todisco, and A. Uhl (Eds.): BIOSIG 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022

Evaluating Face Image Quality Score Fusion for Modern Deep Learning Models

Torsten Schlett[1], Christian Rathgeb[1], Juan Tapia[1], Christoph Busch[1]

Abstract: Face image quality assessment algorithms attempt to estimate the utility of face images for biometric systems, typically face recognition, since the performance of these systems can be limited by the image quality. Hand-crafted quality score fusion has previously been examined for a variety of mostly factor-specific quality assessment algorithms. This paper instead examines score fusion for various recent "monolithic" quality assessment deep learning models. The evaluation methodology is based on Error-versus-Reject-Characteristic partial-Area-Under-Curve values, which are used to quantitatively rank quality assessment configurations in a face recognition context. Mean quality score fusion configurations were found to slightly improve performance on the TinyFace database, while the tested fusion types were ineffective on the LFW database.

Keywords: Biometrics, face image quality assessment, fusion, face recognition.

1 Introduction

FIQA (Face Image Quality Assessment) is an active research area predominantly focused on automatic utility [ISO16] estimation for face recognition in the visible spectrum [Sch+21], although the term FIQA can also be applicable for other biometric scenarios. A concrete use case for face recognition and FIQA is automated border control.

For the purposes of this paper, FIQA methods have these properties, which correspond to Figure 1: The input is one 2D image, cropped and aligned via detected facial landmarks. The output is one scalar quality score. Higher quality scores imply better face recognition utility. Quality scores may or may not be restricted to the $[0, 1]$ value range, depending on the method. Either way, quality score distributions for a given set of images may differ between methods. The definition can be less strict outside of this paper [Sch+21], but these properties apply to all of the methods examined herein.

FIQA methods can be conceptually specific to certain human-interpretable "factors", such as blur/sharpness, or they can instead be "monolithic" in the sense that they do not conceptually correspond to such an isolated factor [Sch+21]. The FIQA methods examined herein are deep learning models that belong to the monolithic category. These models are specifically intended to holistically estimate utility for face recognition, whereas a single factor-specific (e.g. blur measurement) FIQA concept may only be partially linked to utility.

[1] Hochschule Darmstadt, Germany, contact: torsten.schlett@h-da.de

Fig. 1: The FIQA (Face Image Quality Assessment) process.

Quality score fusion takes a number of quality scores as input and generates a "fused" quality score as output, as illustrated in Figure 2. Each FIQA method fusion configuration can thus itself be considered as a new FIQA method.

After the following section on related work, this paper presents an evaluation methodology to quantitatively evaluate the performance of FIQA method fusion configurations, which is then applied to various recent monolithic FIQA models across multiple datasets for a number of fusion functions.

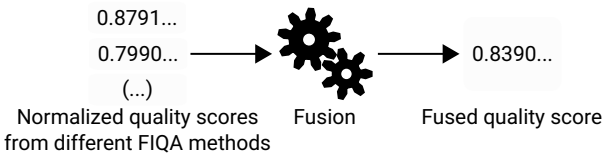


Fig. 2: The quality score fusion process.

2 Related Work

In the FIQA literature, fusion approaches are broadly categorized as "explicit", "trained", or "cascade" in [Sch+21]. Explicit approaches are hand-crafted fusion functions, without automatic database-specific fine tuning if adjustable parameters are involved (e.g. for weighted sum fusion). Trained approaches in contrast involve fine tuning, or machine learning in general (e.g. fusion via artificial neural networks). Both explicit and trained approaches process all input quality scores at once, as in Figure 2. Cascade approaches instead run the involved FIQA methods in multiple stages so that the process can be aborted (e.g. due to very low early quality scores) before all methods are run, to potentially reduce the total computational workload. The cascade approaches surveyed in [Sch+21] could alternatively be considered as a special case of explicit fusion, since the concrete cascade fusion algorithms were hand-crafted. This work is focusing on explicit (non-cascade) fusion of monolithic FIQA methods. Of the works surveyed in [Sch+21], the arguably most closely related one is [Aba+14] by Abaza *et al.*, which evaluated explicit and trained fusion configurations in a binary "good"/"bad" image quality classification scenario for factor-specific (contrast, brightness, illumination) FIQA methods. The best performance was achieved with an artificial neural network, i.e. trained fusion. Refer to [Sch+21] for various more indirectly related instances of fusion in the FIQA literature.

3 Methodology

As described in the introduction and illustrated in Figure 1, the single FIQA methods receive one face image as input. A preprocessing step first detects the facial landmarks within the image, which are then used to align and crop the image. The resulting preprocessed image also has a size (i.e. width/height in pixels) specific to each FIQA method. The quality of this preprocessed image is then assessed by the method as a scalar quality score. In this process (Figure 1), only the facial landmark detection step is allowed to fail depending on the image, in which case the images are completely excluded for the following experiments.

Since the quality score distributions can vary between FIQA methods, a normalization step is used before any fusion function is applied. The normalization parameters are derived from a set of quality scores for each FIQA method, as illustrated in Figure 3. Note that the order of images by their normalized quality scores remains identical to the order by the original quality scores[1]. The normalized quality scores are then fused via a fusion function, as illustrated in Figure 2. A fusion function takes two or more normalized quality scores stemming from different FIQA methods as input, and produces one fused quality score as output. Depending on the fusion function, the number of input quality scores may be further restricted to e.g. exactly two, or to more than two.

To evaluate the performance of the single FIQA methods and the FIQA fusions we employ the "Error-versus-Reject-Characteristic" (ERC) [GT07][Sch+21], using the "False-Non-Match-Rate" (FNMR) [ISO17] as the error, or "FNM-ERC" in short. FNM-ERC performance evaluations have been used in the (recent) FIQA literature to compare the performance of FIQA methods [Sch+21]. Since the fusion of FIQA methods effectively creates new FIQA methods, this evaluation method can be applied directly. In the context

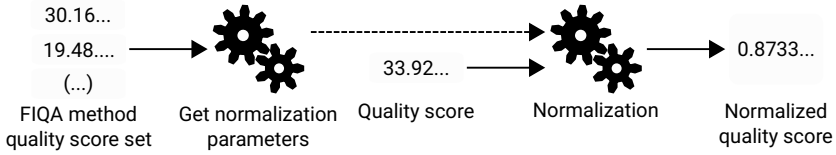


Fig. 3: The quality score normalization process. The normalization parameters are derived from a set of quality scores for the same FIQA method.

of FIQA[2], a FNM-ERC configuration involves one set of mated face image comparison pairs [ISO17], one or more face recognition model(s), each with a face recognition comparison score threshold [ISO17], and multiple FIQA methods. First, the FNMR is computed for the given images and face recognition configuration, independent of the FIQA methods. Next, for each FIQA method, the per-image quality scores are used to derive per-pair quality scores. In this work, lower quality scores imply lower face recognition utility, thus the minimum of the two images' quality scores is selected for each comparison

[1] IEEE 754 floating-point precision errors could technically change the order, but this should not be relevant for the purposes of fusion herein.

[2] The FNM-ERC and pAUC concept can also be used for other modalities.

pair. Pairs with the lowest quality[3] are then progressively "rejected"[4], and the corresponding FNMR is recomputed for the remaining pairs. Each resulting data point contains the FNMR (the error value) and the fraction of comparisons "rejected" by quality. A curve can then be plotted for each FIQA method to show how effectively the FNMR is reduced by the "rejection" of low-quality images (and thus the corresponding comparisons). Finally, "partial-Area-Under-Curve" (pAUC) [OŠB16] values can be computed to quantitatively rank FIQA methods. The pAUC values are the area-under-curve for a chosen "reject"-fraction range, e.g. $[0\%, 1\%]$.

Although pAUC values suffice to rank the FIQA methods without adjustments, the magnitude of the differences might not necessarily be clearly interpretable, since the raw pAUC values depend on the FNM-ERC starting error[5] and the chosen "reject"-fraction range. Olsen *et al.* [OŠB16] proposed to subtract the "area under theoretical best" from the (p)AUC. This refers to the area under the FNM-ERC curve for the "theoretical best case where the decrease in FNMR equals the fraction of samples[6] rejected due to quality" [OŠB16], i.e. the area under $max(0, Error - RejectFraction)$ with the "error" being the FNMR. Note that this is an approximation or lower limit of the theoretical best case, not the actual best case for the given comparison pairs, since the actual best case curve cannot be strictly monotonically decreasing[7]. Subtracting the "area under theoretical best" from the pAUC values does not change the ranking of FIQA methods, since the same value is subtracted for each pAUC "reject"-range configuration in which FIQA methods are ranked. It can however serve as an adjustment to make the pAUC values more easily interpretable, since it removes the effect of the area that cannot possibly be improved. In addition to this theoretical (hard) best case lower error bound curve, a theoretical (soft) worst case upper error bound curve can also be defined: The constant FNM-ERC starting error line approximates the average of infinite curves for random quality scores, and a FIQA method should of course preferably never increase the FNMR above this value regardless of the "reject"-fraction. Therefore, the pAUC values for FIQA methods can be made relative to the pAUC for this upper error bound line[8] for the purposes of interpretability (i.e. the ranking remains unaffected). This work uses the term "relative-pAUC" (rpAUC) for these values. The rpAUC values range from 0% (best case) to 100% or higher (soft worst case), independent of the pAUC "reject"-fraction range and independent of the FNM-ERC starting error, which should improve the interpretability across different evaluation configurations. An rpAUC value above 100% would indicate that the method is not practically useful, or "worse than random quality assessment".

[3] I.e. among the remaining pairs in each step.
[4] While common in FIQA literature [Sch+21], the term "reject" is currently contentious in this context since a different meaning is standardized [ISO17], hence the quotation marks.
[5] I.e. the FNMR for the full set of comparisons without any "rejection".
[6] Or technically the fraction of comparisons.
[7] A real FNM-ERC curve can only change by "rejecting" a non-fractional number of comparisons per data point.
[8] Both preferably adjusted by subtracting the lower bound.

4 Experiments

Two databases are used in the experiments, LFW [Hua+07] and TinyFace [CZG18]. For TinyFace, the images in "Testing_Set/Gallery_Match" and "Testing_Set/Probe" were used. Exact file duplicates were removed within each database[9], removing 184 images for TinyFace and 2 for LFW. The subsequent facial landmark detection step failed for 132 TinyFace images, excluding them from the experiments. All possible (order-independent) mated comparisons are formed between the images for each subject per database. There were no subjects with only one image, so every used image is involved in at least one mated comparison. In total, 8039/13233 images and 19894/242257 mated comparison pairs from TinyFace/LFW are used.

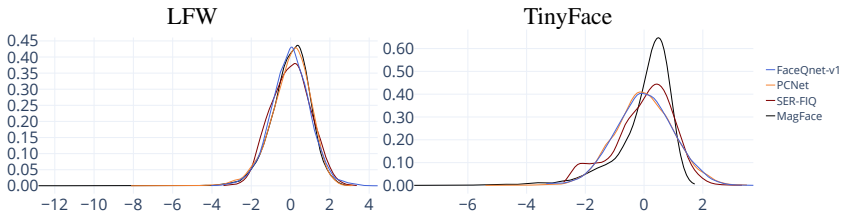
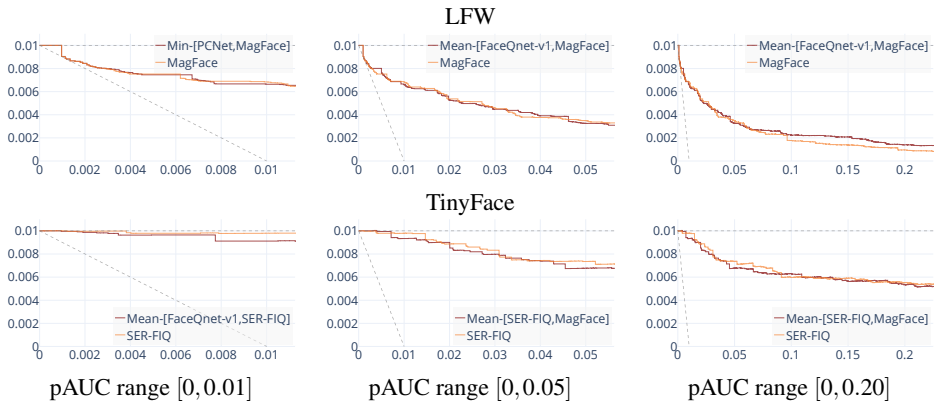


Fig. 4: Normalized quality score (horizontal axis) distribution probability (vertical axis).

Tab. 1: Truncated Z-score normalization parameters for each FIQA method per database.

LFW			TinyFace		
FIQA method	μ	σ	FIQA method	μ	σ
FaceQnet-v1	0.4072	0.1009	FaceQnet-v1	0.3140	0.0942
MagFace	28.8884	1.7604	MagFace	22.2361	1.5408
PCNet	10.0567	2.0058	PCNet	4.8270	1.6985
SER-FIQ	0.8903	0.0042	SER-FIQ	0.8739	0.0127



pAUC range $[0, 0.01]$ pAUC range $[0, 0.05]$ pAUC range $[0, 0.20]$

Fig. 5: FNM-ERC plots with 0.01 starting error with curves for the best fusion configuration results versus the best single FIQA method results for different pAUC range rankings. Horizontal axes correspond to the reject fraction, vertical axes to the FNMR. The lower dashed gray line is the "theoretical best".

[9] The first image was kept for each duplicate set, sorted by path.

Publicly available InsightFace models[10] are used for facial landmark detection and face recognition: The RetinaFace-R50 model [Den+20] is used for facial landmark detection[11], and ArcFace-R100-MS1MV2 [DGZ19] is used for face recognition. The landmark-based image preprocessing used for ArcFace is also used for the FIQA models, only the output size differs, as mentioned in section 3. Four modern FIQA models are used: FaceQnet-v1 [Her+20] (trained on VGGFace2 [Cao+18], 224×224 input image size), PCNet [XBZ20] (trained on VGGFace2 [Cao+18], 224×224 input image size), SER-FIQ [Ter+20] ("same model" variant using ArcFace, 112×112 input image size), MagFace [Men+21] ("iResNet100" backbone trained on MS1MV2 [DGZ19], 112×112 input image size). Z-score normalization is used to adjust the "raw" quality score output of a single FIQA method before fusion: $Q_n = \frac{Q_r - \mu}{\sigma}$, with Q_n/Q_r respectively being the normalized and raw quality score, and μ/σ being the mean and standard deviation of a set of raw quality scores. To mitigate the effect of imperfect normalization on the FIQA fusion evaluation, the set of all raw quality scores for each database is used to compute μ and σ per FIQA method, as listed in Table 1.

Tab. 2: FNM-ERC rpAUC rankings for different pAUC ranges with starting error 0.01 on TinyFace. The "Δ-Best" columns show the rpAUC difference to the best result. F/P/S/M stand for FaceQnet-v1/ PCNet/SER-FIQ/MagFace.

pAUC range $[0, 0.05]$				pAUC range $[0, 0.20]$			
Method	Rank	rpAUC	Δ-Best	Method	Rank	rpAUC	Δ-Best
Mean-[S,M]	1	0.8182	0.0000	Mean-[S,M]	1	0.6514	0.0000
Mean-[F,S,M]	2	0.8268	0.0087	Min-[P,S]	2	0.6649	0.0135
Mean-[F,P,S,M]	3	0.8294	0.0112	SER-FIQ	3	0.6649	0.0135
Mean-[P,S,M]	4	0.8350	0.0168	Min-[S,M]	4	0.6692	0.0178
SER-FIQ	5	0.8421	0.0239	Mean-[P,S,M]	5	0.6729	0.0215
	...				...		
MagFace	24	0.9236	0.1054	MagFace	20	0.7418	0.0904
	...				...		
PCNet	32	0.9681	0.1500	PCNet	34	0.9343	0.2829
	...				...		
FaceQnet-v1	40	1.0034	0.1852	FaceQnet-v1	41	0.9755	0.3241
	...			Max-[F,P]	42	1.0132	0.3618
Max-[F,P]	42	1.0131	0.1949				

Four fusion functions are employed, all without parameters to avoid the question of fine tuning: The minimum (min), maximum (max), arithmetic mean, and median of the input quality scores. The median fusion function is only used with three or more input quality scores, since it would be equivalent to the mean function for two inputs. The other functions use two or more inputs. All possible (order-independent) combinations of the single FIQA methods are used to form fusion configurations for each fusion function. Rankings

[10] `https://github.com/deepinsight/insightface/tree/8857513df4f5fefa4ba6f0ae5cca4de03da74bd3/model_zoo`

[11] Both rectangular face regions and facial landmark points are detected, but the landmarks are the relevant part for the face image preprocessing step.

for the pAUC ranges $[0, 0.01]$, $[0, 0.05]$, and $[0, 0.20]$ are examined for the three ERC starting errors (FNMRs) 0.01, 0.05, and 0.10.

Figure 4 shows the distribution of the normalized quality scores for each single FIQA method for the databases, with more variation being visible for TinyFace than for LFW. The comparison of the best ERC curves among the fusion configurations and single FIQA methods according to the examined pAUC ranges in Figure 5 indicates that the tested fusion approaches are more viable for TinyFace than for LFW. Results are only shown for the 0.01 starting error due to limited space, but similar results were observed for the tested 0.05 and 0.10 starting error. For LFW with 0.01 starting error, MagFace by itself provided the best results, and only one min-fusion with PCNet happened to very slightly improve performance in the $[0, 0.01]$ pAUC range. For TinyFace, SER-FIQ provided the best results for the three pAUC ranges among the single FIQA methods. Table 2 lists a subset of the more detailed results. The only fusion configuration better than SER-FIQ over all three pAUC ranges was the mean-fusion of MagFace and SER-FIQ.

5 Conclusion

Simple fusion configurations using modern monolithic FIQA models can slightly improve performance on some databases using different fusion configurations. It should be noted that fusion will naturally increase the computational workload relative to running only a single FIQA model, so constraints for practical scenarios may or may not justify the small potential performance improvements. Future work could e.g. examine more complex fusion types with tuneable parameters, fuse both factor-specific and monolithic models, or expand the evaluation to other databases.

Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883356. This text reflects only the author's views and the Commission is not liable for any use that may be made of the information contained therein.

References

[ISO16] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 29794-1:2016 Information technology - Biometric sample quality - Part 1: Framework*. International Organization for Standardization. 2016.

[Sch+21] T. Schlett et al. "Face Image Quality Assessment: A Literature Survey". In: *ACM Computing Surveys (CSUR)* (Dec. 2021). ISSN: 0360-0300.

[Aba+14] A. Abaza et al. "Design and Evaluation of Photometric Image Quality Measures for Effective Face Recognition". In: *IET Biometrics* 3.4 (Dec. 2014), pp. 314–324. ISSN: 2047-4938, 2047-4946.

[GT07] P. Grother and E. Tabassi. "Performance of Biometric Quality Measures". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 29.4 (Apr. 2007), pp. 531–543.

[ISO17] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 2382-37:2017 Information technology - Vocabulary - Part 37: Biometrics*. International Organization for Standardization. 2017.

[OŠB16] M. Olsen, V. Šmida, and C. Busch. "Finger image quality assessment features - definitions and evaluation". In: *IET Biometrics* 5.2 (June 2016), pp. 47–64.

[Hua+07] G. B. Huang et al. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. University of Massachusetts, Amherst, Oct. 2007.

[CZG18] Z. Cheng, X. Zhu, and S. Gong. "Low-Resolution Face Recognition". In: *Asian Conf. on Computer Vision (ACCV)*. 2018.

[Den+20] J. Deng et al. "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild". In: *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 5202–5211. ISBN: 978-1-72817-168-5.

[DGZ19] J. Deng, J. Guo, and S. Zafeiriou. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *Conf. on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[Her+20] J. Hernandez-Ortega et al. *Biometric Quality: Review and Application to Face Recognition with FaceQnet*. June 2020. arXiv: 2006.03298.

[Cao+18] Q. Cao et al. "VGGFace2: A Dataset for Recognising Faces across Pose and Age". In: *Intl. Conf. on Automatic Face and Gesture Recognition* (May 2018). arXiv: 1710.08092.

[XBZ20] W. Xie, J. Byrne, and A. Zisserman. "Inducing Predictive Uncertainty Estimation for Face Recognition". In: *British Machine Vision Conf. (BMVC)* (2020).

[Ter+20] P. Terhörst et al. "SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness". In: *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 5650–5659.

[Men+21] Q. Meng et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment". In: *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021, pp. 14225–14234.