SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness

Philipp Terhörst\textsuperscript{1,2}, Jan Niklas Kolf\textsuperscript{2}, Naser Damer\textsuperscript{1,2}, Florian Kirchbuchner\textsuperscript{1,2}, Arjan Kuijper\textsuperscript{1,2}

\textsuperscript{1}Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
\textsuperscript{2}Technical University of Darmstadt, Darmstadt, Germany

Email:{philipp.terhoerst, naser.damer, florian.kirchbuchner, arjan.kuijper}@igd.fraunhofer.de

Figure 1: Visualization of the proposed unsupervised face quality assessment concept. We propose using the robustness of an image representation as a quality clue. Our approach defines this robustness based on the embedding variations of random subnetworks of a given face recognition model. An image that produces small variations in the stochastic embeddings (bottom left), demonstrates high robustness (red areas on the right) and thus, high image quality. Contrary, an image that produces high variations in the stochastic embeddings (top left) coming from random subnetworks, indicates a low robustness (blue areas on the right). Therefore, it is considered as low quality.

Abstract

Face image quality is an important factor to enable high-performance face recognition systems. Face quality assessment aims at estimating the suitability of a face image for recognition. Previous work proposed supervised solutions that require artificially or human labelled quality values. However, both labelling mechanisms are error-prone as they do not rely on a clear definition of quality and may not know the best characteristics for the utilized face recognition system. Avoiding the use of inaccurate quality labels, we proposed a novel concept to measure face quality based on an arbitrary face recognition model. By determining the embedding variations generated from random subnetworks of a face model, the robustness of a sample representation and thus, its quality is estimated. The experiments are conducted in a cross-database evaluation setting on three publicly available databases. We compare our proposed solution on two face embeddings against six state-of-the-art approaches from academia and industry. The results show that our unsupervised solution outperforms all other approaches in the majority of the investigated scenarios. In contrast to previous works, the proposed solution shows a stable performance over all scenarios. Utilizing the deployed face recognition model for our face quality assessment methodology avoids the training phase completely and further outperforms all baseline approaches by a large margin. Our solution can be easily integrated into current face recognition systems and can be modified to other tasks beyond face recognition.

1. INTRODUCTION

Face images are one of the most utilized biometric modalities\textsuperscript{[41]} due to its high level of public acceptance and since it does not require an active user-participation\textsuperscript{[39]}. Under controlled conditions, current face recognition systems are able to achieve highly accurate performances...
[14]. However, some of the most relevant face recognition systems work under unconstrained environments and thus, have to deal with large variabilities that leads to significant degradation of the recognition accuracies [14]. These variabilities include image acquisition conditions (such as illumination, background, blurriness, and low resolution), factors of the face (such as pose, occlusions and expressions) [23, 22] and biases of the deployed face recognition system. Since these variabilities lead to significantly degraded recognition performances, the ability to deal with these factors needs to be addressed [19].

The performance of biometric recognition is driven by the quality of its samples [4]. Biometric sample quality is defined as the utility of a sample for the purpose of recognition [19, 31, 13, 4]. The automatic prediction of face quality (prior to matching) is beneficial for many applications. It leads to a more robust enrolment for face recognition systems. In negative identification systems, it prevents an attacker from getting access to a system by providing a low quality face image. Furthermore, it enables quality-based fusion approaches when multiple images [6] (e.g. from surveillance videos) or multiple biometric modalities are given.

Current solutions for face quality assessment require training data with quality labels coming from human perception or are derived from comparison scores. Such a quality measure is generally poorly defined. Humans may not know the best characteristics for the utilized face recognition system. On the other hand, automatic labelling based on comparison scores represents the relative performance of two samples and thus, one low-quality sample might negatively affect the quality labels of the other one.

In this work, we propose a novel unsupervised face quality assessment concept by investigating the robustness of stochastic embeddings. Our solution measures the quality of an image based on its robustness in the embedding space. Using the variations of embeddings extracted from random subnetworks of the utilized face recognition model, the representation robustness of the sample and thus, its quality is determined. Figure 1 illustrates the working principle.

We evaluated the experiments on three publicly available databases in a cross-database evaluation setting. The comparison of our approach was done on two face recognition systems against six state-of-the-art solutions: three no-reference image quality metrics, two recent face quality assessment algorithms from previous work, and one commercial off-the-shelf (COTS) face quality assessment product from industry.

The results show that the proposed solution is able to outperform all state-of-the-art solutions in most investigated scenarios. While every baseline approach shows performance instabilities in at least two scenarios, our solution shows a consistently stable performance. When using the deployed face recognition model for the proposed face quality assessment methodology, our approach outperforms all baseline by a large margin. Contrarily to previous definitions of face quality assessment [4, 23, 22, 19] that states the face quality as a utility measure of a face image for an arbitrary face recognition model, our results show that it is highly beneficial to estimate the sample quality with regard to a specific (the deployed) face recognition model.

2. Related work

Several standards have been proposed for insure face image quality by constraining the capture requirements, such as ISO/IEC 19794-5 [23] and ICAO 9303 [22]. In these standards, quality is divided into image-based qualities (such as pose, expression, illumination, occlusion) and subject-based quality measures (such as accessores). These mentioned standards influenced many face quality assessment approaches that have been proposed in the recent years. While the first solutions to face quality assessment focused on analytic image quality factors, current solutions make use of the advances in supervised learning.

Approaches based on analytic image quality factors define quality metrics for facial asymmetries [13, 10], propose vertical edge density as a quality metric to capture pose variations [42], or measured in terms of luminance distortion in comparison to a known reference image [35]. However, these approaches have to consider every possible factor manually, and since humans may not know the best characteristics for face recognition systems, more current research focus on learning-based approaches.

The transition to learning-based approaches include works that combine different analytical quality metrics with traditional machine learning approaches [31, 2, 20, 1, 8].

End-to-end learning approaches for face quality assessment were first presented in 2011. Aggarwal et al. [3] proposed an approach for predicting the face recognition performance using a multi-dimensional scaling approach to map space characterization features to genuine scores. In [43], a patch-based probabilistic image quality approach was designed that works on 2D discrete cosine transform features and trains a Gaussian model on each patch. In 2015, a rank-based learning approach was proposed by Chen et al. [5]. They define a linear quality assessment function with polynomial kernels and train weights based on a ranking loss. In [27], face images assessment was performed based on objective and relative face image qualities. While the objective quality metric refers to objective visual quality in terms of pose, alignment, blurriness, and brightness, the relative quality metric represents the degree of mismatch between training face images and a test face image. Best-Rowden and Jain [4] proposed an automatic face quality prediction approach in 2018. They proposed two methods for quality assessment of face images based on
(a) human assessments of face image quality and (b) quality values from similarity scores. Their approach is based on support vector machines applied to deeply learned representations. In 2019, Hernandez-Ortega et al. proposed Face-Qnet [19]. This solution fine-tunes a face recognition neural network to predict face qualities in a regression task. Beside image quality estimation for face recognition, quality estimation has been also developed to predict soft-biometric decision reliability based on the investigated image [38].

All previous face image quality assessment solutions require training data with artificial or manually labelled quality values. Human labelled data might transfer human bias into the quality predictions and does not take into account the potential biases of the biometric system. Moreover, humans might not know the best quality factors for a specific face recognition system. Artificially labelled quality values are created by investigating the relative performance of a face recognition system (represented by comparison scores). Consequently, the score might be heavily biased by low-quality samples.

The solution presented in this paper is based on our hypothesis that representation robustness is better suited as a quality metric, since it provides a measure for the quality of a single sample independently of others and avoids the use of misleading quality labels for training. This metric can intrinsically capture image acquisition conditions and factors of the face that are relevant for the used face recognition system. Furthermore, it is not affected by human bias, but takes into account the bias and the decision patterns of the used face embeddings.

3. Our approach

Face quality assessment aims at estimating the suitability of a face image for face recognition. The quality of a face image should indicate its expected recognition performance. In this work, we based our face image quality definition on the relative robustness of deeply learned embeddings of that image. Calculating the variations of embeddings coming from random subnetworks of a face recognition model, our solution defines the magnitude of these variations as a robustness measure, and thus, image quality. An illustration of this methodology is shown in Figure 2.

3.1. Sample-quality estimation

More formally, our proposed solution predicts the face quality $Q(I)$ of a given face image $I$ using a face recognition model $M$. The face recognition model have to be trained with dropout and aims at extracting embeddings that are well identity-separates. To make a robustness-based quality estimation of $I$, $m = 100$ stochastic embeddings are generated from the model $M$ using stochastic forward passes with different dropout patterns. The choice for $m$ is defined by the trade-off between time complexity and stability of the quality measure as described in Section 3.2. Each stochastic forward pass applies a different dropout pattern (during prediction) producing a different subnetwork of $M$. Each of these subnetworks generates different stochastic face embeddings $x_s$. These stochastic embeddings are collected in a set $X(I) = \{x_s\}_{s \in \{1,2,\ldots,m\}}$. We define the face quality

$$q(X(I)) = 2\sigma\left(-\frac{1}{m^2} \sum_{i<j} d(x_i, x_j)\right),$$

of image $I$ as the sigmoid of the negative mean euclidean distance $d(x_i, x_j)$ between all stochastic embeddings pairs $(x_i, x_j) \in X \times X$. The sigmoid function $\sigma(\cdot)$ ensures that $q \in [0,1]$. Since Gal et al. [12] proved that applying dropout repetitively on a network approximates the uncertainty of a Gaussian process [33], the euclidean distance is a suitable choice for $d(x_i, x_j)$. A greater variation in the stochastic embedding set $X$ indicate a low robustness of the representation and thus, a lower sample quality $q$. Lower variations in $X$ indicate high robustness in the embedding space and is considered as a high sample quality $q$. The quality prediction strategy is summarized in Algorithm 1.

3.2. Properties

The aim of SER-FIQ is to estimate the face image quality from the perspective of utilisation in recognition tasks, which might be different than estimating the notion of image quality. An image that produces relatively stable identity-related embeddings despite various variations (here caused by dropout) is an image with high utilisation in a recognition task, given that the recognition network training aims at being robust against intra-identity variations.
Face recognition algorithms are trained with the aim of learning robust representations to increase inter-identity separability and decrease intra-identity separability. Assuming that a face recognition network is trained with dropout and the quality of a sample correlates with its embedding robustness, different subnetworks can be created from the basic model so that they possess different dropout patterns. The agreement between the subnetworks can be used to estimate the embedding robustness, and thus the quality. If the \( m \) subnetworks produce similar outputs (high agreement), the variations over these random subnetworks (the stochastic embedding set \( X \)) are low. Consequently, the robustness of this embedding, and thus the quality of the sample, is high. Conversely, if the \( m \) subnetworks produce dissimilar representations (low agreement), the variations over the random subnetworks are high. Therefore, the robustness in the embedding space is low and the quality of the sample can be considered low as well.

Our approach has only one parameter \( m \), the number of stochastic forward passes. This parameter can be interpreted as the number of steps in a Monte-Carlo simulation and controls the stability of the quality predictions. A higher \( m \) leads to more stable quality estimates. Since the computational time \( t = \mathcal{O}(m^2) \) of our method grows quadratically with \( m \), it should not be chosen too high. However, our method can compensate for this issue and can easily run in real-time, since it is highly parallelizable and the computational effort can be greatly reduced by repeating the stochastic forward passes only through the last layer(s) of the network.

In contrast to previous work, our solution does not require quality labels for training. Furthermore, if the deployed face recognition system was trained with dropout, the same network can be used for determining the embedding robustness and therefore, the sample quality. By doing so the training phase can be completely avoided and the quality predictions further captures the decision patterns and bias of the utilized face recognition model. Therefore, we highly recommend utilizing the deployed face recognition model for the quality assessment task.

### 4. Experimental setup

#### Databases

The face quality assessment experiments were conducted on three publicly available databases chosen to have variation in quality and to prove the generalization of our approach on multiple databases. The ColorFeret database [32] consists of 14,126 high-resolution face images from 1,199 different individuals. The data possess a variety of face poses and facial expressions under well-controlled conditions. The Adience dataset [9] consists of 26,580 images from over 2,284 different subjects under unconstrained imaging conditions. Labeled Faces in the Wild (LFW) [21] contains 13,233 face images from 5749 identities. For both datasets, large variations in illumination, location, focus, bluriness, pose, and occlusion are included.

#### Evaluation metrics

To evaluate the face quality assessment performance, we follow the methodology by Grother et al. [16] using error versus reject curves. These curves show a verification error-rate over the fraction of unconsidered face images. Based on the predicted quality values, these unconsidered images are these with the lowest predicted quality and the error rate is calculated on the remaining images. Error versus reject curves indicates good quality estimation when the verification error decreases consistently when increasing the ratio of unconsidered images. In contrast to error versus quality-threshold curves, this process allows to fairly compare different algorithms for face quality assessment, since it is independent of the range of quality predictions. The curve was adapted in the approved ISO working item [25] and used in the literature [4, 37, 15].

The face verification error rates within the error versus reject curves are reported in terms of false non-match rate (FNMR) at fixed false match rate (FMR) and as equal error rate (EER). The EER equals the FMR at the threshold where \( \text{FMR} = 1 - \text{FNMR} \) and is well known as a single-value indicator of the verification performance. These error rates are specified for biometric verification evaluation in the international standard [24]. In our experiment, we report the face verification performance on three operating points to cover a wider range of potential applications. The face recognition performance is reported in terms of EER and FNMR at a FMR threshold of 0.01. The FNMR is also reported at 0.001 FMR threshold as recommended by the best practice guidelines for automated border control of Frontex [11].

#### Face recognition networks

To get face embedding from a given face image, the image is aligned, scaled, and cropped. The preprocessed image is passed to a face recognition models to extract the embeddings. In this work, we use two face recognition models, FaceNet [34] and ArcFace [7]. For FaceNet, the image is aligned, scaled, and cropped as described in [26]. To extract the embeddings, a

---

**Algorithm 1** Stochastic Embedding Robustness (SER)

**Input:** preprocessed input image \( I \), NN-model \( M \)

**Output:** quality value \( Q \) for input image \( I \)

1: procedure SER(\( I \), \( M \), \( m = 100 \))
2: \( X \) ← empty list
3: for \( i \) ← 1, ..., \( m \) do
4: \( x_i \) ← \( M \).pred(\( I \), dropout = True)
5: \( X \) = \( X \).add(\( x_i \))
6: \( Q \) ← \( q(X) \)
7: return \( Q \)

---

5654
pretrained model\textsuperscript{1} was used. For ArcFace, the image pre-
processing was done as described in [17] and a pretrained model\textsuperscript{2} provided by the authors of ArcFace is used. Both
models were trained on the MS1M database [18]. The output
size is 128 for FaceNet and 512 for ArcFace. The identity
verification is performed by comparing two embeddings
using cosine-similarity.

**On-top model preparation** To apply our quality assess-
ment methodology, a recognition model that was trained
with dropout [36] is needed. Otherwise, a model contain-
ing dropout need to added on the top of the existing model.
The direct way to apply our approach is to take a pretrained
recognition model and repeat the stochastic forward passes
only in the last layer(s) during prediction. This is even ex-
pected to reach a better performance than training a custom
network, because the verification decision, as well as the
quality estimation decision, is done in a shared embedding
space.

To demonstrate that our solution can be applied to any
arbitrary face recognition system, in our experiments we
show both approaches: (a) training a small custom network
on top of the deployed face recognition system, which we
will refer to as \textit{SER-FIQ (on-top model)}, and (b) using the
deployed model for the quality assessment, which we will
refer to as \textit{SER-FIQ (same model)}.

The structure of \textit{SER-FIQ (on-top model)} was optimized
such that its produced embeddings achieve a similar EER
on ColorFeret as that of the FaceNet embeddings. It con-
sist of five layers with \(n_{\text{emb}}/128/512/n_{\text{emb}}/n_{\text{ids}}\) di-
ensions. The two intermediate layers have 128 and 512 di-
- mensions. The last layer has the dimension equal to the
number of training identities \(n_{\text{ids}}\) and is only needed during
training. All layers contain dropout [36] with the recom-
mented dropout probability \(p_d\) = 0.5 and a tanh activation.
The training of the small custom network is done using the
AdaDelta optimizer [44] with a batchsize of 1024 over 100
epochs. Since the size of the in- and output layers (blue and
green) of the networks differs dependent on the used face
embeddings, a learning rate of \(\alpha_{\text{FN}} = 10^{-1}\) was chosen for
FaceNet and \(\alpha_{\text{AF}} = 10^{-4}\) for the higher dimensional Ar-
cFace embeddings. As the loss function, we used a simple
binary cross-entropy loss on the classification of the training
identities.

**Investigations** To investigate the generalization of face
quality assessment performance, we conduct the experi-
ments in a cross-database setting. The training is done
on ColorFeret to make the models learn variations in a
controlled environment. The testing is done on two un-
constrained datasets, Adience and LFW. The embeddings
used for the experiments are from the widely used FaceNet

To put the experiments in a meaningful setting, we eval-
uated our approach in comparison to six baseline solutions.
Three of these baselines are well-known no-reference im-
age quality metrics from the computer vision community:
Brisque [28], Niqe [29], Piqe [40]. The other three baselines
are state-of-the-art face quality assessment approaches from
academia and industry. COTS [30] is an off the shelf in-
dustry product from Neurotechnology. We further compare
our method with the two recent approaches from academia:
the face quality assessment approach presented by Best-

Training the solution presented by Best-Rowden was done
don ColorFeret following the procedure described in [4]. The
generated labels come from cosine similarity scores using
the same embeddings as in the evaluation scenario. For all
other baselines, pretrained models are utilized.

Our proposed methodology is presented in two set-
tings, the \textit{SER-FIQ (on-top model)} and the \textit{SER-FIQ (same model)}. \textit{SER-FIQ (on-top model)} demonstrates that our
unsupervised method can be applied to any face recognition
system. \textit{SER-FIQ (same model)} make use of the deployed
face recognition model for quality assessment, to show the
effect of capture its decision patterns for face quality as-
essment. In the latter case, we apply the stochastic forward
passes only between the last two layers of the deployed face
recognition network.

![Figure 3: Face quality distributions of the used databases: Adi-
ence, LFW, and ColorFeret. The quality predictions were done
using the pretrained models FaceQnet [19], COTS [30], and
the proposed SER-FIQ (same model) based on FaceNet and ArcFace.](image)

**Database face quality rating** To justify the choices of
the used databases, Figure 3 shows the face quality distri-
butions of the databases using quality estimates from four
pretrained face quality assessment models. ColorFeret was
captured under well-controlled conditions and generally
shows very high qualities. However, it contains non-frontal
head poses and for COTS and SER-FIQ (on FaceNet) (Fig-
ure 3a) this is considered as low image quality. Because
of these controlled variations, we choose ColorFeret as the
training database. Adience and LFW are unconstrained
databases and for all quality measures, most face images

\textsuperscript{1}https://github.com/davidsandberg/facenet
\textsuperscript{2}https://github.com/deepinsight/insightface
are far away from perfect quality conditions. For this reason, we choose these databases for testing.

5. Results

![Sample face images from Adience with the corresponding quality predictions from four face quality assessment methods. SER-FIQ refers to our same model approach based on ArcFace.](image)

The experiments are evaluated at three different operation points to investigate the face quality assessment performance over a wider spectrum of potential applications. Following the best practice guidelines for automated border control of the European Border and Coast Guard Agency Frontex [11], Figure 4 shows the face quality assessment performance at a FMR of 0.001. Figure 6 presents the same at a FMR of 0.01 and Figure 7 shows the face quality assessment performance at the widely-used EER. Moreover, Figure 5 shows sample images with their corresponding quality predictions. Since the statements about each tested face quality assessment approach are very similar over all experiments, we will make a discussion over each approach separately.

**No-reference image quality approaches** To understand the importance of different image quality measures for the task of face quality assessment, we evaluated three no-reference quality metrics Brisque [28], Niqe [29], Piqe [40] (all represented as dotted lines). While in some evaluation scenarios the verification error decrease when the proportion of neglected images (low quality) is increased, in most cases they lead to an increased verification error. This demonstrates that image quality alone is not suitable for generalized face quality estimation. Factors of the face (such as pose, occlusions, and expressions) and model biases are not covered by these algorithms and might play an important role for face quality assessment.

**Best-Rowden** The proposed approach from Best-Rowden and Jain [4] works well in most scenarios and reaches a top-rank performance in some minor cases (e.g. LFW with FaceNet features). However, it shows instabilities that can lead to highly wrong quality predictions. This can be observed well on the Adience dataset using FaceNet embed-
Figure 6: Face verification performance for the predicted face quality values. The curves show the effectiveness of rejecting low-quality face images in terms of FNMR at a threshold of 0.01 FMR. Figure 6a and 6b show the results for FaceNet and ArcFace embeddings on Adience. Figure 6c and 6d show the same on LFW.

- **FaceQnet** [19], proposed by Hernandez-Ortega et al., shows a suitable face quality assessment behaviour in most cases. In comparison with other face quality assessment approaches, it only shows a mediocre performance. Although FaceQnet was trained on labels coming from the same FaceNet embeddings as in our evaluation setting, it often fails in predicting well-suited quality labels on these embeddings, e.g. in Figure 4c on LFW. Also on Adience (e.g. Figure 6a and 7a), the performance plot shows a U-shape that demonstrates that the algorithm cannot distinguish well between medium and higher quality face images. Since the method is trained on the same features, these FaceNet-related instabilities might result from overfitting.

- **COTS** The industry baseline COTS [30] from Neurotechnology generally shows a good face quality assessment when the used face recognition system is based on FaceNet features. Specifically on LFW (see Figure 4c, 6c, and 7c) a small U-shape can be observed similar to FaceQnet. While it shows a good performance using FaceNet embeddings, the face quality predictions using the more recent ArcFace embeddings are of no significance (see Figure 4b, 4d, 6b, 6d, 7b, and 7d). Here, rejecting face images with low predicted face quality does not improve the face recognition performance. Since no information about the inner workflow is given, it can be assumed that their method is optimized to more traditional face embeddings, such as FaceNet. More recent embeddings, such as ArcFace, are probably intrinsically robust to the quality factors that COTS is trained on.

- **SER-FIQ (on-top model)** On the contrary to the discussed supervised methods, our proposed unsupervised solution that builds on training a small custom face recognition network shows a stable performance in all investigated scenarios (Figure 4, 6, and 7). Furthermore, our solution is always close to the top performance and outperforms all baseline approaches in the majority of the scenarios, e.g. in Figure 4a, 4d, 6a, 6b, 6d, 7a, 7b, and 7d. Our method proved to be particularly effective in combination with recent ArcFace embeddings (see Figures 6b, 6d, 7b, and 7d). The unsupervised nature of our solution seems to be a more accurate and more stable strategy.
SER-FIQ (same model) Our method that avoids training by utilizing the deployed face recognition systems is build on the hypotheses that face quality assessment should aim at estimating the sample quality of a specific face recognition model. This way it adapts to the models’ decision patterns and can predict the suitability of face sample more accurately. The effect of this adaptation can be seen clearly in nearly all evaluated cases (see Figure 4, 6, and 7). It outperforms all baseline approaches by a large margin and demonstrates an even stronger performance at small FMR (see Figures 4a, 4b, 4c, and 4d at the Frontex recommended FMR of 0.001). This demonstrates the benefit of focusing on the face quality assessment to a specific (the deployed) face recognition model.

6. Conclusion

Face quality assessment aims at predicting the suitability of face images for face recognition systems. Previous works provided supervised models for this task based on inaccurate quality labels with only limited consideration of the decision patterns of the deployed face recognition system. In this work, we solved these two gaps by proposing a novel unsupervised face quality assessment methodology that is based on a face recognition model trained with dropout. Measuring the embeddings variations generated from random subnetworks of the face recognition model, the representation robustness of a sample and thus, the sample’s quality is determined. To evaluate a generalized face quality assessment performance, the experiments were conducted using three publicly available databases in a cross-database evaluation setting. We compared our solution on two different face embeddings against six state-of-the-art approaches from academia and industry. The results showed that our proposed approach outperformed all other approaches in the majority of the investigated scenarios. It was the only solution that showed a consistently stable performance. By using the deployed face recognition model for verification and the proposed quality assessment methodology, we avoided the training phase completely and further outperformed all baseline approaches by a large margin.

Acknowledgement 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. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.
References


