Gallery-Query Protocol for Evaluating Face Image Quality Metrics

Haoyu Chen, Praneet Singh, Edward J. Delp, Amy R. Reibman
Elmore Family School of Electrical and Computer Engineering, Purdue University

***Abstract*—As more automated face recognition systems are integrated into society, face image quality estimators (QE) become important. These QEs help quantify whether an input face image contains reliable information necessary for face recognition. However, to assess the effectiveness of face QEs, it is essential to have reliable evaluation protocols. Current face QE evaluation protocols require long computation times and do not have explicit real-world implications. In this paper, we propose a novel face QE evaluation protocol named "Gallery-Query (GQ) Protocol". The GQ protocol is significantly faster in evaluating face QEs when compared to previous approaches. Furthermore, it has a very explicit real-world use case in constructing an optimal gallery set for face recognition tasks. In addition to this, we used these evaluation protocols to investigate the generalizability of face Quality Estimators (QEs) across various face recognition models, which also has implications for real-world use cases.**

***Index Terms*—task-based image quality, evaluation of quality estimators, face recognition**

I. Introduction

Quality estimation of images and videos has been an important component of realistic systems, often involved with data compression, transmission, and storage. Conventionally, quality estimation focuses on perceptual quality for human viewers. With the advancement of machine vision and automated systems, quality estimation for specific machine vision tasks become more and more necessary, and many researchers have investigated various ways of estimating image quality for certain tasks.

While methods of estimating quality are being created, it is also important to have a reliable way of knowing how effective a method is; therefore, we need evaluation protocols. In this paper, we will focus on evaluating the face quality estimators that are designed for face recognition systems.

In our previous work [1], we thoroughly examined two existing evaluation protocols for face QEs (quality estimators) — the "Error-versus-Reject (EvR)" protocol and the "Best-Middle-Worst (BMW)" protocol, and proposed additional stress tests under more extreme conditions to gain meaningful insights on the effectiveness of face QEs. However, both EvR and BMW protocol are computed based on the task of face verification (1:1 match); hence, the face QEs' effectiveness on the alternative task of face identification (1:N match) was not explicitly explored.

In this paper, we propose an alternative evaluation protocol, namely the "Gallery-Query (GQ) protocol", that focuses on the underlying task of face identification. This evaluation protocol questions the effectiveness of a face QE when used to construct the gallery set in a face identification setting; this has direct implication for a real-world use case. In addition to understanding a QE's effectiveness at the alternative task of face identification, another benefit of this evaluation protocol is that it is significantly faster when compared to the existing evaluation protocols.

Moreover, we also investigate the generalizability of face Quality Estimators (QEs) across various face recognition models. Recent learning-based face QEs are developed based on strong face recognition models, and their performance evaluation has involved using the same recognition models to calculate similarity scores for face image pairs. However, in our research, we decouple the QE from the recognition model to assess the consistency of QE performance when different models are employed for face recognition. By doing so, we gain insights into the effectiveness of face QEs across diverse recognition models. These insights are very useful in real-world scenarios where the user can utilize face QEs to determine how the performance of the recognition system will vary on a given face dataset without actually having direct access to the recognition model.

II. Background and Related Work

A. Face Recognition

In general, face recognition involves two major tasks — face verification and face identification. Face verification (1:1 match) compares two images and determines if they belong to the same identity; its application includes checkpoints like airports or Customs. Face identification (1:N match) compares a query image to a set of gallery images. Application cases often assume that the gallery images are labeled and associated to identities, and the queries are incoming unknown images that need to be identified. Many face benchmark datasets establish experiments to reflect such scenarios; for example, both Face Recognition Grand Challenge (FRGC) [2] or Face Recognition Vendor Test (FRVT) [3] considered cases where images taken in uncontrolled environments are compared to images taken in controlled environments.

Conventional face recognition algorithms such as Eigenfaces [4] and Fisherfaces [5] relied on manually designed features and techniques. Recently, deep learning has revolutionized face recognition by leveraging Convolutional Neural Networks (CNNs) to automatically learn discriminative feature representations from raw face images in an end-to-end manner. In this section, we briefly summarize the three popular learning-based face recognition models that we use in our experiments: **MobileFaceNets** [6], **QMagFace** [7], and

GhostFaceNets [8]. These recognition models encompass a wide range of deep-learning architectures and techniques. In Section IV, we evaluate the same QEs across these different models to understand how consistent a QE's performance is across multiple recognition models.

MobileFaceNets [6] are lightweight recognition models specifically designed for face recognition on mobile devices. They achieve high accuracy while maintaining small model sizes and low computational requirements, making them suitable for resource-constrained platforms. MobileFaceNets utilize Global Depthwise Convolutions and Residual Bottleneck Blocks [9] to achieve excellent recognition performance with fewer learning parameters.

QMagFace [7] is a face recognition solution that effectively combines a quality-aware comparison score and a recognition model using a magnitude-aware angular margin loss, MagFace [10]. It incorporates face image qualities specific to the model in the comparison process, enhancing recognition performance in unconstrained scenarios. Including quality awareness during training consistently improves face recognition performance. This approach also excels in challenging scenarios such as cross-pose, cross-age, or cross-quality face recognition tasks.

GhostFaceNets [8] are a family of lightweight architectures which utilize GhostNets [11] as backbones for face recognition tasks. Similar to MobileFaceNets, these models utilize carefully designed Global Depthwise Convolutions. Furthermore, these models use Squeeze and Excitation modules to enhance the discriminative power between face images and replace Rectified Linear Units (ReLU) activation with Parametric Rectified Linear Units (PReLU) to improve recognition accuracy.

B. Quality Estimators

*1) **QEs Designed For Face Matching**:* Compared to conventional perceptual quality estimators, task-specific image quality must consider additional factors. Two major factors for biometric sample quality mentioned in ISO/IEC 29794-1 standard [12] are fidelity and character. Fidelity describes the similarity between a sample and its source, while character refers to the inherent traits, features, and distinctiveness of a biometric sample. The latter is not a component in perceptual quality estimation.

Early explorations of face quality typically take the approach to directly predict performance. Dutta et al. [13] use a Bayesian model to predict the false reject rate (FRR) for query images in a face recognition system. Kim et al. [14] use a AdaBoost-based classifier to predict whether a face image can produce a correct result or not. Best-Rowden and Jain [15] use a CNN to predict either a human-labeled quality score or a proposed face quality measure computed from face matchers.

In this paper, we consider three more recent no-reference approaches for face image quality estimation. They are **FaceQNet** [16], **SER-FIQ** [17] and **SDD-FIQA** [18]. FaceQNet [16] is a supervised approach proposed to correlate the quality of a face image to its expected accuracy for face recognition. It creates a ground truth for image quality by computing the ICAO compliance level [19]. State-of-the-art deep learning frameworks are then trained to predict image quality scores.

In contrast, SER-FIQ [17] is an unsupervised approach that uses feature vector robustness to assign a quality score to face images. Here, face images are passed several times through a recognition network like ArcFace [20] with dropout enabled. Dropout introduces randomness into the feature vectors generated for a given image. The SER-FIQ quality score of a face image is the Euclidean distance between the different feature vectors. A lower variation indicates more consistency in the feature space, implying the corresponding image can be considered higher quality. Because SER-FIQ relies on an underlying network, the training details regarding this network are necessary for interpreting the results.

SDD-FIQA [18] is another unsupervised face image quality estimator. It relies on the same basic principles used to design recent learning-based face recognition systems, namely that a high-quality face image should be similar to its intra-class samples and dissimilar to its inter-class samples. To compute a ground truth quality for each face image, SDD-FIQA first computes a similarity distribution distance, using the Wasserstein Distance, between its intra-class and inter-class distributions. Similar to SER-FIQ, SDD-FIQA relies on an underlying recognition network to create intra-class and inter-class distributions that help generate quality scores. In addition, SDD-FIQA also depends on a fixed database to define the intra-class and inter-class members used to create its ground truth.

Within the context of this paper, our implementation of both SER-FIQ and SDD-FIQA use the same underlying face recognition network to generate their quality scores, specifically ArcFace [20] with a ResNet [21] backbone trained on MS1-MV2 [22] dataset. To ensure fair evaluation and create some separation between the QEs and the matching system, for the face recognition model we start by using ArcFace [20] with MobileFaceNet [6] as the backbone, trained on the MS1-MV1 dataset [23].

We also investigate each QEs' performance when used with alternative face recognition models, namely QMagFace and GhostFaceNet. The goals of these experiments are to reduce dependencies between the face matching system and the QEs, and assess whether they can be consistently useful with different face matching systems.

*2) **QEs Designed For Human Perception**:* We also consider three conventional no-reference image quality estimators: **BRISQUE** [24], **NIQE** [25], and **PIQUE** [26]. These QEs are designed to assess perceptual, not task-related quality. For example, they address the question "do people think this image has high quality?" Including these in our experiments provides a useful contrast to illustrate the effectiveness of face QEs for face matching.

Both BRISQUE and NIQE use statistical features to quantify the naturalness of an image. BRISQUE is trained with collection of natural and distorted images, whereas NIQE is solely trained with a collection of natural images. PIQUE does

not require training, but instead extracts block-based spatial features to decide whether distortion is present.

C. Evaluation Protocols for Face QEs

*1) **Error Versus Reject (EvR) Protocol**:* This protocol was introduced by [27] and has been extensively used to evaluate biometric and face quality measures [13]–[15], [17], [27]. This experiment characterizes whether a quality measure can effectively rank images by their usefulness and potential reliability to a system. In this experiment, we rank-order the face images according to their QE. We reject a fraction of face images based on each QE, and evaluate the performance on the dataset with those face images removed. Note that this is an evaluation protocol, so it may not be implemented in an actual system. However, as a protocol, it allows us to observe both how well the QE orders low-quality images (by reading from the left of the plot), as well as how well it orders high-quality images (by reading from the right of the plot). These align with potential systems goals where a user might want to prioritize high-quality (more reliable) images, or to request human review before acting on a potential recognition result using a low-quality input.

It is critical when evaluating the EvR protocol to use the exhaustive set of pairs in a dataset to obtain an accurate assessment as shown in [1]. Unfortunately, this can be prohibitive in understanding QE effectiveness in the case of larger face recognition datasets like Glint360k [28], since the number of pairs required to compute the EvR curve grows rapidly as the number of face images increases.

*2) **Best-Middle-Worst (BMW) Protocol**:* Another protocol that is commonly used to evaluate face-based QE performance is the "Best-Middle-Worst" (BMW) performance protocol. In this protocol, the dataset is partitioned into three non-overlapping sets based on the quality score; the best, middle and worst sets each contain 33% of the dataset's total images. Then, we demonstrate that the subsets (ideally) create ordered performance.

This protocol has been used in FaceQNET [16] as well as the NIST fingerprint quality project [29]. Relative to the EvR protocol, this protocol requires the comparison of fewer pairs, because it only considers pairs within each partition. However, the method of splitting the dataset into three sets without any constraints lacks any real-world use case implications.

*3) **Targeted Stress Test**:* It is also possible to construct targeted stress tests [1] to examine if a QE is robust to explainable changes to the input image. One approach is to ask the question — can a QE consistently predict when performance degradation happens across multiple perturbations and across multiple subjects? This approach exposed a weakness for FaceQNet when images are subjected to compression or noise [1].

III. Gallery-Query (GQ) Protocol

The existing face QE evaluation protocols have drawbacks, and they do not capture the underlying essence of the face recognition task. Typically, in face recognition, **gallery** images are carefully selected under controlled environments to be of high quality; a drivers license dataset is an example of a good quality gallery in real life. Then, **query** images captured in the wild are matched against these gallery images. The new face QE protocol we propose here, the **Gallery-Query (GQ) Protocol**, directly addresses the question — "if we use a face quality estimator to automatically select the gallery set, how effective is it?".

Pseudo code describing details of the steps involved to assess QE effectiveness using the GQ protocol can be found in Algorithm. 1. The first half of the figure describes an upper bound, or ideal selection of the gallery set, while the second half describes creating a gallery set using a QE.

Gallery-Query Protocol

For any Face Dataset with $\mathbf{k}$ identities, $\mathbf{n}$ images per identity

IDEAL Case:
Gallery Set = [ ], Query Set = [ ]
for $\mathbf{k}'$ in $\mathbf{k}$ face identities:
  Feature Vectors ← [ ], Similarity Score ← [ ]
  for $\mathbf{n}'$ in $\mathbf{n}$ images of $\mathbf{k}'$ face identity:
    Feature Vectors ← Face Recognizer[$\mathbf{n}'$]
  for $\mathbf{n}'$ in $\mathbf{n}$ images of $\mathbf{k}'$ face identity:
    Similarity Score ← Average Cosine Similarity[Feature Vectors[$\mathbf{n}'$,**n-1**]]
  Gallery Set ← $\mathbf{n}'$ ← max(Similarity Score[$\mathbf{n}'$])
  Query Set ← Remaining **n-1** images

QE Case:
Gallery Set = [ ], Query Set = [ ]
for $\mathbf{k}'$ in $\mathbf{k}$ face identities:
  QE Score ← [ ]
  for $\mathbf{n}'$ in $\mathbf{n}$ images of $\mathbf{k}'$ face identity:
    QE Score ← QE[$\mathbf{n}'$]
  Gallery Set ← $\mathbf{n}'$ ← max(QE Score[$\mathbf{n}'$])
  Query Set ← Remaining **n-1** images

Algorithm 1: Gallery-Query Protocol for QE evaluation.

To obtain a theoretical upper bound of performance (Ideal), we first obtain feature vectors for all the face images in a dataset using the face recognition model. For each image, we compute the average cosine similarity between the features of the given image and all the other images of the same identity. For each identity, we choose the image with highest average cosine similarity to be admitted to the gallery set, and all remaining images of the identity are added to the query set. Similarly, when using a QE to select a gallery set, we select the one image with highest quality (according to the QE) to be admitted to the gallery.

The benefits of the GQ Protocol are twofold. First, it has a real-world use case in various face identification systems (1:N matching). All previous evaluation protocols are specifically focused on the task of face verification (1:1 matching), so this protocol fills an important role by focusing on face identification (1:N matching). Second, the GQ protocol drastically decreases the computation time needed to evaluate face QEs. As seen in our previous work [1], the EvR protocol requires comparing exhaustive face pairs in order to get an accurate result, which consumes a lot of time. Although the BMW protocol reduces the number of pairs required to evaluate face

QEs, it has its own drawbacks as stated in Section II-C2. On the contrary, the GQ protocol requires fewer face pair comparisons to determine the performance of face QEs. In this protocol, only face image pairs between gallery and query sets are compared; in other words, we no longer need to compute similarity scores for exhaustive face pairs nor do we need to create non-overlapping sets.

Mathematically, for a dataset with N images from M identities, computing the EvR curve requires at least $\frac{N\times(N-1)}{2}$ similarity comparison between face pairs, and the GQ protocol requires $(N - M) \times M$ comparisons, given that we pick one image per identity to construct the gallery. The IJB-C dataset [30] that we use in this paper consists of 17,474 images of 3,464 identites. Therefore, EvR requires 152 million pair comparisons, while GQ only requires 48 million. As the face datasets get larger, the advantage of utilizing the GQ protocol becomes more prominent. For example, with a huge dataset like Glint360k [28] which consists of 17 million images from 360,232 identities, EvR requires 1.46×10^{14} pair comparisons, while GQ only requires 6.03×10^{12} comparisons. On the other hand, the BMW protocol requires $\frac{3}{2} \times (\frac{N}{3}) \times (\frac{N}{3} - 1)$ pairs, which means 4.87×10^{13} pairs in Glint360k dataset, and also requires significantly more comparisons than the GQ protocol.

IV. Experiments and Results

In this Section, we demonstrate that the GQ protocol provides valuable information about the QE effectiveness, while using fewer face image pairs compared to the EvR and BMW protocols. The GQ protocol also provides additional insights for the task of 1:N face identification, whereas the EvR and BMW protocols are tailored towards 1:1 face verification task.

For our experiments, we utilize the face recognition models from Section II-A. We consider MobileFaceNets ($0.99M$ parameters), QMagFace-iResNet-100 ($44M$ parameters), and GhostFaceNets ($7M$ parameters). All these models are trained using the ArcFace [20] loss function. We evaluate these models on the IJB-C dataset [30], specifically using only the image subset of IJB-C that contains 17,474 images of 3,464 subjects. We detect faces in these images using the MTCNN detector [31] and then align them using similarity transformations before passing them onto the recognition model. All quality scores are estimated on the aligned IJB-C face images.

Performance of face identification task (1:N) is quantified by rank-1 accuracy, which is determined by whether the closest match from the gallery is indeed the correct identity. Performance of face verification task (1:1) is quantified by an ROC curve between the True Positive Rate (**TPR**) and False Positive Rate (**FPR**). For easier interpretation, researchers often report the TPR value given a fixed FPR threshold; in this paper, we report the maximum TPR obtained when FPR is less that 1e-4, abbreviated as TPR@FPR=1e-4.

To begin with, we report the MobileFaceNets performance of QE-based gallery selections in Table I and Table II, using rank-1 accuracy for identification setting (1:N match) and TPR @ FPR=1e-4 for verification setting (1:1 match), respectively. We establish a baseline using randomly selected galleries and compute an average performance and the 95% confidence interval (middle column); if a gallery selected by a certain QE results in better performance than randomly selected galleries, we can say that the QE is somewhat effective at the task of selecting a gallery set.

	Rank-1 Accuracy (%)		
Selection	Best	Random	Worst
Upper Bound	95.07	91.46 ± 0.15	76.18
SER-FIQ	93.83		79.01
SDD-FIQA	94.07		80.79
FaceQNet	92.99		85.85
BRISQUE	92.29		87.93
PIQUE	91.22		88.53
NIQE	90.09		92.63

TABLE I: Rank-1 accuracy (identification setting) when using QEs to construct a "best" and a "worst" gallery set. Results from randomly generated gallery sets are also included as reference.

	TPR @ FPR=1e-4 (%)		
Selection	Best	Random	Worst
Upper Bound	94.75	90.91 ± 0.17	71.38
SER-FIQ	93.49		75.11
SDD-FIQA	93.77		76.84
FaceQNet	92.35		84.58
BRISQUE	91.69		86.22
PIQUE	90.87		86.99
NIQE	89.27		92.03

TABLE II: TPR @ FPR=1e-4 (verification setting) when using QEs to construct a "best" and a "worst" gallery set. Results from randomly generated gallery sets are also included as reference.

In both identification and verification settings, four QEs (SER-FIQ, SDD-FIQA, FaceQNet, BRISQUE) exhibit their effectiveness in selecting a higher quality gallery set so that the overall identification/verification performance improves, while the other two fail to show any improvements. Among these four, SER-FIQ and SDD-FIQA significantly outperform the other two.

The tables also show results on how well a QE can select low-quality images and construct a "worst" gallery set. It does not necessarily have real-world implications, but helps us understanding how effective are the QEs at ranking image usefulness in a identification/verification system.

Again, we see that both SER-FIQ and SDD-FIQA are most effective, since these QEs chose their worst gallery that is closest to the theoretical worst gallery of the dataset. FaceQNet, BRISQUE and PIQUE perform similarly, showing some effectiveness, and NIQE shows no effectiveness at all. Interestingly, we can see that SDD-FIQA show better performance than SER-FIQ at picking the best-quality gallery set, as its pick achieves closer rank-1 accuracy to the theoretical upper bound; on the other hand, SER-FIQ is better at picking the low-quality gallery set than SDD-FIQA.

We can look further into the detailed matching performance across all FPR ranges for the best gallery selected by each QE in Fig. 2. First, we observe that SER-FIQ and SDD-FIQA still

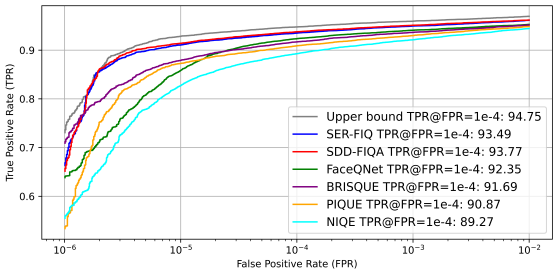


Fig. 2: Verification performance using "best" gallery sets selected by each QE, for MobileFaceNet.

remain on top, closest to the theoretical upper bound. However, we also observe that at lower FPR ranges, BRISQUE and PIQUE achieve better TPR than FaceQNet, despite the fact that FaceQNet is designed for face images. One way to think about this is that, in situations when the users want a more strict face matcher (with lower FPR), then BRISQUE might be more effective at constructing the gallery than FaceQNet.

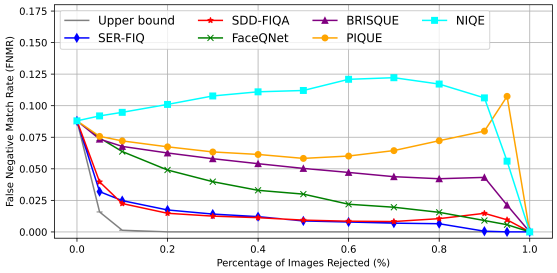
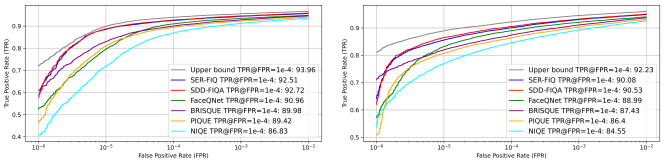


Fig. 3: EvR curves on the full set of pairs from IJB-C image dataset, using MobileFaceNet.

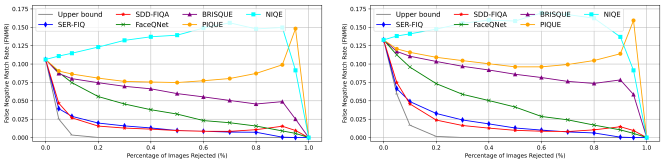
To verify that we were able to draw the same conclusions about the QEs using the GQ protocol, we also evaluate all the QEs using the commonly used EvR protocols, shown in Fig. 3. By comparing Table I and Table II regarding how GQ ranks the QEs to Fig. 3 regarding EvR's ranks, we can see that GQ obtained the same ordering of QE effectiveness using the EvR curve. SER-FIQ and SDD-FIQA are significantly better than the others; FaceQNet is worst than these two but better than perceptaul QEs; BRISQUE and PIQUE are somewhat effective, and NIQE is not effective at all. Similar observations can be seen from the BMW protocol, but these results are omitted due to space constraints; the BMW results of the same QEs can be found in [1].

This verifies that the GQ protocol can be used to assess the relative performance of the quality estimators. In addition, the GQ protocol requires significantly fewer face image pair comparisons to determine QE effectiveness, and provides extra insights for the task of face identification. For example, under the exhaustive verification setting used by EvR curve, SER-FIQ performs better than SDD-FIQA for the high-quality images (right portion of plot). However, as observed earlier, under a query-gallery identification setting, SDD-FIQA performed slightly better at selecting the best gallery. Gaining such insights helps potential users to select the most suitable QE for the application scenario it will be used on.



(a) QMagFace (b) GhostFaceNet

Fig. 4: Verification performance using "best" gallery sets selected by each QE, using QMagFace and GhostFaceNet.



(a) QMagFace (b) GhostFaceNet

Fig. 5: EvR curves on the full set of pairs from IJB-C image dataset, using QMagFace and GhostFaceNet.

Next, we explore the QEs performance on an unseen system, namely QMagFace and GhostFaceNet. This could be very useful in scenarios when the user does not have access to the exact model in a deployed system; i.e., when the model used in the QE is different from the model used in the deployed face matcher.

In this experiment, for SER-FIQ, SDD-FIQA, and FaceQNet, the trained models remain the same, and we use GQ and EvR to evaluate their effectiveness on a new system. For the three perceptual QEs, the computation is deterministic and does not rely on an underlying network.

GQ evaluation results using QMagFace and GhostFaceNet are shown in Fig. 4, and the results using MobileFaceNet are shown previously in Fig. 2. The first interesting finding is that MobileFaceNet outperforms both QMagFace and GhostFaceNet by achieving better TPR at FPR=1e-4, despite having the fewest parameters. By comparing Fig. 2 and Fig. 4, we can see that the ranking of the QEs remain the same at FPR=1e-4. In addition, the finding that BRISQUE performs better than FaceQNet at lower FPR ranges also holds true. This demonstrates that the GQ Protocol is consistent when ranking the relative effectiveness of the QEs in an unseen system.

Similarly, the EvR results in Fig. 5 show the same ranking of QEs as the original case, using MobileFaceNet. These results show that these QEs' performance are consistent in different face recognition systems.

V. Conclusion

In this paper, we proposed an alternative evaluation protocol for assessing a face QE's effectiveness, namely the Gallery-Query (GQ) protocol. As opposed to existing protocols (error-versus-rejection curve or best-middle-worst partitioning) that

focuses on the underlying task of face verification (1:1 match), the GQ protocol focuses on the task of face identification (1:N match). Consequently, the results for the GQ protocol can be interpreted as how effective the QE is when used to automatically construct a gallery set for a face identification system. Another important benefit of the GQ protocol is that it requires much less computation power than the previous evaluation protocol error-versus-rejection (EvR) curve.

We have shown that if the user wants to rank effectiveness of several face QEs, the GQ protocol obtains the same result as the more computationally-intense EvR protocol. Nevertheless, we still recognize that each evaluation protocol provides its own insights on a face QE's usefulness in certain aspect.

Having multiple slightly different evaluation protocols helps to reduce the incentive for a QE designer to optimize their QE for one protocol. Also, having multiple slightly different evaluation protocols allows a system designer the flexibility to choose a protocol that best aligns with the needs of their applications.

In addition, we have explored the QEs across multiple face recognition frameworks using different model, to explore the generalizability of the face QEs. This provides valuable information for real-world cases where the user may not have direct access to the models used for a certain deployed system, yet wants to use an off-the-shelf face QE.

REFERENCES

[1] P. Singh, H. Chen, E. J. Delp, and A. R. Reibman, "Evaluating Image Quality Estimators for Face Matching," in *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 204–209, 2022.

[2] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 947–954, 2005.

[3] P. Grother, M. Ngan, and K. Hanaoka, ""NISTIR 8271 DRAFT SUPPLEMENT: Face Recognition Vendor Test (FRVT) Part 2: Identification"," tech. rep., National Institute of Standards and Technology, 2023.

[4] M. A. Turk and A. P. Pentland, "Face Recognition using Eigenfaces," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–587, IEEE Computer Society, 1991.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[6] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for Accurate Real-time Face Verification on Mobile Devices," in *Chinese Conference on Biometric Recognition*, pp. 428–438, Springer, 2018.

[7] P. Terhörst, M. Ihlefeld, M. Huber, N. Damer, F. Kirchbuchner, K. Raja, and A. Kuijper, "QMagFace: Simple and Accurate Quality-Aware Face Recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3484–3494, 2023.

[8] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi, "GhostFaceNets: Lightweight Face Recognition Model From Cheap Operations," *IEEE Access*, vol. 11, pp. 35429–35446, 2023.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[10] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A Universal Representation for Face Recognition and Quality Assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14225–14234, 2021.

[11] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More Features from Cheap Operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1580–1589, 2020.

[12] "ISO/IEC 29794-1:2016: Information technology — biometric sample quality — part 1: Framework," standard, International Organization for Standardization, Geneva, CH, Jan. 2016.

[13] A. Dutta, R. Veldhuis, and L. Spreeuwers, "A Bayesian Model for predicting Face Recognition performance using Image Quality," *IEEE International Joint Conference on Biometrics*, 2014.

[14] H. Kim, S. H. Lee, and M. R. Yong, "Face image assessment learned with objective and relative face image qualities for improved face recognition," *IEEE International Conference on Image Processing*, pp. 4027–4031, 2015.

[15] L. Best-Rowden and A. K. Jain, "Learning face image quality from human assessments," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 3064–3077, Dec 2018.

[16] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "FaceQNet: Quality Assessment for Face Recognition based on Deep Learning," *Proceedings of the International Conference on Biometrics*, 2019.

[17] P. Terhörst, J. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness," *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[18] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang, "SDD-FIQA: Unsupervised Face Image Quality Assessment with Similarity Distribution Distance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[19] D. Maltoni, A. Franco, M. Ferrara, D. Maio, and A. Nardelli, "Biolabicao: A new benchmark to evaluate applications assessing face image compliance to ISO/IEC 19794-5 standard," *Proceedings of the IEEE International Conference on Image Processing*, pp. 41–44, 2009.

[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[22] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*, pp. 87–102, Springer, 2016.

[23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "DeepGlint: Face feature test/trillion pairs.." http:// trillionpairs.deepglint.com/overview. Accessed: 2022-05-01.

[24] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[25] A. Mittal, R. Soundararajan, and A. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, pp. 209–212, 03 2013.

[26] Venkatanath N, Praneeth D, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind Image Quality Evaluation using Perception based Features," *Twenty First National Conference on Communications (NCC)*, 02 2015.

[27] P. Grother and E. Tabassi, "Performance of Biometric Quality Measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 531–543, April 2007.

[28] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, *et al.*, "Partial FC: Training 10 Million Identities on a Single Machine," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1445–1449, 2021.

[29] E. Tabassi, M. Olsen, O. Bausinger, C. Busch, A. Figlarz, G. Fiumara, O. Henniger, J. Merkle, T. Ruhland, C. Schiel, and M. Schwaiger, "Nist fingerprint image quality 2," 2021.

[30] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus Benchmark - C: Face Dataset and Protocol," in *International Conference on Biometrics (ICB)*, pp. 158–165, 2018.

[31] J. Xiang and G. Zhu, "Joint Face Detection and Facial Expression Recognition with MTCNN," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 424–427, 2017.