Improving Performance of Facial Biometrics With Quality-Driven Dataset Filtering

Iurii Medvedev[1] and Nuno Gonçalves[1,2]
[1] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal
[2] Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal

***Abstract*— Advancements in deep learning techniques and availability of large scale face datasets led to significant performance gains in face recognition in recent years.**

Modern face recognition algorithms are trained on large-scale in-the-wild face datasets. At the same time, many facial biometric applications rely on controlled image acquisition and enrollment procedures (for instance, document security applications). That is why such face recognition approaches can demonstrate the deficiency of the performance in the target scenario (ICAO-compliant images). However, modern approaches for face image quality estimation may help to mitigate that problem.

In this work, we introduce a strategy for filtering training datasets by quality metrics and demonstrate that it can lead to performance improvements in biometric applications that rely on face image modality. We filter the main academic datasets using the proposed filtering strategy and present performance metrics.

I. INTRODUCTION

Automatic face recognition has evolved significantly in the last decade with the help of deep learning tools. Currently, face recognition models effectively learn a highly discriminative and low dimensional feature domain of the considered biometric data. Biometric samples, which are transformed into this domain to a form of a template, can be then distinguished with computationally simple similarity metrics.

Deep networks for face recognition are trained on the large datasets of labeled face images. These collections are usually based on the wild images of celebrities from the web, even for applications where such variation of the acquisition conditions is not needed. Such data choice is caused by the hard availability of ICAO-compliant (International Civil Aviation Organization[19]) images collections for research and development due to privacy and legislation issues. For instance, face images are considered as sensitive personal data by the European GDPR (General Data Protection Regulation) [9]. These aspects sometimes lead to the withdrawal of public face datasets, which now can only be obtained through redistribution. [12]. That is why collecting a dataset of ICAO-compliant face images with a size comparable to the popular wild training datasets is complicated.

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics - the University of Coimbra for the support of the project Facing. This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020.

In modern systems, the differentiating of face images by their identities is done by the comparison of their respective biometric templates with generic similarity measures (like euclidean distance or dot product). The evaluation procedures and metrics are defined by ISO/IEC 19795-1:2021 [1]. There are several scenarios of benchmarking, which characterize the face recognition method from different perspectives. For instance, a template may be used for 1-1 verification, when the comparison of testing and the trusted genuine samples is performed. This scenario can be utilized in the validation of the ID documents in the match-on-document applications [23], [24]. Some applications, where a template is enrolled into a database for performing further search requests, follow a 1-N identification scenario.

The majority of face recognition benchmarks follow the 1-1 verification scenario [17], [16], [22]. Some particular benchmarks are adapted for estimating the performance under variation of a particular characteristic like age, pose and quality [50], [51], [20].

However, the results in the particular benchmark indeed should be expanded carefully to the overall face recognition performance. Namely, focusing on of achieving the best results in novel and sophisticated benchmarks, the performance in more simple scenarios can drop (as an example for ICAO-compliant images in document security applications). This aspect is usually avoided in face recognition research, which is directed toward improving performance under unconstrained conditions.

This becomes especially important for biometric applications (for instance, for document security), which usually require a certain minimal image quality level or follow the protocols with controlled image acquisition and enrollment.

In this work, we address the above problem and propose a technique to reduce this negative effect. We introduce our novel strategy of filtering the wild face datasets for training deep networks with the use of image quality metrics. Several sample-specific loss function modifications [42], [25] were addressed to this problem, however, to the best of our knowledge, it was not yet approached from the dataset filtering perspective.

We demonstrate that careful training data filtering can help adapting the deep network for a particular scenario, namely to improve the 1-1 verification performance for ID document compliant images, while slightly sacrificing the results in wild scenarios.

As additional contribution we provide the extracted quality

metrics data for the main academic face datasets for training deep networks (CASIA-WebFace[47], VGGFace2[5], MS-Celeb-1M[12], Glint360K[2], WebFace260M[52]) and also the results of proposed filtering *(The result metadata of this work will be published in case of acceptance).*

A. Collaboration Statement

This work raises the question of correspondence between the training data and the benchmark results interpretation within the scopes of facial biometrics in specific scenarios. For instance, it can facilitate the development and evolution of face recognition applications for ID and travel documents. The broader impacts of this work could help revisit the issue of quality sampling in image pattern recognition.

II. RELATED WORK

To introduce our methodology we need to discuss recent advances in face recognition and face image quality assessment.

A. Face Recognition

The great success of deep learning tools in solving pattern recognition problems [29] was expanded to many areas of biometrics including face recognition. Among various deep learning techniques, which allow learning highly discriminative features from unconstrained images, the convolutional neural network (CNN) has become one of the most efficient tools.

Modern strategies are focused on extracting low-dimensional facial biometric template, which is based on deep features of a backbone network and maximizing the discriminative power of that template under required conditions. The recent mainstream research is usually related to various kinds of modifications of the loss function, which drives the training process. Conceptually the strategies of learning the deep network for face recognition can be divided into contrastive and classification approaches.

Contrastive methods (or metric learning methods) utilize the target similarity metric (for instance euclidean distance) to straightforwardly optimize the distance between deep features by matching face image pairs during the learning process [7], [31]. However, these methods are usually characterized by the high demands of dataset size and diversity for reliable convergence of the training process.

Another opportunity to train a face recognition network is to solve the multi-class closed-set classification problem for the existing training dataset of face images. The discriminative data of the result trained network is encapsulated in its hidden feature layer and may be further used for open-set identity discrimination purposes. These approaches usually utilize softmax loss and its variants for performing the classification [39], [38], [40]. They find their use in biometric applications related to document security. For instance, in matching live portraits to Identification Document (ID) photos [33], [34].

Basing on the softmax loss, there were introduced numerous modifications, which apply additional restrictions to the deep features. They are focused on increasing intra-class compactness and maximizing inter-class discrepancy by different means. For instance, it may be achieved by additional compacting features of intra-class samples to their mean [45], or by penalizing inter-class variance with marginal constraints in feature domain [21], [44], [8], [43], [37].

Modern approaches usually consider sample-specific strategies, which allow a better control of a feature domain for achieving higher intra-class compactness and inter-class separation. For example, sample labeling may be performed by its hardness [48], [18], additional data augmentation applied [36] or even by treating its deep features in distributional manner (by specifying sample *uncertainty*) [32].

Some of such methods indeed follow a motivation, which is similar to ours and try to adapt the deep network for better quality images using various image characteristics.

For instance, recent reports mentioned the correlation between the deep feature magnitude and sample quality [43]. MagFace [25] uses the magnitude of the feature vector during the training and explicitly associates it with the quality of samples to regularize the training process.

QualFace approach proposes to control the marginal penalization by image quality characteristics in a sample-specific way. The approach is conceptually similar to the MagFace, but the adaptation is performed with a more explicit and diverse set of quality metrics. This allows controlling the distribution of deep features during the training process.

B. Face Image Quality Assessment (FIQA)

The procedure of biometric enrollment becomes more standardized and controlled nowadays and apply a number of constrains on the quality of the result biometric samples. That is why automatic face image quality assessment have become an important area in modern facial biometrics with its specific metrics and benchmarks [11]. Most of the recent face image quality metrics are discussed in a survey by Schlett et al. [30]

The quality of a digital face image can be estimated from different perspectives. For instance, generic image quality assessment can be used in some cases. Such indicator as an image blur can help to reject some images as it is included into the list of properties for ICAO-compliance. The estimation of image blur can be performed by convolving it with a Laplacian filter and computing the variance of the result [3].

Another generic image quality assessment tool is BRISQUE (Blind/Referenceless Image Spatial QUality Evaluator), which can quantify the ”naturalness” an image with use of its statistics.

The image acquisition attributes can also serve for quality estimation purposes. For example, the face illumination is one of such important attributes for compliance with ID-Document. Zhang et al. predicted a face illumination quality with CNN [49]. The network is trained on the Face Image Illumination Quality Database (FIIQD), which is labeled with illumination quality score.

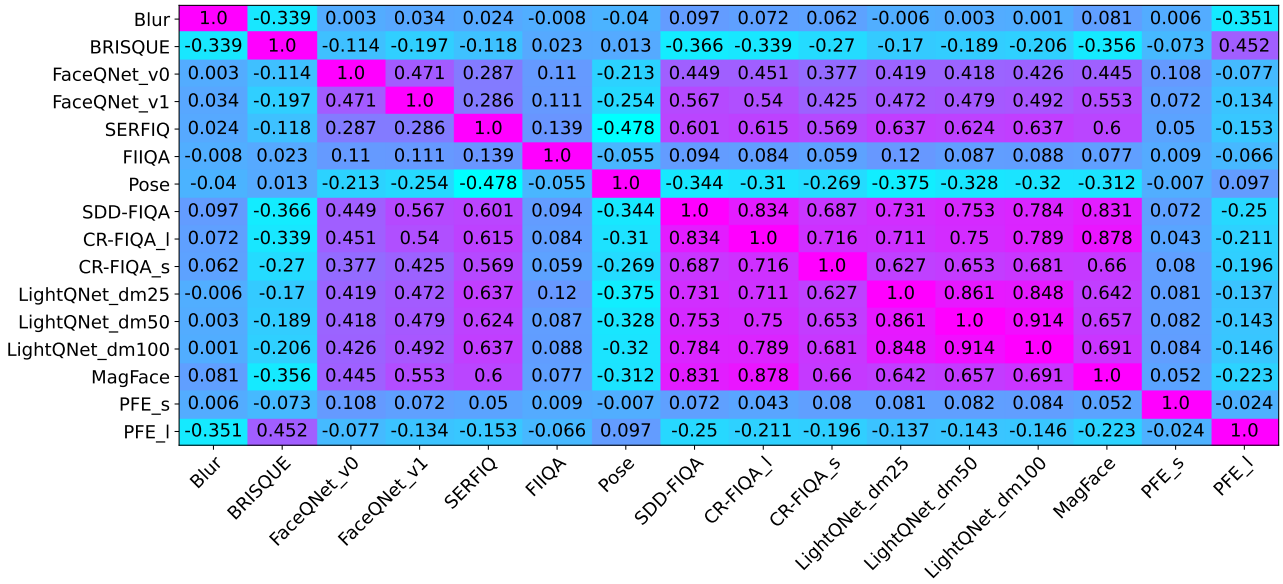


Fig. 1. The correlation of various quality metrics for the VGGFace2 dataset.

Another face specific attribute is the head pose since ICAO standards require a frontal face image for the application. Ruiz et al. [28] performed the estimation of a face pose by three angles (yaw, pitch and roll).

Some works adopted the prediction of a single quality score, which indicates compliance with the full list of ICAO requirements. For instance, FaceQnet [15] is a CNN, that is trained on the images, which are labelled with ICAO compliance scores by a third party software. FaceQnet does not rely on the face recognition performance explicitly, but its scores are highly correlated with biometric verification performance, for several off-the-shelf face recognition systems.

The above methods are heavily influenced by the quality attributes, which are related to ICAO standards and thus defined basing on human perception of an image. Several modern methods consider face image quality from the performance of a face recognition system and use its output for quality estimation. That is why they are usually criticized for their weak explainability since the resulting scores do not have a clear physical sense and are hardly interpreted from the perspective of standard ICAO requirements.

One of such methods is SER-FIQ [41] which uses the robustness of an image representation as a quality indicator. Namely, the quality of a sample is defined by the stability of its feature embeddings in different sub-networks. The similarity of the outputs for different sub-networks indicates the higher quality of the sample. To achieve that SER-FIQ relies on applying dropout during the training of a network.

Probabilistic Face Embedding (PFE) propose to encode a measure of uncertainty in the face feature embedding. In contrast to common deterministic embedding, PFE consists of two output vectors, which correspond to Gaussian mean (features) and variance (features uncertainty) [35]. In this formulation, the uncertainty vector can be associated with an implicit measure of image quality. The method is learned via the similarity scores of both genuine and impostor pairs.

SDD-FIQA [27] estimates the quality of an image by predicting quality pseudo-labels while performing face recognition. This is done by mapping the inter-class and intra-class similarity scores to the pseudo-labels by using a distribution distance metric.

LightQnet approach is also based on the pairwise binary quality pseudo-label generated by the face similarity. It treats the quality assessment following a binary classification problem, focusing on difficult samples near the classification boundary. LightQnet is accompanied by a branch-based quality distillation method and achieves state-of-the-art performance maintaining the small model size and low computation complexity.

PCNet [46] proposes a scheme for learning predictive confidence which is associated with the quality of the samples. The training of PCNet is performed with the use of pairwise verification scores, which are then disentangled to single images.

CR-FIQA method learns the quality estimation of a face image implicitly by predicting its relative classifiability [4]. During training, the feature representation of a sample is optimized in angular space with respect to its class center and the nearest negative class center.

III. METHODOLOGY

Following our initial motivation, we intend to adapt the deep network to the scenario of ICAO-compliant face images by filtering the training dataset with quality data. However, suitability for using in ID Documents applications is not guaranteed by using a particular single metric. Some metrics rely on generic image properties, while others define a sample quality by the recognition system response, and usually does not carry a clear physical sense related to a particular

image characteristic. This can be reduced by accumulating data from a number of quality indicators, which describe face image samples from different perspectives.

That is why we focus on the joint usage of a number of various quality metrics to benefit from the combined usage of various quality scores. We aim to correlate those metrics with the human acceptance indicator. Summing up we are looking for the adaptive thresholding procedure, which evenly accounts for the various quality metrics and allows to filter the face dataset by criteria of *suitability for using in ID Documents*.

A. Quality Sampling

In this work we employ the following list of quality metrics for further filtering: Blur[3], BRISQUE [26], FaceQNet(v0 [15] and v1 [14]), SERFIQ [41], FIIQA [49], Pose [28], SDD-FIQA [27], CR-FIQA (***s*** - ResNet-50 trained on CASIA-Webface and ***l*** - ResNet-100 trained on MS-Celeb-1M) [4], LightQNet (3 models with different size) [6], MagFace [25], PFE (***s*** - trained on CASIA-Webface and ***l*** - trained on MS-Celeb-1M) [35].

The pose is indeed estimated by three angles (yaw, pitch, roll). In our work, we use the mean of those angles as the quality indicator. MagFace provides the quality measure as the mean of magnitudes of its deep face features. PFE is implicitly related to the face image quality since it indeed indicates to *uncertainty* of deep features. In our work, we use the harmonic mean of the uncertainty vector as a quality measure.

We extract the above metrics for the mentioned wild face datasets and estimate the correlation between them by computing Pearson product-moment correlation coefficients (see the correlation for VGGFace2 in Fig. 1) [10].

We observe that metrics, which rely on face recognition performance are highly correlated. The only exception is PFE, which is even weakly correlated between its versions. Also, these metrics are better correlated with face pose, and naturalness rather than with the blur and illumination.

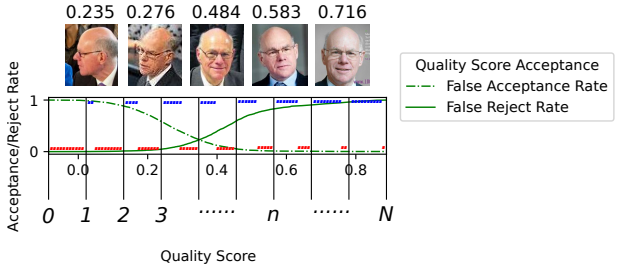


Fig. 2. Schematic of samples labeling by their quality score. Blue(true)/red(false) points illustrates the accepted/rejected samples, which are demonstrated to the user. FaceQNet_v1 is taken as example.

B. Joint Quality Filtering

Dataset filtering requires the definition of a threshold value for each quality metric. In order to do that it is necessary to map the full dynamic range of each metric with acceptance score rates. We achieve this with the following semiautomatic

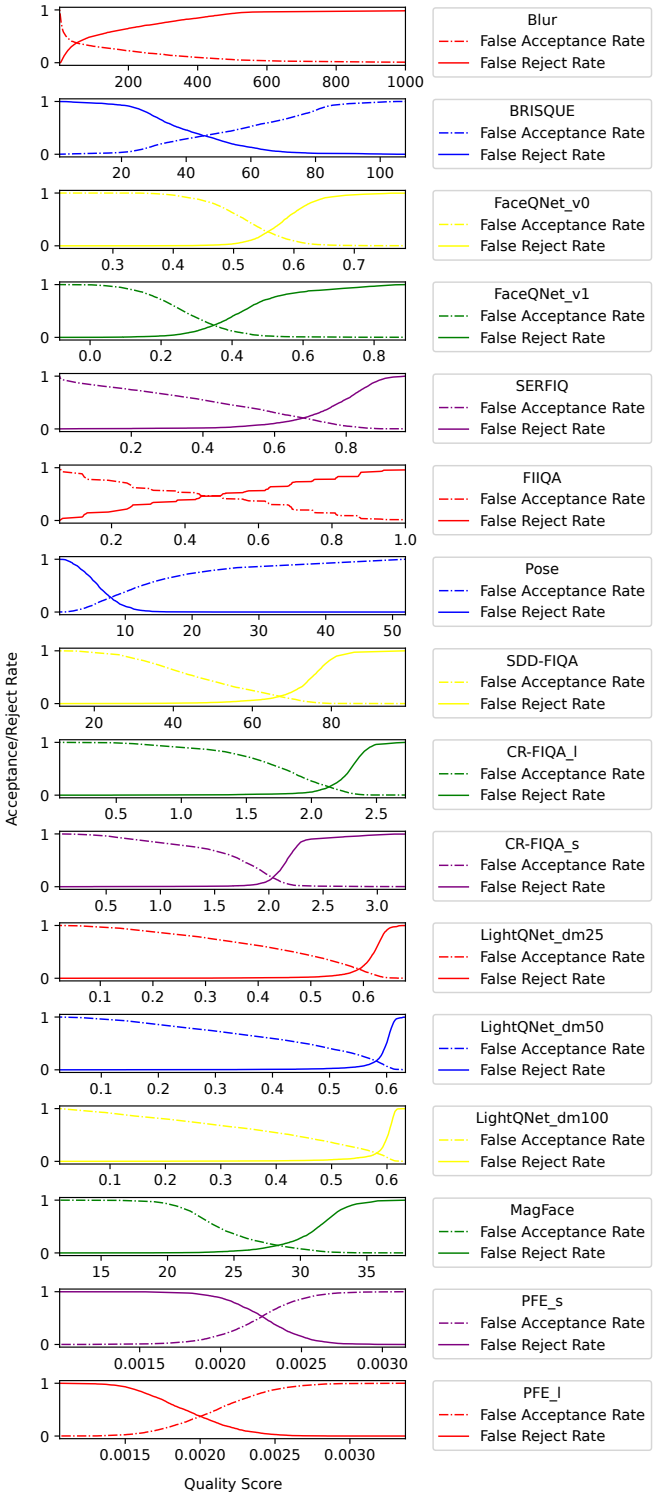


Fig. 3. Crossover Error Rates for various quality scores of images from the VGGFace2 dataset. The filtering parameters are the following: $N = 21$, $P = 50$. In total $\sim$ 5k samples were labeled for plotting the curves.

technique (see Fig. 2). The dynamic range of each quality score is split into several equal segments. Then we define a minimum number of samples P that is required to fill each segment with the mapped samples. Next, we randomly select samples to be demonstrated to an instructed human participant(informed about ICAO face image requirements), who labels the sample manually by interacting with a keyboard. The instruction for the participant simply requires to ask a question in the following form for each received face image: "*Is this face image suitable for an ID-Document? (The following characteristics can be neglected: background, straight look at the camera, inexpressive emotions)*". The procedure is repeated until the segments are filled with at least P samples. We also skip images, that belong to filled segments, which significantly accelerate the process and allow to complete it in a reasonable time. Such process allows to correlate the heuristics of human choice with the the quality metrics that can be obtained automatically. By using multiple quality metrics the better generalization of that heuristics can be achieved.

This technique allows to gather the image acceptance in full quality metric dynamic range and to build a Crossover Error Rates (CER) plot, which is the joint dependency of False Acceptance Rates (FAR), the False Reject Rates (FRR) from the quality score threshold (see Fig. 3) These curves were plotted basing on a labeled collection of $\sim$ 5k samples.

Further filtering should be performed evenly along with the quality metrics. We achieve this from the perspective of equal limiting of an undesirable effect - false discarding images, which can be estimated by False Reject Rate (see Fig.3) and should be minimized.

That is why to define the level of filtering we set the limiting value of FRR for the quality scores to constrain the negative effect of removing samples with acceptable quality. The score value which corresponds to the defined FRR acts as a threshold for filtering. In this perspective, FRR becomes the only global filtering coefficient in our technique.

$$\{X'\} = \{X_i \mid q_j > th_j, \forall q_j \in Q_i\} \tag{1}$$

where X and X' are original and filtered datasets, q_j is a score of j^{th} quality metric, th_j is a threshold value for j^{th} quality metric and a defined FRR and $Q_i \in \mathbb{R}^k$ (where k is a number of quality scores) is a quality scores set of i^{th} image in the dataset.

With this formulation, we generate several training protocols with different FRR values and make a number of experiments in Section IV.

IV. EXPERIMENTS

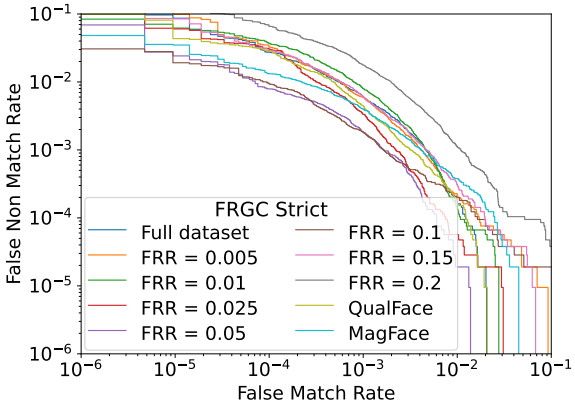
To analyse the impact of our strategy, we train the deep networks via identity classification task on various filtered datasets and compare their performance in different benchmarks for several scenarios.

As a source dataset for our experiments, we choose the VGGFace2 (only its train part). Train image are aligned by method in [8]. Due to the large initial "image per class" value, it is well suited to demonstrate the effect of our technique. We filter this dataset with a defined set of FRR values $[0.0, 0.005, 0.01, 0.025, 0.05, 0.1, 0.15, 0.2]$ (see Table I).

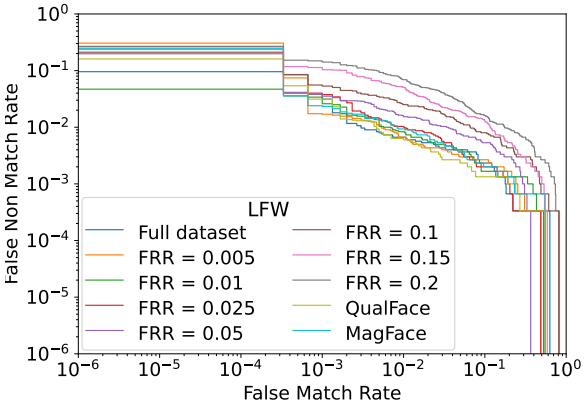
TABLE I

PARAMETERS OF THE FILTERED VGGFACE2 DATASETS.

Filtering Level	Number of Images	Number of Classes	Images per Class
Full Dataset	3074k	8631	356
FRR = 0.005	2413k	8630	279
FRR = 0.01	2155k	8630	249
FRR = 0.025	1634k	8630	189
FRR = 0.05	1158k	8628	134
FRR = 0.1	667k	8586	77
FRR = 0.15	401k	8492	47
FRR = 0.2	235k	8234	28



(a)

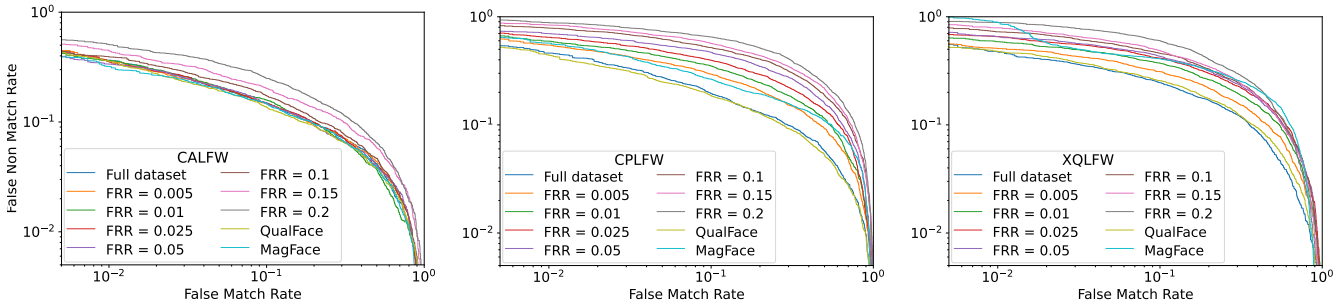


(b)

Fig. 4. DET curves of ResNet-50 trained on various filtered VGGFace2 tested on *Strict* benchmark(a) and *LFW* benchmark(b)

Deep learning classification is usually performed with the softmax loss function, which now serves as the basis for most of the recently developed loss functions in face recognition. The softmax loss is usually formulated as follows:

$$L_{Softmax} = \frac{1}{N}\sum_i -\log\left(\frac{e^{f_{y_i}}}{\sum_j^C e^{f_{y_j}}}\right). \tag{2}$$



(a) (b) (c)

Fig. 5. DET curves of ResNet-50 trained on various filtered VGGFace2 tested on *CALFW* benchmark(b), *CPLFW* benchmark(b) and *XQLFW* benchmark(c)

TABLE II

PERFORMANCE METRICS OF RESNET-50 TRAINED ON VARIOUS CONFIGURATIONS OF VGGFACE2 DATASET.

Filtering Level	FNMR@FMR = α, EER, AUC of ROC									
	Protocol Strict					LFW				
	α=10^{-3}	α=10^{-4}	α=10^{-5}	EER	AUC	α=10^{-2}	α=10^{-3}	α=10^{-4}	EER	AUC
Full Dataset	0.00604	0.02733	0.08667	0.0024	0.999972	0.00667	0.03133	**0.09567**	**0.0073**	0.99919
FRR=0.005	0.00591	0.03254	0.09892	0.0022	0.999969	0.00667	**0.01700**	0.30600	0.0080	0.99909
FRR=0.01	0.00804	0.03686	0.06207	0.0026	0.999965	0.00833	0.02600	0.04700	0.0086	0.99906
FRR=0.025	0.00336	0.03024	0.06031	0.0016	0.999983	0.01000	0.03767	0.26767	0.0100	0.99909
FRR=0.05	**0.00182**	**0.00790**	0.02417	**0.0013**	**0.999992**	0.01533	0.03533	0.19833	0.0146	0.99831
FRR=0.1	**0.00182**	0.00931	**0.01917**	**0.0013**	0.999986	0.02500	0.05333	0.21100	0.0203	0.99687
FRR=0.15	0.00596	0.02860	0.08473	0.0024	0.999968	0.05067	0.10367	0.23567	0.0256	0.99579
FRR=0.2	0.01773	0.06574	0.11455	0.0041	0.999904	0.06933	0.14133	0.26433	0.0360	0.99293
QualFace	0.00443	0.02807	0.04340	0.0019	0.999979	**0.00600**	0.03133	0.16033	**0.0073**	**0.99931**
MagFace	0.00399	0.01318	0.03496	0.0020	0.999978	0.00933	0.02333	0.24733	0.0093	0.99898

Here the C is the number of identities (classes), N is the number of samples in a batch, y_i is the numerical index of the class of the $i-th$ sample and f_{y_j} is the $y_j - th$ element of the logits vector $\mathbf{f}$ in the final layer.

L2 normalization of the weights $\mathbf{w_j}$ and deep features $\mathbf{x_i}$ is usually performed to constrain them on the hypersphere in $\mathbb{R}^d$ space (where d is the size of $\mathbf{f}$). Then f_{y_j} can be represented as: $f_{y_j} = w_j^T x_i = \cos(\theta_j)$ and biometric templates of samples are conveniently discriminated with angular similarity metric.

Clean softmax is usually modified with additional penalization, which applies constraints on feature distributions of classes. For instance, to obtain ArcFace loss, it can be reformulated with the feature normalization and additional angular marginal penalization parameter m to the positive logit:

$$L_{ArcFace} = \frac{1}{N}\sum_i -\log\left(\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}\right) \tag{3}$$

The popularity of this formulation is caused by the high discriminability and class compactness of the result deep features and at the same time, robust convergence [8], [52]. In our experiments in this section we use the ArcFace and maintain its margin $m = 0.5$, and the scaling constant $s = 30$.

A. Benchmarking

The purpose of the benchmarking in our work is to demonstrate the performance difference by the relative comparison of the results under various conditions, namely in the Wild scenario and in the ICAO compliant scenario. For the last one, we employ the *Strict* protocol from [42], which is constructed from the FRGC_V2 dataset that is manually filtered by the property of ID Documents compliance. We consider this protocol as the target one and aim to improve its performance in it with our quality filtering strategy. For the Wild Scenario, we use a well-known LFW benchmark [17], [16].

To analyse the performance difference under variation of age, pose, and quality, we also employ a set of benchmarks CALFW [51], CPLFW [50] and XQLFW [20], which are constructed similarly to LFW.

We report the performance by FNMR@FMR = α and also include additional metrics such as Equal Error Rate (EER) of Detection Error Trade-off (DET) and Area Under Curve (AUC) of Receiver Operating Characteristic (ROC).

B. Training Settings

In all our experiments we use the following training settings. The backbone network architecture is the widely used ResNet-50 [13] with the input image size 299×299. The backbone is followed by the pooling, flattening dense feature, batch normalization and dense classification layers.

The feature layer contains 512 nodes. For additional regularization, we add the dropout layer between the flattened and dense feature layers with the dropout rate 0.25.

The networks are initialized with the ImageNet weights and then trained for 10 epochs with linear learning rate scheduling, which started at 0.01 in the beginning and decayed to 0.00001 in the end. The momentum value is equal to 0.9. The batch size is set to 36. We also apply the weight decay 0.0005.

C. Learning On Filtered Datasets

With the above settings, we train a set of ResNet-50 models with ArcFace loss on the listed configurations of filtered VGGFace2 (see Table I). We also train QualFace (which is controlled by the blur quality metric) and a MagFace models to perform the comparison with similar approaches. In our work these strategies are based on sample specific adaptation of the ArcFace loss.

The results on the *Strict* benchmark demonstrate that careful dataset filtering with our strategy allows to outperform the baseline model (which is trained on the full dataset) and the adaptive strategies (QualFace and MagFace) on several levels. However, the high filtering levels lead to an overall performance drop, which is related to the insufficient dataset size. We observe that the optimal evaluation metrics are achieved at the level of filtering, which corresponds to FRR=0.05 (see Table II).

Regarding the LFW benchmark, the dataset filtering does not affect the performance significantly up until the filtering level at $FRR = 0.025$. After this value evaluation metrics reduces evenly with the increase of the filtering level (see Fig. 4b). Adaptive strategies (QualFace and MagFace) as expected demonstrate the superiority over the baseline ArcFace for the *Strict* scenario and performs on par with the ArcFace in the *Wild (LFW)* scenario.

Our results on CALFW, CPLFW and XQLFW demonstrate the dataset filtering effect against various face attributes variation (namely age, pose, quality). We observe that the variation of age starts to affect the performance only at the high levels of filtering ($FRR > 0.1$) when pose and quality affect the curves evenly with the increase of the filtering level (see Fig. 5). This is the expected result of our strategy since we reduce variations of pose and quality in our training data. QualFace behaves similarly to the baseline ArcFace model in these tests. MagFace demonstrates the significant drop of the performance under the variations of pose and quality.

V. Conclusion

In this work, we revisit the scenarios of benchmarking a face recognition deep network and show that if it is trained on sophisticated and diverse data, the network can lose the performance in simple (but usually the target) scenarios. We raise this question, particularly for facial biometrics and propose a solution for reducing such negative effects by using quality-driven dataset filtering.

We demonstrate that such careful filtering (removing the worst samples) of training data with quality metrics can help to adapt the deep network for a particular scenario, for instance, to improve the 1-1 verification performance for ID document compliant images, while slightly sacrificing the results in wild scenarios. We propose our novel strategy of filtering the wild face datasets by a number of various quality metrics. These results may be important for biometric applications, which deal with ICAO-compliant face images (namely ones, which are related to document security).

We also provide the extracted quality metrics data for the main academic face datasets for training deep networks (CASIA-WebFace, VGGFace2, MS-Celeb-1M, Glint360K, WebFace260M) and the results of our filtering strategy.

VI. Acknowledgements

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics - the University of Coimbra for the support of the project Facing. This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020.

References

[1] ISO/IEC 19795-1:2021. Information technology — Biometric performance testing and reporting — Part 1: Principles and framework. ISO/IEC JTC 1/SC 37 Biometrics, 05 2021.

[2] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu. Partial FC: Training 10 Million Identities on a Single Machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1445–1449, October 2021.

[3] R. Bansal, G. Raj, and T. Choudhury. Blur image detection using Laplacian operator and Open-CV. In *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, pages 63–67, 2016.

[4] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer. CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability, 2021.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[6] K. Chen, T. Yi, and Q. Lv. LightQNet: Lightweight Deep Face Quality Assessment for Risk-Controlled Face Recognition. *IEEE Signal Processing Letters*, 28:1878–1882, 2021.

[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:539–546 vol. 1, 2005.

[8] 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)*, pages 4685–4694, 2019.

[9] European Commission. General Data Protection Regulation. Official Journal of the European Union, 2016.

[10] D. Freedman, R. Pisani, and R. Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[11] P. Grother, A. Hom, M. Ngan, and K. Hanaoka. Ongoing Face Recognition Vendor Test (FRVT). Part 5: Face Image Quality Assessment, 2021.

[12] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Proceedings of ECCV*, volume 9907, pages 87–102, 10 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[14] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay. Biometric quality: Review and application to face recognition with faceqnet. *ArXiv*, abs/2006.03298, 2020.

[15] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[16] G. B. Huang and E. Learned-Miller. Labeled Faces in the Wild: Updates and New Reporting Procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[17] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[18] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] ISO/IEC JTC1 SC17 WG3. Portrait Quality - Reference Facial Images For MRTD. https://www.icao.int/Security/FAL/TRIP/Documents/TR - Portrait Quality v1.0.pdf, 2018. Version: 1.0 Date – 2018-04, Accessed: 2021-04-04.

[20] M. Knoche, S. Hoermann, and G. Rigoll. Cross-Quality LFW: A Database for Analyzing Cross- Resolution Image Face Recognition in Unconstrained Environments. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5, 2021.

[21] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017.

[22] 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 *2018 International Conference on Biometrics (ICB)*, pages 158–165, 2018.

[23] I. Medvedev, N. Gonçalves, and L. Cruz. Biometric System for Mobile Validation of ID And Travel Documents. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2020.

[24] I. Medvedev, F. Shadmand, L. Cruz, and N. Gonçalves. Towards facial biometrics for id document validation in mobile devices. *Applied Sciences*, 11(13), 2021.

[25] 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 (CVPR)*, pages 14225–14234, June 2021.

[26] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, June 2021.

[28] N. Ruiz, E. Chong, and J. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 06 2018.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[30] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch. Face Image Quality Assessment: A Literature Survey. *ACM Computing Surveys (CSUR)*, 2022.

[31] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference CVPR*, pages 815–823, 2015.

[32] Y. Shi and A. Jain. Probabilistic Face Embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910, 2019.

[33] Y. Shi and A. K. Jain. DocFace: Matching ID Document Photos to Selfies*. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2018.

[34] Y. Shi and A. K. Jain. DocFace+: ID Document to Selfie Matching. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1:56–67, 2019.

[35] Y. Shi, A. K. Jain, and N. Kalka. Probabilistic Face Embeddings. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910, 2019.

[36] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. Jain. Towards Universal Representation Learning for Deep Face Recognition. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6816–6825, 06 2020.

[37] J. Sun, W. Yang, J. Xue, and Q. Liao. An Equalized Margin Loss for Face Recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2020.

[38] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. In *NIPS*, 2014.

[39] Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation from Predicting 10,000 Classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[40] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2892–2900, 2015.

[41] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5650–5659, 2020.

[42] J. Tremoço, I. Medvedev, and N. Gonçalves. QualFace: Adapting Deep Learning Face Recognition for ID and Travel Documents with Quality Assessment. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, 2021.

[43] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive Margin Softmax for Face Verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[44] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing.

[46] W. Xie, J. Byrne, and A. Zisserman. Inducing Predictive Uncertainty Estimation for Face Verification. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[47] D. Yi, Z. Lei, S. Liao, and S. Li. Learning face representation from scratch. *ArXiv*, abs/1411.7923, 2014.

[48] D. Zeng, H. Shi, H. Du, J. Wang, Z. Lei, and T. Mei. NPCFace: A Negative-Positive Cooperation Supervision for Training Large-scale Face Recognition. *CoRR*, abs/2007.10172, 2020.

[49] L. Zhang, L. Zhang, and L. Li. Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model. *Lecture Notes in Computer Science*, 10636 LNCS:583–593, 2017.

[50] T. Zheng and W. Deng. Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018.

[51] T. Zheng, W. Deng, and J. Hu. Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments. *CoRR*, abs/1708.08197, 2017.

[52] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou. WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502, June 2021.