IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 May 2022, accepted 6 July 2022, date of publication 15 July 2022, date of current version 28 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3191463

RESEARCH ARTICLE

Face Image Quality Assessment in Electronic ID Documents

ANNALISA FRANCO[1], ANTONIO MAGNANI[1], DAVIDE MALTONI[1], (Senior Member, IEEE), DARIO MAIO[1], (Life Member, IEEE), LEONARDO ODORISIO[2], AND ANDREA DE MARIA[2]

[1]Department of Computer Science and Engineering, University of Bologna, 47521 Cesena, Italy
[2]Istituto Poligrafico e Zecca dello Stato, 00138 Roma, Italy

Corresponding author: Annalisa Franco (annalisa.franco@unibo.it)

This work is part of iMARS. The project received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 883356. Disclaimer: this text reflects only the author's views, and the Commission is not liable for any use that may be made of the information contained therein.

ABSTRACT Face image quality estimation is still an open issue since, unlike what happens for other biometric characteristics such as fingerprints, no standard definitions are available yet. The problem is even harder when the focus of quality assessment is the context of electronic ID documents for which, according to the provisions of ISO/IEC 39794-5, a quality value will be stored in the future in dedicated quality blocks. In case of high-quality images, the general indicators available in the literature tend to assign a flat score that does not contribute to provide significant information. This work documents a study aimed at defining a quality score indicator for high-quality images, able to predict the utility of a specific image for face verification purposes. A quality regressor is proposed, based on a large set of quality elements including ISO/ICAO controls and quality scores provided by deep-learning based solutions. A number of experiments highlight specific issues to be addressed in this scenario and confirm the effectiveness of the proposed approach with different face recognition systems.

INDEX TERMS Face image quality assessment, electronic ID documents, face recognition, ISO/IEC 39794-5.

I. INTRODUCTION

Face has been chosen as the primary globally interoperable biometric characteristic for automated identity verification in electronic ID Documents and it plays an essential role in many governmental applications. The performance of Face Recognition Systems (FRSs) heavily depends on the quality of the biometric samples used in the enrollment and verification process, as confirmed by the diversified results measured in large-scale evaluation campaigns [1] in different evaluation scenarios (e.g. VISA vs. Wild). The ID document application falls of course in the branch of controlled scenarios with cooperative users, where the quality of biometric samples is good, especially concerning the enrollment image stored in the document for which specific requirements have been clearly defined in the ISO/IEC 19794-5 standard [2], successively modified by ISO/IEC 39794-5 [3]. Despite of the general good quality of the images considered, measuring and storing the specific sample image quality may be very useful for multiple purposes:

- to guide image acquisition, providing live feedback on the images acquired by the camera;
- to monitor the document issuance process to guarantee a constant quality level for the circulating documents. This ensures that the biometric samples will be later useful to successfully verify the identity of the document's owner.
- to improve the face verification process, possibly exploiting the quality of the document sample.

The next ISO/IEC 39794-5 [3], in fact, provides for the possibility of storing information about biometric samples' quality in dedicated quality blocks in the new generation passports. The importance of biometric samples' quality motivates the huge number of studies on this topic for which generic image quality measures, designed for traditional image processing tasks, are not sufficiently explanatory

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

and ad hoc quality indicators are generally considered more effective.

While for some other biometric traits standardized quality measures have been established (e.g. NFIQ 2 [4] for fingerprints), for face the standardization process is still ongoing (the current ISO working draft is publicly available [5]). This lack motivates the current study, in combination with the need of defining a significant quality measure for controlled images. Many of the existing studies, in fact, focus on more general application scenarios (please refer to [6] for a recent survey), where the image acquisition is not strictly controlled or even uncontrolled; quality indicators defined in these contexts risk are likely to provide quality scores which are ''flat'' (in a small range of values) or not highly correlated to the utility of a specific sample for face verification. Let's consider, for instance, FaceQNet [7] which is a state-of-the-art technique very effective and widely used in the literature; despite of its general very good performance, our experiments show that this indicator is not sufficient to assign a meaningful quality score to high-quality images. This is comprehensible if we consider that FaceQNet has been trained taking ICAO compliant images as the target for top-quality images. Image quality assessment applied to ID documents is certainly more complex with respect to the case of general quality images and poses many questions related to the development of an effective indicator as well as to its evaluation. Moreover, due to the specific application scenario, quality assessment should present some degree of *explainability*, meaning that the quality score assigned to an image should be justified in terms of specific and human-understandable image features (e.g. blurring, wrong illumination) in order to allow, if needed and possible, the human intervention to rectify the issues highlighted. This explainability property is of course more difficult to achieve when only deep-learning based solutions are adopted due to the well-known limited knowledge about the processes leading a deep network to produce a given result.

Before getting into the details of the quality measure defined in this work, it's important to introduce the general concept of quality which actually refers to different definitions. In particular, the ISO/IEC 29794-1 [8] definition encompasses three different aspects:

- **character**: refers to attributes associated to a biometric characteristic that cannot be controlled during acquisition (e.g. scars in fingerprints);
- **fidelity**: measures the degree of similarity of a biometric sample to its source biometric characteristic;
- **utility**: relates to the adequacy of a biometric sample to accomplish the desired identity verification task.

Among the three, utility is certainly the most relevant aspect in the context of electronic ID documents, being the prediction of the biometric verification accuracy the main objective of the evaluation; utility is also the central one in the quality score definition adopted by NIST in the Face Recognition Vendor Test for face image quality assessment and by ISO [9]. Many factors concur to determine the utility of a face image, ranging from subject-related characteristics (physical or behavioral) to environmental factors influencing the outcome of the acquisition process. The definition of a robust quality indicator must necessarily consider all these different aspects.

To better frame this work in the general panorama of the face image quality indicators proposed in the literature, it is important to underline some points. As already discussed, we focus on high-quality ISO/ICAO compliant images, taken from a fully cooperating subject, and we thus refer to a **controlled** scenario. Moreover, we assume that no reference images are available for quality computation, falling therefore in the category usually known as **no-reference** quality assessment. A further aspect to consider is that, for this specific application, it's also important that the quality score can be easily interpreted by humans, who must be able to understand what images characteristics determined the specific quality score; for this reason, we believe it's important to score **single aspects** related to the general concept of quality (e.g. focus, saturation, facial expression) and eventually combine them with unified quality indicators (e.g. FaceQNet [7] discussed later).

Motivated by previous experiences in biometric samples quality estimation (discussed in Section II), we propose an approach based on the use of a Random Forest Regressor, trained to estimate a quality score based on a set of input features, derived from different image characteristics. One of the most critical steps to solve in order to effectively train a quality predictor is the definition of target quality values, which are not available a-priory. The feature vector provided in input to the regressor combines ICAO compliance scores (both commercial and research SDKs are evaluated), quality indicators derived from the ISO/IEC WD 29794-5:2020 and the FaceQNet general quality score. All the details are provided in section III-A. Several experiments, reported in section IV, have been designed to analyze the relevance of the different features considered in the controlled scenario, as well as their complementarity, using established methodologies and indicators adopted by NIST in the FRVT Quality Assessment [10]. To the best of our knowledge, this is the first work focusing on image quality assessment for very high-quality images and we believe that this discussion might be useful for researchers in this field to face specific issues not observable in the general context of face image quality evaluation.

This work is the results of a collaboration between the University of Bologna and some members of the Italian institute in charge of the emission of electronic ID documents (Istituto Poligrafico e Zecca dello Stato); the outcomes of this joint study will be useful for the introduction of a quality assessment module in the image sample selection or document enrollment processes.

II. RELATED WORKS

The existing literature on face image quality estimation is huge and a number of techniques have been proposed in the

last twenty years, ranging from traditional computer vision approaches to the most recent, deep-learning based solutions. Interested readers can refer to a good recent survey [6] for a comprehensive review.

Focusing our attention on the controlled ID document scenario, the approaches regarding face image compliance to ISO/IEC 19794-5 [5] are particularly relevant for our analysis. Many research works, indeed, focus on the automated verification of the image requirements established in the ISO standard (e.g. natural expression, frontal pose, absence of shadows, etc.) or a subset of them; in [11] the authors propose a face validation system based on a set of hierarchical tests including 17 different requirements related to acquisition process (e.g. resolution), background, subject's pose or presence of shadows. An interesting analysis is carried out in [12] where 27 different factors contribute to the definition of image quality, interpreted by the authors as an estimation of the utility of that biometric sample, in accordance with what we previously discussed; the authors highlight the importance of score normalization and propose a neural network model to maximize the correlation between the overall quality score (derived from the different factors) and the face recognition matching scores. The authors of [13] identify four main categories of defects in face images, related respectively to environment, camera conditions, user's face conditions and user-camera positioning, and define 6 quality indicators to quantify the presence of such defects, mainly based on the evaluation of possible facial asymmetries related to non-frontal lighting or improper facial pose. The same concept of facial symmetry is exploited in [14] where Gabor wavelet features are used to evaluate facial pose and illumination conditions and combined to DCT for the evaluation of camera focus. The authors of [15], [16] (some of them are co-author of the present paper) defined a set of 30 indicators for the evaluation of face image compliance to ISO/IEC 19794-5; while most of the previous approaches above described referred to internally acquired datasets, not available for public use, a major contribution of [15] was the creation of a public benchmark for the evaluation of ISO/ICAO compliance verification algorithms, named BioLab-ICAO and still available at [17]. The evaluation focuses in this case on the capabilities of research or commercial SDKs to evaluate the single requirements, rather than on the definition of an overall quality measure. Finally, two recent works [18], [19] deal with facial image quality in the context of smartphone-based identity verification; a quality measure is derived from 9 different factors in [18] and used to accept/reject biometric samples. Evaluation is based on the Error Reject Curve (ERC) which illustrates the rate of decrease of False Non Match Rate (FNMR) with respect to the rejected images due to low quality.

This literature review cannot ignore recent deep-learning based approaches, even if they are usually designed for image quality estimation in a more general context. In this category, one of the major contributions is certainly given by FaceQNet [7], a well-established approach for face image quality estimation in a variety of conditions. The proposed framework aims at attributing to ISO/ICAO compliant images top quality scores, and adopts the BioLab-ICAO framework [15] to produce the ground truth quality score used for model training. The good performance of FaceQNet is confirmed by the results obtained in the NIST FRVT Quality Assessment evaluation [10]. For its relevance, FaceQNet will be included in our approach as one of the quality indicators used to estimate our quality score for the supervised scenario. As an alternative (or in addition) other recent proposals could be considered; for instance, [20] describes the Simplified Face Quality Assessment (SFQA) approach, where a hashing-based deep learning model is used for the prediction of face quality from the features of a related FR algorithm, while the recent work [21] introduces SER-FIQ, where the quality score is established in an unsupervised fashion, based on the relative robustness of deeply learned embeddings of that image, rather than on a predefined ground truth derived from human labeling or face matching scores.

Finally, extremely relevant for this work is NFIQ (NIST Fingerprint Image Quality), the standard quality score defined for fingerprints in [22] where the target quality score $ns(x_i)$ of a given sample $x_i \in X$ is obtained as follows:

$$ns(x_i) = \frac{s(x_i, x'_i) - \mu_n(s(x_i, x_j), j \neq i)}{\sigma_n(s(x_i, x_j), j \neq i)} \quad (1)$$

where:

- x'_i is another sample of the same identity of x_i;
- (x_i, x'_i) are two samples related to the same identity and $s(x_i, x'_i)$ is the genuine comparison score provided by a given automatic matcher;
- $\mu_n(s(x_i, x_j), j \neq i)$ and $\sigma_n(s(x_i, x_j), j \neq i)$ represent respectively the average and standard deviation obtained from a set of n impostor comparison scores with other subjects ($j \neq i$).

The structure of feature vector v_i and the algorithm used to produce the final quality score was different for the two versions of NFIQ. In both cases, quality assessment is formulated as a classification problem (instead of a regression problem). In particular:

- 1.0: fingerprints are represented by a 11-dimensional feature vector (including information about the number of minutiae, minutiae quality, statistics about foreground quality, etc.) and a neural network classifier is trained to classify them into one of the five predefined quality classes (where class 1 means top quality and class five the worst quality) derived from the normalized quality score provided in Eq. 1.
- 2.0: fingerprints are represented by a 69-dimensional feature vector including a wide range of image features; fingerprints are classified as low (0) or high (1) utility based on the NFIQ 1.0 class and statistics about the comparison scores and a Random Forest Classifier is trained to perform this classification task. The output quality value is obtained as the probability of input being class 1, properly quantized in the range [1,100].

The NFIQ approach is the starting point for our proposal, but many aspects of that methodology are widely revised to take into account specific peculiarities of the controlled face domain, as explained in the next sections, leading to the definition of an original approach for image quality assessment.

III. DEFINITION OF THE IMAGE QUALITY SCORE

In this paper, we focus on defining a quality metric related to the biometric sample's utility. Therefore, the quality metric is interpreted as a predictor of the performance of an FRS: a high-quality image should lead to better identification of the individual. In this context, the definition of groundtruth should be accomplished through a performance-based approach that is able to describe the correlation between the biometric sample, which in our reference scenario is stored in an electronic machine-readable travel document (e-MRTD), and the verification capabilities of a recognition system, such as an automatic border control (ABC) gate.

In the above scenario, the quality assessment of the image provided by the citizen or live captured (depending on the issuing country's regulations) should take place at the end of the enrolment process. This phase, which only involves using the input image (no reference), should provide a numerical score (e.g. in the range between 0 and 100) able to predict the recognition performance for the given face image. In compliance with the ISO/IEC 39794-5 [3] standard, both the quality score and the compressed face image should be stored in the e-MRTD's second data group (DG2).

Of course, the biometric samples considered are of high quality: besides referring to inherently controlled images, it should be stressed that the quality assessment occurs after the required ISO/ICAO compliance verification. This aspect has a significant impact on the quality assessment process: on the one hand, many low-quality images are preemptively discarded due to non-compliance concerning some static properties (e.g. the subject's glasses are equipped with dark lenses, optical distortion presence) or dynamic properties (e.g. the subject's expression is accentuated, brightness variations), but on the other hand, determining a consistent score can be noticeably more complex. In fact, in a more typical ''in-the-wild'' scenario, significantly less controlled conditions occur, thus offering inherently more variability and making it easier to discriminate between low and high-quality images. From this perspective, ISO/ICAO checks are reasonably tight, and the challenge here is to develop an approach able to effectively discriminate the quality of face images fully compliant with such standards.

A. A REGRESSOR FOR IMAGE QUALITY ESTIMATION

Our work is inspired by the quality assessment process adopted in the NIST Fingerprint Image Quality (NFIQ), the de facto standard for fingerprint images quality evaluation. NFIQ recalls the concept of utility previously described, defining the quality of a biometric sample as a prediction of a matcher performance (i.e. good quality fingerprints are likely to produce high matching scores). This approach presents good properties in terms of efficacy, efficiency and objectivity: the ground truth is strictly related to matching performance and can be determined according to an automated procedure, not requiring subjective human visual inspections often difficult to accomplish. However, NFIQ1 only represents a starting point; many peculiarities of the face domain required ad hoc solutions to be defined.

In our approach, a Random Forest (RF) Regressor is used for image quality assessment, rather than a Random Forest Classifier as in [4]. We believe that this is a natural choice that avoids the complex task of defining precise criteria to categorize the images as high or low quality; the regressor is effective in providing a continuous quality score, which is the desired outcome of our evaluation. An easy normalization procedure can return quality values in the desired range.

A fundamental component of the quality assessment system is the target quality score, used to train the quality regressor. We consider here two different target values. The first one is analogous to the one in Eq. 1, extended to take into account multiple genuine scores as follows:

$$qs_1(x_i) = \frac{\mu_m(s(x_i, x'_i)) - \mu_n(s(x_i, x_j), j \neq i)}{\sigma_n(s(x_i, x_j), j \neq i)} \tag{2}$$

where:

- x'_i is another sample of the same identity of x_i;
- $s(x_i, x'_i)$ is the genuine comparison score provided by a given automatic matcher and (x_i, x'_i) are two samples related to the same identity;
- $\mu_m(s(x_i, x'_i))$ is the average score obtained from a set of m genuine comparison scores computed from different images of the same subject;
- $\mu_n(s(x_i, x_j), j \neq i)$ and $\sigma_n(s(x_i, x_j), j \neq i)$ represent respectively the average and standard deviation obtained from a set of n impostor comparison scores with other subjects ($j \neq i$).

In this case, for each image, the genuine scores are normalized taking into account subject-specific impostor values. It is worth noting that, according to this definition, two different factors can determine low qs_1 values:

- low genuine scores, that might determine *false non matches*, often related to image features (e.g. pose or illumination) or other factors (e.g. aging);
- high impostor scores, that might lead to *false matches*, typically determined by a high similarity with other subjects.

For this reason, we argue that this definition fits well for those applications where both the two kinds of errors play a relevant role. However, since from the image quality perspective the main focus is on the control of false non-matches, a better-suited score definition can be defined as:

$$qs_2(x_i) = \frac{\mu_m(s(x_i, x'_i)) - \mu_p(s(x_j, x_k), j \neq k)}{\sigma_p(s(x_j, x_k), j \neq k)} \tag{3}$$

where $\mu_p(s(x_j, x_k), j \neq k)$ and $\sigma_p(s(x_j, x_k), j \neq k)$ represent respectively the average and standard deviation obtained from p impostor comparison scores computed on the whole population (all the subjects in the training set). The difference w.r.t. Eq. 2 is that subject-specific genuine scores are normalized according to general (not subject-specific) impostor scores. In practice, the same impostor score average $\mu_n(s(x_j, x_k), j \neq k)$ and standard deviation $\sigma_n(s(x_j, x_k), j \neq k)$ values are used for all the images, thus removing the impact of similarity between specific subject pairs from the target quality value. The experiments documented in section IV point out the different impact of these two definitions on the Error Vs. Discard Curves adopted for performance assessment. Other techniques could be adopted to compute the target value; for instance, subject-specific impostor scores could be considered, but discarding part of the top scores (outliers caused by similarity with other subjects). Some preliminary experiments have been conducted to analyze this possibility, but from a practical point of view identifying the right portion of images to discard in order to produce the desired effect is not an easy task, especially when different datasets are used. Furthermore, the target values adopted have a general validity (no parameters to fix) and are easier to compute.

The feature vectors v_i used to encode the image characteristics are built from three sources: ICAO compliance scores provided by research/commercial SDKs, image quality indicators described in ISO/IEC WD 29794-5:2020 and the FaceQNet quality score. All the details are provided in the next section.

B. IMAGE QUALITY ELEMENTS

Several image characteristics are taken into account in this work to describe the face image and compose the feature vector v_i used for quality estimation by the Random Forest Regressor described in the previous section. The elements can be organized into three macro-categories: ISO/ICAO compliance scores, scores pertaining the ISO/IEC WD 29794-5 standard [5] and an evaluation provided by a deep-learning-based solution, specifically FaceQNet [7]; it is worth noting that, even if the controls originate from these different groups, they are used individually, as single quality elements in the proposed quality regressor. Further details are provided in the next sections. It is necessary to anticipate that even if there are overlaps between the three subgroups, we decided to start with an overabundant set of features to exploit the capabilities of the Random Forest (RF) Regressor to identify the most relevant features in relation to the specific target. Considering a wide range of possible values will enable a comprehensive evaluation of the different elements contributing to the definition of the quality score in such a highly controlled scenario. A discussion about the contribution of the single features will be provided in Section IV.

It is worth noting that in this scenario most of the quality elements evaluated derive from hand-crafted features, being FaceQNet the only contribution from the deep-learning world. This choice is motivated by two highly desirable properties that make hand-crafted features particularly suited for this specific application:

- they are ''explainable'' to humans, this is an important feature in this context since the quality score could also be used to drive the image acquisition process, and it's therefore important to provide to the operator precise indications about possible factors determining a low quality score (e.g. non frontal pose, uneven lighting, unnatural skin color);
- training approaches for hand-crafted feature computation do not require a large amount of training data which is instead required for deep learning-based approaches; this is particularly important in this scenario where the available datasets of high-quality ISO/ICAO compliant images are very limited.

1) ISO/ICAO COMPLIANCE SCORES

It is important to stress that in the aforementioned quality assessment process, an essential contribution is provided by the ISO/ICAO compliance checks to which each face image must be exposed. Several available solutions offer a global assessment of compliance, eventually with a single comprehensive score assignment. Others provide a binary response for each characteristic foreseen by the standard, defining face images that do not meet at least one requirement as non-compliant. We have considered two distinct solutions that provide a specific evaluation for each requirement: a commercial SDK (whose name is not disclosed due to a confidentiality agreement) and the BioLab-ICAO-Check [15] tool. Indeed, both provide a set of scores referring to different face image characteristics, both photo-metric (e.g. exposure, sharpness, colour saturation) and subject-related (e.g. pose, presence of accessories, expression). Although both software meet the standard's requirements entirely, the various compliance tests do not directly match the two solutions. For instance, BioLab-ICAO has a dedicated ''blurring presence'' check; in the commercial SDK, this control may be coupled with tests regarding sharpness and focus. Besides, the BioLab-ICAO offers an overall assessment regarding the presence of occlusions in the eyes area (typically due to hair or eyeglass frames), while the other SDK offers separate scores for individual eyes. Thus, the commercial SDK offers the evaluation of 32 different characteristics versus 23 considered by the BioLab-ICAO tool even though the handled requirements are mostly the same. Indeed, Table 1 reports the tests performed by the different quality indicators considered. To ease the comparison, we propose a categorisation of the various controls in eight macro-categories of tests. Of course, several controls are strictly related; we expect therefore that, during the training procedure, a subset of controls will be selected for each category and that the overall importance of some of such categories will be higher. A posterior importance analysis could also reveal unexpected (or counterintuitive) issues.

TABLE 1. Categorisation of the quality controls considered for BioLab-ICAO Check, Commercial ICAO verification SDK and ISO/IEC WD 29794-5.

Category	BioLab ICAO	Commercial SDK	ISO/IEC WD 29794-5
Blurring and mis-focus	Blurred Pixelation	Focus Sharpness	De-focus Sharpness Edge-density
Exposure, variation in lightning and shadows	Unnatural skin tone Too dark/light Washed out Flash reflections on skin Flash reflections on lenses Shadows across face Shadows behind head	Saturation Colour control Dynamics Glare Shadows	Under/over-exposure Illumination uniformity Illumination modulation
Eyes	Looking Away Hair across eyes Eyes closed Red eyes	Gaze Hair covering left/right eye Left/right eye closed Red eyes Intrapupils distance	Eyes visible Eyes open Inter-eye distance
Mouth	Mouth open	Mouth expression	Mouth closed
Face image pose, aspect ratio and other faces	Roll/pitch/yaw rotations Presence of other faces or toys	Face pose Face found Horizontal/vertical face posit. Face image width/height ratio	Pose Number of faces present Horizontal/vertical position
Accessories	Dark tinted lenses Frames too heavy Frames covering eyes Veil over face Hat/cap	Dark lenses Frames too heavy Frames covering left/right eye Face valid Head coverings	
Background	Varied background	Background evaluation	-
Other	Ink marked/creased	-	Compression

2) ISO/IEC WD 29794-5:2020

This standard, currently under development, is taking shape with the primary objective of addressing the need for quality quantification in the context of face images. It specifies a number of quality elements and details the underlying computational approaches. As can be appreciated in Table 1, these elements partially overlap with those required for issuing identity documents, although this standard aims to provide a general methodology for quality assessment. At the time of writing, the draft provides details of a subset of the 23 controls currently envisaged. Indeed, for many of these, no methodology is yet specified, or the scores' acceptability range is defined; for others, the discussion on their possible integration is still open. For our evaluation, we implemented the controls for which a procedure was fully defined. For others, such as the non-detailed *eyes visible* and *eyes closed*, we supplemented them with our implementation. A total of 16 tests were included in our evaluation. However, it is essential to emphasize that some controls may still be subject to modification and evaluation, as this is not the standard definitive version.

- For the *de-focus* test, the standard summarizes the computation procedure proposed in [23]: a 3×3 mean filter is applied to the segmented face region. Subsequently, the difference image between the original face image and the image resulting from the convolution is determined. A small value in the scalar mean of the difference image implies the presence of blurring.
- The computation of the *sharpness* metric is based on a generalization of an eigenvalues problem proposed in [24]. This approach uses Rayleigh quotient optimization by exploiting some information extracted from the image and subsequently represented by a set of eigenvalues. In this case, acceptability thresholds are not yet established.
- The *edge density* is intuitively checked by applying a Sobel edge mask to the image and then calculating the average value over the face region. Even in this case, the draft does not describe value range and acceptability thresholds.
- Concerning the controls for brightness variation, under/over-exposure are defined as the proportion of face pixels whose intensity lies in the extreme ranges (i.e. defined in [0.7] and [247,255]).
- *Illumination uniformity* test evaluates the difference in illumination on the left and right side of the face. Therefore, the mean and standard deviation of the pixels belonging to the right and left side of the face is determined. The aggregate standard deviation is then calculated and used to determine non-uniformity by the difference between the mean pixel values of the two sides of the face.
- The *illumination modulation* check is used to determine the dynamic range of the image: a correctly exposed image must have at least 7 bits of intensity variation in the face region (i.e. a range of at least 128 unique values). The procedure includes luminance recovery and subsequent determination of image entropy. The units of entropy are bits, so the acceptability threshold is set for entropy values greater than or equal to 7 for 8-bit images.

- The *eyes visible* test is not currently documented: we have implemented an eye detection solution based on the Haar Feature-based Cascade Classifiers provided by OpenCV.
- The *eye-opening* control also has no specification. Our implementation, which employs the landmarks provided by Dlib, examines the aspect ratio of both eyes to determine whether they are open or closed. A similar solution is used for the *mouth opening* test.
- Another test that does not have a computational methodology is the head pose test. Our implementation exploits WHENet, recently presented in [25]. Specifically, the angles provided by WHENet are given as input to a sigmoid function to produce a quality value. In our implementation, the lowest score defines the quality value for this test.
- Although it is rather difficult to have an input with more than one face in the reference domain, we have also provided an implementation for the *number of faces present* control. Specifically, we use OpenCV's DNN Face Detector to check whether there are other faces besides the main subject.
- Concerning the tests related to the *horizontal/vertical position* of the face, the draft follows the guidelines expressed by the ISO/IEC 39794-5:2019 standard. Specifically, once the eye centres have been defined using the corresponding landmarks, it is necessary to compute the eyes midpoint (X_c, Y_c). Then, it is necessary to determine the horizontal margin legality product $H(X_c - 0.45 * Img_{width}) * H(0.55 * Img_{width} - X_c)$ where $H(x)$ is the step function whose value is 1 for non-negative x and 0 otherwise. Similarly, the vertical margin legality product is defined as $H(Y_c - 0.3 * Img_{height}) * H(0.5 * Img_{height} - Y_c)$. The acceptability value for both controls is equal to 1.
- Finally, the test regarding compression is well defined. Firstly, it is necessary to store the compressed data size and compute the uncompressed image data size as the product of the image size and the number of colour channels. It is then required to determine the quality parameter Q applied by an encoder to obtain the input compressed dimension. The Q parameter is then used to compress the token image generated following the ISO/IEC 19794-5 standard. Therefore, the compression ratio is determined by the size of the uncompressed token image with respect to the size of the compressed token image.

3) FaceQNet

FaceQNet [7] is a recent deep learning-based quality assessment approach. The solution is built on a ResNet-50 Convolutional Neural Network trained on the VGGFace2 database with the purpose of providing a comprehensive image quality score. For training purposes, the images in the VGGFace2 database have been automatically labeled using the BioLab-ICAO framework [15] with quality information related to their ICAO compliance level. It's worth recalling that FaceQnet is intended to be a general estimator and not restricted to the context of face images for identity documents and that ICAO compliant images are all ideally scored as top level by FaceQNet. We made the assumption that the score provided, derived from features automatically learned from the deep neural network, may complement the evaluations given by other indicators whose controls are entirely based on hand-crafted features. This hypothesis is also supported by the current ISO draft (ISO/IEC WD 29794-5:2020) which suggests to consider both hand-crafted features and holistic quality scores obtained by deep learning-based systems.

IV. EXPERIMENTAL EVALUATION

A. DATABASES AND TESTING PROTOCOL

Several experiments have been carried out to evaluate the effectiveness of the proposed regressor on good-quality images.

The database used for testing the proposed quality regressor has been composed from different public face databases: AR Face Database [26], FERET [27], FRGC [28] and PUT [29]. For each database, we used the commercial ICAO compliance verification software to automatically process and select high-quality compliant images and to tokenize them. More than 63000 images have been processed resulting in a final set of 7628 selected images used for the experiments: 192 images from AR, 640 images from Feret, 5549 images fro FRGC and 1247 images from PUT.

The dataset has been partitioned into a training set (5343 images) and a test set (2285 images).

In order to measure the capability of the proposed quality regressor to predict the utility of the face images for face verification purposes, some face verification experiments need to be carried out. Different face recognition systems have been considered here:

- DLib [30]: a widely used open source and deep learning-based face recognition library that, according to some internal tests is able to provide performance compatible with the typical requirements of the ID document verification scenario;
- VGG-Face and ArcFace, two state-of-the-art deep models for face recognition publicly available in the Deepface library [31];
- VeriLook 12.1 [32]: a commercial face verification tool developed by Neurotechnology.

The verification thresholds used for the four FRSs have been determined on the training set in order to fulfill the operational requirements suggested by Frontex (FMR = 0.1% with a maximum FNMR of 5%).

The genuine and impostor attempts have been all carried out intra-database (i.e. with images taken from the same original dataset) and have been defined as follows:

- All the possible genuine attempts have been carried out using pairs of images of the same subject; the number of attempts varies for the different datasets from one up to 40 per subject.

- The impostor attempts of a given subject have been defined taking at most a single probe image for each other subject of the same dataset, fixing a maximum of 100 impostor attempts for each gallery image.

Overall, the number of verification attempts performed on the images of the testing set is as follows: 44581 genuine attempts and 217156 impostor attempts.

TABLE 2. The different versions of the quality regressor obtained using different subsets of the features: Commercial ICAO Check tool ($ICAO_C$), BioLab ICAO Check tool ($ICAO_B$), ISO/IEC WD 29794-5:2020 (ISO) and FaceQNet.

Feature Combination	$ICAO_C$(C)	$ICAO_B$(B)	ISO (D)	FaceQNet (F)
BCDF	✓	✓	✓	✓
BDF		✓	✓	✓
CDF	✓		✓	✓
BCD	✓	✓	✓	
BF		✓		✓
CF	✓			✓
DF			✓	✓
B		✓		
C	✓			
D			✓	

All the experiments are accompanied by an ablation study aimed at evaluating the effectiveness of the different quality indicators and their complementarity in the different scenarios. Different versions of the quality regressor have been trained based on different combinations of quality indicators subsets used to compose the feature vector representing the images, as described in Table 2. The regressor implementation is based on the scikit-learn library [33]; for each combination, the regressor parameters have been tuned using the RandomizedSearch functionality on a validation set extracted from the training set.

B. METRICS

In the literature, the evaluation of quality assessment approaches is based on the idea that their effectiveness is related to the capability of identifying low-quality images which have a negative impact on the face verification process. For this reason, the approach adopted at NIST in the FRVT Quality Assessment benchmark [10], is to analyze the so-called Error versus Discard Characteristic (EDC), a curve obtained by observing the FNMR variations as a function of the fraction r of low-quality images discarded from a reference dataset used for face verification experiments. An error reduction (decreasing trend) is desirable since it confirms that the quality indicator used to select images is actually related to the utility of the specific samples.

When coming to the ID documents scenario, the application of these metrics is not straightforward as it appears to be due to some very specific issues related to the use of high-quality images:

- the failure of a genuine verification attempt is certainly influenced by the quality of the two compared images, but if one of the two images is high quality (the one we are evaluating), the errors are mainly determined by the quality of the probe image. This makes it difficult to relate the quality score assigned to an ICAO compliant image to the face verification errors typically caused by the bad quality of the probe.
- the obvious solution to the previous point is to select high-quality probe images. This is the approach we followed here, but it is not the ultimate solution since when only high-quality images are used for face verification, the errors made by FRSs are very low. Indeed, in our experience, some errors (both false matches and false non-matches) are observed for the open source FRSs tested, while state-of-the-art commercial tools such as VeriLook almost do not commit false non-match errors, which are the most relevant ones from the image quality perspective. The absence of errors makes inapplicable the computation of EDC. This is another specific issue for the high-quality image scenario which is not typically observed in the general context of image quality assessment.

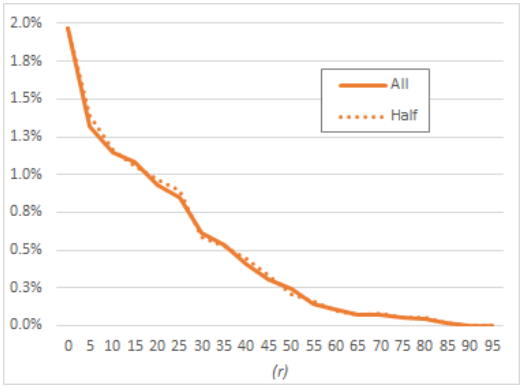


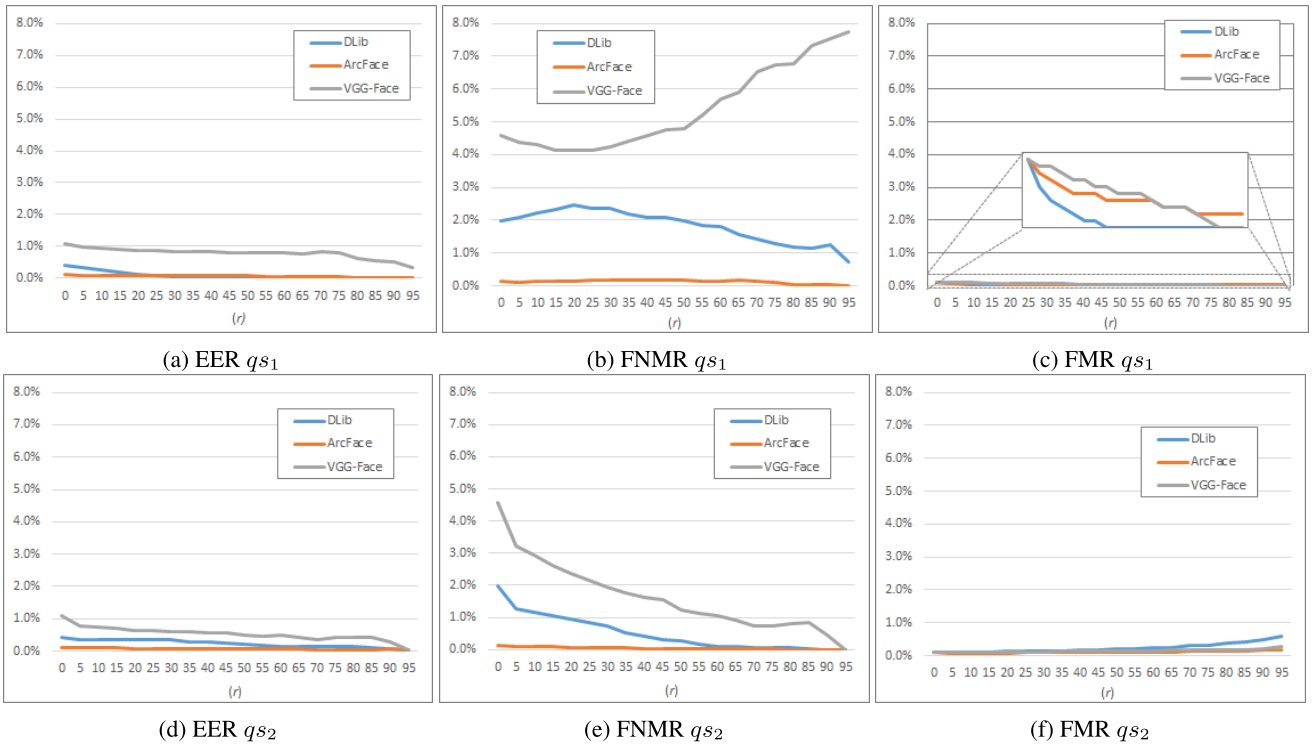
FIGURE 1. FNMR vs. Discard curves obtained using respectively all the training impostors scores and only half of them for the computation of the target values according to Eq. 3.

In order to deal with these issues, we decided to perform face verification attempts only based on high-quality images, and to report the EDC graphs for the open source FRSs. We will analyze the EDC curve for both FMR and FNMR, to better evaluate and compare the two alternative definitions of the target quality score used for regressor training.

C. EXPERIMENTAL RESULTS

Several experiments have been carried out to evaluate the effectiveness of the proposed approach, taking into account different aspects:

- the impact of the target quality value definition;
- the correlation between the estimated quality score and the target value computed from face verification scores;
- the different feature subsets used by the regressor and their impact on EDC;
- the importance of the single quality features used to encode images.



(a) EER qs_1 (b) FNMR qs_1 (c) FMR qs_1

(d) EER qs_2 (e) FNMR qs_2 (f) FMR qs_2

FIGURE 2. Error vs. Discard curves computed for EER, FNMR and FMR using the two alternative definitions of the target quality value for regressor training: qs_1 in the first row and qs_2 in the second row.

1) IMPACT OF THE TARGET QUALITY VALUE

We proposed in section III-A two alternative definitions of the target quality score computed for regressor training (Eq. 2 and Eq. 3), on the basis of a different computation of the impostor score values (subject-specific in qs_1, general over the whole population in qs_2). Some experiments have been conducted using the open source FRSs to evaluate its impact, using BDF as feature combination for regressor training.

Figure 2 reports the EDC graphs for FNMR, FMR and EER obtained using the quality scores predicted by two regressors trained using the target values defined in Eq. 2 and Eq. 3, respectively. In both cases, the EDC trend for the EER is decreasing (Figure 2a and Figure 2d), indicating that the quality scores predicted by the regressor are effectively related to the utility of the images for face verification and that discarding images based on the predicted quality score produces an overall error reduction. However, the trend for FMR and FNMR curves are noticeably different, even opposite, for qs_1 and qs_2. For qs_1 the FNMR EDC (Figure 2b) does not exhibit the desired decreasing trend, which characterizes on the contrary the FMR EDC (Figure 2c). This is not the expected optimal behaviour since, as already mentioned in Section III-A, it is mainly due to the high impact of subject-specific impostor scores. The FNMR EDC graph obtained with qs_2 (Figure 2e) is much closer to the desired result since a smoothly decreasing trend is observed for FNMR. The FMR trend (Figure 2f) is stable or slightly increasing for ArcFace and VGG-Face, while for DLib a larger increment is reported (even if in general the FMR changes are one order of magnitude smaller than the FNMR changes); this indicates that high-quality scores are assigned to images producing false matches. However, in this case, the face verification errors are related to an effective similarity between subjects (and DLib seems to be particularly sensitive to this issue) and not attributable to image quality features. To summarize, we believe that the target value definition provided in Eq. 3 is more effective to produce the desired behaviour in terms of error reduction.

From a practical point of view the adoption of qs_2 (Eq. 3) would require a continuous updating of the average impostor score ($\mu_p(s(x_j, x_k), j \neq k)$) and of its standard deviation ($\sigma_p(s(x_j, x_k), j \neq k)$). However, being the average impostor score only a normalization factor, we argue that it can be determined on a limited training set and subsequently used without further updates. To confirm our hypothesis we performed an experiment where the average impostor score and its standard deviation have been determined using only half of the scores available rather than all of them. We used the scores to compute the target values for training our quality regressor and we measured the resulting EDC curves (FNMR vs. discard) reported in Figure 1. The graph clearly shows that the difference between the two approaches is negligible and confirms our hypothesis that the reference value can be computed from an initial training set and used unaltered for the subsequent computations.

The target quality score definition of Eq. 3 will be used in the rest of the experiments.

2) ESTIMATED QUALITY SCORE VS. TARGET VALUE

A set of experiments has been carried out to evaluate the capability of the regressor to predict quality values highly correlated to the target quality scores, computed according to Eq. 3. To this aim, the correlation between the target values and the quality values predicted by the different versions of the regressor on the test set has been computed for DLib, ArcFace, VGG-Face and VeriLook. The results are given in Table 3. Overall the correlation values are always positive and quite good in some cases, even if differences can be observed between the different versions of the regressor; the combinations BCDF, BDF, CDF and BCD generally achieve higher values w.r.t. other combinations. As a term of comparison, the correlation between the target quality values and the quality scores provided by FaceQNet is lower (0.3283).

TABLE 3. Correlation values between the target quality values, computed with all the FRSs, and the quality scores predicted by the different versions of the quality regressor.

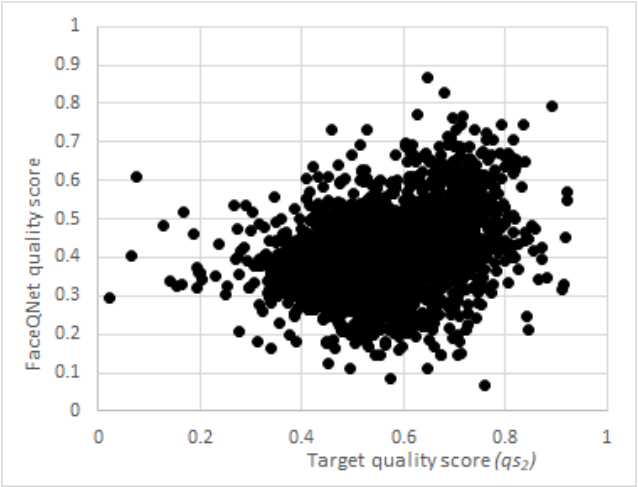
	BCDF	BDF	CDF	BCD	BF	CF	DF
DLib	0.68	0.66	0.56	0.66	0.62	0.50	0.53
ArcFace	0.57	0.54	0.51	0.49	0.37	0.40	0.37
VGG-Face	0.47	0.50	0.44	0.47	0.40	0.40	0.33
VeriLook	0.54	0.51	0.47	0.53	0.50	0.44	0.34

This different behaviour is visualized in the two scatter plots of Fig. 3; Fig. 3a compares the target quality values computed from Eq. 3 (using DLib for face verification) and the values predicted by the BDF regressor, while Fig. 3b compares target values and FaceQNet quality scores. The graph in Fig. 3a shows a good correlation between the two values and a quite smooth increasing trend. Considering that the test images are all well controlled and of good quality, we can conclude that the estimated quality measure is able to effectively capture the limited variability present in the dataset. In the graph of Fig. 3b the correlation is less evident, as clearly highlighted by the "flat" trend line. This means that the FaceQNet quality values range between about 0.1 and 0.8 with just a weak relationship with the target quality score. This is quite understandable if we consider that FaceQNet has been designed and trained to deal with a more general concept of image quality; it is however worth of attention the quite low value (<0.2) measured for some samples in the dataset, which is a bit counter-intuitive with reference to the specific scenario.

3) FEATURE SUBSETS

Further experiments have been conducted to compare the effectiveness of the different versions of the quality regressor, obtained using different feature subsets; this analysis is based on the EDC graphs, obtained by measuring the FNMR in subsequent face verification tests where part of the images are gradually removed from the test dataset, discarding at each step the samples with worst quality score.

(a) Target quality scores vs. Quality score predicted by the BDF regressor using DLib for face verification.

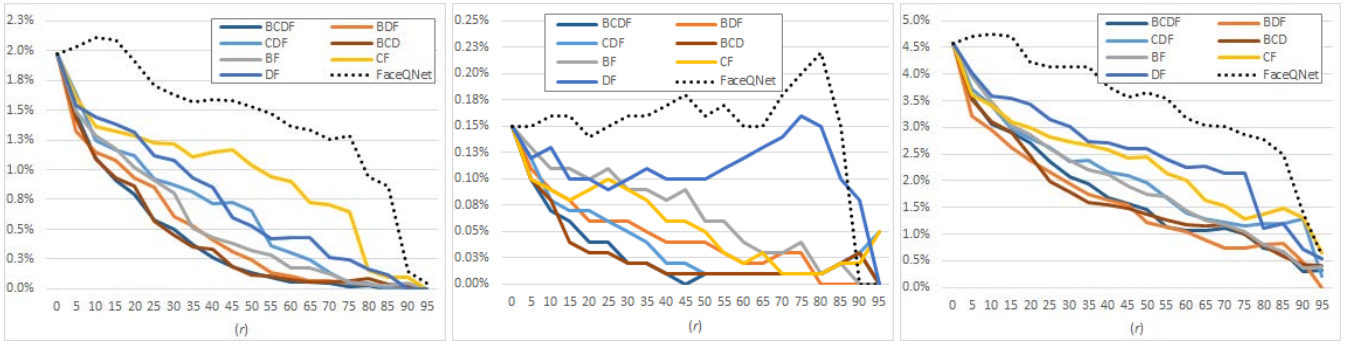


(b) Target quality scores vs. FaceQNet quality score using DLib for face verification.

FIGURE 3. Scatter plot representing the correlation between target quality scores (computed according to Eq. 1) and the quality values provided by the BDF regressor and FaceQNet, respectively.

The results obtained are given in Fig. 4 for the different versions of the proposed quality regressor. In particular, the graphs show, the FNMR value as a function of the fraction (r) of low-quality images discarded, for the different versions of the quality regressor (see Table 2) corresponding to different feature subsets, using DLib (Fig. 4a), ArcFace (Fig. 4b) and VGG-Face (Fig. 4c) for face verification, respectively.

Overall, a good behaviour is observed; in all cases discarding low-quality images based on the proposed indicator allows to reduce the error rate. Of course, the error rates measured on this set of high-quality images are quite low (all FRSs comply with the Frontex guidelines), nevertheless the quality score computed by the regressor shows to be effective in capturing the small differences between images that impact to some extent the verification results. The decreasing trend is common to the different versions of the regressor, even if



(a) DLib (b) ArcFace (c) VGG-Face

FIGURE 4. FNMR vs. Discard curves obtained on the digital images for the different versions of the quality regressor (different feature subsets - BDF, CDF, BF, CF, DF), using respectively Dlib, ArcFace and VGG-Face for face verification. As a term of comparison, the FNMR vs. Discard curve obtained using the FaceQNet quality score for selecting the images to discard is reported (dashed line).

(a) Low quality score

(b) High quality score

FIGURE 5. Examples of images that received low (a) and high (b) quality scores using the BDF quality regressor and DLib for face verification.

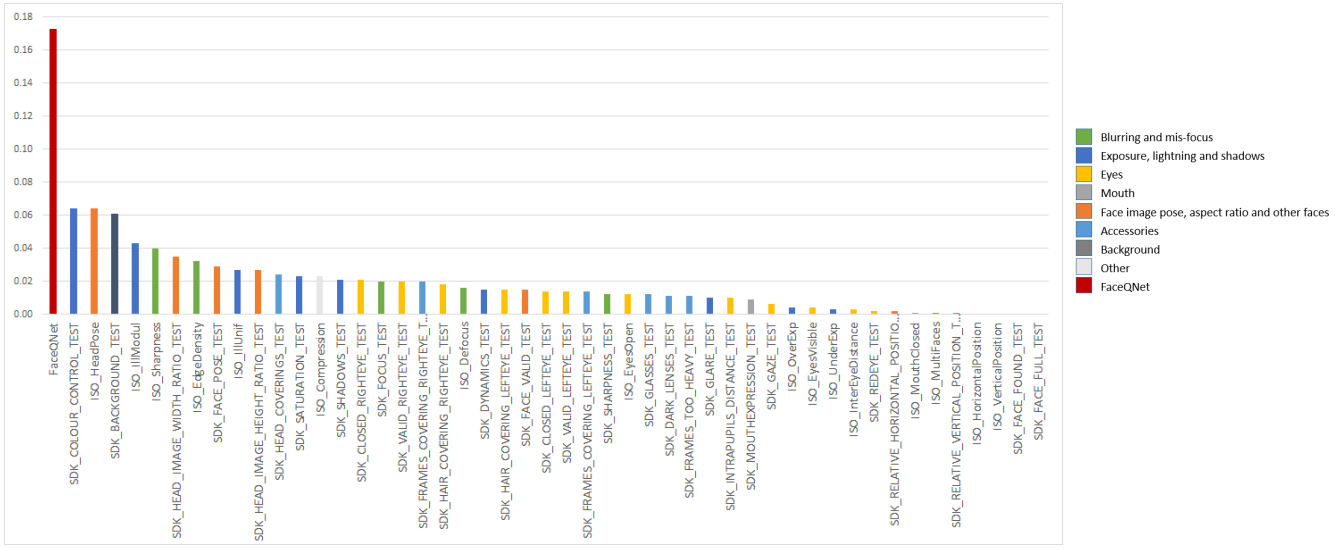
(a) Hat/cap (b) Pose (c) Contrast

FIGURE 6. Examples of images that passed the automatic ICAO check despite of the presence of non-compliant elements: (a) hat/cap, (b) non-frontal pose, (c) low contrast.

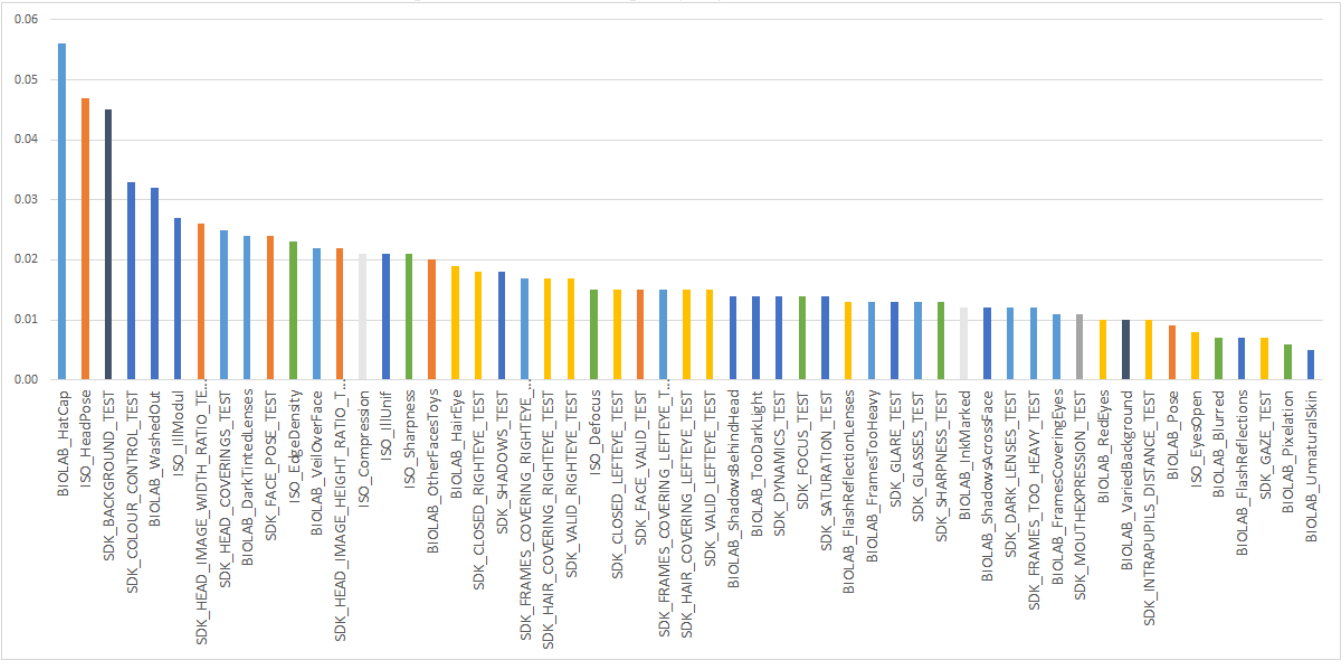
some differences can be appreciated and some combinations achieve better results; in particular, BCDF, BCD, BDF and CDF are the most promising for the tested face verification systems. For some combinations the contribution of the FaceQNet quality score is relevant, but it can somehow be substituted by a set of single quality controls; for instance, the BDC combination that doesn't include FaceQNet achieves the best general performance. A slightly anomalous behaviour is visible in Fig. 4b for the feature combination DF where the curve has a sudden increment after an initial decrement. This trend seems to indicate that the images discarded in that range of experiments are not responsible for false rejections; the reduction of the number of test images and a constant rejection rate determines the small increment observable in the graph. However, we believe that this is due to the fact that the test set only contains high-quality images, with small differences in terms of quality, and that FRSs are generally quite robust to this small factors.

To establish a term of comparison, we computed the same FNMR vs. Discard curves for the different FRSs gradually discarding the images in the test set according to the quality score provided by FaceQNet (dashed black lines in Fig. 4). It is clearly visible that the error decrement is slower and less accentuated for FaceQNet w.r.t. all the different combinations of the quality regressor. This analysis confirms that, for the specific task of quality assessment in eMRTD documents, existing general and unified quality scores are not sufficient and must be combined and complemented with other indicators able to better capture small image details affecting image quality.

Fig. 5 provides some examples of images that received low (Fig.5a) and high (Fig. 5b) quality scores. The outcome of quality prediction is reasonable from a visual point of view; low quality images are affected by some limited issues in terms of resolution, pose, lighting or shadows that might contribute to produce lower genuine scores during the verification attempts. On the contrary, the images scored with a high value are all well controlled, and do not present relevant defects.



(a) Feature importance for the CDF quality regressor.



(b) Feature importance for the BCD quality regressor.

FIGURE 7. Feature importance for the regressors trained using the CDF and BCD feature combinations and ArcFace for face verification. The bar colors refer to the feature categorization given in Table 1 (see legend).

4) IMPORTANCE OF THE SINGLE QUALITY FEATURES

Of course, not all the single quality indicators used to train the regressor are equally relevant for the purposes of quality estimation. A useful output of the regressor training process is also the importance of the single features with respect to the predictability of the target variable, computed by considering the relative rank (i.e. depth) of a feature used as a specific node in a tree. Features used at the top of the tree contribute to the final prediction of a larger fraction of the input samples. The expected fraction of the samples they contribute to can thus be used as an estimate of the relative importance of the features. In scikit-learn [33], the fraction of samples a feature contributes to is combined with the decrease in impurity from splitting them to create a normalized estimate of the predictive power of that feature. The estimates of predictive ability are generally averaged over several randomized trees; this is known as the mean decrease in impurity, or MDI. Refer to [34] for more information on MDI and feature importance evaluation with Random Forests. Fig.7 shows the feature importance values obtained by the different quality scores for the regressors trained using CDF and BCD feature combinations and ArcFace for face verification. For a more immediate

visual evaluation, the histogram bars have been assigned a colour code referring to the indicators categorization given in Table 1 (e.g. green stands for indicators related to blurring and mis-focus). For the CDF combination (Fig. 7a) it is interesting to note the major role played by FaceQNet as a synthetic indicator; however, other categories of indicators are fundamental such as exposure, face image pose, blurring and misfocus which capture most of the small defects characterizing the images in the testing set. For instance, although all the images have a negligible level of blurring, some of them are sharper than other (e.g. FRGC images vs. AR ones). Another set of relevant indicators is related to exposure, lighting variations and shadows. Even in this case the variations in the test set are very limited and negligible, however small differences can be appreciated; the related indicators are therefore very useful in the quality estimation process. For the BCD regressor (Fig.7b), where FaceQNet score has not been used, an important feature is represented by the check related to hat/cap. This is quite surprising, but a visual inspection of the images automatically selected as ICAO compliant revealed that a number of images with hat passed the control (see Fig. 6a). It is therefore reasonable that this check assumes a certain importance in quality assessment. Other important features are also related to small variations observable in the dataset such as limited pose deviations from frontal (Fig. 6a) or low contrast (Fig. 6c). Overall the most relevant features are those related to exposure, lighting, pose and presence of accessories.

In both cases, the features obtaining a lower relevance score fall in one of the following two cases: i) features highly correlated to other features that received a higher score (e.g. SDK_SHARPNESS_TEST belongs to the blurring class and other similar features such as ISO_SHARPNESS are probably more effective); ii) features related to specific variations that do not appear in the dataset due to the preliminary selection of high-quality images (e.g. mouth closed or anomalous head position).

An additional experiment has been carried out to evaluate the importance of the single quality features within each group (ICAO_C, ICAO_B, ISO) thus excluding the impact of overlaps between different feature groups. More specifically, we trained different quality regressors starting from the features of single feature subsets (ICAO_C, ICAO_B, ISO) and we comparatively analyzed two different aspects: the FNMR vs. discard curves and the feature importance. The FNMR vs. discard curves obtained using ArcFace for face recognition are reported in Figure 8. The results show that the feature subsets ICAO_C (C) and ICAO_B (B) perform quite well producing the desired decreasing trend, while the ISO feature set (D) performs worse.

Regarding feature importance, we tried to estimate the overall feature importance of the different categories of quality controls reported in Table 1: 1) Blurring and mis-focus, 2) Exposure, lighting variations and shadows, 3) Eye, 4) Face pose, aspect ratio, other faces, 5) Mouth, 6) Accessories, 7) Background, 8) Other (e.g. compression). For this evaluation we proceeded as follows:

1) we considered the feature importance obtained by regressor training for the three different regressors trained on single feature subsets (ICAO_C, ICAO_B, ISO);
2) we adopted a Borda count approach to assign a score to each feature for each regressor; in particular, we ranked the single features in decreasing order of their importance and assigned them a decreasing score (from 1 to $1/N$, where N is the number of features used to train the specific regressor).
3) the features scores for each regressor are then summed over the different categories of Table 1, and normalized according to the number of controls per category, thus obtaining the results reported in Table 4.

The bolded values represent the three more relevant categories for each feature subset. The results obtained are aligned with those reported in Figure 7 and confirm that very relevant factors for quality assessment in this context are those related to exposure/lighting, blurring, presence of accessories and mouth (expressions).

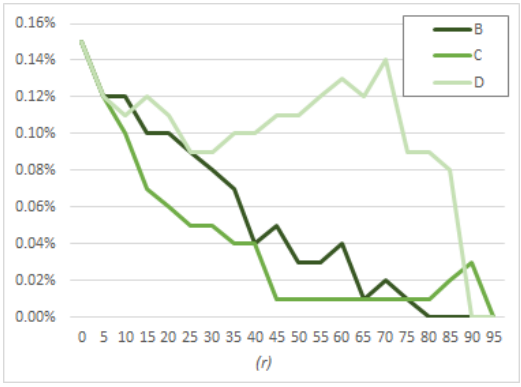


FIGURE 8. FNMR vs. Discard curves obtained using obtained using different quality regressors trained of different feature subsets (ICAO_C, ICAO_B, ISO); ArcFace has been used here for face recognition.

TABLE 4. Overall category importance for different versions of the quality regressor (B, C and D). The value of a given category for a given regressor is computed summing up the Borda count scores obtained by the single features of that category.

Class	Overall importance score		
	B	C	D
Blurring and mis-focus	0.056	**0.139**	**0.233**
Exposure, lighting variations and shadows	**0.166**	**0.165**	**0.189**
Eye	0.119	0.134	0.142
Face pose, aspect ratio, other faces	0.013	0.057	0.078
Mouth	**0.213**	0.109	0.107
Accessories	**0.208**	0.135	–
Background	0.125	**0.261**	–
Other	0.100	0.000	**0.252**

V. CONCLUSION

In this paper an approach for image quality assessment in electronic documents has been proposed. The method is

based on a Random Forest regressor, trained starting from a large set of features including ISO/ICAO controls, quality indicators described in the draft ISO/IEC WD 29794-5 and the quality score provided by FaceQNet. The regressor has been trained with the objective of learning to predict the utility of a specific image for face verification purposes. Indeed, the target quality value used for training has been derived (following a novel approach) for each image from a set of genuine and impostor verification attempts.

The proposed regressor has been tested in combination with different face recognition systems (DLib, ArcFace, VGG-Face and VeriLook) to get a broad overview of its effectiveness. In line with the procedure adopted at NIST [10], performance evaluation is based on Error vs. Discard curves that analyse the FNMR trend as a function of the proportion of testing images discarded based on the quality score assigned.

The results obtained are quite encouraging; the different versions of the proposed quality regressor provide all interesting results confirming their capability of predicting a significant quality score from a set of quality indicators related to ISO/ICAO compliance controls and FaceQNet output. An analysis of the feature importance shows that the most relevant factors for quality assessment in this context are those related to exposure/lighting, head pose and blurring/focus. Indeed, even ICAO compliant images can present a (very limited) amount of such alterations and the proposed quality regressor is able to capture such small variations for image quality score prediction.

The experimental validation carried out in this paper highlighted a major issue that specifically emerges in the high-quality image scenario. Indeed, in this context the error rates measured for state-of-the-art FRSs are very low and make it difficult to adopt the standard Error vs. Discard curves to evaluate the effectiveness of quality assessment approaches; for instance, no FNM errors have been observed when a state-of-the-art FRS such as VeriLook has been applied to our high-quality images. On the one hand, this may indicate that quality assessment in this context is not so critical or relevant; on the other hand, the next ISO/IEC 39794-5 allows to store information about biometric samples quality in dedicated quality blocks in the next generation passports, thus suggesting the need for reliable quality estimation approaches on ISO/ICAO compliant images. Therefore we believe that the problem of performance assessment for high-quality images should be further studied.

Several possible extensions will be considered in our future research activity. A desirable feature for a quality assessment approach is to produce a result that is clearly explainable to humans, in terms of what factors mostly determined the predicted quality score; the method should be therefore extended to output this kind of information, that could be obtained by comparing the single quality features (input to the regressor) and the feature importance obtained during the training stage. Moreover, further investigations on printed and scanned images are certainly needed to analyze another relevant application scenario. Finally, of uttermost importance is an experimentation on real data that will be carried out in a near future.

ACKNOWLEDGMENT

This study has been carried out under a Framework Agreement between Alma Mater Studiorum - Università di Bologna and Istituto Poligrafico e Zecca dello Stato.

REFERENCES

[1] (2022). *FRVT 1:1 Verification*. [Online]. Available: https://pages.nist.gov/frvt/html/frvt11.html
[2] *Information Technology—Biometric Data Interchange Formats—Part 5: Face Image Data*, Standard ISO/IEC 19794-5, International Organization for Standardization, 2005.
[3] *Information Technology—Extensible Biometric Data Interchange Formats—Part 5: Face Image Data*, Standard ISO/IEC 39794-5, International Organization for Standardization, 2019.
[4] 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," NIST, Nat. Inst. Standards Technol., Tech. Rep. NISTIR 8382, 2021. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=920087
[5] *Information Technology—Biometric Sample Quality—Part 5: Face Image Data*, Standard ISO/IEC WD 29794-5, International Organization for Standardization, 2020.
[6] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," *ACM Comput. Surv.*, Jan. 2022.
[7] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, R. Haraksim, and L. Beslay, "FaceQnet: Quality assessment for face recognition based on deep learning," in *Proc. Int. Conf. Biometrics (ICB)*, Crete, Greece, Jun. 2019, pp. 1–8.
[8] *Biometrics, Information Technology—Biometric Sample Quality—Part 1: Framework*, Standard ISO/IEC TR 29794-1:2016, ISO/IEC JTC1 SC37, International Organization for Standardization, 2016.
[9] *Information Technology—Vocabulary—Part 37: Biometrics*, Standard ISO/IEC 2382-37:2017, ISO/IEC JTC1 SC37, International Organization for Standardization, 2017.
[10] M. N. P. Grother, A. Hom, and K. Hanaoka, "Ongoing face recognition vendor test (FRVT). Part 5: Face image quality assessment," Nat. Inst. Standards Technol. (NIST), Gaithersburg, MD, USA, Tech. Rep., 2020.
[11] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec, "Face image validation system," in *Proc. ISPA. Proc. 4th Int. Symp. Image Signal Process. Anal.*, 2005, pp. 30–33.
[12] R.-L.-V. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *Proc. Biometrics Symp., Special Session Res. Biometric Consortium Conf.*, Sep. 2006, pp. 1–6.
[13] X. Gao, S. Z. Li, R. Liu, and P. Zhang, "Standardization of face image sample quality," in *Advances in Biometrics*, S.-W. Lee and S. Z. Li, Eds. Berlin, Germany: Springer, 2007, pp. 242–251.
[14] J. Sang, Z. Lei, and S. Z. Li, "Face image quality evaluation for ISO/IEC standards 19794-5 and 29794-5," in *Advances in Biometrics*, M. Tistarelli and M. S. Nixon, Eds. Berlin, Germany: Springer, 2009, pp. 229–238.
[15] M. Ferrara, A. Franco, D. Maio, and D. Maltoni, "Face image conformance to ISO/ICAO standards in machine readable travel documents," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 4, pp. 1204–1213, Aug. 2012.
[16] M. Ferrara, A. Franco, and D. Maltoni, "Evaluating systems assessing face-image compliance with ICAO/ISO standards," in *Biometrics and Identity Management*, B. Schouten, N. C. Juul, A. Drygajlo, and M. Tistarelli, Eds. Berlin, Germany: Springer, 2008, pp. 191–199.
[17] *FVC-Ongoing: Face Image ISO Compliance Verification*, BioLab, Chennai, India, 2021.
[18] A. Khodabakhsh, M. Pedersen, and C. Busch, "Subjective versus objective face image quality evaluation for face recognition," in *Proc. 3rd Int. Conf. Biometric Eng. Appl. (ICBEA)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 36–42.
[19] P. Wasnik, K. B. Raja, R. Ramachandra, and C. Busch, "Assessing face image quality for smartphone based face recognition system," in *Proc. 5th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2017, pp. 1–6.
[20] S. Bhattacharya, C. Kyal, and A. Routray, "Simplified face quality assessment (SFQA)," *Pattern Recognit. Lett.*, vol. 147, pp. 108–114, Jul. 2021.

[21] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5651–5660.
[22] E. Tabassi, C. Wilson, and C. I. Watson, "Fingerprint image quality," NIST, Nat. Inst. Standards Technol., Tech. Rep. NISTIR 7151, Aug. 2004. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=905710
[23] F. Weber, "Some quality measures for face images and their relationship to recognition performance," in *Proc. Biometric Quality Workshop*. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2006.
[24] C. Y. Wee and R. Paramesran, "Measure of image sharpness using eigenvalues," *Inf. Sci.*, vol. 177, no. 12, pp. 2533–2552, Jun. 2007.
[25] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," in *Proc. 31st Brit. Mach. Vis. Conf. (BMVC)*. Durham, U.K.: BMVA Press, Sep. 2020.
[26] A. M. Martinez, "The AR face database," Ohio State Univ., Columbus, OH, USA, CVC Tech. Rep. 24, 1998.
[27] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, Apr. 1998.
[28] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 947–954.
[29] A. Kasinski, A. Florek, and A. Schmidt, "The PUT face database," *Image Process. Commun.*, vol. 13, nos. 3–4, pp. 59–64, 2008.
[30] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.
[31] *DeepFace*. Accessed: Jul. 22, 2022. [Online]. Available: https://github.com/serengil/deepface
[32] Neurotechnology. (2021). *Verilook Face Identification Technology*. [Online]. Available: https://www.neurotechnology.com/verilook.html
[33] (2022). *Scikit-Learn: Machine Learning in Python*. [Online]. Available: https://scikit-learn.org/stable/index.html
[34] G. Louppe, "Understanding random forests: From theory to practice," 2014, *arXiv:1407.7502*.

ANNALISA FRANCO received the Ph.D. degree in electronics, computer science and telecommunications engineering from the DEIS, University of Bologna, Italy, in 2004, with a focus on multidimensional indexing structures and their application in pattern recognition. She is currently an Associate Professor with the Department of Computer Science and Engineering, University of Bologna. She is a member of the Biometric System Laboratory (computer science), Cesena. She has authored several scientific papers and served as a referee for a number of international journals and conferences. Her research interests include pattern recognition, biometric systems, image databases, and multidimensional data structures. Her recent research activity is mainly focused on face recognition in the context of electronic identity documents.

ANTONIO MAGNANI received the Ph.D. degree in computer science and engineering from the University of Bologna, Italy, in 2020. He is currently a Postdoctoral Researcher with the European Centre for Living Technologies, University of Venice Ca' Foscari, Italy. He has published on the IEEE INTERNET OF THINGS JOURNAL, *Smart Cities*, and *Distributed Simulations*. His research interests include computer vision and machine learning techniques for ambient intelligence solutions.

DAVIDE MALTONI (Senior Member, IEEE) is currently a Full Professor with the Department of Computer Science and Engineering—DISI, University of Bologna. He is the coauthor of the *Handbook of Fingerprint Recognition* (Springer, 2009) and holds three patents on Fingerprint Recognition. His research interests include the area of pattern recognition, computer vision, machine learning, and computational neuroscience. He is the Co-Director of the Biometric Systems Laboratory (BioLab), which is internationally known for its research and publications in the field. Several original techniques have been proposed by BioLab Team for fingerprint feature extraction, matching and classification, for hand shape verification, for face location and for performance evaluation of biometric systems. He was elected as an International Association for Pattern Recognition (IAPR) Fellow, in 2010.

DARIO MAIO (Life Member, IEEE) has been a Full Professor in information processing systems with the University of Bologna, Italy. Since November 2021, he has been an Alma Mater Honorary Professor. He has been the Founder and the Co-Director of the Biometric Systems Laboratory. He has devoted his research work to various aspects of computer science, including distributed computer systems, computer performance evaluation, database design, information systems, neural networks, autonomous agents, artificial vision, and biometric systems, ICT for smart cities. He is the coauthor of about 200 research papers in peer-reviewed journals, conference proceedings and books, and has collaborated with several research institutions and industries worldwide.

LEONARDO ODORISIO received the degree in physics, in 1999. After an experience in the semiconductor industry, he was worked with the Istituto Poligrafico e Zecca dello Stato, the Italian Secure Printing House, for smart card Government applications, where he is currently working with the Innovation Department. Actually, he is working on the improvement of the issuing process for identity documents and on the improvement of production processes and quality.

ANDREA DE MARIA is currently an Innovation Manager with the Istituto Poligrafico e Zecca dello Stato, the Italian Secure Printing House. With a background in electronic engineering, he has worked on smart cards and security in fields ranging from GSM to banking and Government applications. He is also working on digital identity and on the improvement of the issuing process for identity documents.