A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Image Quality Assessment on Identity Documents

Claudio Yáñez[1], Juan Tapia[2]

Abstract: This paper developed a method for performing Image Quality Assessment (IQA) on ID-card images. First, we built the dataset, consisting of 204 images from Chilean ID cards, containing real and tampered images with varying quality levels. Then, we evaluated different features, obtaining the best results using the BRISQUE features and a newly trained SVR, with an R^2 score of 0.5868. This proposed method is called BRISQUE-ID. The IQA on ID cards can be used as a pre-processing stage for discarding lousy quality images and helping the subsequent steps in the processing pipeline.

Keywords: Image quality assessment, IQA, identity documents, ID cards, biometric sample quality.

1 Introduction

Identity Document (ID) cards are used nowadays in a wide variety of remote services, such as digital banking, government services and e-commerce, to verify the identity of customers. Ideally, access to this document would be obtained through Near Field Communication (NFC) —which would guarantee the information is read correctly and increase the difficulty of tampering—, or by scanning the document using dedicated hardware. These options are not always feasible and can be expensive. For instance, in South America, a country such as Brazil has a population of more than 200 million inhabitants, and their national ID card is chipless. However, the widespread availability of smartphones with cameras facilitates remote access to ID cards by photographing them while opening new challenges to ensure proper reading and use of the information.

In order to process an ID card remotely, the first step is to capture a digital image of it using a camera. These images are captured remotely in non-controlled scenarios, with different backgrounds, illumination, distances, and hardware qualities. Additionally, different smartphones have unique camera models. These conditions present many difficulties in the process of getting the information from the ID card. For example, if a blurry image were captured, the Optical Character Recognition (OCR) algorithm could fail reading some important data, like the person's name or the national ID number. Therefore, a method to verify the quality of the capture must be implemented, ensuring the subsequent processes operate on an image with enough quality.

According to the literature [Sc20, ZM20] image quality algorithms have focused in two main branches: Face Image Quality Assessment (FQA) which analyses face images fo-

[1] R+D Center TOC Biometrics, claudio.yanez@tocbiometrics.com
[2] da/sec-Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, juan.tapia-farias@h-da.de

cused on biometrics applications, and Image Quality Assessment (IQA) oriented to general-purpose images for perceptual quality. This perceptual quality can be objective or subjective.

This paper focuses on performing IQA on ID-card images developing a method based on the BRISQUE features [MMB12], called BRISQUE-ID. Performing IQA on ID-card images early in the pipeline can help save computational resources if low quality images are discarded, and at the same time, improve the results in subsequent stages, such as Presentation Attack Detection systems [GMF14] or OCR.

The objective of this paper is to develop a system for predicting subjective image quality on ID-card images. This is accomplished by studying multiple features for ID image quality assessment and using them to predict subjective image quality scores. An ID-card subjective IQA dataset was generated by surveying 15 subjects on the quality of 204 images, which enabled us to evaluate IQA performance.

The rest of this paper is organized as follows: Section 2 describes previous related work. Section 3 describes the dataset used and the protocol employed for obtaining subjective quality scores. Section 4 describes the experiments performed. Section 5 reports the results obtained. Finally, conclusions and future work are reported in Sections 6 and 7 respectively.

2 Related work

Image 'quality' can mean fidelity of an image to its source, utility to perform tasks related to the image —such as facial recognition or OCR—, or perceived subjective quality based on the previous meanings [AFFOG12]. Much of the work done on IQA focuses on Face Quality Assessment (FQA) or general-purpose IQA [Sc20, ZM20]. Some standards describe how biometric sample quality assessments, including FQA, should be performed [IS16]. FQA is used to ensure face pictures in ID documents have been adequately taken, ensuring high sample quality [GNH19].

In [Sc20], over 50 works on FQA were surveyed. Some of the methods shown employ measurements that are specific for evaluating faces, *i.a.* pose, location of facial features and facial expression. Additionally, the scores yielded by FQA algorithms are usually intended to predict facial recognition performance. These two factors may prevent FQA algorithms from being usable for other types of IQA.

Objective blind or No-Reference (NR) IQA refers to automatic quality assessment of an image through an algorithm that only requires the image to be assessed as input information [MMB12]. In contrast, Full-Reference (FR) or Reduced-reference (RR) IQA algorithms require a 'clean,' new reference image in the case of FR IQA, or some information about the reference image (such as a watermark or template) in the case of RR IQA [SB12]. In this sense, NR IQA methods have the advantage that they can be used in scenarios where a reference image cannot be obtained beforehand.

Mittal *et al.* developed a NR IQA model called Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [MMB12]. BRISQUE was developed considering certain regular statistical properties present in natural images —images were taken with an optical camera as opposed to synthetic images— that change by the presence of distortions and can be measured. These properties were measured by extracting locally normalized luminance coefficients, as well as pairwise products of these coefficients, and modeling them using natural scene statistics [Ru94]. These properties resulted in 18 parameters, which were extracted at two scales for a total of 36 features. Finally, a Support Vector machine Regressor (SVR) was trained using the BRISQUE features extracted from the LIVE IQA database [SSB06].

The LIVE IQA database was presented by Sheikh *et al.* as one of the most significant subjective image quality studies at the time [SSB06]. The database has 779 distorted images, generated from 29 high-resolution reference images, which included faces, people, animals, nature scenes and artificial objects, among others. Five distortion types were used to generate the distorted images: JPEG2000 compression, JPEG compression, white noise, Gaussian blur, and simulated fast fading Rayleigh channel. These distorted images were evaluated in 7 sessions using a double-stimulus methodology [Se12]. The images were evaluated by an average of 23 subjects per session, on a scale of 0–100, where 0 is the lowest possible quality and 100 the highest. The scale was divided into five equal portions with the adjectives 'Bad', 'Poor,' 'Fair,' 'Good,' and 'Excellent'. In every session, both the original and the distorted version of an image appeared, which enabled the authors to obtain a differential score for every distorted image. These scores were subject-normalized and realigned using the responses of an 8th session, yielding a Difference Mean Opinion Score (DMOS) that indicate the subjective quality of the images.

3 IQA dataset

For this proposal, an IQA dataset was constructed using 204 images taken from the database used in [GVT21]. The dataset is comprised of real Chilean national ID cards ('digital'), printed ID cards ('printed') and ID cards displayed on screens ('screen'), with 68 images per class. The images are of varied quality, ranging from completely out of focus or tampered, to well-focused or real-looking forgeries. The background was removed automatically, and then the images where resized to 320×240 pixels. Example images are shown in Figure 1.

Fig. 1: Left: High quality example. Right: Low quality example. Sensible information was covered with black boxes.

Subjective quality assessments were obtained by surveying 15 subjects. The number of subjects was due to availability, and no subject selection procedure was carried out, besides ensuring the subjects could follow the survey instructions. Out of these subjects, 6 had prior experience in ID card quality assessment or image processing, while the other 9 had no relevant prior experience.

In our attempt to perform ID-card IQA using BRISQUE, we loosely based our survey protocol on [SSB06]. The purpose and the instructions of the survey were explained to the subjects before starting. The survey was separated into three sessions, each with only one class of images to reduce inter-class biases. In each session, the subjects had to evaluate the images one at a time, on a 1-to-5 scale; the five opinion scores (OS) were labeled 'Bad,' 'Poor,' 'Fair,' 'Good,' and 'Excellent'. Subjects were instructed to consider aspects such as focus, lighting and integrity for their evaluation. In most cases, the three sessions were completed in succession, with a short rest between them. Each subject evaluated each of the 204 images, resulting in 3,060 individual quality judgments.

The OS were processed in order to obtain a single quality score for each image. First, a Normalized Opinion Score (NOS) is calculated for each image i and each subject j according to the following equation:

$$NOS_{ij} = \frac{OS_{ij} - \mu_j}{\sigma_j}$$

where OS_{ij} is the OS given by subject j to image i, and μ_j and σ_j are the mean and standard deviation of all scores given by subject j. The resulting NOS are averaged for each image, resulting in a Mean Normalized Opinion Score (MNOS):

$$MNOS_i = \frac{\sum_{j=1}^{N} NOS_{ij}}{N}, \quad N = 15$$

These scores are based on the DMOS scores of [SSB06]. However, since our survey sessions were few and performed over a short time, no realigning step was performed.

The main difference between [SSB06] and our work is in the scope of the work. The dataset generated by [SSB06] was intended for detecting specific distortions, and had images of various kinds. Our dataset contains only images of Chilean ID cards, and includes distortions due to source (photos of real ID cards, photos of printed ID cards and photos of ID cards displayed on screens) and due to image capture conditions (lighting and focus variations). Thus, our dataset is application-specific for training a Chilean ID card IQA model.

4 Experiments

Two experiments on the ID-cards dataset were conducted. First, the BRISQUE score "out-of-the-box" was used as a measure of image quality, obtaining poor results. Then, we trained an SVR using BRISQUE features and compared its performance to other features.

4.1 Experiment 1 – Out-of-the-box BRISQUE tests

Initially, we evaluated BRISQUE "out-of-the-box" (OOB) ability to predict subjective image quality on ID cards. For this, PyBRISQUE[3] implementation was used. This implementation includes the 36 luminance-based BRISQUE features, and the pre-trained SVR. The implementation results closely resemble those obtained by the original BRISQUE paper [MMB12].

The quality scores given by OOB BRISQUE (features and SVR) showed no correlation with the subjective image quality of ID cards. This result was evidenced by a Pearson Correlation coefficient of 0.0985, calculated between the OOB BRISQUE scores and the MNOS subjective scores of our dataset (scores shown in Figure 2). This is further discussed in Section 5.

4.2 Experiment 2 – Feature comparison

The following hand-crafted feature types were selected for comparison: raw pixel intensity, BRISQUE, Local Binary Patterns (LBPs)/quadrant-LBPs (QLBPs) [OPM02], Histograms of Oriented Gradients (HOG) [DT18], and discrete Fourier transform (DFT) [LLJ08]. The number of features per type is shown in Table 1. In all cases, images were transformed to grayscale and scaled to 320×240 pixels using OpenCV before extracting the features. Some of the features have multiple parameters, which are mentioned in the following section. In those cases, only the parameters that yielded the best results are reported.

Feature type	N. of features
Raw image	$76,800$
BRISQUE	36
LBPs	$1,024$
QLBPs	$4,096$
HOG	$12,000$
DFT	$76,800$

Tab. 1: Number of features per feature type.

As a baseline, the raw grayscale intensity values of the dataset images were used as features. These images correspond to automatically cropped IDs, which resulted in different sizes. For that reason, all ID cards were resized to 320×240 pixels and flattened to a $1 \times 76,800$ vector. This size kept the image ratio closest to the cropped images.

PyBRISQUE implementation allows for the raw BRISQUE features to be used instead of just obtaining an image quality score. As described previously, BRISQUE extracts 36 features —18 at two different scales— which describe the distributions of locally normalized luminance coefficients. Further on, we call the combination of BRISQUE features with the newly trained SVR BRISQUE-ID, to differentiate it from OOB BRISQUE.

[3] `https://github.com/bukalapak/pybrisque`

LBP features were selected as an attempt to use texture descriptors for predicting image quality. The sklearn Python library was used to extract LBP features. Among all LBP variants, the default and the nonrotation-invariant uniform LBPs yielded the best results. The former was tested using ten neighbors, and the latter (due to size constraints) was tested using 8, 16, and 24 neighbors. In both cases, a radius of one, two, four, and six were used. The best results were obtained using default LBP with ten neighbors and a radius of one among these combinations.

In order to preserve spatial information when using LBPs, the images were divided into quadrants, and LBPs were extracted separately for each quadrant. This is referred to as quadrant-LBPs (QLBPs). The resulting feature vector is obtained by concatenating the four resulting LBP vectors; thus, QLBPs have four times as many features as their corresponding LBPs. The same variants and number of neighbors and radii used in LBPs were explored. The best results were also obtained using default LBP with ten neighbors and a radius of one.

HOG features were selected as an attempt to use shape descriptors for predicting image quality. HOG features were extracted using the sklearn Python library. When using 8 orientations, cells of 8×8, 10×10, 12×12 and 16×16 pixels were used. When using 10 and 12 orientations (due to size constraints), only cells of 8×8 pixels were used. Among these combinations, the best results were obtained using ten orientations and 8x8 cells.

The last feature used was the discrete Fourier transform (DFT) of the ID document image. DFT has successfully been used in the past to detect blur in images [LLJ08]. Lower frequencies of the DFT image are removed by shifting the zero-frequency component to the center of the spectrum, setting the 120×120 pixels area surrounding it to zero, and shifting the zero-frequency back. This feature slightly improved the results obtained when using DFT. Additionally, every feature was scaled to the [0–1] range.

4.3 Model training

An SVR with a Gaussian kernel was used to model and predict MNOS values for ID cards. In order to reduce the chance of biases due to partitioning, accuracy was averaged over ten trials with stratified random 80/20 train/test partitions; stratification keeps the image classes balanced in every trial. Furthermore, the SVR parameters were set using five-fold cross-validation in each trial. The C and γ parameters were selected from the following options: $C : \{10^x,\, x \in [-3,1]\}$, and $\gamma : \{10^x,\, x \in [-3,0]\}$.

4.4 Metrics

Mean Squared Error (MSE), Mean Absolute Error (MAE), coefficient of determination (R^2 score), and Pearson correlation index are used to evaluate the regression model performance. The MSE metric penalizes outliers more heavily compared to the MAE metric. The R^2 score is a good indicator of how well new samples are likely to be predicted by the model, with a value of one indicating a perfect model and a value of 0 indicating a model that always outputs the expected value of y. The Pearson correlation index is used to compare OOB BRISQUE results with the features that were evaluated.

5 Results and discussion

The results of the IQA feature comparison are shown in Table 2. BRISQUE-ID yielded the best results using all three metrics. Results using BRISQUE-ID were consistently better across all ten trials, with a minimum R^2 score of 0.4754 and a maximum of 0.7870.

HOG yielded results very close to BRISQUE-ID. However, BRISQUE consists of only 36 features, whereas HOG, with ten orientations and 8×8 cells, consists of 12,000. This makes model training and prediction with HOG much slower when compared to using BRISQUE-ID.

Baseline results using the raw image as a feature yielded the worst results. Results using LBP and QLBP were the least consistent, varying drastically across trials; LBP maximum and minimum R^2 scores were 0.0603 and 0.4985, respectively. QLBP shows improvements when compared to LBPs as a result of preserving more spatial information.

	Metric			
Feature type	R^2 score	MSE	MAE	Pearson c.c.
OOB BRISQUE				0.0985
Raw image	0.1127 ± 0.0510	0.4300 ± 0.0809	0.5363 ± 0.0502	0.3862 ± 0.1018
BRISQUE-ID	$\mathbf{0.5868 \pm 0.0885}$	$\mathbf{0.1972 \pm 0.0511}$	$\mathbf{0.3278 \pm 0.0296}$	$\mathbf{0.7703 \pm 0.0748}$
LBPs	0.1927 ± 0.1228	0.3978 ± 0.1049	0.4728 ± 0.0589	0.4516 ± 0.1336
QLBPs	0.3069 ± 0.1226	0.3412 ± 0.0918	0.4312 ± 0.0610	0.5574 ± 0.1010
HOG	0.5083 ± 0.0534	0.2404 ± 0.0590	0.3891 ± 0.0468	0.7677 ± 0.0305
DFT	0.2385 ± 0.0697	0.3675 ± 0.0649	0.4995 ± 0.0433	0.5545 ± 0.0853

Tab. 2: Comparison of IQA regression results using different features.

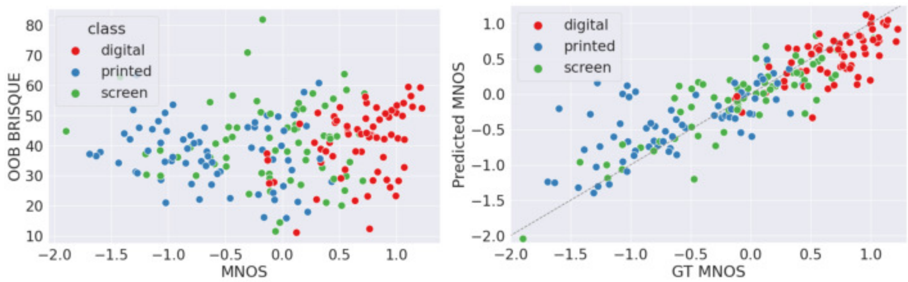

Fig. 2: Left: Comparison between OOB BRISQUE and MNOS scores on the ID-Cards dataset, showing no correlation between them. Right: MNOS, ground truth vs. predicted, using BRISQUE-ID. The dashed line represents perfect prediction. All dataset images are displayed here, although metrics were calculated only on the test partitions.

As mentioned in Section 4.1, the OOB BRISQUE scores show little correlation with the subjective quality MNOS scores. This is reflected both on the low Pearson correlation coefficient (Table 2) and on Figure 2. Note that, while MNOS scores ranged from -2.0 to 1.5 and OOB BRISQUE scores ranged from 10 to 80, correlation does not depend on the scale of the scores.

Because both OOB BRISQUE and BRISQUE-ID use the same features, the difference of results was due to the SVR. As described in Section 3, the dataset used to train the OOB BRISQUE SVR (the LIVE IQA dataset) and our dataset have a different scope. The SVR learns the features that give an image high or low quality index (DMOS or MNOS), but these features may be different in each case —i.e., a JPEG-compressed image of a tree has bad quality for different reasons than a picture of an ID card being displayed on a screen.

This could be extrapolated to more closely-related fields. ID cards of different countries, or even Chilean IDs distorted in a way that was not accounted for, may be evaluated incorrectly by our model. Our model has not yet been tested on images with different content or distortions. In those cases, training a new application-specific model may be required.

6 Conclusions

In this work, we studied image quality assessment on ID-card images. We were able to perform No Reference IQA on Chilean ID cards using BRISQUE-ID. This method performed significantly better than other hand-crafted features, such as LBPs, while using only 36 features. The BRISQUE-ID methodology could be replicated for performing IQA on other types of images, as long as the proper dataset is present. We also showed that the general-purpose OOB BRISQUE was not adequate for Chilean ID–card IQA.

7 Future work

A better and larger dataset could be constructed by surveying more subjects and a wider variety of images. The limited availability of ID-card images makes it harder to generate a large-scale dataset. However, this would improve results so that they reflect more closely real-world application performance. Additionally, using ID-card images taken in a wider range of conditions would let us assess the robustness of our method.

While we were able to perform subjective image quality prediction on ID cards, its impact on the following steps of the pipeline —such as tampering detection or OCR— remains to be studied. These steps should be done by analyzing how bad quality scores correlate with tampering detection or OCR performance.

Acknowledgments

We would like to thank all the volunteers that participated in this research. TOC Biometrics, R+D Center SR-226, supported this work.

References

[AFFOG12] Alonso-Fernandez, Fernando; Fierrez, Julian; Ortega-Garcia, Javier: Quality Measures in Biometric Systems. IEEE Security and Privacy, 10(6):52–62, 2012.

[DT18] Dalal, N.; Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). volume 1. IEEE, pp. 886–893, may 2018.

[GMF14] Galbally, Javier; Marcel, Sebastien; Fierrez, Julian: Image Quality Assessment for Fake Biometric Detection: Application to Iris, Fingerprint, and Face Recognition. IEEE Transactions on Image Processing, 23(2):710–724, feb 2014.

[GNH19] Grother, Patrick; Ngan, Mei; Hanaoka, Kayee: Face Recognition Quality Assessment Concept and Goals. National Institute of Standards and Technology (NIST), 2019.

[GVT21] Gonzalez, S.; Valenzuela, A.; Tapia, J.: Hybrid Two-Stage Architecture for Tampering Detection of Chipless ID Cards. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3(1):89–100, 2021.

[IS16] ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 29794-1:2016 Information technology — Biometric sample quality — Part 1: Framework. Standard, International Organization for Standardization, Geneva, CH, September 2016.

[LLJ08] Liu, Renting; Li, Zhaorong; Jia, Jiaya: Image partial blur detection and classification. 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008.

[MMB12] Mittal, Anish; Moorthy, Anush Krishna; Bovik, Alan Conrad: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12):4695–4708, 2012.

[OPM02] Ojala, T.; Pietikainen, M.; Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):971–987, 2002.

[Ru94] Ruderman, Daniel L: The statistics of natural images. Network: computation in neural systems, 5(4):517–548, 1994.

[SB12] Soundararajan, Rajiv; Bovik, Alan C.: RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment. IEEE Transactions on Image Processing, 21(2):517–526, 2012.

[Sc20] Schlett, Torsten; Rathgeb, Christian; Henniger, Olaf; Galbally, Javier; Fiérrez, Julian; Busch, Christoph: Face Image Quality Assessment: A Literature Survey. CoRR, abs/2009.01103, 2020.

[Se12] Series, BT: Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-13, 2012.

[SSB06] Sheikh, Hamid R.; Sabir, Muhammad F.; Bovik, Alan Conrad: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing, 15(11):3440–3451, 2006.

[ZM20] Zhai, Guangtao; Min, Xiongkuo: Perceptual image quality assessment: a survey. Science China Information Sciences, 63(11):1–52, 2020.