Deep Learning-Based Automated Detection of Inappropriate Face Image Attributes for ID Documents

Amineh Mazandarani, Pedro Amaral, Paulo Da Pinto, S. Hosseini

Deep Learning-Based Automated Detection of Inappropriate Face Image Attributes for ID Documents

Amineh Mazandarani[1], Pedro Miguel Figueiredo Amaral[1], Paulo da Fonseca Pinto[1], Seyed Jafar Hosseini Shamoushaki[2],

[1] Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisbon, Portugal
[2] PhD Graduate from University of Coimbra, Senior R&D Specialist, Lisbon, Portugal

a.mazandarani@campus.fct.unl.pt, {pfa, pfp}@fct.unl.pt, jafar@isr.uc.pt

Abstract. A face photo forms a fundamental element of almost every identity document such as national ID cards, passports, etc. The governmental agencies issuing such documents may set slightly different requirements for a face image to be acceptable. Nevertheless, some are too critical to avoid, such as mouth closedness, eyes openness and no veil-over-face. In this paper, we aim to address the problem of fully automating the inspection of these 3 characteristics, thereby enabling the face capturing devices to determine, as soon as a face image is taken, if any of them is invalid or not. To accomplish this, we propose a deep learning-based approach by defining model architectures that are lightweight enough to enable real-time inference on resource-constrained devices with a particular focus on prediction accuracy. Lastly, we showcase the performance and efficiency of our approach, which is found to surpass two well-known off-the-shelf solutions in terms of overall precision.

Keywords: Image Quality Verification, Deep Learning, Face Detection, Binary Classification.

1 Introduction

It has always been quite common for an identity document to contain a face image of its holder to facilitate the recognition of his/her identity. Consequently, the face is known to be the most important biometric trait and, in respect to other traits, offers several advantages such as: non-intrusive acquisition; minimal hardware requirements thanks to camera technology advances; and easy capture process without any intervention of the individual. Over the past decade, electronic ID documents started to replace their traditional equivalents and thus information comes embedded into an internal chip, whether it is demographic details or biometric features like the face image. As a result, AI-powered biometrics inspection has been enabled for automated real-time identity verification [1], [2].

In 2002, the International Civil Aviation Organization (ICAO) asked a group of experts to establish a specific physical feature of an individual as a biometric

identifier that can be read by a machine to confirm his/her identity [3]. The group made the decision and selected the face as the primary globally interoperable biometric feature for machine-assisted identity verification (MAIV) in machine readable travel documents. The face typically poses some limitations in comparison with other features (fingerprint, for example) if face images do not fulfill minimum quality requirements, suggesting that for highly successful MAIV the images are required to meet certain strict quality standards. Therefore, following on from that decision, the ISO/IEC 19794–5 standard [4] defines a set of rules for a proper face image acquisition procedure along with scene conditions and proposes some guidelines and plenty of examples to demonstrate acceptable/unacceptable face images to help in interpreting the image quality levels such as Blurred, Looking Away, Too Dark/Light, Eyes-Closed, Washed-Out, Hair-Across-Eyes, Mouth-Open, Unnatural Skin Tone. It is worth mentioning that here the concept of image quality no longer necessarily reflects its classic meaning (i.e. only a noisy image presents bad quality) but also will encompass other criteria such as the mouth must be closed. In this sense, several businesses delivering commercial biometrics technology began to release SDKs that serve to auto-verify the compliance of face images with ISO/ICAO standards for the document issuing process.

Over the years, there has been thorough research on image quality assessment which is often an early processing stage for vision-based applications. However, the research reported in the literature is limited in terms of ISO/ICAO requirements. Authors in [5] define 17 quality requirements and discuss basic approaches to assess them. All in all, their work is mostly concerned with the impact of image quality on the facial recognition accuracy, not with ISO/ICAO compliant face analysis. In [6], the authors propose a different set of requirements together with the corresponding evaluation algorithms and present a series of tests on 189 images from people within their organization despite most of the images being incompliant with ISO/ICAO specs. A web-based system built on 28 ISO/ICAO requirements is proposed in [7]. This paper however lacks a complete experimental study. In [8] a potential framework is briefly described to determine the compliance level of each ISO/ICAO requirement by measuring quality scores that are almost always in line with the human perception, according to the authors.

There exist several other papers in the literature that only cover a few specific requirements. For example, the work presented in [9] exploits geometric attributes of the face. Each attribute is characterized by a numeric score and all the individual scores are then merged into a single global score. [18] tackles the constraints associated with lighting, sharpness and head pose by applying Gabor filters and Discrete Cosine Transform. The work in [10] addresses “eyes closed, red eyes and looking away”, [11] takes into consideration “unnatural skin tone, shadows across the face and flash reflection on skin” and [12] deals with “pixelation, hair across eyes, veil over face and mouth open”.

2 Contribution to Applied Artificial Intelligence Systems

The authorities that issue identity documents such as national ID cards, driving licenses, passports, etc., use printed and/or digital face photos from the citizens. As stated earlier, ICAO compliance guidelines serve as a global reference to be used by these authorities to check photos for quality correctness. Three of these guidelines must always be met, namely, mouth closedness, eyes openness and no veil-over-face. In this paper, we focus on the inspection of these events and attempt to automatically determine if any of them is absent on a face image. For this purpose, we propose a processing pipeline that incorporates multiple modules such as raw face data capturing, face detection, facial landmark extraction, cropping and/or scaling, image quality analysis. This last module is, in fact, the focus of this paper, aiming to verify the three quality attributes mentioned above although we build up and evaluate the entire pipeline to ensure that our testing conditions will closely resemble real-world scenarios. The previous works mostly proceed with such verification by applying classic machine learning methods despite all the size, accuracy and speed constraints of these. We will also propose a classification-based approach but, in our case using deep learning (DL), to solve the multi-classification problem. We define DL network architectures, efficient enough to extract representative features with lighter computations to make fast and accurate predictions on mouth closedness, eyes openness and no veil-over-face. The remainder of the paper is organized as follows: Firstly, an overview of the pipeline is provided. Next, we discuss the DL networks and model specifications for quality analysis; Lastly, we cover the DL relevant parts (e.g. training), and a comprehensive set of experiments is presented to validate the accuracy of the DL models as well as the performance of the overall pipeline based on the results obtained.

3 Quality Analysis Pipeline

We now outline how we approach the problem of face image quality inspection by means of a pipeline designed to perform a complete computer vision workflow. Figure 1 illustrates the structure of our proposed pipeline as a block diagram. An overall description of this pipeline is provided below.

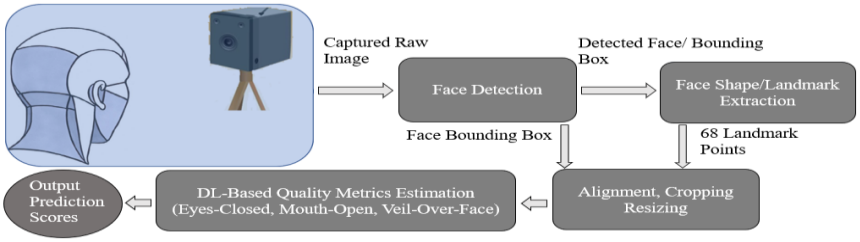

Fig. 1. The sketch of the proposed pipeline.

Our pipeline starts by capturing a person's face through a monocular RGB camera and can flexibly operate in two modes 1- video streaming (i.e. each frame is processed) 2- one single still image at a time. There is no strict restriction to the camera choice, provided that the region tightly spanning the face has a minimum resolution of 100 x 100. The subject being photographed should stand at a range of 0.5 m from the camera lens, with the face being frontal relative to the camera direction. A frame is then recorded and transmitted to the subsequent module in the pipeline. This module is responsible for face detection which involves finding the face in the input image. Face detection was handled in traditional computer vision using classical feature-based techniques such as the cascade classifier [14], however in the past few years the use of deep learning has been preferred, mainly because they achieve improvements in both accuracy and speed when applied on standard benchmark face detection datasets. One of the most popular DL-based face detectors is the Multi-task Cascade Convolutional Neural Network, or MTCNN for short [15]. This network employs a cascade architecture combining three networks: *proposal network* makes candidate proposals of face regions; *refine network* removes false bounding boxes; and the *third network* estimates facial landmarks. Implementing the MTCNN architecture is not easy due to its complexity, but open-source implementations are available for public use and can be used to train a custom DL model of our own on a dataset of face images. The face detection module outputs the coordinates of a bounding box that nicely encloses the face region. The next module deals with the extraction of facial landmarks. For this purpose, we apply the dlib toolkit which is quite popular and widely used in a range of computer vision applications [16]. We select dlib over MTCNN for landmark extraction considering that its face shape, which consists of 68 landmark points, better represents a facial structure and is also estimated with great precision by the toolkit´s shape prediction functionality. All we need to do is to train the dlib shape predictor with the detection output of the MTCNN face detector. You can see the preset shape landmarks in Figure 2. The face bounding box and the detected landmarks are passed together onto another module as input (Figure 2 also shows some shape estimates with only the landmarks in use). This module utilizes this information to identify, crop and rescale the regions of interest to be processed by the quality analysis module which is meant to verify the face image quality with respect to mouth-open, eyes-closed and veil-over-face attributes.

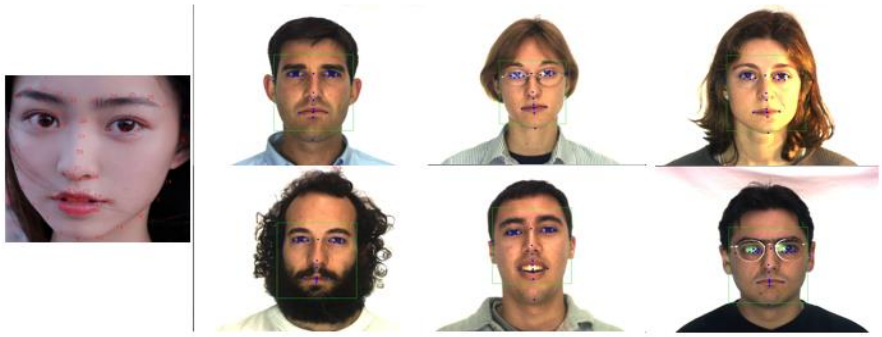

Fig 2. The single face on the left represents the pre-defined dlib shape landmarks. On the right the pipeline´s result is shown i.e. a bounding box and the relevant landmarks.

4 DL-Based Quality Analysis

The quality analysis here is equivalent to one computational metric for each of the 3 attributes. The core algorithm of the metric is implemented as a binary classification. Thus, we will end up with three image classification tasks for eyes-closed, mouth-open and veil-over-face and 3 DL models are derived to extract highly discriminative information depending on the category/object being classified.

4.1 Mouth-Open/Eyes-Closed Classification

Figure 3 shows some example training images for both the categories of mouth-open and eyes-closed. These images illustrate the two possible classes i.e. presence and absence of a category which we call positive and negative samples respectively for simplicity's sake. While the samples may look very similar (as they all contain the mouth or the eyes) subtle differences will arise in texture and pixel-level variations when the mouth or the eyes make the transition from one state to the other. These differences must be precisely captured by the convolutional layers of the deep network we want to construct. VGGNet [17] is particularly common in classification tasks. It makes an improvement over AlexNet [20] by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3X3 kernel-sized filters one after another. Multiple smaller size kernels stacked together do better than one larger size kernel because they lead to an increase in the network´s depth, allowing this network to learn more complex visual features and at a lower computational cost as well. While VGGNet achieves remarkable accuracy it has two major limitations: 1- painfully slow to train; 2- the network weights are quite large in terms of disk/bandwidth; Due to its depth and number of fully-connected nodes, VGG is over 533MB for VGG16 and 574MB for VGG19, meaning that deploying VGG is a tedious task. For this reason, smaller network architectures are often preferable such as SqueezeNet which is a lightweight network that was designed as a more compact replacement for AlexNet. It has almost 50x fewer parameters than AlexNet, yet it performs 3x faster. This architecture was proposed in [13] and features a mini-architecture called fire module that is composed of *squeeze* and *expand* layers. A squeeze convolutional layer has only 1 × 1 filters. These are fed into an expanded layer that has a mix of 1 × 1 and 3 × 3 convolution filters. Our aim is to build a network that best matches the specificities of our classification task: a) we are dealing with such small images that the choice of a very deep network like Resnet [18] (or Inception [19]) can´t be justified; b) the computer vision pipeline must be capable of being smoothly executed on low-power, resource-constrained devices, and be flexible enough for deployment to such devices without any struggle to get it running at the edge. For this purpose, we need to obtain high-performance models; c) Also the final models must produce relatively precise predictions and so there is a need to make a trade-off between inference speed and prediction accuracy.

As a result of the above, we propose a deep neural network architecture that combines the strengths that the networks VGGNet and SqueezeNet offer. In this sense, we propose to make a major adjustment to the output layers by removing the

fully connected layers from VGGNet7 (which is the lightest variant of VGGNet) and instead add a 1 × 1 convolutional layer and a global average pooling over the 2D feature maps just like in SqueezeNet plus a sigmoid activation. Aside from this, small tweaks will be adopted, for example the application of dropouts across the network. Note: integration of the fire module into other architectures can be considered a potential design strategy to attain better hybrid networks.

4.2 Veil-Over-Face Classification

Veil-over-face category has a somewhat different nature from the other two categories, as an external object (i.e. non-body part) is also part of the positive samples. A veil can be thought of as a form of occlusion given that the face gets partially concealed or at least it is barely visible with a wedding veil for example. Refer to Figure 3 for some example images. Consequently, veil-over-face classification is analogous to classifying a specific type of occlusion. In this case, we will have a larger region of interest than in the previous classification tasks and there is also a greater data variability in the samples. This somehow forces us to adopt a deeper network and thus we will pick VGGNet16 and follow the same correction we made for eyes-closed and mouth-open classification above. The architecture we will apply here is similar but will have two extra building blocks (i.e. SqueezeVGG). A face mask is also an obvious conflict with a correct ID photo and may be assumed as a sort of veil in some cases. We have developed a tool to generate synthetic masks laid over real faces and can benefit from this at least during the training process by increasing our training samples with synthetic images.

5 Experiments and Validation

The proposed pipeline is implemented as a software package in connection with an RGB camera and can run on an embedded board (Raspberry Pi, for example). Here, we avoid irrelevant details about this package and will concentrate specifically on describing the machine learning models developed and trained for use throughout the pipeline. At first, we trained a DL model for the MTCNN face detector. Next, we trained the dlib shape predictor by using as input the detection bounding boxes predicted by the face detector. Then, we need to train 3 more DL models for the quality analysis module. The specifications of the training process are given in Table 1.

Table 1. Overview of the training specs for the classification tasks: mouth-open, eyes-closed and veil-over-face.

	Mouth-Open	**Eyes-Closed**	**Veil-Over-Face**
Platform / Framework	Python/Keras (Tensorflow backend)	Python/Keras (Tensorflow backend)	Python/Keras (MXNet backend)
Model Architecture	SqueezMiniVGG	SqueezeMiniVGG	SqueezeVGG
Dataset	~10000 Pos, ~10000 Neg	~8000 Pos, ~8000 Neg	~20000 Pos, ~20000 Neg
Validation	Yes (0.1 of the dataset)	Yes (0.1 of the dataset)	Yes (0.15 of the dataset)

Input Image Size	72×72	56×56	128×128
Image Type	Grayscale	Grayscale	Grayscale
Channels First	Yes	Yes	Yes
Classification Type	Binary decision	Binary decision	Binary decision
Number of Epochs	100	100	100
Batch Size	32	16	8
Use of Augmentation	Yes (Rescale, Rotation, Flip, etc.)	Yes (Rescale, Rotation, Flip, etc.)	Yes (Rescale, Rotation, Flip, etc.)

5.1 Data Pipeline

We collected a sufficiently large number of face images from different sources for each classification task. For our data collection pipeline, we can simply reuse the proposed pipeline after making minor adjustments to it as follows: the input image is loaded from disk rather than using a camera and obviously the quality analysis module must be removed. The remaining part can act as our data collection pipeline, through which the image is passed to obtain cleanly cropped/resized portions with the region of interest. For mouth-open, eyes-closed and veil-over-face, we collected a training dataset of about 20000, 16000 and 40000 images respectively with a balanced distribution between positive and negative samples (see Figure 3 for sample images). The veil-over-face dataset contains, in addition to the real images, images of real faces with a synthetic mask acquired by a custom tool intended to operate as mask augmenter. We performed model training separately for the 3 classification networks while applying data augmentation and validation set.

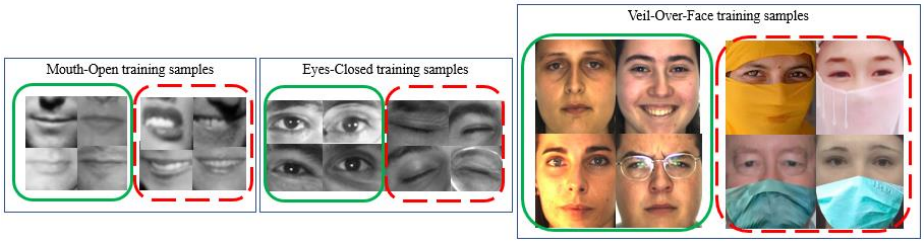

Fig 3. Examples of the training samples.

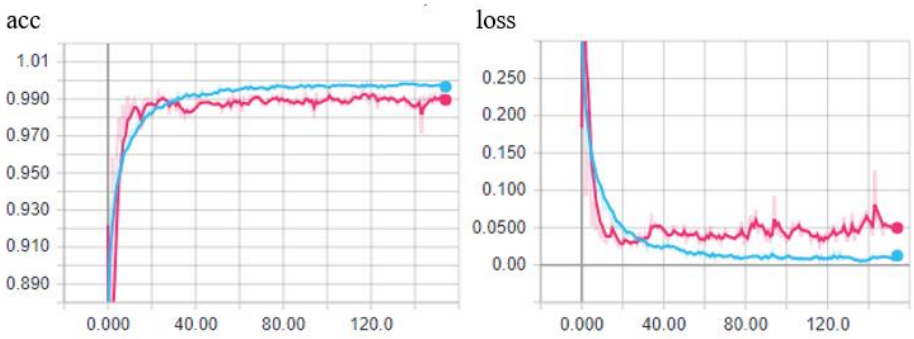

Fig 4. Accuracy/Loss plots for eyes-closed (Blue curve refers to training and Red curve to validation).

5.2 Training Results

We successfully finished the training process without facing issues like overfitting and the model has shown to generalize well. The Table 2 shows the key measures of the training output, such as final loss, accuracy, Precision and Recall.

Table 2. Presentation of key training measures.

	Mouth-Open		**Eyes-Closed**		**Veil-Over-Face**	
	Training	Validation	Training	Validation	Training	Validation
Loss	0.003	0.080	0.001	0.050	0.002	0.06
Accuracy	0.965	0.934	0.995	0.990	0.990	0.980
Precision	93%		94%		94%	
Recall	95.5%		97%		97%	

In Figure 4, we included the training graphs for eyes-closed as an illustration of the training evolution. As you can see in the Accuracy plot, the validation curve smoothly follows the training curve and both curves show a similar behavior, which is an indication of good generalization and no overfitting problem.

5.3 Model Testing and Evaluation Experiments

To assess the performance of the three trained models, we conducted a set of benchmarks on test datasets which are described in Figure 5. The test images do not overlap with the training datasets. For testing purposes, we execute our entire pipeline after excluding the camera and alternatively, we read stored images from disk. The images go through each module to the next until the three-quality metrics are calculated by performing inference with the models. Then, by running our evaluation framework, we compute all the relevant performance measures (see Figure 5).

	Mouth-Open		**Eyes-Closed**		**Veil-Over-Face**	
Type of samples	Positive samples	Negative samples	Positive samples	Negative samples	Positive samples	Negative samples
Number of samples	2148	19302	156	21334	1882	19612

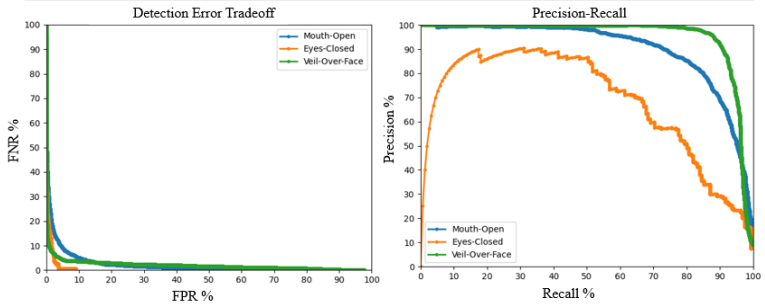

Fig 5. DET and Precision-Recall Plots – A table with the distribution of test images

Basically, there is a large proportion of negative samples with respect to positive samples. The prediction scores span the 0 to 1 range and can be used to generate detection-error-tradeoff (DET) curves which represent the relationship between the false positive rate (FPR) and false negative rate (FNR) as they vary from 0% to 100%. We are mainly concerned with both the lowest FPR and the lowest FNR and the point where we can get both at the same time is referred to as prediction threshold or aka

operating point. The lower these rates the better the prediction accuracy will be. The Precision-Recall curves are presented too. Given fewer positive samples than negative samples, having a lower Precision at the operating point totally makes sense. For mouth-open, for an operating point of 0.072 we obtained 5.12% and 5.68% for FNR and FPR along with a Recall of 94.05%, which sounds satisfactory. For eyes-closed, each eye is classified independently of the other and we only select the higher of the two scores. The operating point is estimated to be 0.0213 for an FNR of 2.12%, an FPR of 2.58% and a Recall of 98.08% which again seems good. For veil-over-face, we obtained an FNR of 4.21% and an FPR of 4.68% at the operating point of 0.246 as well as a Recall of 96.25%.

We also compare the prediction accuracy of our models against metric implementations from two major biometrics solution suppliers (Cognitec and Neurotechnology [21]). However, the veil-over-face metric implementation was not available for evaluation. Table 3 shows the comparison details indicating the ability of our models to produce quite competitive results.

Table 3. Comparison of FAR/FRR with off-the-shelf solutions.

	Cognitec		**Neurotechnology**		**Our Solution**	
	FPR	FNR	FPR	FNR	FPR	FNR
Eyes-Closed	1.53	1.52	3.99	3.87	2.58	2.12
Mouth-Open	22.60	22.64	14.38	14.14	5.68	5.12

5.4 Pipeline Assessment

The failure of any of the three models to make a correct prediction is solely caused by the limitations of the model itself on the assumption that the rest of the pipeline is performing quite well. An important issue that arises here is the fact that if the other modules, including face detection/landmark extraction, exhibit a malfunction and do not produce the desired results for the analysis module to use, this will introduce an extra error in the model performance, letting us conclude that the misclassification cases are partially influenced by the model input. From our benchmark results, we realized that this situation could be only serious for no veil-over-face classification where, for example, the face detector is at risk of missing a face if the face is heavily covered, and the face missing rate was estimated to be about 0.1% while for mouth-open and eyes-closed this is so negligible (0.005%).

6 Conclusions

In this paper, we presented a real-time computer vision pipeline for inspecting a captured face photo meant for use in ID documents. The automated inspection is aimed at verifying if any of the three attributes *mouth-open*, *eyes-closed*, *veil-over-face* is present on the image or not. This pipeline is functional and can potentially be integrated into any face capturing system. We have particularly focused on the pipeline´s last module which consists of a deep-learning multi-classification approach to the design of three individual metrics for the three attributes in question. Lastly, our benchmark results demonstrated the effectiveness of our proposed approach at dealing

with face image quality verification which clearly has yet to be further developed with future improvements.

References

1. Bourlai, T., Ross, A., Jain, A.K.: On matching digital face images against passport photos. in *Proc. IEEE Int. Conf. Biometrics, Identity and Security*, Tampa, FL, (2009)
2. Bourlai, T., Ross, A., Jain, A.K.: Restoring degraded face images for matching faxed or scanned photos. *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 371–384, (2011)
3. Biometric Deployment of Machine Readable Travel Documents, ICAO, (2003)
4. ISO International Standard ISO/IEC JTC 1/SC 37 N506, Text of FCD 19794–5, Biometric Data Interchange Formats—Part 5: Face Image Data, (2004).
5. Hsu, R.L.V., Shah, J., Martin, B.: Quality assessment of facial images. in *Proc. Biometric Consortium Conf.*, pp. 1–6, (2006)
6. Subasic, M., Loncaric, S., Petkovic, T., Bogunovic, H., Krivec, V.: Face image validation system. in *Proc. Int. Symp. Image and Signal Processing and Analysis*, pp. 30–33, *(2005)*
7. Gonzalez-Castillo, O.Y., Delac, K.: A web based system to calculate quality metrics for digital passport photographs. in *Proc. 8th Mexican Int. Conf. Current Trends in Computer Science*, pp. 105–112, (2007)
8. Ferrara, M., Franco, A., Maio, D., Maltoni, D.: Face Image Conformance to ISO/ICAO Standards in Machine Readable Travel Documents. in *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 4, pp. 1204-1213, (2012)
9. Han, Q., Gonzalez, Y., Guerrero, J.M., Niu, X.: Evaluating the content-related quality of digital ID images. in *Proc. Congress on Image and Signal Processing*, pp. 440–444, (2008)
10. Borges, E.V.C.L., et al.: Analysis of the Eyes on Face Images for Compliance with ISO/ICAO Requirements. *SIBGRAPI*, Sao Paulo, pp. 173-179. (2016)
11. Andrezza, I.L.P., et al.: Facial Compliance for Travel Documents. *SIBGRAPI*, Sao Paulo, pp. 166-172, (2016)
12. Parente, R.L., et al.: Assessing Facial Image Accordance to ISO/ICAO Requirements. *SIBGRAPI*, Sao Paulo, 2016, pp. 180-187, (2016)
13. Iandola, F., et al.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and textless1MB model size. In ArXiv, (2017)
14. Viola P., Jones, M.: Rapid object detection using a boosted cascade of simple features. *Proceedings of CVPR.* (2001)
15. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. in *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, (2016)
16. dlib: A toolkit for making real world machine learning and data analysis applications in C++, http://dlib.net/
17. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, pp. 730-734, (2015)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, (2016)
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. in CVPR pp. 2818-2826, (2016)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097-1105, (2012)
21. Cognitec: https://www.cognitec.com/, neurotechnology: https://www.neurotechnology.com/