BioLab-ICAO: A new benchmark to evaluate applications assessing face image compliance to ISO/IEC 19794-5 standard

Conference Paper *in* Proceedings / ICIP ... International Conference on Image Processing · November 2009

DOI: 10.1109/ICIP.2009.5414000 · Source: DBLP

5 authors, including:

Davide Maltoni
University of Bologna

Annalisa Franco
University of Bologna

Matteo Ferrara
University of Bologna

Dario Maio
University of Bologna

BIOLAB-ICAO: A NEW BENCHMARK TO EVALUATE APPLICATIONS ASSESSING FACE IMAGE COMPLIANCE TO ISO/IEC 19794-5 STANDARD

Davide Maltoni, Annalisa Franco, Matteo Ferrara, Dario Maio, Antonio Nardelli

C.d.L. Scienze dell'Informazione - Università di Bologna, via Sacchi 3, 47023 Cesena, ITALY
DEIS – Viale Risorgimento, 2 – 40126 Bologna, Italy.
E-mail: {davide.maltoni, annalisa.franco, matteo.ferrara, dario.maio, antonio.nardelli}@unibo.it

ABSTRACT

This work focuses on performance assessment of software applications designed to evaluate the compliance of a face image to the ISO/ICAO standards for machine readable travel documents. In this paper we describe the new large database (of compliant and non-compliant images) we gathered, the associated testing protocol and the preliminary results measured on some existing algorithms.

Index Terms — ICAO, ISO/IEC 19794-5, face, machine readable travel documents.

1. INTRODUCTION

Recent evaluation campaigns have clearly shown that face recognition algorithms provide satisfactory results only in the presence of images acquired under controlled pose and illumination. The achievement of reliable results in large-scale identification is thus subject to the availability of standardized, high-quality images. For the application of face recognition to Machine Readable Travel Documents, some rules and encoding formats were initially proposed by the International Civil Aviation Organization (ICAO) and successively adopted by the International Standard Organization (ISO). In fact, the ISO/IEC 19794-5 standard [1] specifies a record format for encoding, recording and transmitting the facial image information and defines scene constraints, photographic properties and digital image attributes of facial images. Although the standard provides some generic guidelines and several examples of acceptable/unacceptable face images, a clear, detailed and unambiguous description of all the requirements is still not available.

In view of the widespread adoption of the new standard, some vendors of biometric technologies started to develop and distribute software applications (SDKs) able to automatically verify the compliance of a face image to the ISO standard. At today, no independent and systematic evaluation of these algorithms has been performed.

A preliminary study conducted by some of the authors of this paper has been proposed in [2] where a set of 30 well defined characteristics, related to geometric (e.g., eye location and distance) and photographic (e.g., focus, contrast) properties of the face image has been defined(see Table 1). However, the database used in that work was still incomplete due to the lack of non-compliant images for many characteristics. Starting from the same set of characteristics and requirements already introduced in [2], a more in-depth evaluation is here carried out; in particular, two commercial software and a prototype developed in our laboratory have been tested on the new large database where all the characteristics are now represented.

2. DATABASE AND TESTING PROTOCOL

2.1. Database

One of the main limitations of [2] was related to the database used for testing that did not include images non-compliant to the ISO standard for some characteristics (e.g., hat/cap, veil over face). That database has been significantly extended and all the possible image "defects", as listed in Table 1, are now adequately covered. The new database (called BioLab-ICAO) consists of 7740 images gathered from different sources:

1) the whole AR Face dataset [3] (3314 images);
2) part of the images of the FRGC database [4] (2665 images);
3) some images internally acquired (943 images);
4) some "artificial" images (818 images).

The artificial images have been generated, through image processing operations starting from some full compliant images, to cover "pixelation" and "washed out" defects. This was necessary because these defects were not present in the databases considered and these kind of images are difficult to acquire with a "working" standard camera.

Throughout the database the image size varies from 762×564 to 2272×1704. Each image has been manually labeled and some additional information is added: in particular, the compliance to each characteristic is manually indicated by using a three state logic (compliant, non-compliant and dummy[1]) and some markers indicating the

[1] The dummy state is used in case of uncertainty for a specific characteristics.

position of the main facial features (eye center and corners, nose center, left and right nostrils, mouth corners and upper and lower lip center) are defined with the help of an ad-hoc graphical tool.

Table 1 - Tests defined to evaluate systems for ISO/IEC 19794-5 compliance check. The last column of the table (Section) denotes the section of **[1]** from which the test was derived.

N°	Description of the test	Section
	Feature extraction accuracy tests	
1	Eye Location Accuracy	
2	Face Location Accuracy (other points)	
	Geometric tests (Full Frontal Image Format)	
3	Eye Distance (min 90 pixels)	8.4.1
4	Relative Vertical Position (0.5B<=BB<=0.7B)	8.3.3
5	Relative Horizontal Position (no tolerances)	8.3.2
6	Head Image Width Ratio (0.5A<=CC<=0.71A)	8.3.4
7	Head Image Height Ratio (0.7B<=DD<=0.8B)	8.3.5
	Photographic and pose-specific tests	
8	Blurring	7.3.3
9	Looking Away	7.2.3
10	Ink Marked/Creased	A3.2.3
11	Unnatural Skin Tone	7.3.4
12	Too Dark/Light	7.3.2
13	Washed Out	7.4.2.1
14	Pixelation	A3.2.3
15	Hair Across Eyes	A3.2.3
16	Eyes Closed	7.2.3
17	Varied Background	A2.4
18	Roll/Pitch/Yaw Greater 5	7.2.2
19	Flash Reflection on Skin	7.2.10
20	Red Eyes	7.3.4
21	Shadows Behind Head	A3.2.3
22	Shadows Across Face	7.2.7
23	Dark Tinted Lenses	7.2.11
24	Flash Reflection on Lenses	7.2.11
25	Frames too Heavy	A4.3
26	Frame Covering Eyes	7.2.3
27	Hat/Cap	A3.2.3
28	Veil over Face	A3.2.3
29	Mouth Open	7.2.3
30	Presence of Other Faces or Toys too Close to Face	A3.2.3

The database contains 272 images compliant to all the characteristics (full compliant, see Figure 1); the number of non-compliant images for each characteristic is given in Figure 2 and some examples are shown in Figure 3. At least 200 non-compliant images are available for each characteristics in 8..30.

2.2. Testing Protocol

The image database is partitioned into 24 sub-databases, each related to a specific characteristic. The sub-database used to evaluate eye localization accuracy contains all the images where both the pupils center are visible (6176 images). For all the other tests (8-30), the number of images n_i included in the sub-database related to the characteristic i is: $n_i = 2 \times \min(nc_i, fc)$ where nc_i is the number of images non-compliant to the characteristic i and fc is the number of full compliant images. The choice of using the same number of compliant and non-compliant images is aimed at defining a balanced test [5].

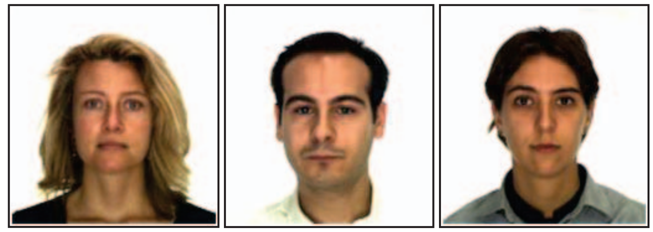

Figure 1 – Three examples of full compliant images (cropped for lack of space).

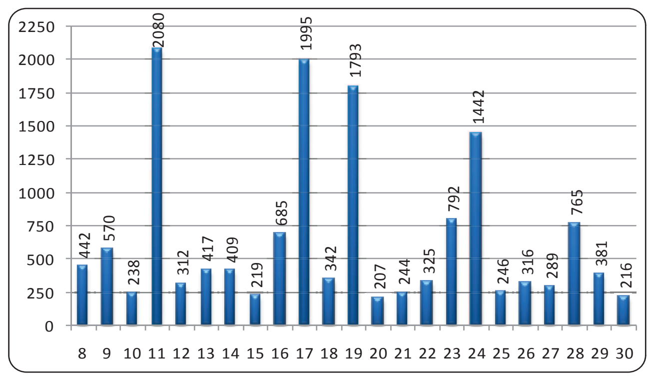

Figure 2 - Number of non-compliant images available in the database for the characteristics 8-30.

Analogously to a biometric verification systems the SDKs can make two types of errors:

1) declaring compliant with respect to a given characteristic an image that is non-complaint (False Acceptance);
2) declaring non-compliant an image that is compliant (False Rejection).

Images labeled as dummy for a given characteristic are excluded from the corresponding test. According to this protocol, the results are reported for each characteristic in terms of EER and rejection rate. A rejection occurs when either the SDK is not able to process an image or the image is processed but the SDK is not able to evaluate the specific characteristic. According to the best practices the rejection is here included in the calculation of EER [5]: this is implicitly done by, assuming that a 0 compliance degree (for the given characteristic) is returned in case of rejection. This choice is aimed at discouraging the software to reject the most uncertain cases thus improving the performance over processed images. It is worth noting that each sub-database is related to a specific characteristic and that consequently only the SDK response for that characteristic is considered for EER calculation.

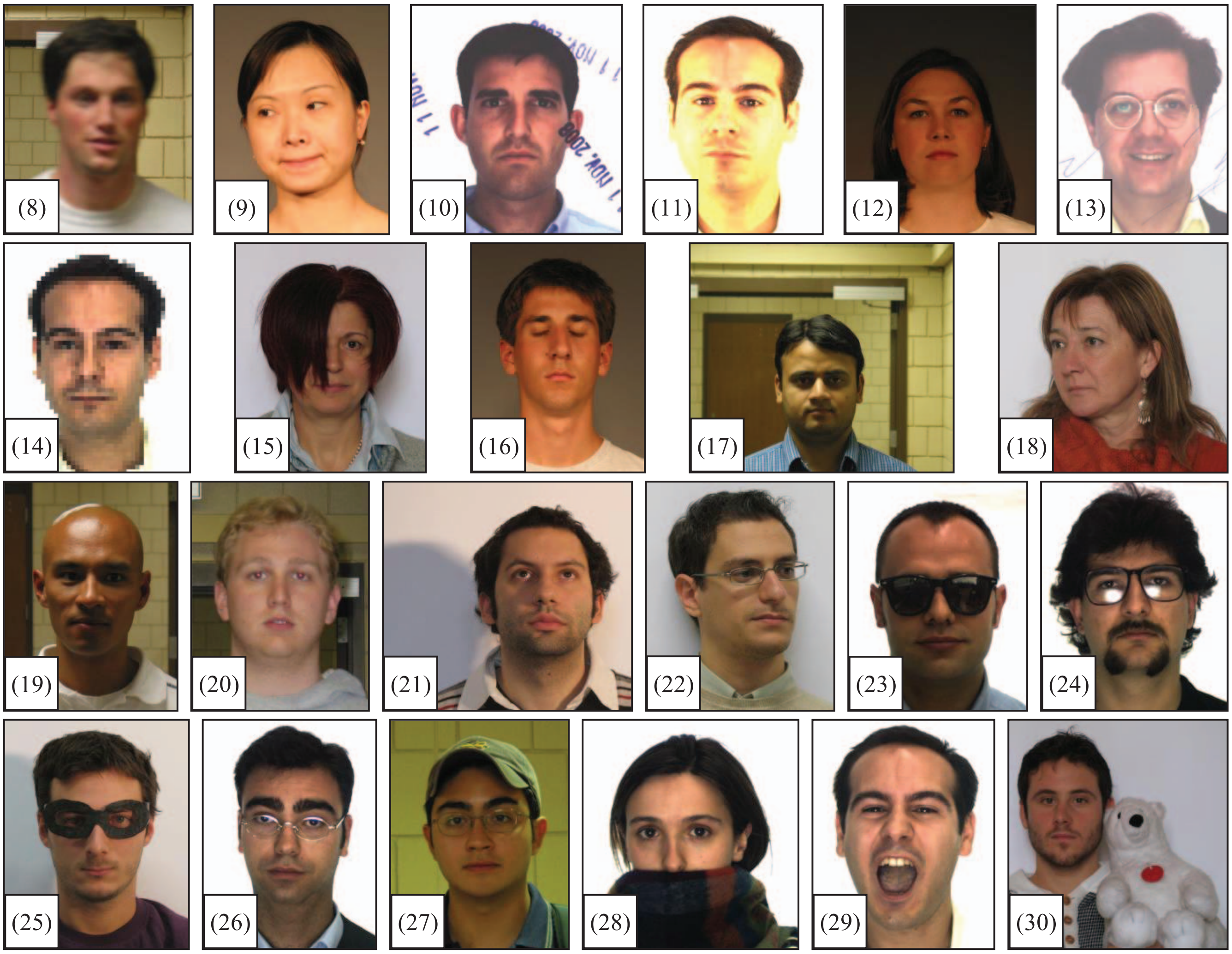

Figure 3 – An example of non-compliant image for the characteristics 8..30 listed in Table 1: the labels indicate the number of the related characteristic. The images are here cropped and zoomed to better show the details of the non-compliant characteristic.

3. EXPERIMENTS

Two commercial SDKs, whose names cannot be disclosed at this stage, and a prototype system internally developed in our laboratory have been evaluated on BioLab-ICAO database. The three software will be referred to as SDK1, SDK2 and SDK3 in the rest of the paper.

The experiments are focused on characteristics 1 (eye location accuracy) and 8..30 (see Table 1); the geometric tests 2..7 are not included in this study, because of the non-uniform way the different SDKs provide in output details about the location of internal face-feature.

As to the geometric requirements, the eye localization accuracy of the SDKs is shown in Table 2. The columns refer to increasing intervals of localization errors (percentage value with respect to the real distance between the two eyes). The column "Correct" includes all the cases where the maximum error (over the two individual errors for the two eyes) is lower than 5%; the column "Reject" refers to the images not processed by a SDK. The result for SDK 2 is not reported since it does not output the eye position. The two SDKs achieve a good localization accuracy; most of the errors are related to specific image characteristics that make critical the eyes localization (e.g. subject looking away).

Table 2 - Eye localization accuracy given both as number of images and percentage with respect to total number of images (6176) used in this test. The values defining the intervals are expressed in percentage with respect to the real distance between the two eyes.

SDK	Rej.	Correct	[5;9[	[9;13[	[13;17[	[17;21[	[21;∞[
1	62 (1.00%)	5052 (81.80%)	289 (4.68%)	50 (0.81%)	28 (0.45%)	19 (0.31%)	676 (10.95%)
3	3 (0.05%)	4444 (71.96%)	363 (5.88%)	263 (4.26%)	183 (2.96%)	178 (2.88%)	742 (12.01%)

The results obtained by the three SDKs on tests 8..30 are reported in Table 3 where the EER and rejection rate are given. The rejection rate is in most cases quite low, but it is worth noting that this value for SDK1 is noticeable for some characteristics (e.g. hair across eyes, frames covering eyes, mouth open). For a further comparison of the three SDKs, the results in terms of EER shown in Table 3 are summarized in Figure 4 where the EER distribution for the three SDKs is reported. Four EER intervals have been defined and each bar of the graph represents the number of tests that a given SDK is able to manage with an accuracy value included in the corresponding range.

Table 3 - EER and Rejection Rate of the three SDKs evaluated.

Characteristic	SDK1		SDK2		SDK3	
	EER	Rej.	EER	Rej.	EER	Rej.
8 Blurred	17.3%	9.8%	48.3%	0.4%	**1.9%**	0.2%
9 Looking Away	**15.1%**	3.1%	-	-	-	-
10 Ink Marked/Creased	-	-	-	-	-	-
11 Unnatural Skin Tone	18.0%	4.0%	48.7%	1.1%	**3.9%**	0.0%
12 Too Dark/Light	-	-	4.8%	0.0%	**4.0%**	0.0%
13 Washed Out	-	-	33.6%	0.2%	**2.6%**	0.0%
14 Pixelation	-	-	**0.0%**	0.2%	0.2%	0.0%
15 Hair Across Eyes	**50.0%**	81.0%	-	-	-	-
16 Eyes Closed	**0.4%**	2.2%	-	-	13.6%	0.0 %
17 Varied Background	**9.4%**	0.6%	15.4%	0.0%	-	-
18 Roll/Pitch/Yaw > 5	-	-	**24.1%**	2.8%	27.4%	0.0%
19 Flash Refl. on Skin	**8.3%**	0.2%	50.0%	0.0%	-	-
20 Red Eyes	**1.7%**	1.4%	34.3%	0.0%	16.9%	0.0%
21 Shadows Behind Head	-	-	-	-	-	-
22 Shadows Across Face	**8.1%**	12.1%	-	-	-	-
23 Dark Tinted Lenses	-	-	-	-	**3.9%**	0.0%
24 Flash Refl. on Lenses	-	-	-	-	-	-
25 Frames too Heavy	-	-	-	-	-	-
26 Frame Covering Eyes	**50.0%**	60.5%	-	-	-	-
27 Hat/Cap	-	-	-	-	-	-
28 Veil over Face	-	-	-	-	-	-
29 Mouth Open	**2.0%**	48.7%	-	-	-	-
30 Objects Close to Face	-	-	-	-	-	-
- the SDK does not support the test for this characteristic The bolded values indicate the best performance for each characteristic						

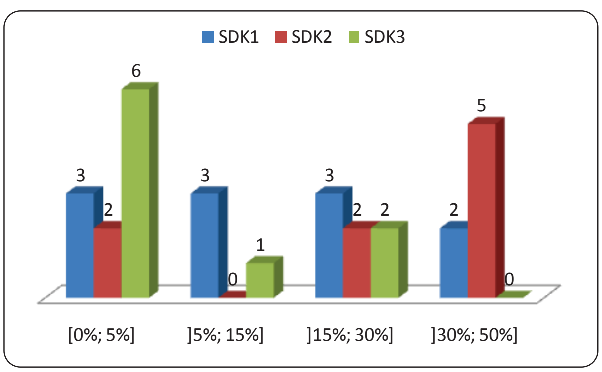

Figure 4 - Distribution of the three SDKs accuracy in four EER intervals. The x-axis reports the EER ranges, and the y-axis indicates the number of tests on which a SDK reaches an EER included in that range.

It is worth noting that the number of characteristics evaluated by the three SDKs is different: in particular, SDK1 verifies 11 requirements while SDK2 and SDK3 evaluate only 9 requirements. SDK1 achieves medium-high results ($EER \leq 15\%$) for most of the characteristics; SDK2 provides unsatisfactory results for many tests and finally SDK3 obtains a good accuracy ($EER \leq 5\%$) for 6 tests over 9 characteristics evaluated.

4. CONCLUSIONS

In this paper a new database and associated evaluation protocol have been proposed for performance assessment of software applications designed to evaluate the compliance of a face image to the ISO/ICAO standard for machine readable travel documents. The database contains 7740 face images of which 7468 are non-compliant to at least one characteristic. All the images have been manually labeled and a set of marker of the main facial features is provided with the database. Moreover, for each image, its compliance to each characteristic is given.

A first round of experimentation has been conducted on two commercial applications and an internally developed prototype. The results obtained show that some characteristics are easier to be evaluated with a good accuracy (blurred, pixelation, eyes closed and red eyes). For other tests a medium accuracy level has been measured for some SDKs (looking away, unnatural skin tone, too dark/light, washed out, varied background, roll/pitch/yaw>5, flash reflection on skin, shadow across face, dark tinted lenses and mouth open). For another set of characteristics the performance is very poor (hair across eyes and frames covering eyes) since a high EER (50%) is obtained, in conjunction with a high rejection rate (60-80%). Finally many characteristics are currently not evaluated by any of the tested SDKs (ink marked/creased, shadows behind head, flash reflection on lenses, frames too heavy, hat/cap, veil over face, objects close to face). These experiments clearly show that, at present, the available software tools can only support a human expert and provide alarms in case of uncertainty, but that for a total automation of this process a lot of research is needed.

Further (and more extensive) tests will be performed by our laboratory in the future; in particular, we are considering to organize on open evaluation campaign (FVC-like [5]) around this topic. After the competition our intention is to make both the database and related information (labels, markers, etc.) available to the scientific community for further studies and advances in the field.

5. REFERENCES

[1] ISO - International Standard Organization, "Text of FCD 19794-5, Biometric Data Interchange Formats – Part 5: Face Image Data," 2004, Available at http://isotc.iso.org.

[2] M. Ferrara, A. Franco, and D. Maltoni, "Evaluating systems assessing face-image compliance with ICAO/ISO standards," in *(BIOID 2008)*, pp. 191-199, 2008.

[3] A. M. Martinez and R. Benavente, "The AR face database," Computer Vision Center, CVC Technical Report 1998.

[4] P. J. Phillips et al., "Overview of the face recognition grand challenge," *in proc. IEEE Computer Vision and Pattern Recognition,* vol. 1, pp. 947-954, 2005.

[5] R. Cappelli, D. Maio, D. Maltoni, J. L. Wayman, and A. K. Jain, "Performance Evaluation of Fingerprint Verification Systems," *IEEE Transactions on PAMI*, pp. 3-18, 2006.