Border control: From Technical to Operational evaluation.

V. Despiegel, S. Gentric, JC. Fondeur.
Morpho, Safran Group

One of the major advantages of face recognition, compared to other biometrics, is the ease-of-use: no contact with the sensor, fast, in-line with the human way of recognizing people. In a border control scenario, fully automated gates using face recognition have now been deployed. As a passenger may use an automated solution only once in his life, the deployed system must operate with as little active cooperation from the passenger as possible.
One way of building an ergonomic, real time, face recognition system is the use of video cameras. In such operational systems, there are different methods to process video streams, using an "image based", 1:1 face recognition algorithm. In the normal automated scenario, the facial information of the person trying to pass the gate is compared to a stored reference (in the case of facial recognition, it is usually stored in the passport chip).

In this presentation, we will first describe and explore two basic methods of processing video images and then discuss operational evaluation capabilities.

1/ two basic methods

The "score driven" method: Images are continuously processed and matched against the reference until a matching score is above a pre-determined threshold.
At a given frame rate, on each image, a face is found, encoded and matched against the reference. If the score is above a matching threshold, the door opens, the image is logged, the acquisition process stopped. If no score reaches the threshold, at a given timeout, the passenger is rejected.
The "quality driven" method: Images are processed until a quality intrinsic to the considered image is above a pre-determined threshold.
At a given frame rate, on each image, a face is found and a quality is computed. If the quality is above a threshold, the image is encoded, and matched against the passport image. The image is logged, the acquisition process is stopped. If the matching score is above a threshold, the door opens, if not, the passenger is rejected.

In order to compare these methods, we have stored video streams, during a period of 10 seconds, at 5 fps, for 180 subjects, in a semi cooperative behavior, moving in front of a camera. All tests were made offline.

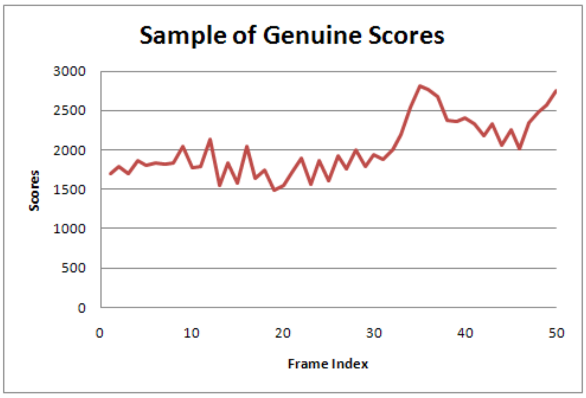

Figure 1

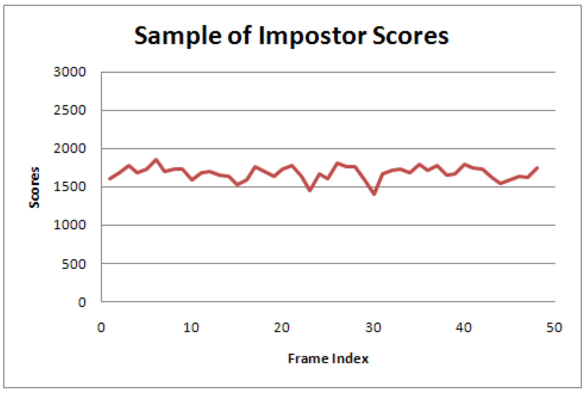

Figure 2

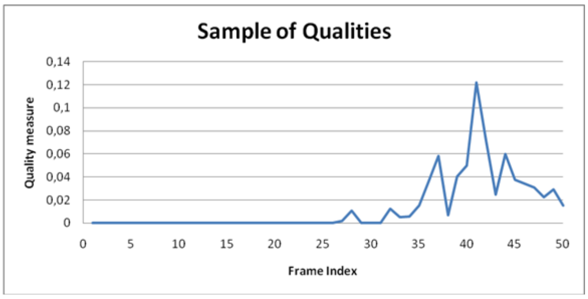

Figure 3

Figures 1 and 2 shows a sample of the evolution of the matching score for a genuine test and for an impostor test (the reference ID picture is different but the associated video frame is the same. Figure

3 presents the associated quality measure associated to each frame of the video). In a test (either impostor or genuine), none of the different frames produce the same matching score versus the reference, or produce independent scores, despite the fact that the person in front of the camera is the same and the time link leads to a strong correlation between the comparison score. However, the pose (yaw, pitch, roll, distance to the camera) the lighting, the expression, the behavior of the person are all slightly different for each frame. The correlation between scores from a video sequence depends on the behavior of the subject but also on the on the properties of the comparison algorithm used. Figure 4 shows the evolution of the FAR for a given matching threshold, computed on a database of thousands of ICAO images (4800 images, 10Ko, jp2, full frontal). In the "score driven" method, the FAR is directly linked to the number of frames tested. In this test, the FAR is 30 times higher (see Figure 4) with the full video stream than when using only the first image for comparison.

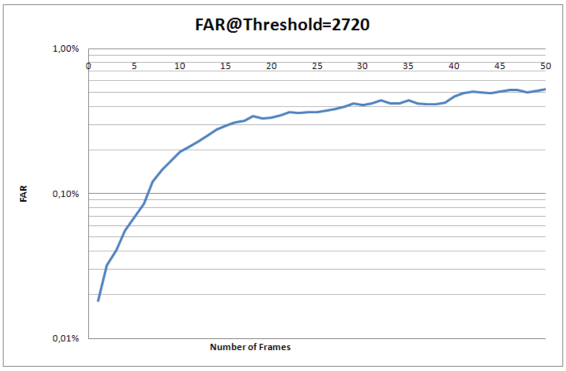

Figure 4

Even if the comparison is made with the image selected during the corresponding genuine test, the FAR measured on the log image would be much lower than the one measured with video stream (which is the one corresponding to the operational scenario).

To see if this estimation error is linked to the intrinsic properties of the face recognition algorithms, we have tested multiples algorithms. Table 1 shows than the error factor is huge and mostly independent of the type of algorithms.

Algorithm	FAR on the logged image	Threshold	Operational FAR	Error factor
1	0.5%	2530	7.2%	14.4
2	0.5%	2510	7.4%	14.8
3	0.5%	2560	6.9%	13.8
4	0.5%	2540	7.1%	14.2
Eigenfaces*	0.5%	2510	9.1%	18.2

**Eigenfaces : FaceFinder from OpenCV, PCA learn on Yale database*

Table 1

In order to compare the relative performance of those two basic methods, we choose a common targeted FAR=0.5% and then a different threshold for each method. Note that "Threshold for the score driven method" is computed a posteriori for the impostor tests, on the full video stream.

<table>
<tr><th></th><th></th><th>threshold</th><th>FAR</th><th>FRR</th><th>Genuine mean crossing time</th></tr>
<tr><td rowspan="2">score driven method</td><td>logged image</td><td rowspan="2">2720</td><td>0,03%</td><td rowspan="2">7,6%</td><td></td></tr>
<tr><td>Operational</td><td>0,50%</td><td>5,2s</td></tr>
<tr><td rowspan="2">quality driven method</td><td>logged image</td><td rowspan="2">2550</td><td rowspan="2">0,50%</td><td rowspan="2">4,4%</td><td></td></tr>
<tr><td>Operational</td><td>7s</td></tr>
</table>

Table 2

Table 2 demonstrates that the difference between the thresholds due to different FAR computations yields an improvement in terms of FRR. Of course, this depends of the information included in the quality and its accuracy in predicting a relevant matching score.

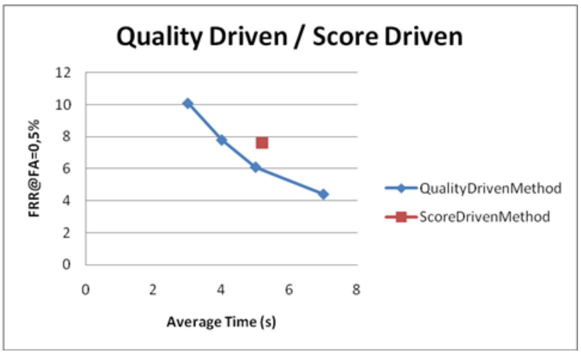

Figure 5

In the border control scenario, processing time is the other important criteria. In the “score driven” method, duration is short when we have a HIT (this is the case for most of the genuine tests) and goes to timeout in case of rejection (for most of the impostor tests). For the “quality driven” method, the duration depends only on the quality of the input video stream.
Table 2 shows a comparison between mean crossing times for passengers with their own passport. This includes the time spent walking into the field of vision for the camera. As presented in Figure 5, it is possible to build a quality which is a good oracle of the matching potential of an image and a quality driven method could lead to better performances than a score driven selection for a similar average time (for genuine test).

2/ *Operational evaluation*

These two basic methods can be improved or mixed. However we have seen that when a “score driven” method is used, there is a bias between the FAR computed on one logged image and the operational FAR. By choosing a “quality driven” method, it is possible to achieve better results in terms of FRR. Depending on the timeout set to the system, it could however lead to an increase of the total acquisition duration.
As it is for border control application and security is of key importance, the operational question is: “is it possible to guarantee and check the FAR of the system?”

For “quality driven” methods, as the logged image is independent of both the passport image, and the matching threshold, we can compute an offline complete DET curve corresponding to operational performances. This also allows for precise control of the impact of a threshold modification on FRR and FAR. This makes analyzing the impact of different external factors on performance (passport origin, airport environment, frequent users, passport ageing) including evolution across time, possible.
More practically, this enables a customer to perform these evaluations through an independent audit.

For “score driven” methods, as the acquisition process takes the reference as an input, the operational FAR cannot be computed offline. According to ISO/IEC 19795-1 (Annex B.1.2) the number of independent tests to claim 0.5% of FAR is several thousands. Having so many people using the system with someone else’s passport to perform impostor tests would carry huge costs. In addition, multifactor analysis is prohibited.

Another option is to compute FAR based on logged data and to estimate the maximum bias between this measurement and the operational performances. As presented above, the bias can be estimated on an internal database or could be evaluated with a few impostor templates. However, the imprecision of such a measurement could hardly replace a proper evaluation as done for a quality driven method.

3/ *Conclusion*

When one wants to build an ABC system from elementary biometric black boxes (an image feature extractor and a template comparator), it seems natural to code and match incoming images from the acquisition system against the template generated from the passport image and to open the doors as soon as the matching score is above given threshold. The threshold is set so that in a 1:1 image versus image scenario it guarantees a given FAR for the biometric black box. However, the overall FAR of the system is neither this one nor can be measured directly from the log image of the system.

Being able to guarantee the security of the border control is a key issue. This should be obtained by regular measurement of the operational FAR. The fact that no unbiased evaluation of the operational FAR is possible for score driven ABC as there is a gap between the FAR measured from the stored genuine template and the operational FAR. This creates a serious weakness for this type of gate.

For Quality Driven ABC, the image is selected by direct quality measurements on the images from the gate acquisition system. This mean of building templates is not influenced by measure on the picture stored in the ID document (neither comparison measurement nor quality extracted from the image). This guarantees that selected image would be the same for impostor or genuine tests. Consequently, a real evaluation of the FAR of the border control from the log images is possible. This could be obtained as presented in this paper at no cost in term of time with a well defined quality.