Evaluation of Face Recognition Technologies for Identity Verification in an eGate based on Operational Data of an Airport

Andreas Opitz and Andreas Kriechbaum-Zabini
AIT Austrian Institute of Technology
Donau-City-Straße 1, 1220 Vienna, Austria
{andreas.opitz, andreas.kriechbaum-zabini}@ait.ac.at

Abstract

Face recognition technologies play an important part in the field of automated border control (ABC). The demand for the reliable verification of a passenger's identity is based on efficiency and security aspects of ABC systems which have to be taken into account. Face recognition technologies allow for an automation of the traditional manual inspection by a border guard. This paper presents an evaluation of the face verification performance of commercial face recognition technologies based on operational data from an existing eGate at an airport. In addition to comparisons of the overall performance, more detailed considerations such as the influence of the remaining validity of the passport are made. Moreover, the performance implications of the nationality of the passports, which may be the result of the different quality of the passport photos or increased cooperativeness of the passengers based on previous experiences with eGates, are considered as well.

1. Introduction

The expected rise in the number of passengers at airports from 400 million in 2009 to 720 million in 2030 increases the demand for more efficient and secure processes and technologies in the field of border control [3]. Existing border crossing points are confronted with space limitations which prohibit the approach of simply increasing the amount of border guard booths. Establishing ABC systems such as eGates enables for spatial separation of the border control process and its remote supervision by border guards. The usage of eGates introduces the requirement for effective face recognition technologies which play an important role in the border process. In the field of ABC they are used for the verification of a passenger's identity based on the passport photo and live image data acquired by the eGate.

When putting face recognition technologies into operation, an inevitable trade-off decision between two error measures has to be made as part of a more comprehensive risk assessment. One measure is the false rejection rate (FRR) which is the ratio of the number of genuine verification attempts erroneously rejected by the system to the total number of genuine verification attempts. The other measure is the false acceptance rate (FAR) which is the ratio of the number of imposter verification attempts erroneously accepted by the system to the total number of impostor verification attempts. Lowering the FRR in order to increase the throughput at the eGate since less passengers are erroneously rejected inevitably results in an increased FAR and vice versa. Further risk assessment considerations such as the detection of attack vectors like presentation attacks are out of the scope of this work.

The latest study in a series of large-scale evaluations of face recognition technologies with published results for the task of face verification is the Multiple-Biometric Evaluation (MBE) 2010 [4]. The type of data on which it is based on ranges from laboratory images, over visa images to law enforcement mugshots. The performance of the evaluated technologies improved by an order of magnitude between subsequent evaluations. In a real-world scenario evaluation by Spreeuwers *et al*. [6] several face recognition technologies from different vendors have been tested on data obtained from two eGates at Schiphol Airport.

The aforementioned works have in common that they report the FRR for a fixed FAR value. Based on operational data acquired from an existing eGate, an evaluation of commercial face recognition technologies is presented in the following. It covers general performance comparisons on the entire dataset for whole range of possible error measure combinations. Additionally, more detailed aspects such as the influence of the remaining validity of the passport are investigated. Another factor is the influence of the nationality of the passports. Its impact on the performance may be the result of different quality of the passport photos or a better cooperation of the passenger due to previous experience with similar technologies. The evaluation is conducted offline using three face recognition technologies from different

vendors hereinafter referred to as system A, B and C.

2. Data Acquisition

The data this evaluation is based on was obtained from an eGate, an integrated two-step system, at the Vienna International Airport in the course of the FastPass project [1]. The eGate basically has the layout of a short corridor, consists of a passport reader in the front of the entrance door and a face camera behind the exit door on the right. A photo from the interior also depicting a magnified detail view of the face camera is shown in Figure 1.

Figure 1. eGate from the inside with a detail view of the face camera

The entire border crossing process, hereinafter referred to as session, including the capturing of the evaluation data works as follows. First, the passenger places the passport on the passport reader, which extracts the information from the chip. This data encompasses the passport number, date of expiry, nationality and a face image of the passport holder, hereinafter referred to as reference image. Immediately after the successful readout of the data, the entrance door opens and the face camera starts with the acquisition of live images, hereinafter referred to as probe images. While capturing the images, the face verification system integrated into the eGate tries to verify the passenger's identity. This is done by computing the match scores between the reference image and the probe images. After passing the entrance door, it closes and the passenger proceeds to the exit door which remains closed until one of the following criteria is matched. Either the face recognition system verifies the identity by computing a match score exceeding the predefined threshold or a timeout occurs which leads to a manual inspection by a border guard.

The data acquired during a session consists of the passport data which is required for the preprocessing and subset selection, the reference and the probe images as well as timing measurements and information about possible manual interventions of the border guard. This data was only stored for the evaluation of the system and during standard operation it is immediately discarded after each session.

The length of the acquisition period of one year allows for an increased heterogeneity of the dataset. One of the contributing factors is the variation of the lighting conditions. There is a large window in the camera's field of view which leads to images with varying quality, e.g. contrast of face region, during certain weather conditions.

3. Data Preprocessing

The acquired data also contains information of border crossing attempts which need to be excluded before conducting the actual evaluation. Most importantly, it has to be ensured that it only consists of genuine border crossing attempts, i.e. only those cases are included where the user is the legitimate passport holder. Instead of manually annotating the dataset for ensuring this precondition, the following automated method was chosen based on two decision criteria which have to be fulfilled: First, the face recognition system of the eGate succeeded in matching the reference image with one of the probe images and exceeded a predefined threshold for the match score. Second, the border guard who was overseeing the process did not intervene by manually overriding the decision of the face recognition system. Furthermore, if multiple border crossing attempts of the same person were present in the dataset, the evaluation would be biased towards the results of this individual. Therefore it has to be ensured that each passenger may appear only once in the dataset. The resulting preprocessed dataset consists of 3224 sessions with an equivalent number of reference images and a total of 176956 probe images with an average of 55 images for each session and a standard deviation of 20.

4. Evaluation Methodology

The main goal of the evaluation is to obtain measures which characterize the verification accuracy of the face recognition systems based on the input data of the specific eGate scenario at the airport. The detection error tradeoff (DET) was chosen as as an established method for representing and comparing the verification performance. It consists of the FRR and the FAR for different thresholds applied to match scores. Each match score originates from the comparison of two images, a reference image and a probe image. The scores are determined offline once the data acquisition phase is finished. The face recognition system is presented two images, tries to detect faces in them and computes digital representations, the so-called face templates. If it succeeds, then the match score for the comparison of two face templates can be determined. This match score serves as a similarity measure. If the face template extraction fails, i.e. a failure to enroll (FTE) occurs during the processing of one of the images, then the score of the comparison is assumed to be zero.

In a scenario where for each person in the dataset only

one reference and one probe image exist, the computation of the match scores for all possible pairwise combinations is evident. On the contrary, the dataset described in section 2 does not contain only one, but multiple probe images for each session. The simplest approach would be to use the last probe image of each session. This image either reached the predefined threshold when matching it with the reference image and thereby resulted in the successful verification of the face recognition system integrated in the operational eGate or is the last image that was captured before reaching the timeout. However, following this approach would enforce a bias towards the aforementioned integrated system, as other face recognition solutions might possibly achieve more accurate results with one of the other probe images. We devised the following methodology in order to tackle this issue.

First, all possible match scores s_{i,j_k} are computed using the face recognition system's matching function f.

$$s_{i,j_k} = f(r_i, p_{j_k}) \tag{1}$$

where r_i refers to the reference image from session i and p_{j_k} refers to the probe image k from session j.

For each session j the number of probe images m varies. In order to determine a single representative match score for the comparisons of one reference image from session i with multiple probe images from session j an aggregation needs to be performed.

The decision of choosing the maximum of all match scores within a session

$$s_{i,j} = max(s_{i,j_0}, \ldots, s_{i,j_m}) \tag{2}$$

is motivated by the following: In the described scenario a user of the eGate implicitly increases the chance of achieving a high match score by having more probe images captured. This is due to the fact that a face template is not a perfect representation of the distinctive characteristics of a face and varies among different images of the same person. It depends, amongst other factors, on the head pose, the lighting conditions and the image resolution. This is true for both the genuine as well as impostor verification attempts .

After computing all possible combinations of reference and probe images resulting in corresponding scores, the FRR and FAR measures can be determined. The FRR is the fraction of the number of sessions where the reference and the probe image originate from the same person but the system erroneously classifies them as not matching compared to the total number of genuine comparisons. The FAR is the fraction of the number of sessions where the reference and the probe image originate from different persons but the system erroneously classifies them as matching compared to the total number of imposter comparisons.

5. Results

Based on the methodology described in section 4 the results of the evaluation are presented in the following. Frontex published recommendations which target authorities responsible for border control in the member states of the European Union. For the task of face verification in eGates the best practice guidelines for ABC systems [2] are of special interest. In these guidelines a FRR of 0.05 at a FAR of 0.001 are considered as the maximally tolerable error rates. The coordinate for the corresponding value in the DET charts is indicated by the label "Frontex".

5.1. Overall Performance

The performance of the tested face recognition systems on the full dataset described in section 3 is shown in Figure 2. The best performance is achieved by the systems A and B which fulfill the recommendations of Frontex mentioned before. When choosing a FAR of 0.001, then system A performs better with a FRR of 0.009 compared to a FRR of 0.02 of system B. When fixing the FRR at 0.05 the FAR of system A is about 6×10^{-5} whereas system B performs better with a value of only 4×10^{-5}. System C performs worse over the largest part of the value range and fails to achieve error rates below the recommended levels.

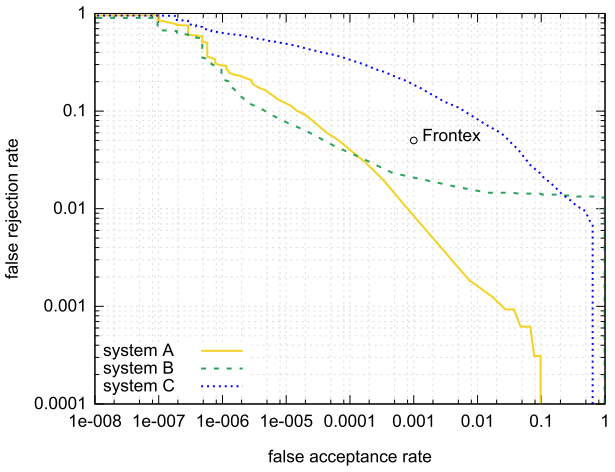

Figure 2. DET chart of entire dataset

The charts depicted in Figure 3, 4 and 5 contain more detailed information about the performance of system A, B and C, respectively. Each chart consists of three graphs, the FRR, the FAR and the average number of seconds until the successful verification of the passenger. The verification duration is measured from the completion of the passport readout which corresponds to the opening of the eGate entrance doors and the beginning of the image acquisition by the camera. Additionally, the minimal requirements for the FRR and the FAR by Frontex are indicated by corresponding markers in the chart.

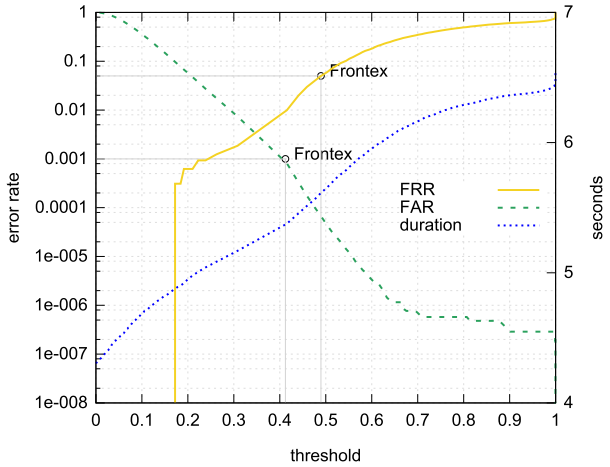

Figure 3. FRR, FAR and verification duration for system A

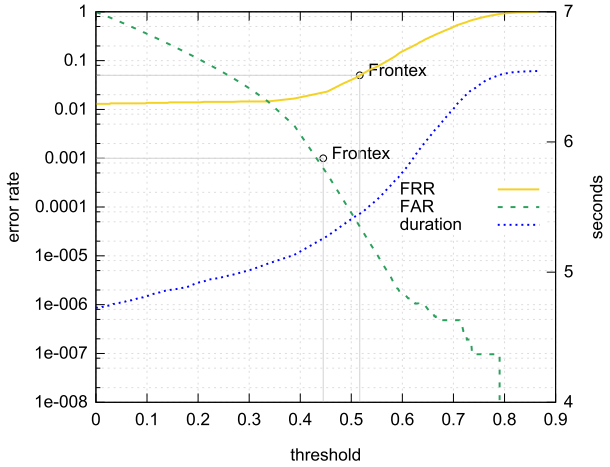

Figure 4. FRR, FAR and verification duration for system B

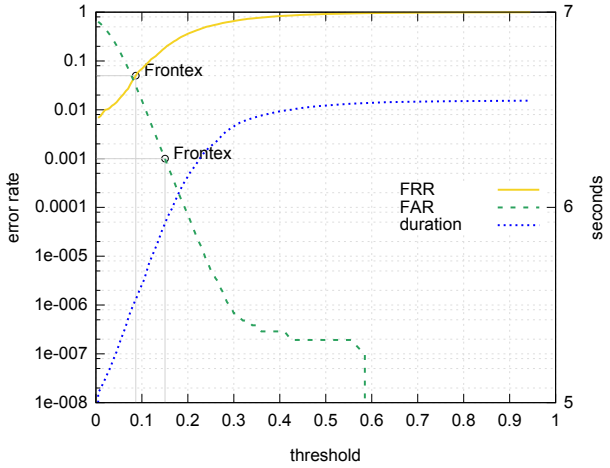

Figure 5. FRR, FAR and verification duration for system C

Increasing the threshold results in the expected increase of the FRR, the decrease of the FAR and an increase of the time it takes to acquire an image which reaches the verification threshold. The resulting time span in the interval of maximal tolerable error rates spanned by the Frontex requirements is only a fraction of a second for all three systems. The choice of the threshold plays only a minor role regarding the speedup of the verification process compared to the the total time required for the verification.

The reason for the minimal verification duration of about 5 seconds is based on the fact that the passenger needs a certain amount of time from the opening of the entrance door of the eGate until entering the face camera's field of view.

5.2. Passport Date of Expiry

The performance related to the remaining validity of the passport, i.e. the difference between the passport date of expiry and the session date is another investigated aspect. This measure can be considered as an indicator for the age of the reference image. It was used because the date of issue of the passport was not available for the evaluation. Only digital passports from member states of the European Union and the Schengen area are processed which have a maximum validity of about 10 years. For this evaluation the dataset was divided into two equal-sized subsets each consisting of 1612 sessions. The corresponding median of the remaining passport validity of all sessions is about 5.8 years leading to the rounded $0-6$ and $6-10$ years validity period labels of the resulting DET chart depicted in Figure 6. The choice for this type of subdivision was preferred over the subdivision into two equal-sized periods regarding the remaining validity in order to ensure the same statistical lower boundaries for the error rates.

It can be observed, that the relative change in the performance of the systems is not homogeneous. While system C generally copes better with longer validity periods, the results for the other systems depend on the chosen threshold. For higher FAR values (greater than 5×10^{-4} for system A and greater than 0.001 for sytem B) a lower validity period achieves better FRR values whereas the opposite is true for FAR values below the aforementioned thresholds.

5.3. Nationality

Using the nationality as a criterion for the evaluation yields the results depicted in Figure 7. The number stated in brackets in the legend is the number of sessions for the corresponding nationality. The largest group is formed by passengers with passports from Austria followed by United Kingdom and Czech Republic with 1799, 509 and 196 sessions, respectively. Since this number amongst the nationalities, the conclusions which can be drawn from the corresponding DET charts have to be considered carefully. According to

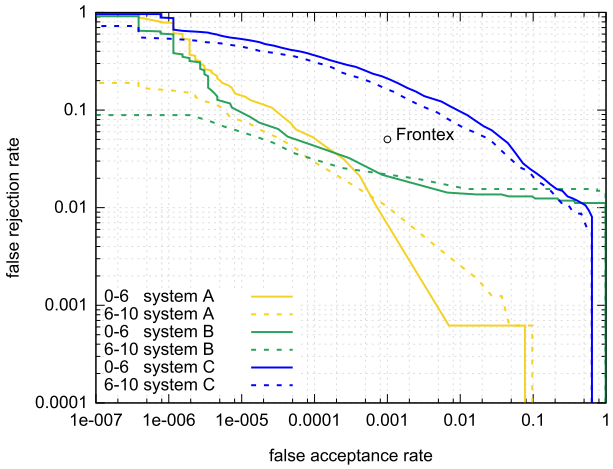

Figure 6. DET chart of passport validity period subsets in years

the "rule of 3" [5] the lowest error rate which which can be determined with a 95% confidence interval based on N comparisons is $3/N$. For passports *e.g.* from Czech Republic only FRR values above 0.02 $(3/196)$ and FAR values above 8×10^{-5} $(3/(196 * 195))$ are considered. Sessions with passports from the United Kingdom or Austria allow for consideration of lower error rates due to the higher number of comparisons. All other nationalities contained in the dataset form groups with lower session counts and are therefore not considered in the evaluation. As can be seen in the chart, system A achieves lower error rates for Austrian passports than for the those from the United Kingdom. For system B and C the nationality has only a small impact on the performance.

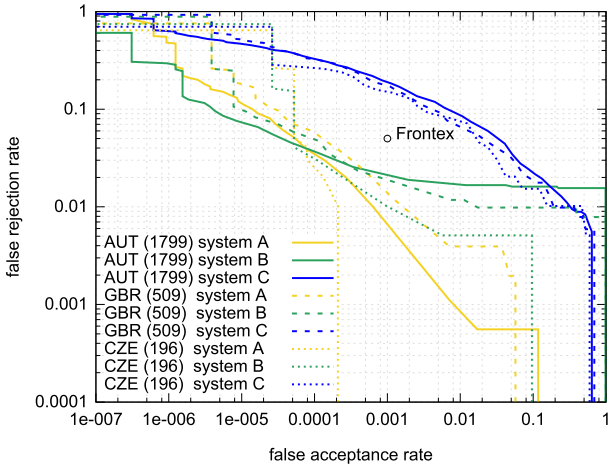

Figure 7. DET chart of nationality subsets

6. Conclusions

The evaluation of commercial face recognition technologies presented in this paper was conducted using data from an eGate at an airport. After preprocessing the data which was acquired over the period of one year, the face recognition technologies were evaluated offline. Two out of the three systems manage to achieve error rates below the limits recommended by Frontex. Varying the threshold within those limits, results in differences of the time required for the verification procedure of less than a second. The remaining validity of the passport has an influence on the performance which depends on the considered error rates and the applied face recognition technology. The same is valid for the nationality where the performance differs with varying degrees amongst the nationalities, the considered error rates and the selected face recognition technologies. Future work includes an evaluation based on data acquired by a different face image acquisition system and an analysis of methods for the detection of spoofing attempts.

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 312583.

References

[1] FastPass - a harmonized, modular reference system for all european automated border crossing points. http://www.fastpass-project.eu. accessed on 2.7.2015.

[2] Best practice technical guidelines for automated border control (ABC) systems. Technical report, Frontex, 2012.

[3] European Commission. 'Smart borders': enhancing mobility and security. Press Release, February 2013.

[4] P. J. Grother, G. W. Quinn, and P. J. Phillips. Multiple-biometric evaluation (MBE) report on the evaluation of 2D still-image face recognition algorithms. Technical report, National Institute of Standards and Technology, 2010.

[5] A. J. Mansfield and J. L. Wayman. *Best practices in testing and reporting performance of biometric devices*. Centre for Mathematics and Scientific Computing, National Physical Laboratory, 2002.

[6] L. J. Spreeuwers, A. Hendrikse, and K. J. Gerritsen. Evaluation of automatic face recognition for automatic border control on actual data recorded of travellers at schiphol airport. In *BIOSIG*, pages 1–6, 2012.