A Method for Estimating Authentication Performance over Time, with Applications to Face Biometrics

Norman Poh, Josef Kittler, Ray Smith, and J. Rafael Tena

CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK
{norman.poh,j.kittler,r.s.smith,j.tena}@surrey.ac.uk

Abstract. Underlying biometrics are biological tissues that evolve over time. Hence, biometric authentication (and recognition in general) is a *dynamic* pattern recognition problem. We propose a novel method to track this change for each user, as well as over the whole population of users, given only the system match scores. Estimating this change is challenging because of the paucity of the data, especially the genuine user scores. We overcome this problem by imposing the constraints that the user-specific class-conditional scores take on a particular distribution (Gaussian in our case) and that it is continuous in time. As a result, we can estimate the performance to an arbitrary time precision. Our method compares favorably with the conventional empirically based approach which utilizes a sliding window, and as a result suffers from the dilemma between precision in performance and the time resolution, i.e., higher performance precision entails lower time resolution and vice-versa. Our findings applied to 3D face verification suggest that the overall system performance, i.e., over the whole population of observed users, improves with use initially but then gradually degrades over time. However, the performance of individual users varies dramatically. Indeed, a minority of users actually improve in performance over time. While performance trend is dependent on both the template and the person, our findings on 3D face verification suggest that the person dependency is a much stronger component. This suggests that strategies to reduce performance degradation, e.g., updating a biometric template/model, should be person-dependent.

Keywords: Biometric authentication, performance assessment, face recognition.

1 Introduction

In general, pattern recognition can be categorized as either static or dynamic [1]. A static pattern does not tend to change dramatically over time whereas a dynamic one does. The latter is problematic because as the variability of dynamic patterns in the same class becomes gradually larger, a classifier that does not update itself will have tremendous difficulty when discriminating between dynamic patterns belonging to different classes.

Biometrics can be considered as a dynamic pattern principally because underlying the metrics are living tissues that tend to modify themselves either as a result of muscle movements or tissue growth (aging). In the former case, the change can take place in seconds whereas in the latter case, the change can be gradual. Apart from this change, variation in patterns can also be caused by an imperfect biometric acquisition process, e.g., in the way a biometric sample is presented and the environmental conditions. These

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 360–369, 2007.
© Springer-Verlag Berlin Heidelberg 2007

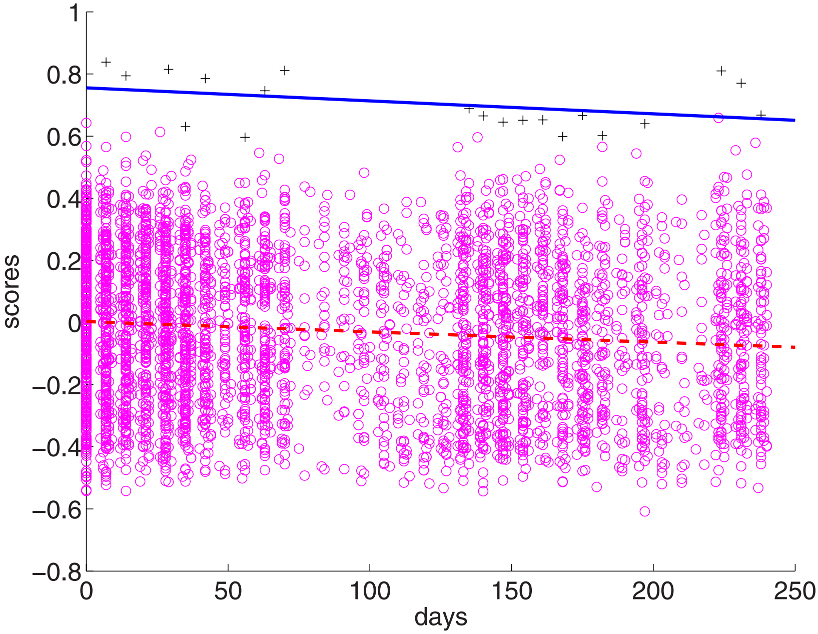


Fig. 1. Scatter plot of genuine user (“+”) and impostor (“○”) match scores for a single user’s template over 250 days (the X-axis). Higher match scores imply genuine user class. The interruption in genuine match scores around the 100-th day is due to no observations being made during the term break. The straight lines are the regression fits on the data (continuous line for the genuine user match scores and dashed line for the impostor ones).

factors cannot often be decoupled but their effects can readily be observed from the resulting match scores.

To give a further motivation, we plotted the class-conditional match scores in Figure 1 of a user selected at random from a face verification system applied to the Face Recognition Grand Challenge (FRGC) database. This database contains images collected over 250 days. Two clusters of scores are available, namely genuine user match scores and impostor match scores. The genuine user match scores are the results of comparing a reference template with query images of the same user. The impostor match scores are the results of comparing the reference template with query images of other users. In this figure, one can observe that genuine user scores are very sparse whereas the impostor match scores, as a result of comparing a sequence of query images from many persons, are very dense.

The ability to track the dynamic change of biometric patterns in terms of performance is valuable because it can determine whether or not a biometric system degrades over time. If it does then preventive measures will have to be taken to maintain the performance. One of the pilot studies in this direction is reported in [2], whereby the performance of four face recognition systems coupled with two face detection algorithms (hence altogether eight systems) were assessed on the FRGC database. This database contains 250 users whose images were captured over a period of two years. It was observed that all the face identification systems decrease in performance (in terms of rank-1 false rejection) with time-lapse. However, time-lapse is not the only factor; in [2], it was noted that the precision of eye localization is another important factor.

This paper differs significantly from [2] because our concern is with the individual user performance. According to [3], the users in a database can exhibit very different performance. In particular, some users are more easily recognized than the others. As a result, it is reasonable to expect that the performance change will be different from one user to another. We argue that our approach is more useful because it can calculate the person-specific performance. This enables one to sort the users according to their current performance, thereby identifying the weak users in this process. If the performance of these users can be corrected, for instance, by updating the user model, one can potentially improve the overall system performance. Deciding when and how to update a biometric template/model will be investigated in the future.

This paper is organized as follows: Section 2 explains how the user-dependent error over time can be calculated using the proposed procedure; Section 3 describes the database used; Section 4 shows the results and Section 5 presents the conclusions.

2 Modeling Performance over Time on a Per-person Basis

Suppose that each user j in a database has two sequences of scores over time: one from the genuine user set of scores and the other from its impostor counterpart. We denote the two sequences by $\mathbf{y}_j^k = [y_{j,1}^k, \ldots y_{j,N_k}^k]'$ for genuine user and impostor classes, $k = \{G, I\}$, respectively, and each sequence has N_k number of scores. For clarity, we drop the user index j everywhere. In this study, the impostor scores with respect to the reference user are generated by the rest of the users exhaustively. Therefore, the constraint $N_G \ll N_I$ is true in this case. Note that each sequence of scores has a corresponding time delay sequence $\mathbf{d}_j^k = [d_{j,1}^k, \ldots d_{j,N_k}^k]'$ or simply $\mathbf{d}^k$ (omitting j).

For the genuine user scores, this time delay sequence is just the time difference between the template and the query image associated with the respective score. Suppose that these images have the following time stamps: $t_0, t_1, \ldots, t_{N_G}$. We reserve the first image with time t_0 as a template. This template is then compared to the remaining images in the sequence. The resulting genuine match scores will have the following *relative* time stamps: $\mathbf{d}^G \equiv [d_1^G, d_2^G, \ldots, d_{N_G}^G]' \equiv [t_1 - t_0, t_2 - t_0, \ldots, t_{N_G} - t_0]'$.

For the impostor sequence, this time delay sequence is with respect to the *relative time difference* between the first impostor attempt and the subsequent impostor attempts by the same impostor. Suppose the image sequence of an impostor has the following time stamps: $t_1, t_2, \ldots t_{N_I}$. We define its relative time sequence by $\mathbf{d}^I \equiv [d_1^I, \ldots, d_{N_I}^I]' \equiv [t_1 - t_1, t_2 - t_1, \ldots t_{N_I} - t_1]'$, i.e., taking the difference between the time stamp of an image in the sequence with the first one. Note that the first element in this list has a time stamp of 0. By so doing, we assume that the time difference between the first impostor attempt and the template has no importance. This is a reasonable assumption given the fact that the two feature sets under impostor matching are not from the same persons.

The goal is to estimate the performance in terms of False Match Rate (FMR) and False Non-Match Rate (FNMR)[1] at a given time d_t for $t = 0, 1, \ldots$ and for each user to an arbitrary precision. This implies that FMR and FNMR are themselves *smooth* functions over time. This is clearly a difficult task since the conditional sequence $\mathbf{y}^k$ has very few data points, especially for the genuine user sequence.

For each sequence k, let us fit a regression function to $(\mathbf{d}^k, \mathbf{y}^k)$. Regression functions are also called smoothers because they give in general a smoothed output of $\mathbf{y}^k$. Some examples are kernel, running mean, running-line, locally weighted running-line, running spline and regression spline smoothers [4, Chap. 3]. We will use a polynomial regression model of order D for this purpose so that we obtain the regression parameter $\mathbf{p} = [p_D, \ldots, p_0]'$. By evaluating the parameter $\mathbf{p}$, we obtain a smoothed conditional score $\mu_t^k = p_0 + p_1 d_t + \ldots + p_D d_t^D$ at time d_t along with standard deviation σ_t^k. By tracing (d_t, μ_t^k) for $t = 0, 1, \ldots$, one obtains a smoothed curve with 95% confidence bound $(d_t, \mu_t^k \pm 2\sigma_t^k)$ for each $k \in \{G, I\}$. In summary, for a given instance of time d_t, we have the parameters $\{\mu_{j,t}^k, \sigma_{j,t}^k\}$ for each class k and for each user j (note that the index j is reintroduced here).

If the conditional regression fit is adequate, then the error residual should be approximately normally distributed. Unfortunately, given a limited number of data points of size N_k, especially for the genuine user sequence, in practice, one has no way of assessing whether the fit is adequate or not. This can be determined subjectively (visually). Another way to proceed is to use a polynomial model with a low degree of freedom D, based on the fact that we have few data points. The consequence is that the fit will lead to a large bias but a low variance. A more in-depth discussion of the bias-variance trade-off in regression can be found in [4, Chap. 3].

Once the regression parameters are found, we can then model instantaneous FMR and FNMR by:

$$\mathrm{FMR}_{j,t}(\Delta) = \Phi\big(\Delta | \mu_{j,t}^I, (\sigma_{j,t}^I)^2\big) \tag{1}$$

and

$$\mathrm{FNMR}_{j,t}(\Delta) = 1 - \Phi\big(\Delta | \mu_{j,t}^G, (\sigma_{j,t}^G)^2\big) \tag{2}$$

for a given threshold Δ in the score space, where $\Phi\big(\Delta | \mu, (\sigma)^2\big)$ is a cumulative normal density function with mean μ and standard deviation σ. Under such condition, a result from [5] shows that at Equal Error Rate (EER), i.e., FMR=FNMR, the user-specific EER is:

$$\mathrm{EER}_{j,t} = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\text{F-ratio}_j}{\sqrt{2}}\right), \tag{3}$$

where

$$\text{F-ratio}_j = \frac{\mu_{j,t}^G - \mu_{j,t}^I}{\sigma_{j,t}^G + \sigma_{j,t}^I}, \tag{4}$$

and

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-x^2\right] dx. \tag{5}$$

[1] Also called False Acceptance Rate and False Rejection Rate, respectively when evaluating the overall system performance, as opposed to algorithmic-level performance.

The end results are sequences of user-specific FMR and FNMR over the desired time period d_t estimated to an arbitrary accuracy.

The next issue to be dealt with is to calculate the population performance given the parameters $\{\mu^k_{j,t}, \sigma^k_{j,t}\}$ for each class $k = \{G, I\}$ and *all* the users $j = 1, \ldots, J$ at the desired time d_t. In order to calculate this quantity, we first need to calculate the class-conditional score distributions of the population. From the Gaussian assumption, the user-specific version of this distribution (for a given user j) is $\mathcal{N}(\mu^k_{j,t}, (\sigma^k_{j,t})^2)$. The population's conditional score distribution must be then a mixture of user-specific score distributions weighted by their respective prior probabilities, i.e., $\sum_{j=1}^{J} \mathcal{N}(\mu^k_{j,t}, (\sigma^k_{j,t})^2)$ $p(j|k)$. Therefore, the population's FMR is

$$\mathrm{FMR}_t(y) = \sum_{j=1}^{J} \Phi\big(y|\mu^I_{j,t}, (\sigma^I_{j,t})^2\big) P(j|I). \tag{6}$$

Similarly, the population's FNMR is:

$$\mathrm{FNMR}_t(y) = 1 - \sum_{j=1}^{J} \Phi\big(y|\mu^G_{j,t}, (\sigma^G_{j,t})^2\big) P(j|G). \tag{7}$$

The population's EER point, i.e., $\mathrm{FMR}_t(y) = \mathrm{FNMR}_t(y)$ can be found numerically.

The section that follows will discuss the database used before applying the proposed procedure on the real data.

3 Experimental Approach

The publicly available FRGC Experiment 3 data [6] is divided into two parts, training and test sets. Each part contains a set of 3D scans together with the corresponding 2D color intensity images. Additionally the 3D coordinates of landmark points located at the eye corners, the tip of the nose and the tip of the chin are also provided for each scan. The data was captured in near frontal pose using a Minolta Vivid 900 range scanner at a resolution of 640×480 and it includes males and females in approximately equal numbers, covering a range of ages and ethnic backgrounds. The training set consists of 943 face scans and images of 270 different subjects, with the number of samples per subject varying from 1 to 8. 410 subjects were included in the test set; with the number of samples per subject ranging from 1 to 22 for a total of 4007 scans and images. It is worth mentioning that 31 samples of the training set were discarded for our experiments, because the provided landmarks were off their mark by more than 50mm.

For the purpose of these experiments we use all of the training data to train face matching algorithms. To study the effects of changes over time we choose a subset of 285 users from the test data such that each one has a sequence of more than 6 accesses within the observed 250 days. Instead of just using the first image as template, we also used the second and third images as templates. When the second image is used for this purpose, the first image is not used to construct the genuine user sequence of match scores. This makes sense because one cannot compare a template with a sample acquired *before* the template is constructed.

Three sets of face verification experiments are described in this study. These are the PCA baseline system [6] supplied by FRGC (3D-baseline), 3D face verification with an error-correcting output-code based matcher (3D-ECOC) and 2D face verification with a local binary pattern based matcher (2D-LBP).

The 3D-ECOC method follows that described in [7]. Angular linear discriminant analysis is used to establish a low-dimensional feature space in which individuals are reasonably well separated. An error-correcting output code ensemble of Gaussian SVM classifiers is then trained within this feature space and the outputs from this ensemble are used to define a new feature space in which separation is further improved. A final similarity measure between pairs of 3D scans is obtained based on the Manhattan distance in this second feature space.

For the 2D-LBP matcher each face image is subdivided into a 7×6 grid of rectangular non-overlapping regions and a local binary pattern histogram [8] computed for each region. A similarity measure between pairs of images is then computed based on the mean Manhattan difference between corresponding histograms.

The 3D verification experiments require accurate registration and this is performed using the method of dense correspondence with a 3D model as described in [9].

4 Performance Trend Analysis

We first examined if the user-specific performance is template dependent or not. For this purpose, we selected a user at random from the 2D-LBP experiment. Using the first three images in the time-stamped sequence as templates, we plotted the fitted regression function with time being the input (independent) variable and score being the output (dependent) variable (see Figure 2). Their corresponding EERs are also shown at the bottom of each sub-figure. As can be observed, the user-specific performance is template dependent.

We then proceeded to compare the EER trends of different persons but used the first image as a template for all users. The purpose is to examine if the user-specific performance is person dependent or not. The results are shown in Figure 3. As can be observed, different users can exhibit dramatically different EER trend even though the same verification system is used. While most users decrease in performance, there are users who actually improve in performance over time. In any case, the user-specific performance is unlikely to be constant. This experimental result supports our conjecture that biometric authentication (and recognition in general) is a *dynamic* pattern recognition problem. Furthermore, the user-specific performance is both person *and* template dependent. Between the two, the choice of template seems to play a less important part in determining the trend.

Lastly, we plotted the system performance, using DET curves, over the whole population of users for the three different templates used. The results are shown in Figure 4. A DET curve [10] is a plot of false rejection rate (FRR) or FNMR versus false acceptance rate (FAR) or FMR. As can be observed, the DET curve also changes over time. In particular, when we analyzed the EER point in Figure 4(d), we observe that there is a

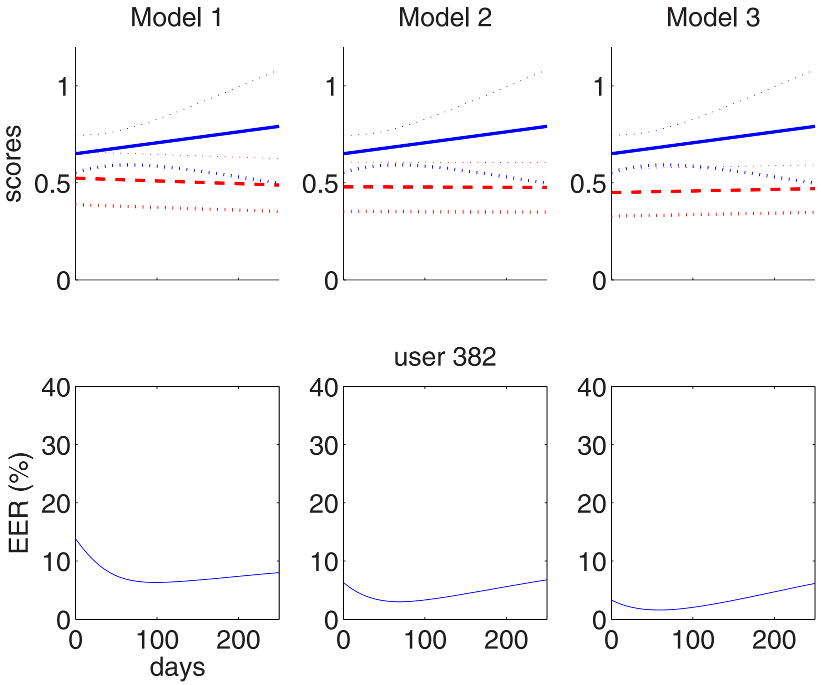


Fig. 2. The evolution of scores as estimated by regression (in the top row of figures) and their corresponding EER trend (bottom row) when using the first (column one), second (column two), third (column three) images according to the time-stamped sequence of a given user. The system used here is the 2D-LBP system. In the top figures, thick continuous lines are the expected trends of the genuine user match scores over time and thick dashed lines are that of the impostor match scores. Around these lines are their corresponding ± two standard deviations (shown in dotted lines).

general decrease in error rates over time before increasing again. It can be argued that, in general, biometric users become more acquainted with the system. As a result, the system performance may increase with use. However, because biometrics may change over time, the query images may gradually differ from the reference template. As a result, the system may degrade in performance. The system-level performance can be regarded as the *average performance* across users and so the above explanation cannot be readily observed from the set of individual user performance.

5 Conclusions

In this paper, we proposed a method to estimate user-specific performance. This is a difficult problem mainly due to the paucity of the genuine score samples. The availability of scores in time depends very much on how regular a biometric system is used. In the FRGC database, the most frequent interval is 7 days, followed by 14 days. By using an empirical error estimation approach, it is thus possible to estimate the error rate on a

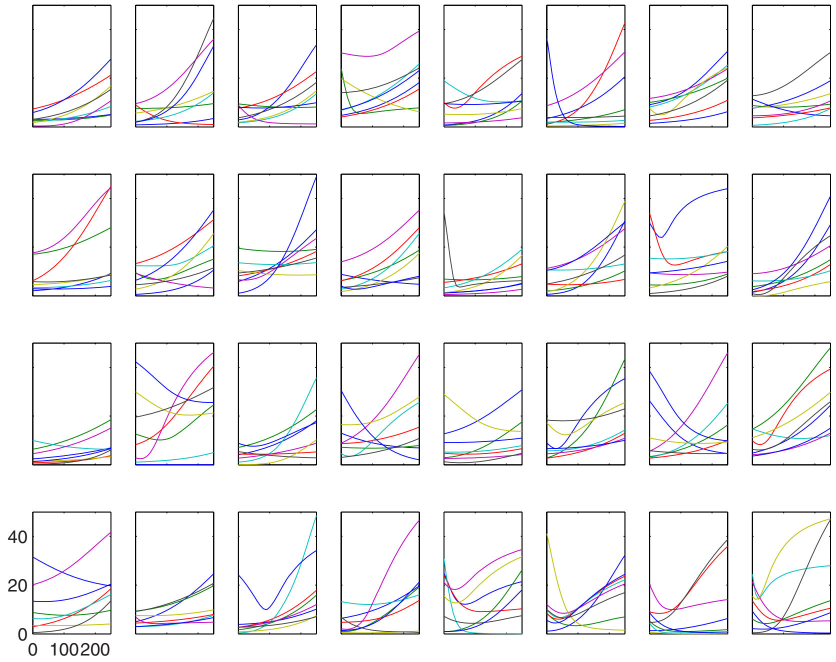


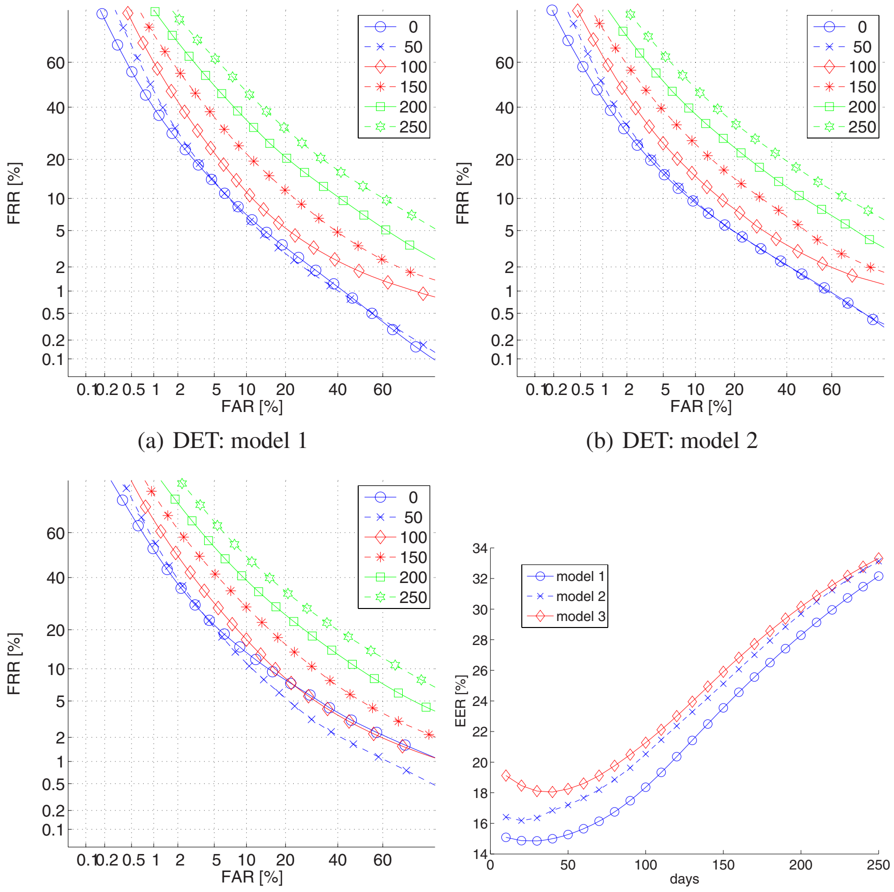
Fig. 3. The EER trend of all 256 users. Each of the 4×8 figures shows the trend of 8 users. The X-axis shows the number of days in $[0, 250]$ and the Y-axis is EER (%) in $[0, 50]$.

per day basis. By imposing the constraints that the user-specific class-conditional score sequence is continuous in time and that it takes on a particular distribution (Gaussian in our case), we demonstrated that our method can estimate the error rate on a per day basis. While the use of Gaussian assumption can be appropriate in our case, we do not claim that this is, in general, the case. The methodology, however, should be equally applicable on other data sets with a sensible choice of distribution.

Our experiments highlight the importance of user-specific performance analysis. This may open up a new research avenue towards customized biometric verification system, i.e., a system that is designed to adapt to the individual characteristic of a user. The proposed method can serve as an evaluation tool for this purpose. Customized biometric system is fascinating because learning with user-specific samples is a difficult task due to the small training sample size.

To the best of our knowledge, our study may be the first attempt to uncover person-dependent performance in a more principled way.

Our experiments show that the impostor score sequence does not need to evolve with time due to the aggregate effect of considering multiple impostor score sequences from a pool of impostors. As a result, modeling the genuine user sequence is of critical importance. Although a polynomial regression was used in this study, it may be logical to replace it with one that does not assume equal variance over the entire score sequence. Another obvious improvement is to replace the Gaussian assumption with a more realistic one.



(a) DET: model 1

(b) DET: model 2

(c) DET: model 3

(d) EER trends: all three models

Fig. 4. The evolution of the entire DET curve over the population of users (285 in total) on a 50-day interval given that the (a) first, (b) second and (c) third images in the time-stamped sequence are used as templates. Figure (d) shows the EER trend of the three models over 250 days. The system used here is the 3D-baseline system. The other two systems give similar trends, although their absolute performance differs slightly.

Acknowledgments

This work was supported partially by the prospective researcher fellowship PBEL2-114330 of the Swiss National Science Foundation, by the BioSecure project (www.biosecure.info) and by the Engineering and Physical Sciences Research Council (EPSRC) Research Grant GR/S46543. The main author also thank Prof. Anil Jain and Dr. Choonwoo Ryu for fruitful discussion on a similar topic except that it was applied to fingerprint recognition. This publication only reflects the authors' view.

References

1. Chen, K.: On the Dynamic Pattern Analysis, Discovery and Recognition, IEEE SMC Society eNewsletter (September 2005)
2. Flynn, P.J., Bowyer, K.W., Phillips, P.J.: Assessment of Time Dependency in Face Recognition: An Initial Study. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 44–51. Springer, Heidelberg (2003)
3. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Woves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: ICSLP, Sydney (1998)
4. Hastie, T.J., Tibshirani, R.J.: Generalized Additive Models. Chapman and hall (1990)
5. Poh, N., Bengio, S.: Why Do Multi-Stream, Multi-Band and Multi-Modal Approaches Work on Biometric User Authentication Tasks? In: ICASSP, Montreal, vol. V, pp. 893–896 (2004)
6. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the Face Recognition Grand Challenge. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 947–954 (2005)
7. Smith, R.S., Kittler, J., Hamouz, M., Illingworth, J.: Face Recognition Using Angular LDA and SVM Ensembles. In: Proc. 18th Int'l. Conf. on Pattern Recognition, pp. 1008–1012 (2006)
8. Ahonen, T., Hadid, A., Pietikainen, M.: Face Recognition with Local Binary Patterns. In: Proc. European Conference on Computer Vision, Prague, pp. 469–481 (2004)
9. Tena, J.R., Hamouz, M., Hilton, A., Illingworth, J.: A Validated Method for Dense Non-Rigid 3D Face Registration. In: AVSS 2006, p. 81 (November 2006)
10. Martin, A., Doddington, G., Kamm, T., Ordowsk, M., Przybocki, M.: The DET Curve in Assessment of Detection Task Performance. In: Eurospeech 1997, Rhodes, pp. 1895–1898 (1997)