Challenges and Research Directions for Adaptive Biometric Recognition Systems

Norman Poh[1], Rita Wong[1], Josef Kittler[1], and Fabio Roli[2]

[1] University of Surrey, Guildford, GU2 7XH, Surrey, UK
{n.poh,s.wong,j.kittler}@surrey.ac.uk
[2] Department of Electrical and Electronic Engineering,
University of Cagliari Piazza d'Armi 09123 Cagliari, Italy
roli@diee.unica.it

Abstract. Biometric authentication using mobile devices is becoming a convenient and important means to secure access to remote services such as telebanking and electronic transactions. Such an application poses a very challenging pattern recognition problem: the training samples are often sparse and they cannot represent the biometrics of a person. The query features are easily affected by the acquisition environment, the user's accessories, occlusions and aging. Semi-supervised learning – learning from the query/test data – can be a means to tap the vast unlabeled training data. While there is evidence that semi-supervised learning can work in text categorization and biometrics, its application on mobile devices remains a great challenge. As a preliminary, yet, indispensable study towards the goal of semi-supervised learning, we analyze the following sub-problems: model adaptation, update criteria, inference with several models and user-specific time-dependent performance assessment, and explore possible solutions and research directions.

1 Introduction

Portable electronic devices such as mobile phones and PDAs are becoming important means to provide wireless access to the Internet and other telecommunication networks anytime, anywhere. Very often, such access requires the verification of the user's identity in order to ensure that the person is really whom he/she claims to be. While knowledge-based authentication such as PINs or passwords can be used, they can be forgotten, or easily compromised when shared, copied or stolen. In comparison, biometrics is a more effective alternative because it is by far a more natural, reliable and friendly means of authentication. Thanks to the availability of cameras and microphones in today's mobile devices, audio- and visual-based biometrics such as face and speech can be readily used for this purpose.

In this context, we aim to develop and evaluate new mobile services that are secured by bi-modal speech and face biometrics. We shall call this problem "mobile biometry" (Mobio). Due to the device mobility, the problem of biometric authentication is much more challenging for at least two reasons: First, it has to deal with changing and often uncontrolled environments, e.g., external noise and varying illumination conditions. Under such conditions, a biometric query data can appear very differently from the one

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 753–764, 2009.

acquired during enrollment (template). As a consequence, the device performance can degrade drastically. Second, mobile devices have limited memory and CPU resources. This provides a natural constraint on the size of biometric template/model[1] and the type of processing algorithms (which favor those of low computation).

One possible way to improve the device authentication performance is by automatically labeling the abundant query (test) data and incorporating them as part of the training data [3,9]. Such a strategy belongs to a category of processes known as *semi-supervised learning* [14]. While initially developed for text classification, its application is now extended to speech verification [6] (termed incremental enrollment by the authors), face recognition [8] (called "Eigenspace updating") and multimodal biometrics [13] (known as template co-updating).

In [13], two distinct methods have been examined: self-training and co-training. In self-training, an initially trained classifier based on labeled data attempts to give labels to an unlabeled data set. The newly labeled data are then incorporated into the training set. The algorithm iterates until a convergence criterion is satisfied. In co-training, two initially trained classifiers based on labeled data attempt to provide labels to an unlabeled data set. Since two classifiers are involved, two labeled data sets are produced. The union of these two newly labeled data sets are then used as part of the training set for each of the two classifiers.

The study in [13] is particularly relevant to our problem here, i.e., mobile biometry. Indeed, it has been shown that both self-training and co-training are promising solutions to overcome the lack of biometric training samples at enrollment. However, mobile biometry presents a significantly different challenge, which can be summarized by the following four issues.

First, mobile biometry requires that data be labeled online (or in small batch of data) instead of offline (i.e., processing in batch), as is always practiced in experiments involving co-training/self-training. The online or small-batch processing is the consequence of the limited storage of mobile devices.

Second, there is no distinctive co-training and test phases as mobile devices are continuously updated as needed. This implies that using the usual three-partition experimental designs – involving training or enrollment (with only labeled data), co-training (with unlabeled data) and test data sets – is inappropriate. An ideal assessment would reflect the actual application scenario, which should contain only two partitions of data set: one for enrollment and a separate one for both co-training and testing (where error is estimated). As a result, there is a need to design an error estimator capable of tracking the error dynamics.

Third, there is a need to use quality measures in model adaptation[2]. Until now, procedures on co-training depend uniquely on the confidence of classifier decision, which is

[1] We use the term "template" when referring to the stored features representing a person's biometric trait whereas the term "model" as a more general concept in order to refer to the parameters of a discriminative classifier or a statistical model, or that of an intermediate feature extraction process such as eigenspace analysis.

[2] Model adaptation is a more general concept than semi-supervised learning such as co-training. The latter refers uniquely to training or adaptation on self-labeled test data. In general, adaptation refers to the update of *model parameters* in various contexts, e.g., quality of samples.

often derived directly from the classifier output. In the biometric community, an important and rapidly developing research area is concerned with the use of quality measures. For example, if a quality measure is designed to annotate the head orientation of a face image (giving pitch, tilt and yaw angles), it would make sense to cluster the images according to different head pose and create a new template for each cluster. Conversely, without quality measures, a direct update to the old model, without considering head orientations can be catastrophic. The same argument applies to other quality measures annotating the presence of glasses, the type of emotional states and lighting conditions. This motivates us to use quality-dependent model adaptation, in addition to the confidence of the classifier output, as a criterion to create a new model or update an existing one.

Last but not least, when there are several models/templates (recalling that the number is bounded by the memory of mobile devices), inference with several models requires a special treatment. Drawing on the work in [10], a possible framework is also proposed in this paper for mobile biometry.

Section 2 begins with a more formal presentation of the concept of co-training according to the original paper of Blum and Mitchell [3]. The remaining sub-sections analyze the above mentioned four issues related to applying semi-supervised learning to biometric authentication in general, but using mobile biometry as a test field. For each issue, a possible solution is also elaborated. Section 3 then concludes the paper.

2 Open Issues and Research Directions for Adaptive Biometric Systems

We wish to learn the concept $f : X \rightarrow Y$ given some labeled L and unlabeled U data set drawn from $P(X)$, where X is a feature vector and Y is a class label. The features describing X can be partitioned into $X^{(1)}$ and $X^{(2)}$, i.e., $X = X^{(1)} \times X^{(2)}$, such that f can be computed from either $X^{(1)}$ or $X^{(2)}$. Let g_1 and g_2 be the trained classifier from $X^{(1)}$ and $X^{(2)}$, respectively. Then, the objective of co-training can be expressed as finding g_1 and g_2 such that:

$$\exists_{g_1,g_2}(\forall_x \in X)g_1(x^{(1)}) = f(x) = g_2(x^{(2)}).$$

from U and L. An example of co-training procedure for a bimodal face and speech biometric authentication problem, which is a binary classification task, is shown in Algorithm 1.

An important result from [3] is that if $X^{(1)}$ and $X^{(2)}$ are *conditionally independent* given Y, and the concept f is "PAC learnable" from noisy labeled data, then, f is PAC learnable from weak initially labeled data plus unlabeled data. Probably approximately correct learning or PAC learning refers to the fact that one can train a classifier in order to learn the concept f, giving low generalization error in finite time and space.

In biometric authentication, $X^{(1)}$ can be a feature vector extracted, for instance, from a face image, whereas $X^{(2)}$ can be a feature vector extracted from a speech recording. Y is a person's identity. Not only that $X^{(1)}$ and $X^{(2)}$ are *conditionally independent* given Y, a condition that must be fulfilled in order for co-training to work, they are simply (unconditionally) independent in our application. Because of this independence,

Algorithm 1. The co-training algorithm

- Given: labeled data L and unlabeled data U
- Loop:
  - Train $g1$ (face classifier) using L
  - Train $g2$ (speech classifier) using L
  - Allow $g1$ to label p positive, n negative examples from U
  - Allow $g2$ to label p positive, n negative examples from U
  - Add these self-labeled examples to L

for instance, it is not possible to predict a person's face features given his/her speech features without any prior information. We therefore have a strong case here supporting the conjecture that co-training will work for bimodal person authentication. However, in practice, updating the wrong samples, i.e., from other users, may result in degraded performance [6]. Therefore, despite the sound theory, over-updating a model with the wrong person's biometric data will eventually be counter-productive.

A second obstacle to the successful deployment of semi-supervised learning in biometrics is that the intra-person variability is larger than inter-person variability. For instance, the lighting conditions can cause two face images of the same person appear much more differently than two images of different persons taken in the same lighting condition. This implies that trying to incorporate all types of variability into a single model will only reduce its discriminative power, hence, decreasing the recognition performance. A simple, yet, effective solution is to maintain several models, each capturing only local variation as gauged by quality measures (lighting conditions being one example here).

The following four sections will address open issues related to adaptive biometric systems. The first section provides a Bayesian interpretation of model adaptation. An example based on fingerprint minutia-based classifier is also shown. The second section addresses training and inference with several models. Each model differs from the others as it captures an aspect of a biometric template based on some observed quality measurements. The third section proposes a simple model creation/update criterion, also based on quality measurements. Finally, the last section addresses the issue of performance estimation as models adapt through time.

2.1 A Bayesian Interpretation of Biometric Model Adaptation

There are basically three operations that are essential to manipulating biometric models/templates; they are defined as follow:

- Model creation
$$Create : data \rightarrow new\ model$$
- Model adaptation:
$$Update : model, data \rightarrow updated\ model$$
- Model deletion
$$Delete : model \rightarrow \emptyset$$

While adding and deleting models is straightforward, it is not so for model adaptation because the latter requires combination of several biometric samples. Classifiers based on statistical models can often be implemented in the form of online learning, i.e., "old" model parameters can be updated with new ones after observing a training sample. This implies that learning can be done incrementally. A state-of-the-art classifier in speaker verification known as Gaussian mixture model (GMM) with maximum *a posteriori* (MAP) adaptation is a good example:

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

where $p(x|\theta)$ is the likelihood of the data (given a model with parameter θ) and $p(\theta)$ is the prior probability over the parameter θ. One maximizes the right hand side in order to find the most probable value of θ, i.e., θ_*. This value becomes an initial estimate that can be updated when a new sample becomes available.

If there are $x_1, \ldots, x_T$ observations, also denoted as $x_1 : x_T$, one finds the value of θ that maximizes $p(\theta|x_1 : x_T)$, as follows (ignoring the normalizing factor in each step since we are only interested in maximizing the function with respect to θ):

$$\begin{aligned} p(\theta|x_1 : x_T) &\propto \prod_{i=1}^{T} p(x_i|\theta)p(\theta) \\ &\propto \prod_{i=2}^{T} p(x_i|\theta)p(\theta|x_1) \\ &\propto \prod_{i=3}^{T} p(x_i|\theta)p(\theta|x_1, x_2) \\ &\propto \vdots \\ &\propto p(x_T|\theta)p(\theta|x_1 : x_{T-1}) \end{aligned} \tag{1}$$

where $p(\theta|x_1) \propto p(x_1|\theta)p(\theta)$. The recursive formulation of (1) implies that in order to calculate the optimal value of θ given all previously observed T samples, one only needs to use the parameter calculated up to $T-1$ to do so. This dispenses with the need to keep all previous training samples, which is a memory demanding requirement. The above recursive formulation of MAP implies that density-based classifiers can benefit from continuous training as this leads to finding the optimal value of θ. In the terms used in [4, Chap. 3], the recursive formulation of (1) is known as *true recursive Bayesian learning*, its right hand term, $p(\theta|x_1 : x_T)$, as a *reproducing density*; and the term $p(\theta)$ as a *conjugate prior*.

For non-statistical model, adaptation may seem not obvious at first. For instance, Jiang and Ser [7] showed that it is possible to combine several minutiae-based fingerprint templates to form a "super-template". It turns out that the super-template approach can be seen as a *simplified* statistical model taking the form of a multivariate Gaussian distribution with isotropic covariance. Let x_m^t be a minutia at the m-th location of the t-th fingerprint. We require here that all minutia locations of different templates are

aligned first. A minutia can contain location and/or directional information. Assuming that the minutiae of a fingerprint is independent, a fingerprint template can then be represented by $p(X|\theta) = \prod_m p(x_m|\theta)$ where $p(x_m|\theta)$ is a Gaussian and $\theta = \epsilon$ is common to all observations t and locations m, i.e.,

$$p(X|\theta,t) = \prod_{m=1}^{M} p(x_m|\theta,t) = \prod_{m=1}^{M} \mathcal{N}(x_m|\mu_m^t,\epsilon)$$
$$= \frac{1}{(2\pi\epsilon)^{M/2}} \exp\left\{-\frac{1}{2\epsilon}\sum_{m=1}^{M} ||x_m - \mu_m^t||^2\right\}$$

where μ_m^t is the location (and orientation) of a minutiae taken from the t-th template. Note that ϵ corresponds to the variance of x_m. The posterior is then:

$$p(\theta|X_1 : X_T) \propto p(X_1 : X_T|\theta)p(\theta) \tag{2}$$
$$= \prod_{t=1}^{T} p(X_t|\theta)p(\theta)$$
$$\propto \exp\left\{-\frac{1}{2\epsilon}\sum_{t=1}^{T}\sum_{m=1}^{M} \| \underbrace{x_m - \mu_m^t}\|^2\right\} \tag{3}$$

with equality only if we consider the normalizing factor $\frac{1}{(2\pi\epsilon)^{MT/2}}$. Here, $p(\theta)$ is a uniform distribution over the θ space.

When $\epsilon \to 0$, the minutia x_m^t that is close to μ_m^t will give a peak value and near zero otherwise. In this limit, we see that the counting approach proposed by Jiang and Ser [7] and (3) converge. According to (3), with a large number of samples T, spurious minutiae will receive low weights, whereas frequently occurred minutiae will receive high weights, hence, playing a more important role during recognition. This behavior is very similar to the k-nearest neighbor algorithm [4].

A related study on face recognition, called "Eigenspace-updating", found in [8] can also be considered a special case of our recursive MAP framework. In this case $p(x|\theta)$ is a multivariate Gaussian whose parameters θ are a mean vector and a covariance matrix. It is well known that mean and covariance can be incrementally updated by observing one example at a time. This is, again, a realization of (1). However, slightly different from this formulation, the authors introduced the concept of weight decay whose aim is to give more weight to the latest test samples rather than giving all samples equal weights.

A second concept called "twin-subspace updating", also found in [8], is to maintain two models (classifiers) instead of one. The motivation is that one model may capture a frontal view whereas another may capture another slightly different view (a profile view would be another extreme). When there are more than one models, it is necessary to decide which model will contribute more to the final match score. This issue is treated in the next section.

2.2 Quality-Based Training and Inference with Multiple Models

In this section, we will first treat the subject of training with several models and then consider how inference conditioned on quality can be approached using a Bayesian framework.

Before doing so, it is instructive to categorize templates/models into the following three primary types:

- single-template, where only a single template is available
- multi-template, where several templates are used
- super-template, where several templates are combined to form a single one.

Higher level types of templates are possible, for example, using multiple super-templates where each super-template captures an aspect of biometric features, acquired in a particular condition (e.g., same head orientation, lighting conditions, etc) or using a particular device.

Let q denote a sample quality measurement and θ_* be the optimal value that maximizes (2), i.e., $\theta_* = \arg\max_\theta p(\theta|X_1 : X_T)$. In this case, it is reasonable to expect that samples of similar quality to be comparable; otherwise, they are not. This suggests that one should choose a model/template that matches a particular acquisition condition measurable by $q \in \mathbb{R}^{\mathtt{q}}$, i.e., a vector containing $\mathtt{q}$ measurements. For instance, for face verification, these measures can be illumination, focus, reliability of face detection, etc. It is possible to cluster q into several states. Let Q be the cluster indices of q. We expect that images of good quality will cluster together and similarly for those of moderate and bad quality. A well known clustering algorithm that can suitably be used for this purpose is a Gaussian mixture model (GMM) [1]. For a pre-specified number of clusters $\mathtt{Q}$, a GMM models the density $p(q) = \sum_{Q=1}^{\mathtt{Q}} p(q|Q)p(Q)$. The parameters of this model are found using the Expectation-Maximization (EM) algorithm. The posterior probability of a cluster Q given an observation q, also known as *responsibility* in EM, is then given by $P(Q|q) = \frac{p(q|Q)P(Q)}{p(q)}$. Now, we can estimate the optimal parameter θ using the recursive MAP framework, but in a quality-dependent manner:

$$p(\theta|x_1 : x_T, Q) \propto p(x_T|\theta, Q)p(\theta|x_1 : x_{T-1}, Q) \tag{4}$$

which is very similar to (1), except that all terms are dependent on the cluster index Q. Thanks to (4), we can estimate:

$$p(\theta|x_1 : x_T, q_1 : q_T) = \sum_Q p(\theta|x_1 : x_T, Q)P(Q|q_1 : q_T) \tag{5}$$

which is the actual goal. One maximizes (5) with respect to θ and this can be done in a recursive MAP framework, i.e., when samples arrive one at a time. Using a derivation very similar to (1), except that all terms are dependent on Q, one can show that (5) is proportional to:

$$p(\theta|x_1 : x_T, q_1 : q_T) \propto \underbrace{\sum_Q p(x_T|\theta, Q)P(Q|q_T)}\cdot p(\theta|x_1 : x_{T-1}, q_1 : q_{T-1}). \tag{6}$$

(6) shows that new parameters (the left hand side of equation) can be updated from old parameters (represented by $p(\theta|x_1 : x_{T-1}, q_1 : q_{T-1})$) using the underbraced term. Therefore, in order to maximize $p(\theta|x_1 : x_T, q_1 : q_T)$ with respect to θ, one does not need to keep the previous samples $(x_1 : x_{T-1}, q_1 : q_{T-1})$ but only needs the last estimate of parameters from these samples. This shows that one can design an online algorithm not only for a conventional model (without quality information), but also for the conditional model based on quality.

Note that the responsibility term $P(Q|q_T)$ only appears in the underbraced term, implying that $P(Q|q_T)$ is only effective on the latest T-th observation. Conditioning the parameter estimation based on Q, as in (4), therefore makes the estimation of (5) (being the real objective) a tractable proposition.

When there are several models/templates, it is reasonable to expect that some models may contribute more in making the accept/reject decision than the others. Let θ_*^Q be the optimal value that maximizes (5), and $Q = 1, \ldots, \mathtt{Q}$. In this case, using the Neyman Pearson theorem, the optimal classifier should give the following output:

$$y = \log \frac{\sum_Q p(x|\theta_*^Q, \mathtt{C}) P(Q|q)}{\sum_Q p(x|\theta_*^Q, \bar{\mathtt{C}}) P(Q|q)} \tag{7}$$

where $p(x|\theta_*^Q, \mathtt{C})$ is the density of a client (the reference user) model whereas $p(x|\theta_*^Q, \bar{\mathtt{C}})$ is the density of an anti-client model, i.e., one that represents all other users. This is a concept borrowed from speaker verification, also known as a universal background model [12]. It can be observed that the contributions of these two terms are appropriately weighted by the responsibility term $P(Q|q)$. The effectiveness of the above formulation was shown in [10], in the context of intramodal fusion involving several face verification systems, with x being a vector of system outputs and q being a vector of some face related quality measures such as reliability of face detection, background uniformity, presence of glasses, head orientation, etc. Note that these factors are used because they are known to affect the performance of a face verification system.

Instead of using (7), in a simplified case, we can also choose to use only a single model that is the *most* appropriate (according to the responsibility term):

$$y = \log \frac{\sum_Q p(x|\theta_*^Q, \mathtt{C}) f(Q, q)}{\sum_Q p(x|\theta_*^Q, \bar{\mathtt{C}}) f(Q, q)},$$

where

$$f(Q, q) = \begin{cases} 1 & Q = \arg\max_Q P(Q|q) \\ 0 & \text{otherwise,} \end{cases}$$

2.3 Quality-Based Model Update and Creation

In the previous section, we showed that it is possible to make inference with several models. Each model is different because it gauges the feature density for a given Q which can be the acquisition condition, or a different presentation. In both cases, an observed biometric feature set will appear very differently from the previously stored biometric template.

A criterion is indeed needed to identify when a model should be updated or created. If x' is a newly labeled sample having q' as its quality measurements, one chooses the quality state Q_* that maximizes the responsibility $P(Q|q')$, i.e.,

$$Q_* = \arg \max_{Q=1}^{\mathtt{Q}} P(Q|q'). \tag{8}$$

However, in reality, the model $p(x'|\theta, Q_*)$ may not exist. In fact, right after enrollment, in a typical scenario where enrollment is done with a single template, there is only one quality state, i.e., $\mathtt{Q} = 1$. In the absence of the model $p(x|\theta, Q_*)$, one should naturally create the model $p(x|\theta, Q_*)$ using the observation x'.

However, if the model $p(x|\theta, Q_*)$ exists, one can update the model using (6). The new parameter can be obtained as follows:

$$\theta_{new}^{Q} = \arg \max_{\theta} \sum_{Q} p(x'|\theta, Q) p(\theta_{old}^{Q}) P(Q|q') \tag{9}$$

A simpler approach, taking the hidden variable Q as an observed one, is to use the following update rule:

$$\theta_{new}^{Q_*} = \arg \max_{\theta} p(x'|\theta, Q_*) p(\theta_{old}^{Q_*}) \tag{10}$$

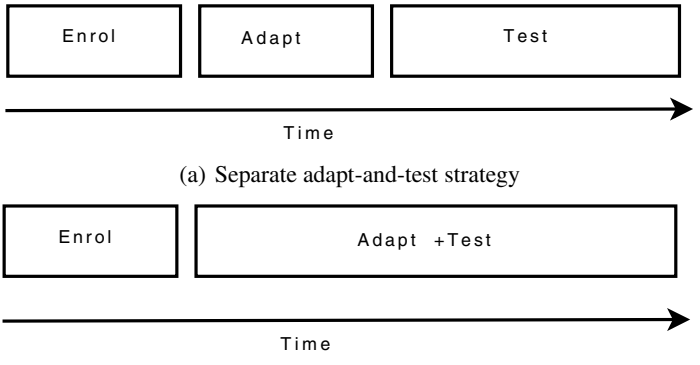
Both (9) and (10) constitute two variants of the maximization step in a typical EM algorithm. In comparison, in both cases, (8) corresponds to the expectation step in an EM algorithm [2].

$P(Q|q)$ represents a prior knowledge of all possible combinations of factors affecting the system performance. This function was obtained by modeling $p(q|Q)$ using a clustering algorithm. This suggests that one should train $p(q) = \sum_Q P(Q)p(q|Q)$ from as much data as possible from a separate development database containing many more persons, environmental conditions, presentation styles, and even different devices. An accurate estimation of $P(Q|q)$ would guarantee the success of model adaptation. Designing a good set of quality measures, therefore, cannot be overemphasized.

2.4 Person-Specific and Time Dependent Performance Evaluation

There are, in general, two approaches to performance assessment with evolving biometric models. We shall refer to one of these approaches as a *separate* adapt-and-test strategy and another as a *joint* adapt-and-test strategy. Both are shown in Figure 1. In the first strategy, a partition of data is used to adapt a biometric model and another disjoint partition is used for evaluating its performance. Such an approach is inefficient in terms of data usage since the data used for adaptation cannot be used for testing, and vice-versa. Worst still, it does not reflect our actual application scenario where a mobile device is allowed to update biometric models on the test data. However, by using a unique test set, one can use the widely accepted error counting approach. In this approach, one first determines a threshold and then calculates the number of false match and false non-match events.

In contrast, the joint adapt-and-test strategy, which allows one to adapt (with unlabeled data) and test (with the known labels), has two advantages. First, it corresponds



(a) Separate adapt-and-test strategy

(b) Joint adapt-and-test strategy

Fig. 1. Two assessment strategies for co-training: (a) is a usual strategy and (b) is our proposal

better to our mobile application scenario. Second, it is more efficient in terms of data usage, since the data used in model adaptation can, at the same time, be used to asses the model performance. Unfortunately, its major disadvantage is the absence of an error estimator.

We shall elaborate on an error estimator that can be used for joint adapt-and-test strategy mentioned above, which is adapted from [11]. This procedure is able to track performance change in terms of false match rate (FMR) and false non-match rate (FNMR) over time and on a per person basis. Estimating this error is difficult because of the paucity of data, especially the genuine user scores. However, it is possible if one imposes the constraint that the user-specific class-conditional (genuine user or impostor) scores follow a particular parametric family of distributions (Gaussian in [11]) and that it is continuous in time. In so doing, one can estimate the performance to an arbitrary time precision. This method compares favorably with the conventional error-counting approach which utilizes a sliding window, e.g., [5], 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.

In the context of [11], it was found that even without any model adaptation, some biometric models can degrade over time, while others improve with use. This phenomenon justifies the adaptation of biometric model by using the test data. However, they also suggest that one should consider the adaptation on a per user basis. In our case, when one regularly adapts a model through time, it is reasonable to expect that the performance of a biometric model will evolve with time.

We shall summarize the procedure to estimate the error trend on a per person basis below [11]:

1. Fit a regression line to each of the genuine and impostor match scores. An example of regression fits for both genuine and impostor scores can be found in Figure 2[3].

[3] In principle, one does not expect any trend for the impostor scores, i.e., the regression line for the impostor should be parallel to the x-axis. Therefore, any deviation from this can be attributed to estimation error.

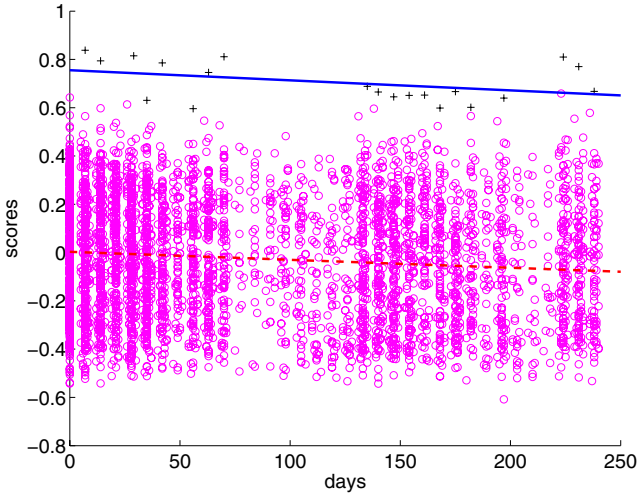


Fig. 2. Scatter plot of genuine user ("+") and impostor ("∘") match scores for a single user's template over 250 days (the X-axis). Higher match scores imply genuine user class. The interruption in genuine match scores around the 100-th day is due to no observations being made during the term break. 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).

2. Obtain the first and second order statistics of the two regression lines for a fixed time interval. Let these parameters be $(\mu^I_{j,t}, \sigma^I_{j,t})$ for the impostor regression line and $(\mu^C_{j,t}, \sigma^C_{j,t})$ for the client counterpart, where j denotes the identity of a user and $t \in [1, T]$ is a time index.
3. Estimate the false match rate and false non-match rate as follows:

$$\text{FMR}_{j,t}(\Delta) = \Phi\big(\Delta|\mu^I_{j,t}, (\sigma^I_{j,t})^2\big) \tag{11}$$

and

$$\text{FNMR}_{j,t}(\Delta) = 1 - \Phi\big(\Delta|\mu^G_{j,t}, (\sigma^G_{j,t})^2\big) \tag{12}$$

for a given threshold Δ in the score space, where $\Phi\big(\Delta|\mu, (\sigma)^2\big)$ is a cumulative density function with mean μ and standard deviation σ for a chosen distribution.

Once $\text{FNMR}_{j,t}(\Delta)$ and $\text{FMR}_{j,t}(\Delta)$ are calculated, they can be plotted for each user j and through time $t = 1 : T$. In this way, one can plot a user-specific detection error trade-off (DET) or receiver operating characteristic (ROC) curves that evolves with time.

3 Conclusions

In this paper, we highlighted four challenges as well as provided solutions to semi-supervised learning in the context of biometric person authentication. The issues are:

online model updating, training and inference with several models, quality-based criteria to control the creation of a new model, and person-specific time-dependent performance evaluation. At the time of writing, an audio-visual database is being collected with mobile devices. Future experiments will be conducted to test each of the proposed solutions.

Acknowledgement

This work was supported partially by the prospective researcher fellowship PA0022_121477 of the Swiss National Science Foundation, and by the EU-funded Mobio project grant IST-214324 (www.mobioproject.org).

References

1. Bishop, C.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1999)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2007)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT: Proceedings of the Workshop on Computational Learning Theory, pp. 92–100. Morgan Kaufmann Publishers, San Francisco (1998)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification and Scene Analysis. John Wiley and Sons, New York (2001)
5. 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)
6. Fredouille, C., Mariéthoz, J., Jaboulet, C., Hennebert, J., Mokbel, C., Bimbot, F.: Behavior of a bayesian adaptation method for incremental enrollment in speaker verification. In: ICASSP 2000 - IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, June 5–9 (2000)
7. Jiang, X., Ser, W.: Online fingerprint template improvement. IEEE Tran. Pattern Analysis and Machine Intelligence 24(8), 1121–1126 (2002)
8. Liu, X., Chen, T., Thornton, S.M.: Eigenspace updating for non-stationary process and its application to face recognition. Pattern Recognition 36(9), 1945–1959 (2003)
9. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.M.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39(2/3), 103–134 (2000)
10. Poh, N., Heusch, G., Kittler, J.: On Combination of Face Authentication Experts by a Mixture of Quality Dependent Fusion Classifiers. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 344–356. Springer, Heidelberg (2007)
11. Poh, N., Kittler, J.: A Method for Estimating authentication Performance Over Time, with Applications to Face Biometrics. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 360–369. Springer, Heidelberg (2007)
12. Reynolds, D.A., Quatieri, T., Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10(1–3), 19–41 (2000)
13. Roli, F., Didaci, L., Marcialis, G.L.: Template co-update in multimodal biometric systems. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 1194–1202. Springer, Heidelberg (2007)
14. Zhu, X.: Semi-supervised learning literature survey, University of Wisconsin-Madison (2008) (online publication)