Combining two strategies to optimize biometric decisions against spoofing attacks

Weifeng Li[a,c], Norman Poh[b], and Yicong Zhou[c,*]

[a]Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China
[b]Department of Computing, University of Surrey, Surrey, United Kingdom
[c]Department of Computer and Information Science, University of Macau, Macau, China

ABSTRACT

Spoof attack by replicating biometric traits represents a real threat to an automatic biometric verification/ authentication system. This is because the system, originally designed to distinguish between genuine users from impostors, simply cannot distinguish between a replicated biometric sample (replica) from a live sample. An effective solution is to obtain some measures that can indicate whether or not a biometric trait has been tempered with, e.g., liveness detection measures. These measures are referred to as evidence of spoofing or anti-spoofing measures. In order to make the final accept/rejection decision, a straightforward solution to define two thresholds: one for the anti-spoofing measure, and another for the verification score. We compared two variants of a method that relies on applying two thresholds – one to the verification (matching) score and another to the anti-spoofing measure. Our experiments carried out using a signature database as well as by simulation show that both the brute-force and its probabilistic variant turn out to be optimal under different operating conditions.

Keywords: Spoof attack, Biometric verification, Biometric authorization

1. INTRODUCTION

1.1 Motivation

A biometric authentication system can often be tempered with easily by using a replicated sample (replica) instead of a live sample. In this scenario, an malefactor has obtained a trace of a genuine sample and produced a biometric replica with high fidelity. This attack is known as a *spoof attack*, as opposed to a *zero-effort* attack. In the latter case, the impostor does not have (or exploit) the knowledge or the trace of the biometric trait of the person he/she attempts to impersonate. Examples of biometric replicas abound: gummy fingers,[1] synthesized voice forgery via transformation,[2] and animated talking faces.[3] Various spoofing attempts for audio-visual person verification have been reported in.[4]

In order to counteract spoof attacks, liveness detection measures have been developed and the research in this direction is gaining momentum. According to,[5] liveness detection measures can be grouped into two categories: (i) the intrinsic properties of the living human tissue and (ii) involuntarily generated signals. The first approach is commonly referred to as *liveness (or spoof) detection*. This solution consists of measuring some some physical properties capable of distinguishing between a living sample and a replica. Examples are the spectral characteristics of the skin, e.g., absorbance, transmittance and reflectance of the electromagnetic radiation of different wavelengths; properties of the body fluids, e.g., blood oxygenation; the electrical properties of the living human skin, e.g., conductance or dielectric constant; and, physical or visual skin properties, e.g., density, elasticity or color of skin. The second approach measures the signals that are spontaneously and uncontrollably generated by a living human body. Examples are pulse, perspiration and temperature. However, this categorization is not necessarily exhaustive. For instance, in an attempt to counteract audio-visual spoofing,[4] one can exploit the audio-visual synchrony information between lips and speech (which are, as a matter of fact, voluntarily generated signals). In this paper, we shall collectively refer to measurements designed to detect spoofing as *evidence of spoofing* (regardless of whether it is a measure of liveness, voluntarily or involuntarily generated signals, or other properties related to living human tissues).

*Yicong Zhou: E-mail: yicongzhou@umac.mo, Telephone: +853 83978458;
Weifeng Li: E-mail: li.weifeng@sz.tsinghua.edu.cn;
Norman Poh: E-mail: normanpoh@ieee.org

Optics and Photonics for Information Processing VIII, edited by Abdul A. S. Awwal, Khan M. Iftekharuddin, Mohammad A. Matin, Andrés Márquez, Proc. of SPIE Vol. 9216, 92161H © 2014 SPIE
CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2061785

Table 1. The four events of a biometric system and the desirable actions to take when considering spoof attacks

action	replication state	comparison	attack type
accept	live	match	non-attack ("genuine")
reject	live	non-match	zero-effort
reject	replica	match	spoof
reject	replica	non-match	impossible

A very common way of making the final decision from the verification score y and the evidence of spoofing e is by thresholding the two measurements:

$$\text{decision}(y, e) = \begin{cases} accept & \text{if } y > \Delta_1 \wedge e < \Delta_2 \\ reject & \text{otherwise,} \end{cases} \quad (1)$$

where Δ_1 is a threshold applied to y and Δ_2 is a threshold applied to e. Setting these two thresholds is a highly empirical exercise and is certainly application dependent.

Although there exist other types of attack in the literature, namely, replay and brute-force attacks,[6] they are not covered here because they are relevant only in the context of biometric authentication *over the Internet*, which involves the communication between a client and a server computer. A *replay* attack involves resubmission of a previously acquired signal at a client computer recorded signal is replayed to the system, bypassing the sensor. The concept of *brute-force* attack has its origin in password-based authentication systems; it involves enumeration of all possible passwords. In the context of biometric security, this attack amounts to enumerating all possible biometric signals or templates. The topic treated here, i.e., spoof attack, is an attack at the sensor level, in which case, an malefactor has no access to the biometric system architecture but possesses a biometric trace of the person whom he/she attempts to impersonate.

1.2 Terminology

Since the subject of liveness or spoof detection is relatively new, we are obliged to define some terms, not necessary adopted elsewhere in the literature, but at least consistent throughout this paper.

A verification score is considered a *match* if it is a result of comparing a biometric reference with a biometric sample, both of which originated from the same person. Conversely, if both the reference and the sample are from two different persons, then, the score is a *non-match*.

A biometric sample is referred to as a *live* one if it has been obtained from a living person; otherwise, it is a *replica* or a spoofed sample. One can use the *evidence of spoofing* to infer how likely it is that a biometric has been tempered with. The term "spoofing" here is synonymous with impersonating, masquerading or mimicking. Thus, the value of an evidence of spoofing should be low if a biometric sample is taken from a living person, and high if it is a biometric replica. We avoid the term "liveness measure" because as mentioned in,[5] a liveness measure does not necessarily quantify liveness, but rather physical properties of a biometric trait (e.g., skin transmittance, absorbance, etc).

For each access, a biometric system can obtain two measurements: the verification (comparison) score and the evidence of spoofing. This joint observation is a result of the following two dichotomies of events: match versus non-match and live versus replicated sample. The desirable course of action to take for each of these four events are shown in Table 1. Thus, we reserve the term "genuine" (in the last column of Table 1) to mean that a biometric sample is both a live one *and* the comparison is a match. In this case the correct decision is to accept the access request. Similarly, we reserve the term "impostor" to cover the remaining three cases which should result in reject decisions, as follows: (i) a *zero-effort attack* in a comparison involving a non-match live sample; (ii) a *spoof attack* under a match comparison involving a replica; and (iii) an *impossible attack* in real life (but irronically conceivable in an experimental setting) involving a non-match replica.

In order to understand the different types of attack, it is instructive to consider the following hypothetical break-in example: John tries to illegally access Smith's notebook that is protected by a fingerprint sensor. In this case, it really does not matter if John uses his own fingerprint or a replicated sample of his fingerprint. The first case is a zero-effort attack and the second is impossible, i.e., John has no incentive to replicate his own fingerprint in order to access Smith's notebook. This is because in either case, the comparison is a non-match and the system is likely reject the access request. In a real world attack, John would use an exact replica of Smith's fingerprint. This constitutes a spoof attack.

1.3 Contributions

We propose two strategies to optimally set the above thresholds: brute-force optimization and probabilistic. The first strategy consists of exhaustively search for an optimal solution, minimizing a performance criterion. The second approach capitalizes on the logic construct of (1) but in *probabilistic sense*. For this reason, we also refer to the first strategy as "double threshold" and the second one as "probabilistic double threshold".

Although both strategies rely on the same logic construct, the brute-force optimization does not commit to an assumption that its probabilistic version does; that is, the latter assumes that both the matching score and the anti-spoofing measure are independent of each other. Experimental results on real and simulated data show that both strategies have their own advantages in different parts of Receiver's Operating Characteristic (ROC) curve. In particular, we found that the double-threshold method (with brute-force optimization) performs better at low false acceptance rate (FAR) whereas its probabilistic version dominates at low false rejection rate (FRR).

The observed superiority of the brute-force optimization suggests that treating both the matching score and the evidence of spoofing (or anti-spoofing measure) as *independent sources of information* is not appropriate. This calls for further investigation of appropriate modelling techniques.

This paper is organized as follow: Section 2 presents our proposal. Section 3 presents a case study, and this is followed by conclusions in Section 4.

2. METHODOLOGY

2.1 Notation

We shall adopt the following notation:

- $y \in \mathbb{R}$ is a verification score. We shall interpret the score as a similarity score, such that a high value implies a match comparison whereas a low value implies a non-match comparison. If a biometric matching module outputs a distance or dissimilarity score, for instance, one can simply invert the sign of the score in order to interpret it as a similarity score.
- $M \in \{1, 0\}$ is the state of comparison (between a reference/template and a query sample), which can either be a match or a non-match.
- the replication event $R \in \{0, 1\}$. This is a binary event, i.e., a sample is either a replica or not.
- $e \in \mathbb{R}$ is the evidence of spoofing. An example of this is a fingerprint liveness detection measure*. Often, the system designer has to design an evidence *specific* to a type of spoof attack. Since there are many ways to spoof a system, in principle, one has to design a measure to seek evidence for each type of attack. As a preliminary study, we shall limit the scope of discussion to a single attack.

2.2 Our Proposals

2.2.1 Double Threshold by Brute-force Optimization

The brute-force approach simply searches for all possible threshold pairs in the space spanned by both the matching score and the evidence of spoofing, $\mathcal{Y} \times \mathcal{E}$ and then search the solution that minimizes a criterion. This procedure is shown in Algorithm 1. In this example, the performance is measured in terms of half total error rate (HTER), defined as the average of false acceptance rate (FAR) and false rejection rate (FRR). Recall that the "genuine" or positive class here is defined by a match comparison and the replication state being "live". The remaining classes which include the zero-effort and spoof attacks, as shown in Table 1 are collectively referred to as a negative class. FAR is, therefore, defined by:

$$\text{FAR} = \frac{\text{\# of wrongly classified negative examples}}{\text{\# of total examples}};$$

*The framework developed here does not restrict e to be a scalar value, i.e., e can be a vector of measurements. We treat e as a scalar value because in the liveness detection literature,[7–9] one often builds a dedicated classifier for this task and so e can be viewed as an output of this classifier.

Algorithm 1 Double Threshold by Brute-force Optimization

HTER$_{min}$ = ∞
for $y \in \mathcal{Y}$ **do**
  for $e \in \mathcal{E}$ **do**
    HTER = Evaluate performance with (y, e)
    if HTER$_{min} \geq$ HTER **then**
      HTER$_{min}$ = HTER
      $y_* = y$
      $e_* = e$
    end if
  end for
end for
return (y_*, e_*)

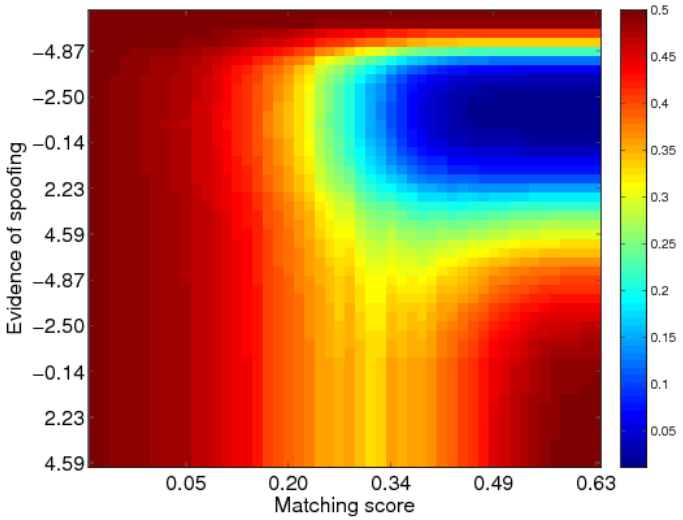

Figure 1. An example of brute-force optimization. The vertical bar shows the value of HTER in the (y, e) space.

whereas FRR by:

$$\text{FRR} = \frac{\text{\# of wrongly classified positive examples}}{\text{\# of total examples}}.$$

Although HTER is used here, other criterion such as the weighted error rate (WER) that weighs FAR and FRR in different proportions can be used. One can recognize that WER is a more general criterion than HTER because the latter weighs the two errors in equal proportions.

2.2.2 Probabilistic Double Threshold

Effectively, one seeks to estimate the posterior probability of a match sample, and that it is not a replica, given the observations y and e. This quantity can be expressed by:

$$P(M = 1, R = 0|y, e) = \underbrace{P(M = 1|y)}\,\underbrace{P(R = 0|e)} \tag{2}$$

assuming that both the posterior probability indicating the identity and the probability of the spoofing attack are two independent quantity.

We resultant posterior map of $P(M = 1, R = 0|y, e)$ using our actual database (to be described in Section 3.1) is shown in Figure 2. The posterior probability map peaks in the lower right corner of the (y,e) space, indicating that genuine samples should have high a matching score as well as low evidence of spoofing, as required.

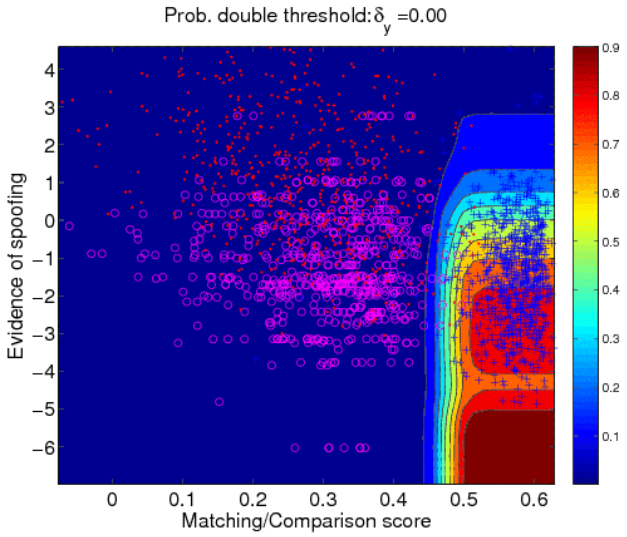

Figure 2. An example of decision boundary obtained by probabilistic double thresholding. Red dots are spoofed (non-zero effort) samples, magenta cicles are zero-effort (impostor) samples, and blue crosses are genuine samples. The vertical bar shows the values of the posterior probability map of the (y,e) space.

In comparison, a conventional biometric system without measuring the evidence of spoofing can be described by one which estimates the posterior probability $P(M = 1|y)$ (not shown here). In both cases, the theoretically optimal decision is to accept an identity claim if the posterior probability exceeds 0.5.

3. A CASE STUDY USING SIGNATURE BIOMETRICS

3.1 Databases

The signature biometrics is arguably the easiest modality subject to forgery. This is evident by the existence of many signature databases supplied with forged signatures. We will use a subset of the Biosecure multimodal biometric database[10†] that contains 105 subjects. For each subject, there are 15 genuine signatures, 5 forged signatures and 10 signatures from uninformed impostors (from other subjects). There are therefore altogether 525 forged signatures, 1575 genuine signatures and 1050 signatures from uninformed impostors, constituting the ground-truth prior probabilities.

Two types of signature classifiers are used, one serving as a signature verifier (classifier) and the other as a forgery detection classifier (distinguishing between a genuine and a forged signature). The signature verifier is classifier based on a dynamic time warping (DTW).

The signature verifier is designed to compare a pair of signature dynamics, taking five raw features, i.e., normalized pen position (to zero mean), pen pressure, azimuth and the altitude of the pen, as well as two derived ones, pen movement direction and pen velocity (both obtained from two consecutive pen locations in time). The DTW is applied to each of the seven features independently and then the resultant distances are combined using a client-specific fusion strategy.[11]

The signature forgery detection classifier developed is very preliminary. It assumes that the forgers do not have access to the temporal dynamics but only the texture signature. This classifier is an ensemble of neural networks trained with the AdaBoost algorithm.[12] It takes ten features, consisting of four features associated with pen pressures and the remaining associated with velocities.

Since in principle, the architecture of both classifiers are not important for the understanding of our topic, the interested readers are referred to.[13]

[†]The database can be downloaded from `http://biosecure.it-sudparis.eu/AB/getdatabase.php`.

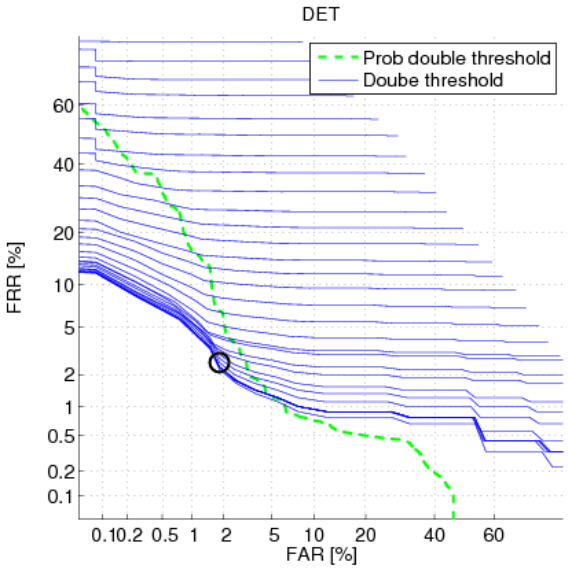

Figure 3. Comparison of performance of the two strategies reported on the signature modality of the Biosecure DS2 database. Blue continuous curves – the brute-force optimisation computed on the test set; green dashed curve – the probabilistic double threshold method; circle – the (FAR,FRR) operating point of the the brute-force method obtained on the training set.

3.2 Empirical Results

Using the data set as shown in Figure 2, we compared the two methods. In order to plot the ROC/DET curves for the double threshold method by brute-force optimization, the test data set was used directly. This is consistent with the approach used in the literature, that is, a DET/ROC curve is always plotted using the test data set. However, there is only one operating point that can validly represent the actual generalisation performance of the brute-force method. This point is shown as a black circle in Figure 3.

As can be observed, the brute-force method is fairly robust at finding the optimal operating point in terms of HTER. This criterion turns out to be very close to EER where, FAR equals FRR. The method also significantly outperforms the probabilistic double-threshold approach at low FAR. Although the probabilistic double-threshold approach appears to be inferior, it can reach low FRR region that the brute-force approach cannot. Therefore, both methods can possibly complement each other, depending on whether FAR or FRR is favoured.

3.3 Simulations

The experiment in the previous section contains two major weaknesses. Firstly, the spoof detection classifier is very weak. Secondly, the impostor skills are too weak to introduce any harm to the verification system; this might wrongly lead to the conclusion that the spoof detection classifier is unnecessary.

To overcome the above two data-dependent weaknesses, we simulate the following experiments.

- **Increase the competency of the spoof detection classifier**: This can be done by adding a small positive constant, γ_e to the evidence of spoof attack, keeping the rest of the data to be the same. Higher values of γ_e will "push" the measure e away from that of the zero-effort attack. In our simulation, the constant γ_e is varied from 0 to 8 in fine steps.
- **Increase the competency of spoof attack**: This can be done by adding a small constant value, γ_y, to the verification score of the spoof attack, keeping the rest of the data unchanged. The result is that the verification scores of the spoof attack are moved closer to that of the genuine attempt. This constant value ranges from 0 to 0.2 in fine steps.

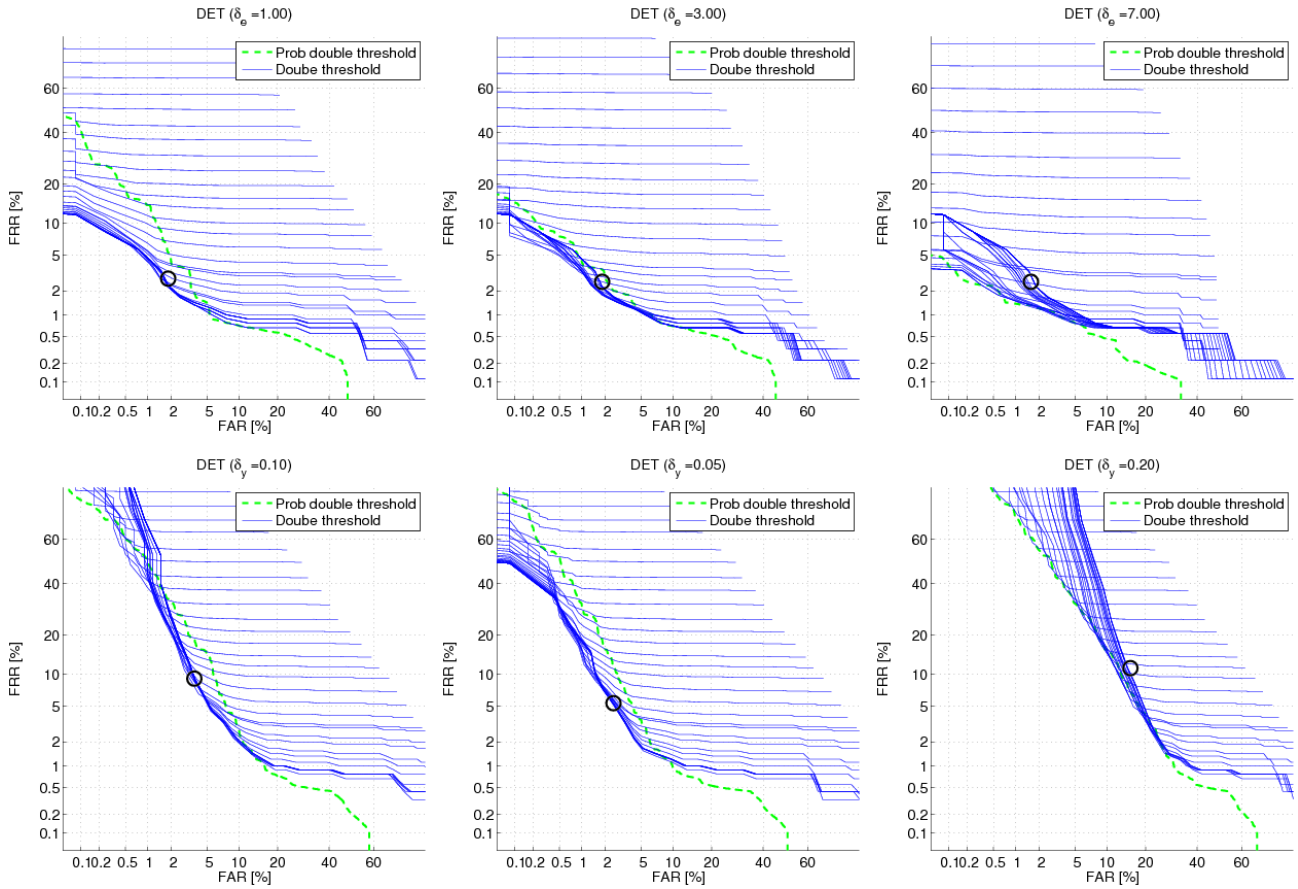

Figure 4. Comparison of performance of the two strategies reported on the signature modality of the Biosecure DS2 database

The results of the two simulation experiments are shown in Figures 4, respectively. As can be observed, under the increased competency of the spoofing classifier as well as under the increased competency the spoofing skill (with both varied independently), the probabilistic double-threshold method will outperform the brute-force method at some point.

4. CONCLUSIONS

In this paper, we study two variants of a method that that seek to combine the verification score (y) and the evidence of spoofing (e), namely a brute-force strategy and its probabilistic version. Experimental results show that both methods competes well with each other at different FAR and FRR operating points. They also complement well each other under the different competency of spoofing classifier and the forgery skill. This study is somewhoat preliminary, possible future research directions include:(1) Measuring and understanding the correlation between y and e; (2) Extension of e to multi-dimensional; (3) Extension to multiple types of attack; (4) Applying the proposed technique to other biometric modalities; and (5) Extension to a discriminative solution.

ACKNOWLEDGMENTS

This work was supported in part by the Macau Science and Technology Development Fund under Grant 017/2012/A1 and by the Research Committee at University of Macau under Grants MYRG2014-00003-FST, MRG017/ZYC/2014/FST, MYRG113(Y1-L3)-FST12-ZYC and MRG001/ZYC/2013/FST.

REFERENCES

[1] Matsumoto, T., Matsumoto, H., Yamada, K., and Hoshino, S., "Impact of artificial gummy fingers on fingerprint systems," in [*Proc. of SPIE 4677: Biometric Techniques for Human Identification*], 275–289 (2002).

[2] Perrot, P., Aversano, G., Blouet, R., Charbit, M., and Chollet, G., "Voice forgery using alisp: Indexation in a client memory," *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on* **1**, 17–20 (18-23, 2005).

[3] Abboud, B. and Chollet, G., "Appearance based lip tracking and cloning on speaking faces," *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on* , 301–305 (Sept. 2005).

[4] Fauve, B., Bredin, H., Karam, W., Verdet, F., Mayoue, A., Chollet, G., Hennebert, J., Lewis, R., Mason, J., Mokbel, C., and Petrovska, D., "Some results from the biosecure talking face evaluation campaign," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* , 4137–4140 (31 2008-April 4 2008).

[5] Drahansky, M. and Lodrova, D., "Liveness detection for biometric systems based on papillary lines," *Int'l J. Security and Its Applications* **2**(4), 29–38 (2008).

[6] Bolle, R., Connell, J., and Ratha, N., "Biometric Perils and Patches," *Pattern Recognition* **35**(12), 2727–2738 (2002).

[7] Coli, P., Marcialis, G., and Roli, F., "Vitality detection from fingerprint images: A critical survey," in [*Advances in Biometrics*], 722–731 (2009).

[8] Parthasaradhi, S., Derakhshani, R., Hornak, L., and Schuckers, S., "Time-series detection of perspiration as a liveness test in fingerprint devices," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **35**, 335 –343 (aug. 2005).

[9] Marcialis, G., Lewicke, A., Tan, B., Coli, P., Grimberg, D., Congiu, A., Tidu, A., and Roli, F., "First international fingerprint liveness detection competition (livdet) 2009," in [*Image Analysis and Processing (ICIAP)*], 12–23 (2009).

[10] Ortega-Garcia, J., Fierrez, J., Alonso-Fernandez, F., Galbally, J., Freire, M. R., Gonzalez-Rodriguez, J., Garcia-Mateo, C., Alba-Castro, J.-L., Gonzalez-Agulla, E., Otero-Muras, E., Garcia-Salicetti, S., Allano, L., Ly-Van, B., Dorizzi, B., Kittler, J., Bourlai, T., Poh, N., Deravi, F., Ng, M. W., Fairhurst, M., Hennebert, J., Humm, A., Tistarelli, M., Brodo, L., Richiardi, J., Drygajlo, A., Ganster, H., Sukno, F. M., Pavani, S.-K., Frangi, A., Akarun, L., and Savran, A., "The multiscenario multienvironment biosecure multimodal database (bmdb)," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**, 1097–1111 (2010).

[11] Muramatsu, D. and Matsumoto, T., "Online signature verification algorithm with a user-specific global-parameter fusion model," in [*Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*], 486–491 (2009).

[12] Freund, Y. and Schapire, R. E., "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences* **55**, 119–139 (1997).

[13] Muramatsu, D. and Matsumoto, T., "Effectiveness of pen pressure, azimuth, and altitude features for online signature verification," in [*LNCS 4642, Proc. Int'l Conf. Biometrics (ICB'07)*], 503–512 (2007).