

TARGETED BIOMETRIC IMPERSONATION

John D. Bustard, John N. Carter, Mark S. Nixon

School of Electronics and Computer Science, University of Southampton

ABSTRACT

When applying biometric algorithms to forensic verification, false acceptance and false rejection can mean a failure to identify a criminal, or worse, lead to the prosecution of individuals for crimes they did not commit. It is therefore critical that biometric evaluations be performed as accurately as possible to determine their legitimacy as a forensic tool. This paper argues that, for forensic verification scenarios, traditional performance measures are insufficiently accurate. This inaccuracy occurs because existing verification evaluations implicitly assume that an imposter claiming a false identity would claim a random identity rather than consciously selecting a target to impersonate. In addition to describing this new vulnerability, the paper describes a novel Targeted_FAR metric that combines the traditional False Acceptance Rate (FAR) measure with a term that indicates how performance degrades with the number of potential targets. The paper includes an evaluation of the effects of targeted impersonation on an existing academic face verification system. This evaluation reveals that even with a relatively small number of targets false acceptance rates can increase significantly, making the analysed biometric systems unreliable.

Index Terms— Forensics, Face Verification, Biometric Evaluation

1. INTRODUCTION

In January 2010 Al-Mabhouh, a co-founder of the military wing of Hamas was assassinated in Dubai. According to Dubai's authorities there are up to 29 suspects, 12 of whom carried British passports, six Irish, four French, one German, four Australian, and two Palestinian. Interpol and the Dubai police believe the suspects stole the identities of real people [1]. This example highlights the risk that sophisticated attackers can undermine existing identification systems by targeting individuals for impersonation. It is therefore important to examine the accuracy of biometric tools when subjected to such targeted attacks.

This paper is concerned with the general vulnerability of biometric verification to targeted impersonation. Verification occurs when a user claims an identity which is then validated by comparing a stored biometric signature against their presented biometric features. Whilst no verification process is in-

fallible, significant progress has been made in improving verification accuracy and there are now many commercial biometric systems in regular use. However, recent research [2] has shown how these systems may be vulnerable to deliberate attempts to subvert them.

Such attacks are conceptually simple: they involve finding an existing person with a similar biometric signature and then fraudulently assuming that identity to spoof a verification check. Traditionally, the security of biometric verification has been measured using false acceptance rates. This provides an estimate of the likelihood that an imposter would successfully be accepted by a biometric system if they randomly claimed a false identity. However, it does not accurately measure the vulnerability of such systems to more deliberate attacks, which is the focus of this paper.

Increases in the use of social networking, online dating and centralized biometric databases have made identity systems more vulnerable to targeted attacks. These large searchable collections of face and other biometric data increase the chance of finding a target that has a closely matching biometric signature. Such attacks are particularly dangerous as they can be effective both against automated biometrics and manual methods of identification, such as visual passport inspection.

The paper starts by surveying the existing literature on the measurement of biometric vulnerabilities. It then examines the effect of targeted spoofing on a face verification system. The investigation uses a publicly available biometric algorithm and dataset. The paper then examines how the effectiveness of attacks increases with the number of potential targets. It concludes by proposing an additional metric for verification performance.

2. BIOMETRIC VULNERABILITIES

Technology evaluations of biometric systems primarily measure verification performance using the false rejection and false acceptance rates of the system under test with different trade-off priorities [3].

Many contextual factors, such as facial pose and lighting, can have a significant effect on verification performance and, as the various biometrics have matured, these factors have been investigated [4]. More recently, deliberate attempts to attack biometric systems have been studied. Uludag et al. [5]

have identified eight different types of attack based on the part of the biometric system being subverted. Attacks from Type 1 are aimed at the sensor and are the focus of this paper. The remaining types are attacks on the electronic systems and enrollment procedures used to set up and perform verification.

In terms of sensor level attacks, three existing methods have been identified [6]:

- *Zero effort attacks*, in which a person claims a random identity and attempts to be incorrectly accepted by the system. Zero effort attacks are the attack type being measured in existing large scale performance evaluations that calculate false accept rates.
- *Brute force attacks*, which repeatedly attempt to access a system, adjusting a biometric feature until a sufficiently close match is obtained [7]. Such attacks generally require unrestricted access to the biometric system (e.g. picking a biometric lock on a stolen laptop). Secure access control scenarios, such as passport control at an airport, make such attacks less feasible as access failures can raise alarms.
- *Artifact attacks*, which use a synthetic biometric feature that has been produced from a genuine user. Such attacks would also cover the attempted use of a surgically removed biometric features and methods which exploit residual features on a sensor [8].

An additional consideration is that not all the users of a system will necessarily have the same level of security. This was highlighted by Doddington et al. [9], who measured the relative recognizability of different users of a speaker recognition system. Here users were classified into four different types: sheep who have normal performance, goats who are difficult to recognize, lambs who are easy to impersonate and wolves who can easily impersonate others. Attackers can exploit this variation to compromise a biometric system. For example, a lamb insertion attack [6] would involve deliberately enrolling a person or synthetic feature that is known to have a similar signature to many subjects. The system containing the lamb subject would then be vulnerable to imposters claiming the lamb identity.

By deliberately selecting a legitimate user with similar biometric features, a targeted attack can turn any imposter into a wolf subject. Targeted attacks are a significant vulnerability as they have no artificial traits that can be recognized, either by an automated system or a human supervisor. They are also possible without control over the enrollment procedure or the need for a confederate whose true identity would be made known, as is the case for twin impersonation or lamb injection attacks. Such attacks are also quite likely, as they are a plausible strategy for even relatively unsophisticated attackers.

3. IMPACT EVALUATION

This section evaluates the effects of targeted attacks on the CSU Baseline Algorithm developed by Bolme et al. [10] for the Good, the Bad and the Ugly face recognition challenge [11]. The system has been trained using images from the NIST Multiple Biometric Grand Challenge dataset [12]. The verification system has partial robustness to lighting variation, expression changes and occlusions. However, its performance is much lower than has been demonstrated with state-of-the-art commercial face verification algorithms [3]. The system was evaluated using the Color FERET face database, which has been available since 1996. The frontal face subset, consisting of files labeled Fa and Fb, has been selected as it is more representative of relatively controlled face verification recordings and is consistent with the original FERET verification testing protocol [13]. The dataset is made up of 1009 subjects of varying age, sex and race. The evaluation assumes the attacker has complete access to the gallery of subjects and the verification algorithms used by the system. In each case, half of the recordings of each subject are randomly selected and used as the gallery to which the attacker has access.

Each subject in the gallery takes the role of an attacker. In each case the gallery data is analyzed to select a target that the attacker will impersonate. In all of the targeted attacks, a target was chosen based on the best match score value of all of the possible combinations of attacker and target recordings within the gallery. The non-gallery recordings of the target are then compared against the attacker to determine imposter scores. Score values are also calculated for all the true matching pairs of users of the system. These score values are used to produce DET curves showing the trade-off of false accept and false reject rates for different verification thresholds. A traditional zero-effort DET curve is also produced to show the relative effect of targeted attacks. The curve is calculated by comparing each of the excluded recordings against each of the gallery recordings to produce a range of scores for both legitimate and zero-effort attacks. It is expected that real deployments may have more challenging input data and in turn may have more sophisticated verification systems; however, the experiments indicate that the relative effect of targeting is sufficient to warrant further investigation.

Figure 1 shows the baseline zero effort attack DET curve and the False Acceptance Rates when targeting is applied at the baseline EER threshold value. The EER of the baseline is 17%. However, when a targeted attack is performed on the same system the false acceptance rate rises to 51%, three times the original value and a significant security risk. If the threshold of the system is selected with the knowledge of targeted attacks, the EER becomes 28%, which reduces the risk but increases the false reject rate to an impractical level.

3.1. Number of Targets

In the baseline experiments the number of targets available to the attacker is necessarily restricted by the size of the datasets. The size of these datasets is consistent with the number of subjects that might access a secure office environment but is much lower than many important identity scenarios such as passport control. To analyze the effect of increasing target numbers, further experiments were performed using the Face verification system. 800 gallery subsets of increasing size were created. These subsets were used in the selection of targets for evaluation. To minimize any potential bias caused by subset selection, for a given size, all non-overlapping subsets within the first 800 subjects were combined to produce average false accept rates across the different subsets. This ensures that a subset size of 1 is virtually identical to the baseline performance. All gallery members took the role of attackers using the subset to generate the imposter scores.

Figure 2 indicates how the false accept rate increases as the size of the target subset increases. The graph shows the false acceptance rate for a threshold that achieves the equal error rate of the baseline system under zero effort attacks. This is a plausible threshold for systems that are unaware of the risks of targeted attacks. As the number of available targets increases, the number of possible subsets decreases, increasing the error in the measured false accept rate. Much of the curve, however, conforms reasonably well to a least squares fit of an $a \cdot \log(x) + b$ model, with $a = 5.2$ and $b = 16.7$.

One difficulty with using a logarithmic fit to predict FAR is that such a curve will produce values below 0% and above 100%. Although FAR values are limited in this way, the difference between individual biometric signatures may not be. There are many different score distributions that could produce 0% or 100% FAR values based on the relative difference between legitimate and imposter score values. As such the logarithmic fit can be seen as expressing the functional shift in the difference between legitimate and imposter score distributions rather than the FAR value itself. As the FAR measurements approach the bounds, excessively distant or close score values will have a diminished effect on the measured FAR. One way to understand this effect is to treat the logarithmic prediction as the centre of a probability distribution over FAR values that can pass outside of the bounds. This probability distribution reflects the likelihood of obtaining any particular FAR when the biometric system is evaluated. When determining the likelihood of 100% or 0% FAR values the entire probability distribution outside of the bounds are combined. In practice this means that when the targeted FAR value reaches 100%, the model predicts that there is a 50% chance of obtaining 100% FAR for any given evaluation of the system. Further research is required to determine the shape of this distribution and to validate these predictions on systems which reach these bounds.

The fitted model can be used to provide estimates of the

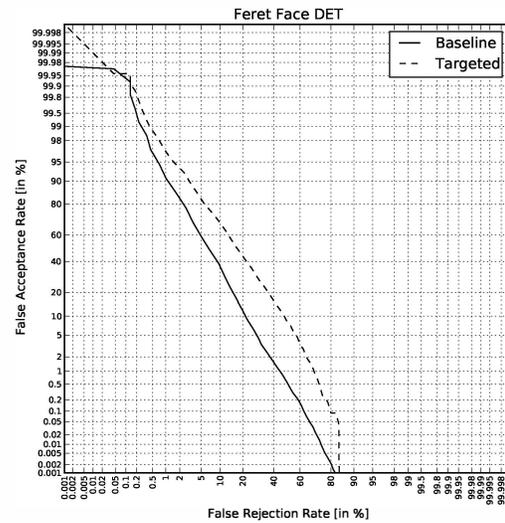


Fig. 1. The effects of a targeted attack on the CSU face verification algorithm. *Baseline* shows the performance of the system under a zero effort attack. *Targeted* shows the increase under targeted biometric attacks.

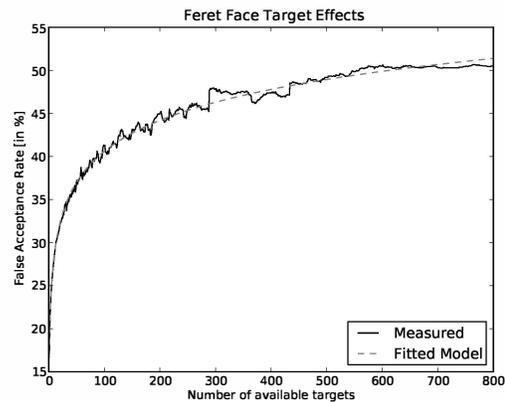


Fig. 2. The effect of target numbers on FAR with a verification threshold set at the EER of the baseline system.

number of targets needed to achieve different false accept rates. For example, using this model, approximately 200,000 targets are required for an 80% FAR, 1,370,000 for 90% and 9,500,000 for 100%. Larger evaluations are needed to confirm these predictions. However, they suggest that for national identity applications with many millions of subjects, such as passport control, there is a greater than 50% chance that this verification system could be subverted by any user.

An additional consideration is how feasible it is for attackers to obtain information about the gallery subjects and the system being attacked. For small scale deployments, surveillance may be sufficient to establish possible targets. How-

ever, some biometrics may be more vulnerable. For example, face, voice and gait are relatively easy to record at a distance while fingerprint, iris and finger vein may require more elaborate social engineering to obtain. For identity applications with a large number of users, such as passports, public information may be sufficient. For example, a number of online dating websites have photographs of millions of users which can be anonymously searched using soft biometric constraints including, age, sex, race, hair color and height [14]. Centralized databases of biometric information are of greater concern. For example, if the US Visit database was hacked, its recordings could be used to identify possible targets for face or fingerprint attacks.

4. CONCLUSION

This paper analyses the effect of targeted attacks, which can reduce the effectiveness of biometric identity verification. It illustrated the problem through the evaluation of a face baseline verification algorithm, revealing that with 800 potential targets, attacks can increase false acceptance rates by a factor of three, reducing security to the point that it is no longer reliable for forensic identification. Further analysis suggests that the false acceptance rate can be estimated using a simple model that is proportional to the logarithm of the number of enrolled subjects. This model provides a means to estimate the vulnerability of systems with many users and shows that for the face verification algorithm analyzed here, national identity schemes could be unreliable under these attacks.

5. ACKNOWLEDGEMENTS

This work was partially funded by the EU FP7 project TABULA RASA (257289). Figure 1 was produced using ScoreToolKit [15].

6. REFERENCES

- [1] "Bbc news," http://news.bbc.co.uk/1/hi/world/middle_east/8522595.stm.
- [2] T. van der Putte and J. Keuning, "Biometrical fingerprint recognition: dont get your fingers burned," in *Conf. Smart card research and advanced applications*, 2001, pp. 289–303.
- [3] A. OToole P. Flynn K. Bowyer C. Schott P. Phillips, W. Scruggs and M. Sharpe, "Frvt 2006 and ice 2006 large-scale experimental results," *IEEE Tran. PAMI*, vol. 32, no. 5, pp. 831–846, 2010.
- [4] P. J. Phillips W. Zhao, R. Chellappa and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computer Surveys*, vol. 35, pp. 399–458, 2003.
- [5] U. Uludag and A. K. Jain, "Attacks on biometric systems: A case study in fingerprints," *Security, Steganography, and Watermarking of Multimedia Contents*, pp. 622–633, 2004.
- [6] T. Dunstone and G. Poulton, "Vulnerability assessment," *Biometric Technology Today*, vol. 5, pp. 5–7, 2011.
- [7] M. Martinez-Diaz, J. Fierrez-Aguilar, F. Alonso-Fernandez, J. Ortega-Garcia, and J. Siguenza, "Hill-climbing and brute force attacks on biometric systems: A case study in match-on-card fingerprint verification," in *Int. Carnahan Conferences Security Technology*, 2006, pp. 151–159.
- [8] H. Matsumoto K. Y. T. Matsumoto and R. L. v. R. S. Hoshino, "Impact of artificial gummy fingers on fingerprint systems," in *SPIE Optical Security and Counterfeit Deterrence Techniques IV*, 2002, vol. 4677, pp. 275–289.
- [9] A. Martin M. Przybocki G. Doddington, W. Liggett and D. Reynolds, "Sheep, goats, lambs and wolves a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," 1998.
- [10] Y. M. Lui B. Draper H. Zhang N. Teli S. OHara D. Bolme, R. Beveridge and J. R. Matey, "Csu baseline algorithms," 2011.
- [11] P. Phillips, J. Beveridge, B. Draper, G. Givens, A. OToole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An introduction to the good, the bad & the ugly face recognition challenge problem," in *Int. Conf. Face Gesture*, 2011, pp. 346–353.
- [12] J. R. Beveridge W. T. Scruggs A. J. OToole D. Bolme K. W. Bowyer B. A. Draper G. H. Givens Y. M. Lui H. Sahibzada J. A. Scallan Iii P. J. Phillips, P. J. Flynn and S. Weimer, "Overview of the multiple biometrics grand challenge," in *Int. Conf. Advances in Biometrics*, 2009, pp. 705–714.
- [13] H. Moon and P. Phillips, "The feret verification testing protocol for face recognition algorithms," in *Int. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 48–53.
- [14] "Plenty of fish," <http://www.plentyoffish.com>.
- [15] André Anjos and Sébastien Marcel, "Scoretoolkit documentation," *Idiap-Com Idiap-Com-02-2012*, Idiap, 4 2012.