Presentation Attack Detection: Towards a structured approach for tailored metrics

Nils Tekampe
TÜV Informationstechnik GmbH
n.tekampe@tuvit.de

Abstract

Mechanisms for the detection of Presentation Attacks against biometric systems have come into the attention of the scientific community and developers over the last few years. In order to describe test results and to evaluate the performance of those Presentation Attack Detection (PAD) mechanisms various metrics have been developed and discussed. This paper summarizes the results of the work that has been performed in the context of the B.E.A.T. project. It acknowledges the fact that there may be the need for more than one metric for PAD mechanisms depending on the concrete test design. It further introduces aspects of a framework for metrics and introduces existing and new metrics in this field. This paper does not claim to be a complete overview of a framework or a complete reference for metrics of PAD mechanisms. It rather aims to trigger further discussions in this field. This paper arises from work that has been performed in the context of the B.E.A.T. project that has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no 284989.

1. Introduction

A useful metric for the description and rating of the capability of biometric system to detect or otherwise withstand presentation attacks is a widely discussed topic today. This paper discusses the need for dedicated metrics in this area and introduces principal aspects of such a metric from different perspectives, such as performance and security.

In classical performance testing of biometric systems it is a well known fact that the performance of a system also depends on the individual who should be recognized. This phenomenon that is also known as Doddingtons Zoo ([3]) however, can be described as one influencing factor. When testing PAD mechanisms the results are often way more depending on the used artefact. It often happens that a systems recognizes various species of artefacts reproducibly while completely failing to recognize other species. With other words: test results of PAD mechanisms are usually way less homogenous compared to tests of the recognition performance of biometric systems. The scope of a test of a PAD mechanisms is way more decisive than in testing biometric recognition performance. It is essential whether a test is carried out with a certain artefact (that may be applied multiple times), a certain species (and multiple artefacts) or with multiple species. This perspective also needs to be reflected by the metric used to describe the test results

On the first sight, it has been obvious to define error rates for PAD mechanisms in an analogue form to the error rates in the area of biometric performance. E.g. [2] defined the **False Non- Presentation Attack Detection Rate (FNPADR)** as the proportion of presentation attacks incorrectly classified as being non-attacks. However, as outlined before, such a definition of a simple metric disregards the exact scope of the test.

The FNPADR from [2] allows an exact description of a simple test. A tester applies a number of artefacts to the system under test and the system failed to correctly recognize a subset of them. However, when it comes to the question, whether the system under test did a good or a bad job, such a simple metric may be misleading. There are various examples for such misleading situations:

1. If a system recognized 999 out of 1000 artefacts, it seemed to have done a good job. However, if the remaining artefact overcomes the system reproducibly and this fact is well known, the system as a whole may be useless as every evil user would simply use this one ”golden” artefact.

2. If a system is accompanied by a repeated tests through development, developers would say that an artefact that the system failed to recognize in a previous test and that the system now recognizes is a greater achievement than to recognize the simple artefact that has always been detected

3. Some sensor technologies do not recognize certain ma-

terials at all. An artefact made from such a material would not have success in overcoming a system under test. However, it remains the question whether this is an achievement of the system and whether such a recognition ”by accident” should count the same as another artefact that has the potential to overcome the system but that has been correctly recognized.

2. Various perspectives to metrics

Metrics for the description of results of tests of PAD mechanisms can be looked at from various different perspectives. One is the perspective of the test corpus that the metric relates to. Three different levels are possible here:

1. A metric can describe the results of a test with a single artefact
2. A metric can describe the results of a test with multiple artefacts that belong to the same species
3. A metric can describe the results of a test with multiple artefacts from multiple species

On the other hand, the metrics themselves can be looked at based on the principles of set theory that they follow. Specifically, metrics can be

- based on an average value,
- based on a weighted average value,
- based on absolute characteristics, such as a minimum or maximum values.

Last but not lease metrics can also be looked at from the perspective of their motivation/scope. E.g. metrics may be used

- to accompany the development process of a PAD mechanism,
- to rank various PAD mechanisms,
- to decide whether a PAD mechanism is sufficiently secure for a certain application case.

3. Overview of metrics

This paper proposes the use of dedicated metrics to rate/benchmark PAD mechanisms. It acknowledges that there may not be one kind of metric suitable for all tests but there may be the need to collect from a variety of metrics based on the concrete test scenario. The main question that may have an influence to the metric to be used is the scope of the test. What is the question that the test is after to answer? Is it a security driven test that wants to answer the question, whether the system under test is secure enough to withstand certain attack scenarios of is it a test that a developer uses to benchmark a PAD mechanism during development?

The rest of this paper introduces a variety of possible metrics. This list of metrics does not claim to be complete; rather it aims to trigger further discussions on the use of those metrics. The introduction of the metrics is structured after the used test corpus (from metrics for a single artefact to metrics for tests with multiple artefacts from multiple species)

3.1. Definitions

The further text in this paper assumes the following symbols to be given

Symbol	Definition
α_n	List of test artefacts
α_{β_n}	Test artefact of species β_n
A	Set of all test artefacts
$\Psi(\alpha_n)$	Weight factor for a certain artefact
β_n	Species of test artefacts
B	Set of all species of test artefacts
O	Number of test attempts executed with one artefact
N	Overall number of test attempts
$\rho(\alpha) \in \{0,1\}$	Result of the system under test. 1 if no attack has been detected, 0 otherwise
γ	Portion of undetected attacks of a certain species leading after which the attack counts as repeatable
$T(\alpha)$	Attack/Threat potential needed to produce and apply artefact α

Table 1. symbols

3.2. Metrics for dedicated artefacts

The following paragraphs introduce various metrics that can be used to describe the test results of a test in which only one artefact is used.

3.2.1 The simple average

The proportion of presentation attacks incorrectly classified as being non-attacks is defined in [2] as the **”False Non-Presentation Attack Detection Rate”** of a certain Artefact (FNPADR(α)) Using the symbols as defined before it can be defined as

$$\text{FNPADR}(\alpha) = \frac{\sum_{n=1}^{O} \rho_n(\alpha)}{O} \tag{1}$$

This average metric can be used to describe the results of a test with one artefact very precisely. However, due to the aforementioned reasons it becomes less precise or even misleading when applied to a test with multiple artefacts from multiple species.

3.2.2 The binary metric

This binary metric answers the question whether an artefact α has been able to repeatably overcome the PAD mechanism under test. Accordingly, it is called **”Binary False Non- Presentation Attack Detection Metric”** and defined as

$$\text{BFNPADM}(\alpha) = \begin{cases} 1, \text{ if } & \frac{\sum_{n=1}^{O} \rho(\alpha)}{O} \geq \gamma \\ 0, & \text{else} \end{cases} \tag{2}$$

3.3. Metrics for a certain species of artefacts

The following paragraphs introduce various metrics that can be used to describe the test results of a test in which multiple artefacts from one species are used.

3.4. The average per species

The proportion presentation attacks incorrectly classified as being non-attacks can be defined as the **”False Non- Presentation Attack Detection Rate of a species (FNPADR(β))”**. It is defined as follows

$$\text{FNPADR}(\beta) = \frac{\sum_{n=1}^{|A_\beta|} \sum_{m=0}^{O} \rho(\alpha_{nm})}{|\beta| * O} \tag{3}$$

The use of an average per species as metric makes sense as long as the results per artefact of the test corpus don’t differ too much. If the variance of the results of the various artefacts it too high, it may make more sense to report the results of each artefact individually.

3.5. Binary per species

Another possibility to express the result of a test of a PAD mechanism goes back to the experience that many PAD mechanism easily recognize certain species of artefacts while having enormous problems in recognizing other. The **”Binary False Non- Presentation Attack Detection Metric”** of a certain species β can be defined as

$$\text{BFNPADM}(\beta) = \begin{cases} 1, \text{ if } & \frac{\sum_{n=1}^{|\beta|} \sum_{m=0}^{O} \rho_{no}}{|\beta| * O} \geq \gamma \\ 0, & \text{else} \end{cases} \tag{4}$$

The use of this metric makes sense if it is of interest whether the system under test can recognize certain species of artefacts.

3.6. System metrics

The following paragraphs introduce various metrics that can be used to describe the test results of a test in which artefacts from multiple species are used. Those metrics aim to characterize the complete system under test.

3.7. The average for a system under test

3.7.1 Simple average

The **”False Non- Presentation Attack Detection Rate”** of a complete system under test (FNPADR) can be defined as follows

$$\text{FNPADR} = \frac{\sum_{i=0}^{|B|} \sum_{n=1}^{|\beta|} \sum_{m=0}^{O} \rho(\alpha_n)}{|\beta| * O * |B|} \tag{5}$$

The use of this simple metric is very limited. It can be meaningful if it can be assumed that the results of the various artefacts of the various species are homogenous (with other words: that the variance of the test results is low). Further, it can provide interested readers of test results with a very rough overview of the test results. However, as the results of the various artefacts are usually very heterogeneous, this metric shall only be used with care.

3.7.2 Weighted average

Many metrics that have been discussed in the past suggest that each test artefact is of the same importance for the performance of the system under test as all others. While this assumption makes the developed metrics more simple, the underlying assumption is generally not true. Some examples for the imbalance include

1. An artefact that is very hard to produce may count more when being correctly recognized by the system under test than a standard artefact that is very easy to produce

2. An artefact that the system under test failed to correctly recognize in earlier tests may count more than an artefact that the system under test reproducibly recognized in the past

This leads to a situation in which a tester may decide to put more attention to the performance of the system under test regarding certain artefacts. In order to reflect such a test design in the results of the test, there is a need for a weighted metric.

In [1] a weighted metric has been proposed from the perspective of security as follows:

$$\text{APCER} = \frac{\sum_{n=1}^{|\beta|} \sum_{m=0}^{O} \rho(\alpha_n) * T_{\alpha_n} * PAISR_n}{|\beta| * O} \tag{6}$$

Where $PAISR_n$ stands for the presentation attack instrument success rate which is close to zero of all sensors can detect this artefact.

While we outline that a security related metric should work after a maximum principle (please refer to chapter 3.8), we acknowledge that there may be other circumstances under which a weighted metric can be useful. Such cases specifically include (but are not limited to):

- A test that is used by a developer to track the progress in development of a PAD mechanism
- A test of a stakeholder that is used to trigger the development of PAD mechanisms into a certain direction

For further discussion we suggest to use a more generic definition of the weighted metric as follows:

$$\text{WFNPADR} = \frac{\sum_{n=1}^{|\beta|} \sum_{m=0}^{O} \rho(\alpha_n) * \Psi_{\alpha_n}}{|\beta| * O} \tag{7}$$

This **"Weighted False Non- Presentation Attack Detection Rate of a system (WFNPADR)"** allows the designer of a test to assign a factor (between 0 an 1) to each artefact in order to express whether the artefact is important for the test or not. There are many different approaches, how such a factor can be developed.

3.8. Security focused metric: The maximum principle

PAD mechanisms are security mechanisms that are developed and used to protect a biometric system against a specific kind of attack. An obvious question that a metric shall be able to answer is therefore: **"Is the PAD mechanism under test secure enough for a certain application case?"**. A metric based on an average value is in general not suitable to answer this question. The reason can be easily illustrated using the following example: A PAD mechanism detects 999 out of 1000 artefacts. However, one artefact (or species of artefacts) does reproducibly overcome the mechanisms. When we further assume that the artefact is easy to produce and the fact that the PAD mechanism can be overcome this way is publicly known, nobody would claim that the PAD mechanism is secure. However, it would get quite good results in all metrics that base on the average value or even weighted average value. Such metrics could even be misused by the developer of a PAD mechanism. An artefact that can reproducibly overcome the PAD mechanism can be burrowed under a large amount of artefacts that are detected by the system.

Such a situation is not specific to PAD artefacts. It's common to all security systems. In general, a security system is only as strong as the weakest part. Or with other words: An attacker will always use the way of least resistance. Therefore, the following subsections introduce metrics that base on the maximum principle.

3.8.1 The real maximum

The absolute and direct reflection of the worst case scenario as outlined before, would be to rate the system under test after the worst error rate of all artefacts in the test corpus. This leads to the definition of the False Non- Presentation Attack System Detection Rate (FNPASDR)

$$\text{FNPASDR} = \max[\text{FNPADR}(\alpha_1)...\text{FNPADR}(\alpha_n)] \tag{8}$$

3.8.2 The maximum per species

It is well known that the ability of a PAD mechanism to detect an artefact depends on those characteristics that allow to group artefacts into different species. With other words: If one artefact of a certain species is recognized by a PAD mechanism, it becomes more likely that other artefacts of the same species are also recognized. In the end the real relevant question for a security evaluation of a PAD mechanism is the question, how likely it is that an attacker can overcome the PAD mechanism by the use of an artefact of a certain species. The answer to this question leads to the definition of the False Non- Presentation Attack System Detection Rate (FNPASDR) that bases on the results of the various species.

$$\text{FNPASDR} = \max[\text{FNSPADR}(\beta_1)...\text{FNSPADR}(\beta_n)] \tag{9}$$

4. Summary

This paper introduced as set of useful metrics to describe test results of a PAD mechanisms. It further presented a structured approach for multiple metrics in this area. This paper does not claim to provide a complete or comprehensive overview of PAD metrics. Instead, it summarizes the work that has been performed in the course of the B.E.A.T project and invites for further discussions. Various complicating aspects of metrics have been ignored for the metrics introduced in this paper. Artefacts that are destroyed during their application (so that they cannot be applied again) or the question whether a biometric system that did not recognize anything of an artefact did sufficiently counter the attack (although done accidentally) are only two to mention. Those aspects will require further discussion along with the metrics. Table 2 summarizes the metrics and their applications that have been introduced in this paper.

References

[1] C. Busch. Presentation attack detection for smartphone finger image recognition. *IBPC 2014 conference, Gaithersburg*, 2014.

[2] ISO/IEC. Text of 5th working draft 30107, presentation attack detection. 2013.

[3] G. D. W. L. A. M. M. Przybocki and D. Reynolds. Sheeps, goats, lambs, and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *Proceeding ICSLP '98*, 1998.

Area	Metric	Summary
Artefact	Simple Average	Precise description of test results for a test with one artefact that is applied multiple times
	Binary metric	Answers the question whether a certain artefact has been able to reproducibly overcome the system under test
Species	Simple Average	Useful if the results of the various artefacts in the test corpus don't differ too much.
	Binary	Answers the question whether a certain species of artefacts is reproducibly recognized by the system under test
System	Simple Average	Useful in case of homogenous results
	Weighted metric	Very flexible; allows to lay focus on certain artefacts
	Real maximum	Direct and strict application of the worst case principle. Useful for security tests
	Maximum per species	Application of the worst case scenario per species of artefacts. Useful for security tests

Table 2. Summary of metrics