# Influence of the datasets size on the stability of the LR in the lower region of the within source distribution

*Rudolf Haraksim, Didier Meuwly*

Netherlands Forensic Institute
Laan van Ypenburg 6, 2497GB, The Hague, The Netherlands
r.haraksim(d.meuwly)@nfi.minvenj.nl

## Abstract

This article focuses on the statistical evaluation of the fingermark evidence using the likelihood ratio (LR) approach. It studies the influence of the quantity of data used to model the within (WS) and between (BS) source variability. The LR system built for the experiment uses an Automated Fingerprint Identification System (AFIS) feature extraction and comparison algorithm, fingermark and fingerprint datasets coupled with a generative approach for modeling the WS and BS variability. This article concentrates on the computation of LRs of the same source in the lower region of the WS distribution. It analyzes the behavior of the LR with an increasing number of entries in the WS datasets while maintaining the constant proportion of the BS dataset in an attempt to estimate the amount of same source scores necessary to achieve consistent LR performance.

## 1. Introduction

While the question of the comparison of complete fingerprints seems to be an issue long solved in the biometric world with many commercial algorithms and applications available, quite some issues arise when analyzing forensic fingermarks (traces). When a fingerprint and a fingermark are subjected to forensic evaluation, the fingermark is almost always partial, its quality severely degraded due to uncontrolled imposition (clarity, distortion) and due to the effects of the development methods.

While the AFIS matching and comparison algorithm is able to achieve great results in terms of performance and speed while producing shortlists of candidates, it is not used in the current practice for the statistical evaluation of fingermarks and fingerprint evidence. Forensic evidence ($E$) in this case is considered the similarity score resulting from the fingermark and fingerprint comparison. In order to quantify the weight of the forensic evidence we start off with a set of mutually exclusive propositions, the one of the prosecution $\boldsymbol{H_p}$ and the one of the defense $\boldsymbol{H_d}$:

- $\boldsymbol{H_p}$ – the fingermark originates from the individual that is also the source of the fingerprint
- $\boldsymbol{H_d}$ – the fingermark originates from an unknown individual, randomly selected

With the propositions defined we can now proceed to the LR calculation which can be derived from the odds form of the Bayes theorem in the following way:

$$LR = \frac{\Pr(E \mid Hp)}{\Pr(E \mid Hd)} \qquad (1)$$

where Pr indicates the probability of observing the evidence $E$ given one of the two hypotheses.

The calculation of the LR implies the modeling of the WS and BS scores distributions using a discriminative, generative or hybrid approach [1]). The main objective of this article is to study the influence of the size of the datasets on the stability of the LR. The influence will be studied using a generative approach[1] for the modeling of the within and between source variability.

An ideal situation would be to dispose of a quantity of score observations large enough to cover the whole range of the BS and WS distributions. However in the tails of these distributions a good estimate of the LR is difficult to obtain, due to the rarity of the scores. In the regions where the number of scores is sufficient to describe reliably the WS and BS the LR value is generally low, and the stability of the LR can be considered as an indicator for the robustness[2] and of the reliability[3] of the method.

In this work we shall analyze the region of the lower tail of the WS score distribution - see figure 1 (similar issues addressed in [6]). We are interested in this region mainly due to the fact that similarity scores in this particular area can “shift” the scales in favor of either of the propositions. Ideally we would like to observe a stabile LR support to either of the propositions, however with the varying number of the WS scores we observe variation in the LRs as well.

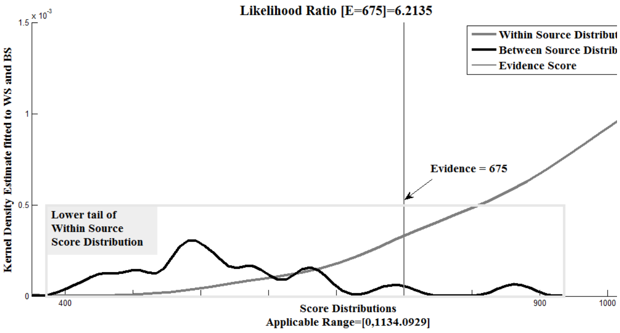

***Figure 1.*** – Area of interest (lower tail of the WS score distribution)

[1] In the generative approach we “generate” the score distributions from the the discrete datasets (similarity scores).

[2] Robustness is defined as the ability of a method to maintain the tendency of its performance when reducing the quality conditions of the data under examination

[3] Reliability defined as the capability of the method of not degrading the trueness of the LR when used in all the possible cases for which it has been designed

In this initial study we will model the similarity scores produced by the AFIS algorithm using the Kernel Density Function (KDF). This choice is based on the fact that we are dealing with discrete datasets and because the comparison algorithm produces multimodal score distributions. Since we are interested in observing the influence of the different sizes of datasets on the LR stability, the over-fitting, which in most of the cases is considered a drawback of the KDF seems to be a desirable side-effect for this particular application.

Before any method developed can be used in a forensic casework, a validation step needs to provide insight about its robustness and reliability (LR > 1 if $\boldsymbol{H_p}$ true, LR < 1 if $\boldsymbol{H_d}$ is true). The aim of this article is to study the stability of the LR produced and in particular the variations due to data on the probability estimates for both the numerator and denominator of the LR. We will show the influence of lowering the quantity of data used for modeling the WS and BS scores on the stability of the LRs. Despite the fact that relatively small number of individuals is used in this study, it provides a valuable insight on the LR stability depending on the decreasing number of WS scores.

## 2. Datasets used

For modeling the BS scores, large quantities of reference fingerprints are available, for example ten-print cards originating from a police fingerprint databases. It is not necessarily the case for WS scores, where a limited number of fingermarks and corresponding fingerprints with the ground truth known is available. Different approaches have been proposed in the literature to handle the data sparsity under $\boldsymbol{H_P}$ [3, 4].

Both methods rely on the use of simulated fingermarks from the suspected individual. In [4] these simulated fingermarks are compared with a set of corresponding fingerprints (multiple fingerprints per finger), when in [3] large quantities of simulated fingermarks are compared with a single fingerprint in order to obtain the WS score distribution.

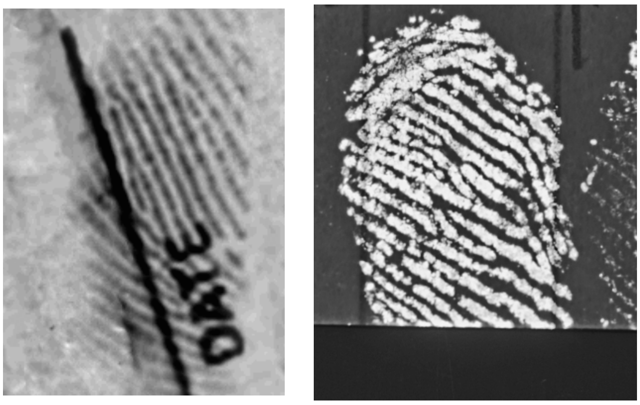

***Figure 2.*** – Simulated fingermark on the left vs. visualized real fingermark from a crime-scene on the right

This method mimics the distortion and provides enough reference material for modeling the WS score distribution. The fingermarks produced by this method are not completely equivalent to real crime-scene fingermarks but for the purpose of this article and based on the results published in [3], their similarity is considered as sufficient (see figure 2). The number of minutiae and the effect of distortion, present in the set of fingermaks used, represent the key elements of variability for the calculation of the evidential value.

Simulated fingermarks with 8 minutiae configurations were chosen for this article, as a majority of the fingermarks recovered as pieces of evidence contains less than 12 minutiae, which is the numerical standard in most countries using a numerical standard. In these countries fingermarks with less than 12 minutiae are currently not considered as evidence that can be presented at court and would primarily benefit from the approach described in this paper.

### 2.1 LR model and size of the dataset used

Figure 3 illustrates the LR model used in this article. The nomenclature used to describe the different datasets refers to the one used in [2].

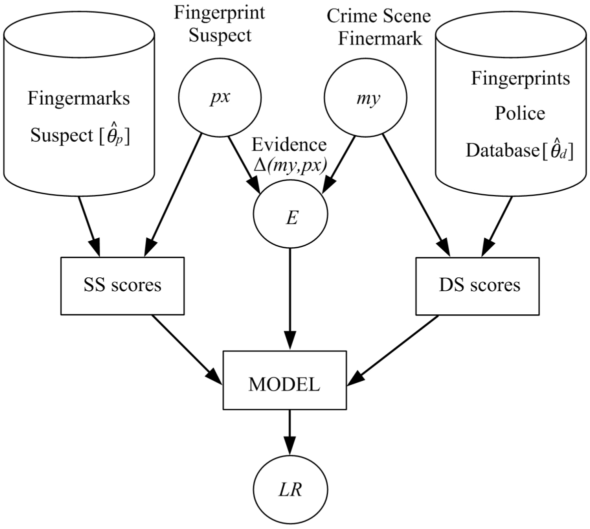

***Figure 3.*** – The LR model

The fingerprint police database consists of electronic copy of ten-print cards. For the purpose of this article we have selected a population of 20.000 individuals (200.000 fingerprints) to represent the BS population.

Since we aim to establish the stability of the LRs in the lower region of the WS score distribution, we will use data from four individuals, for which we have large quantity of simulated fingermarks available – ranging from 2.179 to 8.455. In practice, collecting a WS dataset counting 1000s' of fingermarks for a suspected individual is a time consuming procedure which largely depends on the willingness of the suspect to cooperate (in many cases impossible).

In the following section a forensic evaluation will be described together with the calculation of a likelihood ratio.

## 3. Evidence Evaluation

As indicated in figure 3, we proceed with evidence evaluation in multiple stages:

- Establish the value of the evidence (E) – a similarity score between a fingermark or fingerprint
- Model the WS distribution based on the comparison of the marks and prints of the same individual (same finger)
- Model the BS distribution based on the comparison of the marks and prints of the different individual (different fingers)
- Calculate the Likelihood Ratio

According to [5] the LR is calculated in the following way:

$$LR = \frac{\Pr(E \mid H_p, \Delta_{SS}(m,p))}{\Pr(E \mid H_d, \Delta_{DS}(m,p))} \quad (2)$$

where:

$\Delta_{SS}(m,p)$ is the similarity score of the marks and print of the same source

$\Delta_{DS}(m,p)$ is the similarity score of the marks and prints of the different source

In order to obtain calculate the evidence same source in the same dataset, one of the simulated fingermarks (on a leave-one-out basis) will play the role of the crime scene mark and will be compared to the reference print of the same individual. If the total number of the simulated marks per individual is n, a total of n-1 fingermarks will be available to form the WS score distribution.

As indicated earlier, for WS and BS score distribution modeling we will use the KDF function.

For measuring the stability of the LRs we will vary the number of the WS and BS scores using random sub-sampling. Ideally, with increasing number of the WS scores we should observe more stable LR. More data is in general more informative, especially in the tails of the WS and BS distributions.

In the following section we shall study the influence of the size of the WS and BS datasets on the stability of the LR.

## 4. Method used

Since we aim to examine the lower tail of the WS score distribution, we will focus on the similarity score interval 375 – 900 (shown in figure 1). The similarity scores are dimensionless, which advocates for the use of the LR framework. Simulated fingermarks of 4 individuals are used in this study.

***Table 1*** – Proportion of simulated fingermarks

| | No. of fingermarks |
|---|---|
| Individual 1 | 8455 |
| Individual 2 | 4666 |
| Individual 3 | 3179 |
| Individual 4 | 3758 |

Individual 1 is used as a benchmark (largest number of simulated fingermarks available) to study the influence of the varying size of the simulated marks and police database datasets. We defined 5 experimental conditions:

1. Equal proportion of WS and BS scores (Symmetric)
2. WS[8455] and BS varying (WSmax)
3. WS[500] and BS varying (BSmin)
4. WS varying and BS[500] (BSmin)
5. WS varying and BS[200'000] (BSmax)

These conditions (where available) will be applied to all 4 individuals.

For all scenarios, the smallest number of WS scores tested counts 500 with 500 scores increments until the WSmax (where available). Similarly the smallest number of BS scores configuration counts 500 with 500 scores increments until BSmax. Since we have a lot more scores available for the BS, we will examine the influence of the amount of BS scores on the stability of the LR with 20.000, 50.000, 100.000 and 200.000 scores.

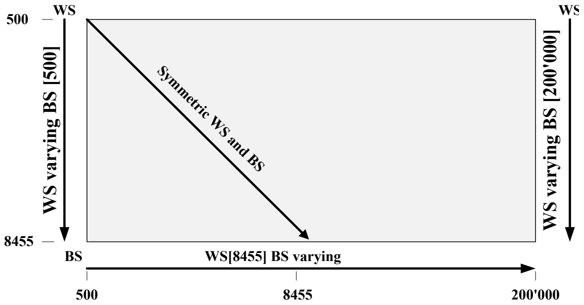

***Figure 4*** – Four scenarios for LR stability analysis

Please recall that we selected the similarity score interval range from 375 to 900 (see Figure 1). Based on the initial assumption that the LRs in this region are of low order of magnitude, we will place the LRs into 4 bins ($10^{-2} <$ LR $< 10^{-1}$; $10^{-1} <$ LR $< 10^{0}$; $10^{0} <$ LR $< 10^{1}$, LR $> 10^{1}$) in order to analyze the LR behavior. We are particularly interested in observing the varying proportions of the LRs crossing the value of the neutral evidence ($LR_E$ = 1), changing the support of $\boldsymbol{H_p}$ to $\boldsymbol{H_d}$ and the actual value of the LRs (observation of the E at a fixed value with changing the experimental conditions). The influence of the varying sizes of the WS and BS datasets on the stability of the LR is presented in the following chapter.

## 5. Results

The experimental setup with most similarity scores (BSmax WSmax) was taken as ideal condition, which we aim to approach with increasing number of the similarity scores. In this sense, we want to get as close to the "best estimate" with the minimum number of scores. Reader should also keep in mind that our aim here is to understand the data rather than draw conclusions of the rather erratic behavior of the LRs produced.

Results are divided into two sections: firstly we will look at the stability of the LR for the individual 1 (counting the most WS scores), while in the second part we will attempt to replicate the results for the remaining individuals.

The sum of all the LRs in the 4 LR ranges is equal (126 – given by the total number of *E* scores for which the LRs have been calculated).

### 5.1 LR stability analysis

In figure 5 one of the populations (BS or WS) is fixed while other one varies from 500 to 8000 (however LRs have been analyzed on the whole range of BS 500 - 200000).

ranges.

Calculated LR values for each piece of evidence E under different experimental conditions are presented in figure 6 on the log-scale. For the experimental condition 1 (symmetric WS and BS) [1000] 85% of LRs support $\boldsymbol{H_p}$, on contrary in the symmetric set WS and BS [4000] only 46% supports $\boldsymbol{H_p}$ (horizontal line in figure 6 indicates LR = 1 and demonstrates the LR shift in support of different hypothesis).

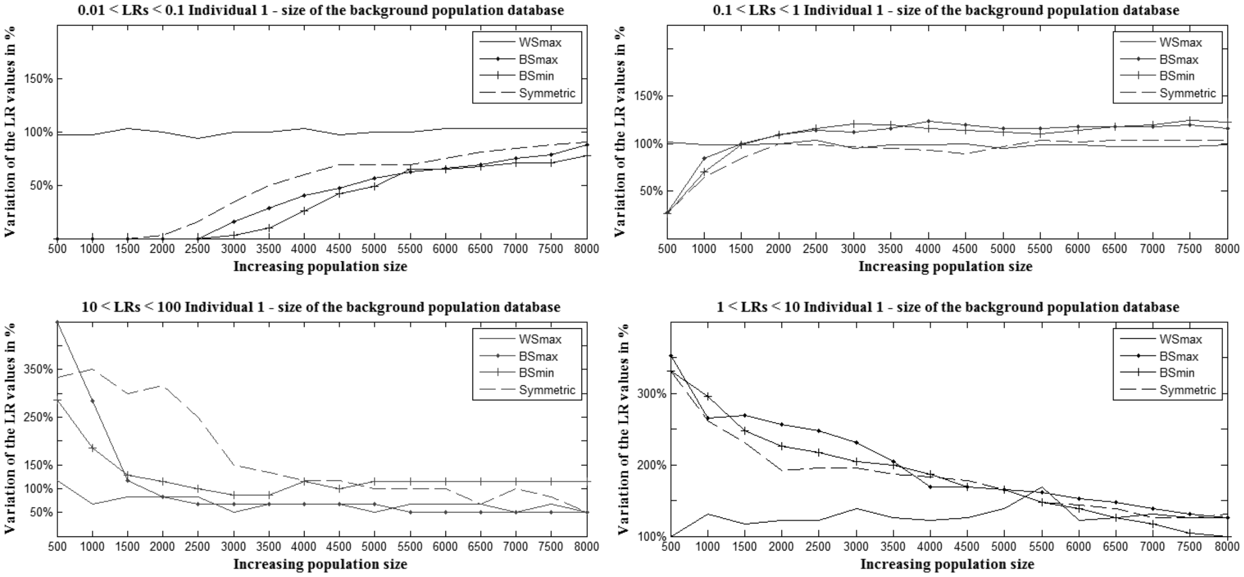

***Figure 5*** – Experimental setups results for individual 1

The stability of the LRs can be observed and compared with varying size of the BS population (BSmin, BSmax…) The experimental results for the individual 1 show that about 4000 scores (WS) are needed to obtain a stabile behavior of ± 10% of the LR values, for the selected LR bin

The size of the BS population does not have a significant influence on the overall stability of the LR. The symmetric experimental condition converges the fastest to the best estimate; therefore this condition will be replicated for the remaining individuals.

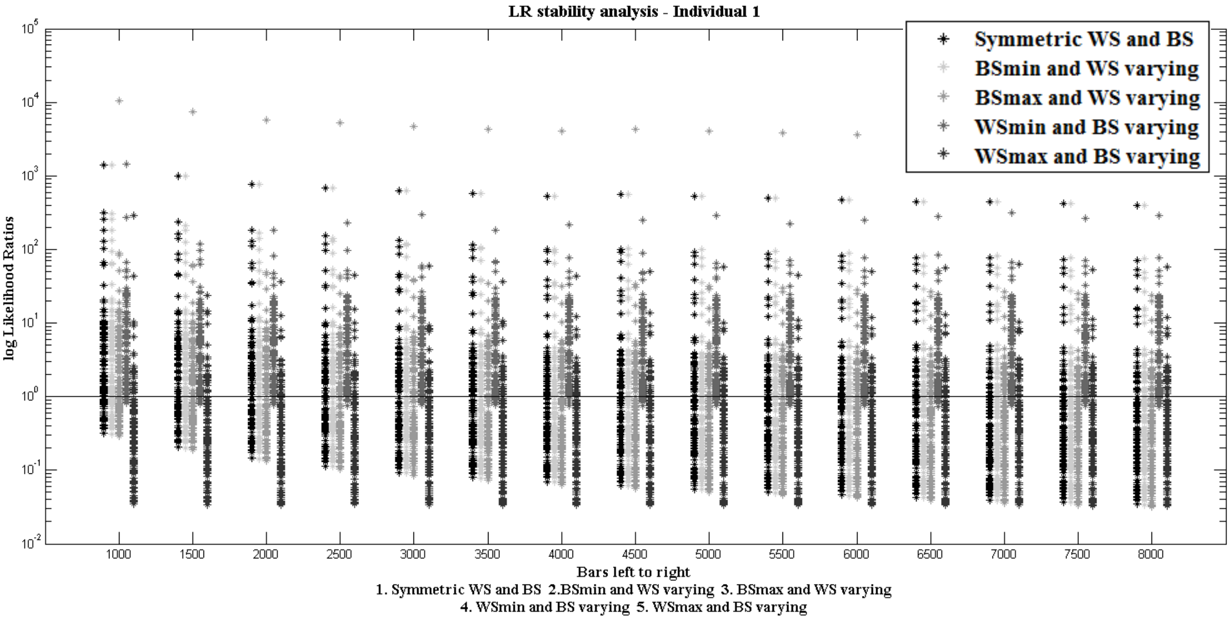

***Figure 6*** – log(LRs) presented with varying BS population

### 5.2 Replication for the remaining individuals

The stability of the LRs is analyzed using the experimental condition 1 (symmetric WS and BS). Figure 7 illustrates the experimental results for the individuals 2, 3 and 4.

The best estimate was calculated from the LRs in the configuration (BSmax and WSmax) of each individual. No LR lower than $10^{-1}$ was recorded for individuals 2 – 4; hence this bin will remain empty.

The results observed advocates for using the LR calculation method as described in [5]. Despite of the different size of the within source dataset for the 4 individuals, the stabilizing effect of increasing the size of the datasets on the LRs (as observed in the benchmark) was replicated with amongst all four individuals. Analyzing the results separately, within source scores dataset counting 4500 seems sufficient to reach stability of ± 10% of the LR values for individual 2, 3000 for individual 3 and population size of 2000 for individual 4. More general conclusions cannot be drawn from such a limited number of individuals.

## 6. Discussion and conclusions

The aim of this article was to study the influence of the size of datasets on the stability of the LR. Judging from the experiments conducted, the increase in the between source population size does not seem to have much influence on the LR stability. The symmetric experimental setup has shown to produce the most stable LRs, while a significant variability was observed between the WSmin and WSmax experiments (see figure 6).

The stabilizing trend of the LR due to the increasing size of within source population was replicated for all four individuals, however the results show differences in the minimum number of the within source scores necessary to obtain a stable LR amongst the different individuals and call for further tests with datasets of comparable sizes before a generic threshold can be set.

The use of simulated fingermarks in the experiments show that they are a valuable evaluation tool, as they are relatively easy to produce in significant quantities and one can be “beyond any doubt” certain regarding their origin.

## 7. Future work

This article is intended as a preliminary study on the stability of the LRs and shows how the LRs behave with varying population sizes. The future work will focus on obtaining similarly large datasets of simulated fingermarks to individual 1 and extend the study for the *E* different source. Following research will be dedicated to non-parametric methods and model-based approaches.

## Acknowledgements

The research has been conducted on behalf of the BBfor2 project of the Marie Curie Actions FP 7 using the NFI data resources.

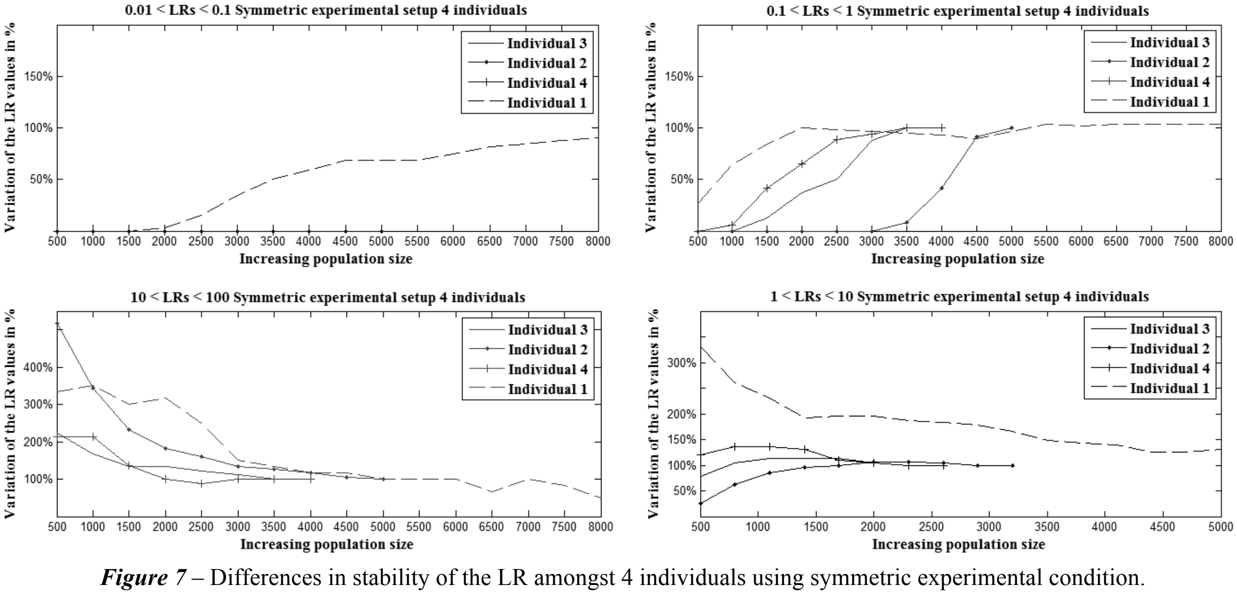

*Figure 7* – Differences in stability of the LR amongst 4 individuals using symmetric experimental condition.

## References

[1] D. Ramos, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, Universidad Autonoma de Madrid, November 2007

[2] C. Neumann, Quantifying the weight of evidence from a forensic comparison: a new paradigm, RSS 175(2), (2011) pp 1 – 26

[3] C. M. Rodriguez, A. de Jongh, D. Meuwly, Introducing a semi-automated method to simulate large numbers of forensic fingermarksfor research on fingerprint identification, JFS 57(2), (2012) pp. 334 – 342

[4] N. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – Modelling within finger variability, FSI 167 (2007) 189 – 195

[5] C. Neumann et al., Computation of Likelihood Ratios in fingerprint identification for configurations of three minutiae. JFS 51(6), (2006) 1255 – 1266.

[6] T. Ali, L.J. Spreeuwers, R.N.J. Veldhuis, A review of calibration methods for biometric systems in forensic applications, In: 33rd WIC Symposium on Information Theory in Benelux, Boekelo, Netherlands, (May 2012), pp. 126-133, WIC. ISBN 978-90-365-3383-6