Forensic Science International 287 (2018) 113–126

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation

H.J. Swofford[a,*], A.J. Koertner[a], F. Zemp[b], M. Ausdemore[c], A. Liu[d], M.J. Salyards[a]

[a] U.S. Army Criminal Investigation Laboratory, Defense Forensic Science Center, USA
[b] School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland
[c] Department of Mathematics and Statistics, University of South Dakota, USA
[d] Department of Statistics, University of Virginia, USA

ARTICLE INFO

Article history:
Received 3 December 2017
Received in revised form 8 March 2018
Accepted 27 March 2018
Available online 3 April 2018

Keywords:
Forensic science
Fingerprints
Strength of evidence
Weight of evidence
Likelihood ratio
Probability

ABSTRACT

The forensic fingerprint community has faced increasing amounts of criticism by scientific and legal commentators, challenging the validity and reliability of fingerprint evidence due to the lack of an empirically demonstrable basis to evaluate and report the strength of the evidence in a given case. This paper presents a method, developed as a stand-alone software application, *FRStat*, which provides a statistical assessment of the strength of fingerprint evidence. The performance was evaluated using a variety of mated and non-mated datasets. The results show strong performance characteristics, often with values supporting specificity rates greater than 99%. This method provides fingerprint experts the capability to demonstrate the validity and reliability of fingerprint evidence in a given case and report the findings in a more transparent and standardized fashion with clearly defined criteria for conclusions and known error rate information thereby responding to concerns raised by the scientific and legal communities.

1. Introduction

Over the last several years, the forensic science community has faced increasing amounts of criticism by scientific and legal commentators, challenging the validity and reliability of many forensic examination methods that rely on subjective interpretations by forensic practitioners [1–7]. Of particular concern, noted in 2009 by the National Research Council (NRC) of the National Academies of Science (NAS) [3] as well as the President's Council of Advisors on Science and Technology (PCAST) as recently as September 2016 [7], is the lack of an empirically demonstrable basis to substantiate conclusions from pattern evidence, thus limiting the ability for the judiciary to reasonably understand the reliability of the expert's testimony for the given case. Consistent with several academic commentators, both the NRC and PCAST strongly encouraged the forensic science community to develop tools to evaluate and report the strength of forensic evidence using validated statistical methods [3,7–8]. While these concerns apply to nearly every pattern evidence discipline, the forensic fingerprint discipline has received most of the attention because fingerprint analysis is one of the most widely used techniques in the criminal justice system. As a result, over the last several years numerous methods and models have been proposed to provide a statistical estimate of the weight of fingerprint evidence using features that are familiar to forensic practitioners, primarily fingerprint minutiae [9–23].

Prior methods can be classified as either (a) feature-based models, which calculate probability estimates from the random correspondence of feature configurations within a pre-determined tolerance or (b) similarity metric models, which calculate the probability estimates from distributions of similarity scores. Among the feature-based models: Zhu et al. proposed a family of finite mixture models to represent the distribution of fingerprint minutiae, including minutiae clustering tendencies and dependencies in different regions of the fingerprint image domain to calculate the probability of a random correspondence [10]; Su and Srihari proposed a model based on the spatial distribution of fingerprint minutiae, taking into account the dependency of each minutiae on nearby minutiae and the confidence of their presence in the evidence, to calculate the probability of random correspondence [14]; Lim and Dass proposed a simulation model based on the distribution of fingerprint minutiae estimated using a Bayesian MCMC framework [15]; Abraham et al. proposed a model based on support vector machines trained with features discovered via

[*] Corresponding author.
E-mail address: henry.j.swofford.civ@mail.mil (H.J. Swofford).

https://doi.org/10.1016/j.forsciint.2018.03.043
0379-0738/Published by Elsevier B.V.

morphometric and spatial analyses of corresponding minutiae configurations for both match and close non-match populations [19]. Among the similarity metric models: Neumann et al. proposed a variety of models based on a similarity metric calculated from feature vectors taking into consideration type, direction, and relative spatial relationships of fingerprint minutiae [9,12,17] as well as taking into account general pattern [18]; Egli [11,13,21], Choi and Nagar [16], and Leegwater et al. [23] proposed a variety of models based on the distribution of similarity scores from Automated Fingerprint Identification Systems (AFIS). Alberink et al. evaluate the effect of different types of conditioning on the impact of the results derived from AFIS-based models [20]. Taking a slightly different approach than those discussed above, Neumann et al. proposed a model relying on an AFIS algorithm to estimate the probability distributions of spatial relationships, directions and types of minutiae rather than directly modeling the distribution of AFIS scores [22].

Although each of the proposed models demonstrated promising performance metrics, none have been widely accessible to the forensic community, thus prohibiting their ability to be further evaluated or implemented into routine casework. Consequently, forensic science laboratories throughout the United States have been unable to adequately address the concerns by the NRC and PCAST by demonstrating the reliability of fingerprint evidence *for the case at hand*. In light of this gap, this paper presents a method, developed as a stand-alone software application, *FRStat*, which measures the similarity between two configurations of friction ridge skin features and calculates a similarity metric. Statistical modeling of the distributions of the similarity statistic values from mated and non-mated impressions facilitates a statistical assessment of the strength of the fingerprint evidence. Although this method builds upon the general concepts of similarity-based models described earlier, this method utilizes a novel approach for quantifying the similarity and strength of fingerprint evidence. Further, having been developed as a stand-alone software application by the United States Government, this method is accessible to the forensic community thereby providing the capability to ensure the strength of fingerprint evidence is evaluated with an empirically grounded basis.

This paper provides a brief overview of the similarity calculations performed by the method followed by more detailed discussions regarding its development, performance and validation. Limitations of the method and considerations for policy and procedure when applied to forensic casework are also discussed.

2. Materials & methods

2.1. Similarity calculations

In general terms, the method measures the similarity between the configurations of friction ridge skin features (often referred to as level 2 detail or minutiae) from two different fingerprint images. The spatial relationships and angles of the features annotated by a forensic examiner are used to calculate a similarity statistic (i.e. score). The similarity statistic is then evaluated against datasets of similarity statistic values derived from pairs of impressions relevant for forensic casework made by mated (same) and non-mated (different) sources of friction ridge skin to calculate a statistical estimate of the strength of the given comparison. The method consists of three overarching steps: (1) feature pairing, (2) feature measurements, and (3) similarity statistic calculations.

In order to perform the similarity calculations, the features must be paired between the two impressions. Features are paired by initially detecting the Cartesian coordinates and angles of the annotated features on each image, which represent the locations and angles of ridge flow for the features. Using those feature details, a series of transformations are performed by iteratively rotating and translating the feature configurations to identify the optimal overlay of features between the two impressions among all possible overlays. Corresponding features are paired between the two images using a well-established combinatorial optimization algorithm to solve for the “optimal assignment” of features within each configuration [24]. Fig. 1 illustrates the overlay and pairing of features. Once paired, the features retain their original Cartesian coordinates and angles as they appear on their respective images.

Feature measurements are performed by applying a series of translation and rotation transformations to the paired features to facilitate anchoring and overlay of feature triplets (sub-configurations of three features). Within the feature triplet, two features serve as primary and secondary anchors while the third feature is measured with respect to the Euclidean distance and angle differences between the paired features. The primary anchor features are aligned on the origin of a coordinate plane and the secondary anchor features are aligned parallel to the x-axis. Fig. 2 illustrates this concept of anchoring and overlaying a feature triplet. Using the measured differences between paired features, a “weight” is calculated for both the distance difference and angle difference between each feature. This process is repeated such that weights for distance and angle differences are calculated for all features using every possible combination of features in each triplet.

The weight functions exploit subtle variations in the measured differences as well as provide context to the significance of those measurements in terms of the plasticity of friction ridge skin. The weight functions were designed such that the following criteria were met:

a. The weight functions are insensitive to common variations of feature location and angle displacements in mated source impressions due to distortion during friction ridge skin deposition under heavy pressure and movement.
b. The weight functions maximize the separation of similarity statistic values between mated and non-mated impressions for a given quantity of features.

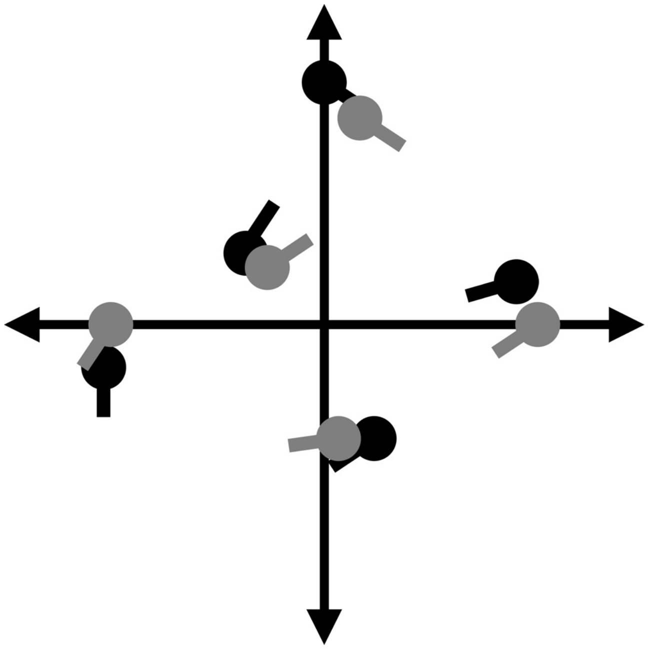

Fig. 1. Conceptual illustration of the overlay and pairing of features. The grey annotations represent features on one impression and the black annotations represent features on the other.

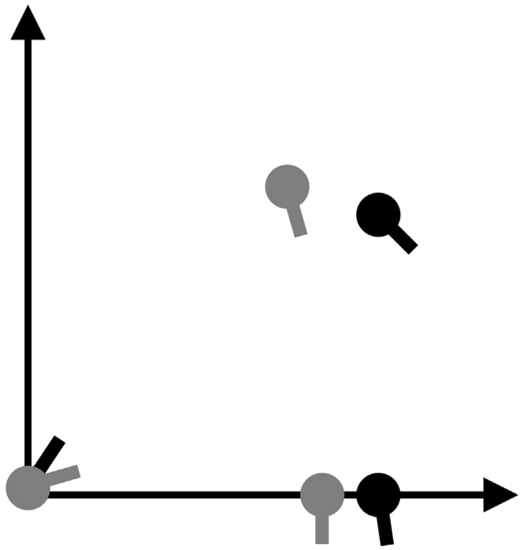

Fig. 2. Conceptual illustration of the anchoring and overlay of a feature triplet. The primary pair of anchor features are on the origin. The secondary pair of anchor points are parallel to the x-axis. The grey annotations represent features on one impression and the black annotations represent features on the other.

c. The weight functions increase the separation of similarity statistic values between mated and non-mated impressions as the number of features increases.

The rules and parameter values for the weight functions are based on the empirical observations by Fagert & Morris [25]. In their study, Fagert & Morris [25] measured the variations of features commonly observed from repeated impressions of mated source fingers under various conditions of lateral pressure with respect to the distance difference and angle differences of features. Using the observations by Fagert & Morris [25] as an initial starting point, manual optimizations of the rules and parameter values for the weight functions were performed using a subset of mated fingerprint samples representing actual casework conditions. Once the measurements and weights for each feature are calculated they are combined into a single statistic and transformed to represent the global similarity of the entire configuration of features (once transformed, higher values indicate higher similarity).

As noted above, the similarity statistic is dependent upon the manual selection and annotation of features by fingerprint experts. Consequently, the precision by which features are annotated introduces uncertainty in the calculated value of the similarity statistic. The method accounts for this uncertainty by applying an iterative random sampling scheme for the annotated details resulting in random displacements of the feature annotations in terms of Euclidean distance and angles. The parameters for the random displacements of feature annotations were determined by modeling the variability of feature annotations in latent impressions and reference impressions across multiple practicing fingerprint experts employed by a federal crime laboratory in the United States. Supplemental Appendix I provides more specific details regarding the evaluation and statistical modeling of the precision of feature annotations by practicing experts.

Following one-hundred iterations of randomly displacing feature annotations and re-calculating the global similarity statistic (using an unseeded random number generator), the final similarity statistic value output to the user is calculated as the lower bound of the 99% confidence interval for the mean. The lower bound of the 99% confidence interval was selected as it provides a conservative estimate of the "true" similarity statistic value for the given annotation.

2.2. Empirical distributions

The empirical distributions of similarity statistic values among mated and non-mated impressions provide the foundation for estimating the strength of the fingerprint evidence. Taking into consideration that this method is intended for use in criminal or civil courts, the empirical distributions are intentionally biased such that the non-mated data are biased to *higher* similarity statistic values and mated data are biased to *lower* similarity statistic values. For non-mated data, this is accomplished by conditioning on (i) the region of friction ridge skin which maximizes the opportunities of observing higher values and (ii) any set of n features determined to be "optimally paired" from a larger set of m possible features with respect to the combinatorial optimization algorithm described in Ref. [24] under any condition of rotation and translation such that the similarity statistic values are maximized. For mated data, this is accomplished by conditioning on lateral pressures and other distortions such that the similarity statistic values are minimized and ensuring that the distributions represent the full range of plausible similarity statistic values that could reasonably be observed in casework when impressions are subject to various distortions during deposition. Keeping in mind that the similarity calculations do not take into account pattern type, feature type, specific feature configurations, or other details which may have biological dependencies, the empirical distributions were not conditioned on those specific criteria. However, because the similarity statistic calculations were designed to account for feature quantity, the distributions are calculated separately for each quantity of features (ranging from 5 to 15).

For the non-mated distributions, conditioning on the delta region was determined to maximize the opportunities of observing higher similarity statistic values. Supplemental Appendix II provides more specific details regarding this determination. The distributions of similarity statistic values characterizing the broader population of non-mated samples for each quantity of features (ranging from 5 to 15) were generated using a subset of impressions from the National Institute of Standards and Technology (NIST) Special Database (SD) 27 [26], cropped to a standard size of 0.5 in. × 0.5 in. (12.7 mm × 12.7 mm) centered on the delta and randomly paired to non-mates. Features were annotated by practicing fingerprint experts beginning with those closest to the delta. Only n number of features under consideration were annotated in "image #1". All visible features, m, in "image #2" were annotated, such that $m » n$ for each comparison. For each quantity of features, a distribution of 2000 similarity statistic values was calculated and conditioned on any set of n features on image #1 determined to be "optimally paired" from the larger set of m possible features on image #2 with respect to the combinatorial optimization algorithm described in Ref. [24]. The two-sample Kolmogorov–Smirnov (K–S) test was used to evaluate the stability of the distributions. This was accomplished by comparing the distribution from one half of the dataset to the distribution from the other half of the dataset (each half distinct from one another) for each quantity of features. The K–S test was selected for this purpose on the basis of its ubiquitous use as a non-parametric test of the equality of continuous probability distributions. For all distributions, the K–S test resulted in a $p » 0.05$ and determined to be sufficiently stable to permit parameter estimation and modeling of the population distributions.

For the mated distributions, a sample of fingerprints were collected from 50 different individuals using a livescan device with extreme distortions deliberately produced. This sample was determined to provide distributions representative of those observed in actual casework. Supplemental Appendix III provides more specific details regarding this determination. For the mated distribution, each individual provided eleven repeated impressions from the right thumb on the livescan device. The thumb was chosen because it results in maximal pliability of skin compared to the other fingers [25]. The repeat impressions consisted of one “non-distorted” impression used as the reference print and the remaining ten were made with lateral distortions applied in the following directions: north, south, east, west, northeast, northwest, southeast, southwest, twist clockwise, and twist counter-clockwise. Pressure was applied in the respective directions until the skin began to lose grip with the livescan surface. Of the 500 pairs obtained (ten distortions each for fifty different individuals), one pair lacked sufficient clarity to permit accurate determination of the corresponding features and therefore was discarded. Fifteen corresponding fingerprint features for the remaining 499 pairs of mated fingerprint impressions were annotated by practicing fingerprint experts in a federal crime laboratory in the United States. The distribution of similarity statistic values for each subset of feature quantities (ranging from 5 to 15) was calculated by randomly selecting (using a random selection algorithm) four combinations of n features out of m available (where $m = 15$). This resulted in 1996 similarity statistic values for each quantity of features (ranging from 5 to 14) and 499 similarity statistic values for 15 features. The stability of the distributions were evaluated using a two-sample K–S test comparing the distribution from one half of the dataset to the distribution from the other half of the dataset (each half distinct from one another) for each quantity of features. For all distributions, the K-S test resulted in a p » 0.05 and determined to be sufficiently stable to permit parameter estimation and modeling of the population distributions.

2.3. Parameter estimation and modeling

The empirical distributions of similarity statistic values described above (non-mated and mated) were modeled to determine plausible probability density functions which may model the similarity statistic values for the relevant populations of non-mated and mated friction ridge skin impressions. Taking into consideration the visual appearance of the empirical distributions and the construct of the weighting functions, the empirical distributions were each modeled using k-component (where $k = 2$ or 3) mixtures of Gaussian distributions. Component weights and parameter estimates were determined using maximum likelihood estimation methods within commercially available statistical analysis software (JMP). Although k-component Gaussian mixtures are more common, logistic distributions were applied on the basis of their heavier tails compared to Gaussian distributions. The heavier tails provide more conservative estimates of probabilities in the extreme ends of the distributions. The parameters for the logistic distribution were approximated using the estimated parameters of the Gaussian distributions. This was accomplished by setting the location parameter of the logistic distribution equal to the mean parameter of the Gaussian distribution as well as applying a coefficient to the standard deviation parameter of the Gaussian to approximate the scale parameter of the logistic distribution such that the difference between the two densities is minimized. Prior to estimating the component weights and parameter values, the empirical distributions were partitioned into two groups. For each bin of feature quantities, three-fourths of the sample was randomly selected using a random selection algorithm and used to estimate the population distribution parameters. The remainder of the sample was used to evaluate the goodness of fit of the estimated parameters for the population distribution. Once the optimal parameters were estimated, a one-sample K–S test was performed to evaluate the goodness of fit between the estimated theoretical logistic mixture distribution and the empirical distribution of the partition of similarity statistic values that was not used to estimate the theoretical distribution parameters. This process was repeated for each quantity of features (ranging from 5 to 15) for both mated and non-mated samples. The parametric models are proposed as plausible estimations of the population distributions for each quantity of features. Supplemental Appendix IV provides more specific details regarding these determinations. Figs. 3 and 4 illustrate the overlays between the theoretical density distributions and the empirical distributions of similarity statistic values for non-mated and mated datasets, respectively.

2.4. Method performance

The overall performance of the method was evaluated in terms of its sensitivity, specificity, within-sample variability, and between-sample variability. The performance of the method may be evaluated in terms of both the similarity statistic (i.e. global similarity statistic, GSS) values alone as well as in terms of the similarity statistic values in the context of the relevant probability distributions of mated vs. non-mated populations.

In terms of the mated distribution, the value of interest is the left tailed probability (the probability of a specific similarity statistic value *or lower*) as depicted in Eq. (1). In other words, the left tailed probability provides an indication of the proportion of similarity statistic values from mated source impressions which are estimated to be *less* than a specified test statistic value for a given case at hand. In terms of the non-mated distribution, the value of interest is the right tailed probability (the probability of a specific similarity statistic value *or higher*) as depicted in Eq. (2). In other words, the right tailed probability provides an indication of the proportion of similarity statistic values from non-mated source impressions which are estimated to be *greater* than a specified test statistic value for a given case at hand.

$$P\left(GSS_n \leq GSS(t)_n | \theta_{n_{mated}}\right)$$

Equation 1: *The left-tailed probability of observing a given similarity statistic, GSS(t), value or lower with respect to the distribution of GSS values from mated impressions, where “t” indicates the test statistic, “n” represents the feature quantity, and θ_n represents the parameters characterizing the distribution of values for a given feature quantity.*

$$P\left(GSS_n \geq GSS(t)_n | \theta_{n_{non-mated}}\right)$$

Equation 2: *The right-tailed probability of observing a given similarity statistic, GSS(t), value or higher with respect to the distribution of GSS values from non-mated impressions, where “t” indicates the test statistic, “n” represents the feature quantity, and θ_n represents the parameters characterizing the distribution of values for a given feature quantity.*

The values derived from Eqs. (1) and (2) may be combined as a ratio, such that the estimated proportion of a given similarity statistic value *or lower* among mated sources is considered relative to the estimated proportion of a given similarity statistic value *or higher* among non-mated sources. Eq. (3) combines Eqs. (1) and (2) as the numerator and denominator, respectively.

$$\frac{P\left(GSS_n \leq GSS(t)_n | \theta_{n_{mated}}\right)}{P\left(GSS_n \geq GSS(t)_n | \theta_{n_{non-mated}}\right)}$$

Equation 3: *Ratio of equations 1 and 2 indicating the relative support of a given similarity statistic, GSS(t), in terms of one proposition (mated) over another (non-mated).*

From equation 3, values greater than 1 indicate a higher probability of the observed similarity statistic value among mated sources compared to non-mated sources and values less than 1 indicate a higher probability of the observed similarity statistic

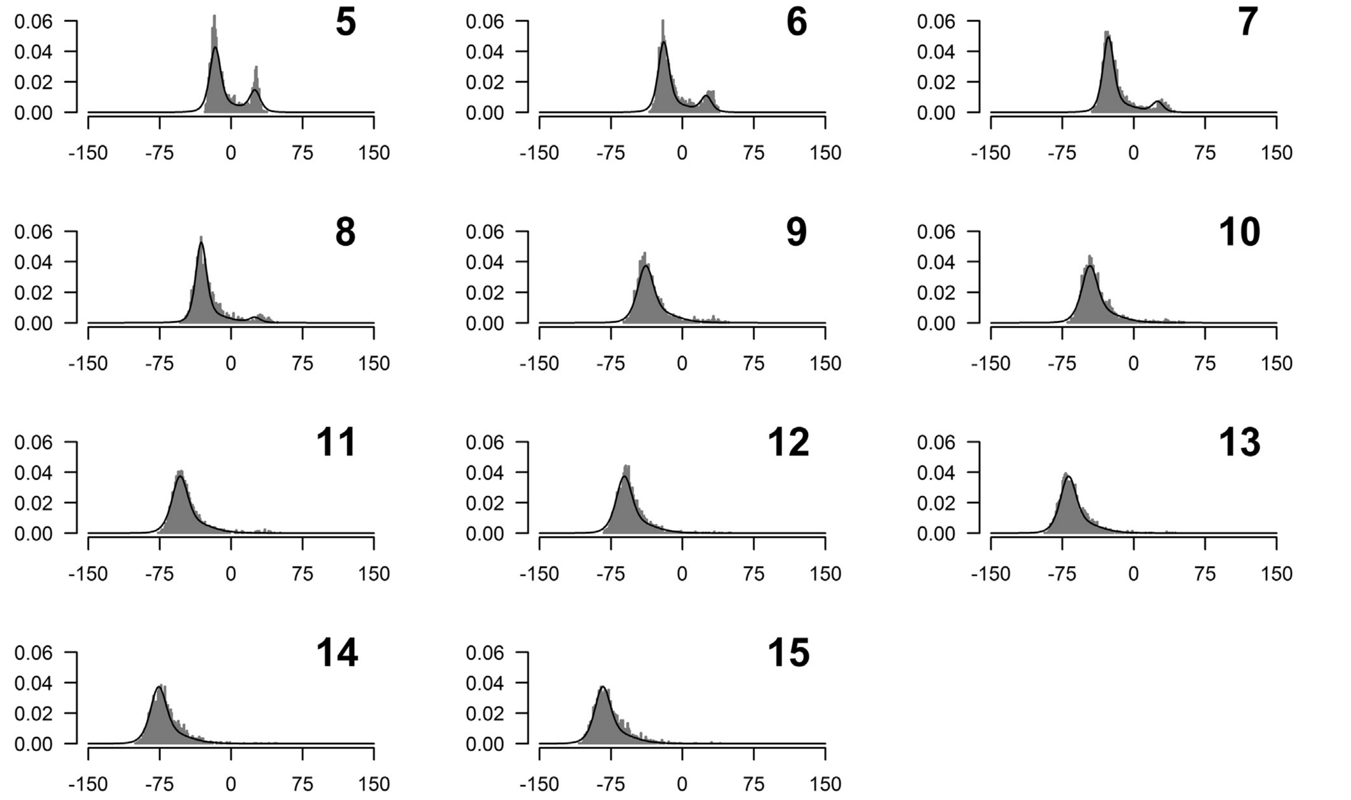

Fig. 3. Empirical density distributions of the similarity statistic values for the non-mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The x-axis represents the global similarity statistic values. The y-axis represents the density.

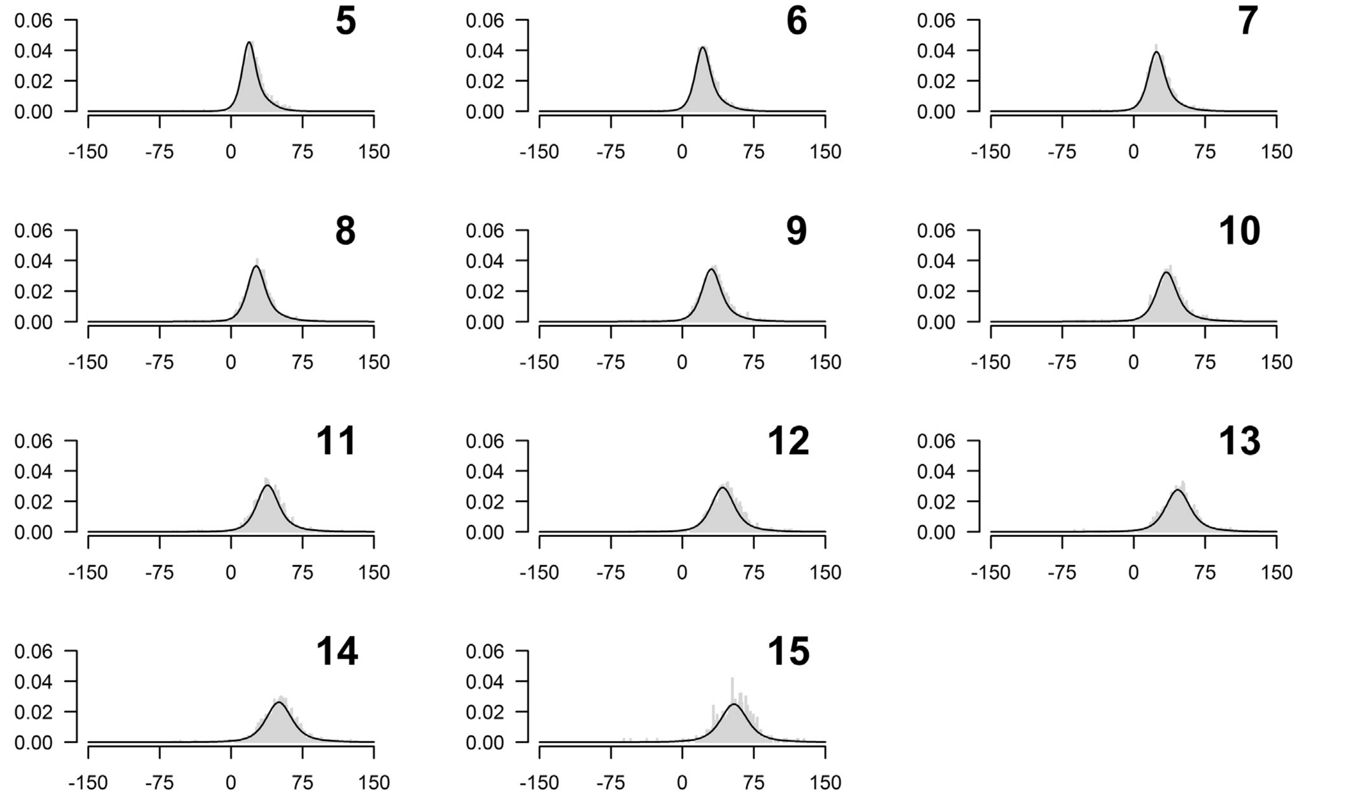

Fig. 4. Empirical density distributions of the similarity statistic values for the mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The x-axis represents the global similarity statistic values. The y-axis represents the density.

values among non-mated sources compared to mated sources. Values equal to 1 indicate equal probability of the observed similarity statistic value among mated and non-mated sources.

It is important to note that Eqs. (1) and (2) are calculated as tail probabilities rather than likelihoods; thus, Eq. (3) is not a true likelihood ratio or Bayes' factor and should not be used as such with the intent of calculating a posterior probability.

2.4.1. Datasets

The performance of the method is evaluated using the following datasets:

1. Mated test dataset #1 (*known* to be mated) —A test dataset of 288 mated latent and reference impressions deposited under semi-controlled, normal handling conditions (to simulate casework) on a variety of different surfaces by 78 different individuals. The purpose of this dataset is to evaluate the performance of the method using latent and reference impressions which are similar to casework in terms of deposition and development, but for which ground truth mated status is known. Latent impressions were developed using a variety of chemical and physical processing techniques commonly used in casework by fingerprint experts, such as cyanoacrylate ester fuming, fluorescent dye stains, ninhydrin, indanedione, 1-8 diazafluoren-9-one, and fingerprint powders. Each set was visually examined and corresponding features (ranging between 5 and 15) were manually annotated by practicing fingerprint experts in a federal crime laboratory in the United States. The overall quality (clarity) of the latent impressions is considered to be representative of casework impressions. This is based on the subjective evaluation by fingerprint experts as well as a comparison of the empirically measured quality scores using LQMetrics software available in the Universal Latent Workstation. A two-sample K–S test was performed comparing the distribution of LQMetric quality (clarity) scores from this dataset to the distribution of LQMetric clarity scores from the publically available dataset of casework impressions (mated test dataset #2 described below). The value of the K–S test statistic ($D_{288,184} = 0.087$) fails to reject the null hypothesis that the two samples originated from the same distribution ($p > 0.05$) based on a p-value decision threshold of 0.01.
2. Mated test dataset #2 (*accepted* to be mated) — A casework dataset of 184 latent and reference impressions publically available by the National Institute of Standards and Technology (NIST) Special Database 27 [26]. Although this dataset is commonly accepted to be mated by the general scientific community, it was collected from adjudicated casework by the Federal Bureau of Investigation and therefore ground truth is not actually known. The purpose of this dataset is to evaluate the performance of the method using latent and reference impressions from actual casework and which has been publically available and commonly used by the general scientific community. Each set was visually examined and corresponding features (ranging between 5 and 15) were manually annotated by practicing fingerprint experts in a federal crime laboratory in the United States. NOTE: The NIST Special Database 27 actually contains 258 latent and reference impressions in total; however, only 184 were able to be evaluated due to a technical issue with the remaining files preventing them from being opened (corrupted image files).
3. Mated test dataset #3 (*believed* to be mated) — A casework dataset of 605 latent and reference impressions collected from casework during the course of routine operations by fingerprint experts in a federal crime laboratory in the United States and reported as "positive associations". The purpose of this dataset is to evaluate the performance of the method using latent and reference impressions from a much larger sample of actual casework impressions as compared to the NIST Special Database 27 alone. The impressions were collected from a wide variety of cases, substrates, and assigned fingerprint experts. The corresponding features (ranging between 7 and 15) were manually annotated by the assigned fingerprint expert during the initial case examination. The selected features were then annotated later in a format suitable for *FRStat* analysis by the same fingerprint expert for purposes of this evaluation.
4. Non-mated test dataset #1 (*known* to be non-mated) — A test dataset of 20 latent print images from the mated test dataset #1 that were selected on the basis of representing the left delta region fingerprint impressions and 25 non-mated reference images obtained from the NIST Special Database 27 [26]. The purpose of this dataset is to evaluate the performance of the method using non-mated impressions for which the impressions were arbitrarily paired and for which the impressions are publically available and commonly used by the general scientific community. For each latent print image, fifteen features were annotated around the delta region. Each reference print was cropped to a standard size of 0.5 in. × 0.5 in. (12.7 mm × 12.7 mm) centered on the left delta. All features visible in the cropped reference images were manually annotated by practicing fingerprint experts. For each comparison of the 20 latent prints to each of the 25 non-mated reference prints, a configuration of n features was randomly selected (using a random selection algorithm) from the latent print and compared against the reference print (each containing m annotated features, where $m \gg n$) resulting in 500 similarity statistic values for each set of n features (ranging from 5 to 15). One similarity statistic value was obtained per image pair. The similarity statistic value was conditioned on any set of n features on image #1 determined to be "optimally paired" from the larger set of m possible features on image #2 with respect to the combinatorial optimization algorithm described in Ref. [24] under any condition of rotation and translation.
5. Non-mated test dataset #2 (*known* to be non-mated; "close non-match" from AFIS database search) — Two separate datasets: (#2a) a test dataset of fingerprint images representing the "delta" region and (#2b) a test dataset of fingerprint images representing the "core" region. The purpose of this dataset is to evaluate the performance of the method using non-mated impressions for which the impressions were paired on the basis of an AFIS similarity algorithm. Each dataset was separated into eleven separate subsets, each containing approximately 100 samples, conditioned on the number of features (n) being compared (ranging from 5 features to 15 features). Features were manually annotated by practicing fingerprint experts such that the features closest to the reference point (core or delta depending on the sample) were annotated first and then the remaining n features were annotated in a radiating fashion outward. Post annotation, each image was cropped by a bounding rectangle such that only those ridges and features that are part of the annotated configuration remain. These images serve as the "query" print. Each query print was then searched using an AFIS against an operational database containing approximately 100 million different fingerprint impressions from approximately 10 million different individuals. The AFIS ranked the top 20 most similar reference fingerprints to the fingerprint image searched. Of the top 20 results, the fingerprint image in rank 1 was confirmed to be a non-mated source with respect to the query print and used for comparison. Supplemental Appendix II provides more specific details regarding the development of this dataset.

2.4.2. Sensitivity & specificity

The sensitivity was measured as the proportion of mated samples which resulted in a probability ratio value greater than a specified threshold ratio value. The specificity was measured as the proportion of non-mated samples which resulted in a ratio value less than a specified threshold ratio. Both the sensitivity and specificity will vary as a function of the ratio value chosen as a threshold. As the threshold ratio value increases, the sensitivity will decrease and the specificity will increase. As the threshold ratio value decreases, the sensitivity will increase and the specificity will decrease. Accordingly, both sensitivity and specificity were measured separately using threshold ratio values of 1, 10, and 100, respectively. In addition to these threshold values, Receiver Operator Characteristics (ROC) curves illustrate the performance of the method across the full range of potential threshold values.

The sensitivity was evaluated using the mated test dataset #1 (*known* to be mated). Mated test dataset #2 (*accepted* to be mated) and mated test dataset #3 (*believed* to be mated) were also utilized to evaluate the consistency between threshold ratio values and experts' interpretation of mated status. The term "consistency" is used here since it is not a true measure of sensitivity because mated status is not truly known. Each dataset was considered separately. Of the total number of available latent and reference impressions in each dataset, up to ten different configurations of n features were randomly selected (using a random selection algorithm) from m available for each quantity of features (ranging between 5 and 15) to evaluate the results across the impressions subject to different conditions of distortion. Each configuration is considered as a separate measurement.

The specificity was evaluated using the non-mated test dataset #1 (*known* to be non-mated) as well as the non-mated test datasets #2a and #2b (*known* to be non-mated, "close non-match" from AFIS database search). The use of both datasets provides two different perspectives of the specificity as a result of prints being paired with non-mated impressions selected arbitrarily (non-mated dataset #1) as well prints being paired with the most-similar non-mated impression selected from a database of approximately 100 million others. In the latter context, "most-similar" is defined as the #1 rank candidate response from a large operational AFIS utilizing blackbox fingerprint search and matching algorithms. It is reasonable to consider the distribution of similarity statistic values from the non-mated test dataset #2 as representing the extreme tail of the distribution of values from the non-mated test dataset #1.

2.4.3. Within-sample variability & between-sample variability

The variability of the method was evaluated separately in terms of the within-sample variability and between-sample variability of the similarity statistic values. The within-sample variability captures the variation as a result of multiple measurements of the *same* features. The between-sample variability captures the variation as a result of multiple measurements of *different* features and prints. Thus, the within-sample variability accounts for variations due to the imprecision and uncertainty of the specific location and angles of the feature annotations and the between-sample variability accounts for variations due to differences in distortions caused by pressure, substrate, etc. from different measurements across different configurations of features and impressions.

By taking into account the imprecision of feature annotations described in Supplemental Appendix I, repeat measurements of the same features (without manual re-annotation) are subject to variation due to the random resampling scheme built into the method. The within-sample variability captures the variation of the similarity statistic values as a result of multiple measurements of the *same* features. The within-sample variability was evaluated using 92 image replicates from the mated test dataset #1 and mated test dataset #2, each of which contained 15 annotated features. Considering the intended use of this method is on impressions believed to be mated by the fingerprint expert, the within-sample variability was not evaluated on the non-mated test datasets. For each image replicate, a configuration of n features was selected at random. Using the *same* configuration of n features for each respective replicate, a series of 25 repeat measurements were taken (where each measurement represents the lower bound of the 99% confidence interval of the k-iterations from the random resampling scheme; and where $k = 100$). The standard deviation of the 25 repeat measurements for each of the 92 image replicates was calculated. Using the standard deviations from each of the 92 image replicates, the combined standard deviation was calculated as the within-sample variability. This was repeated for each bin of feature quantities (ranging from 5 to 15).

The between-sample variability captures the variation of the similarity statistic values as a result of multiple (different) measurements of *different* features across different impressions. While variabilities of the similarity measurements as a result of the imprecision of the feature annotation process are taken into account in the similarity statistic calculations, the variabilities of the similarity measurements as a result of different conditions of distortion across different regions of an impression or across different impressions are not since they are not a consequence of repeat attempts to measure the same feature data. Rather, the between-sample variability is expected to represent a much larger range of similarity statistic values similar to the range of values represented by the estimated parameters of the population distributions discussed in further detail in Supplemental Appendix IV. The between-sample variability was evaluated using all image replicates from the mated test dataset #1 (*known* to be mated), mated test dataset #2 (*accepted* to be mated), and mated test dataset #3 (*believed* to be mated) combined. Considering the intended use of this method is on impressions believed to be mated by the fingerprint expert, the between-sample variability was not evaluated on the non-mated test datasets. For each of the total number of available latent and reference impressions from each mated test dataset (1077), up to ten different k-configurations of n features were randomly selected (using a random selection algorithm) from m available for each quantity of features (ranging between 5 and 15) to evaluate the results across the impressions subject to different conditions of distortion. The standard deviation was calculated as the between-sample variability for each bin of feature quantities (ranging from 5 to 15).

The within-sample variability and between sample variability are both illustrated in terms of the similarity statistic value rather than in terms of the probability ratio because the impact to the probability ratio will vary depending on the location of the similarity statistic value within the distributions — subtle variations of the similarity statistic value in the tail of a distribution will cause a more dramatic change to the probability value compared to the other locations, such as the middle region. Thus, representing the variability in terms of the probability ratio itself would be incomplete and potentially misleading.

3. Results & discussion

The overall performance of the method was evaluated in terms of its sensitivity, specificity, within-sample variability, and between-sample variability. Initially, the expected performance may be evaluated in terms of comparing the empirical distributions of similarity statistic values between mated and non-mated impressions. These distributions served as the empirical

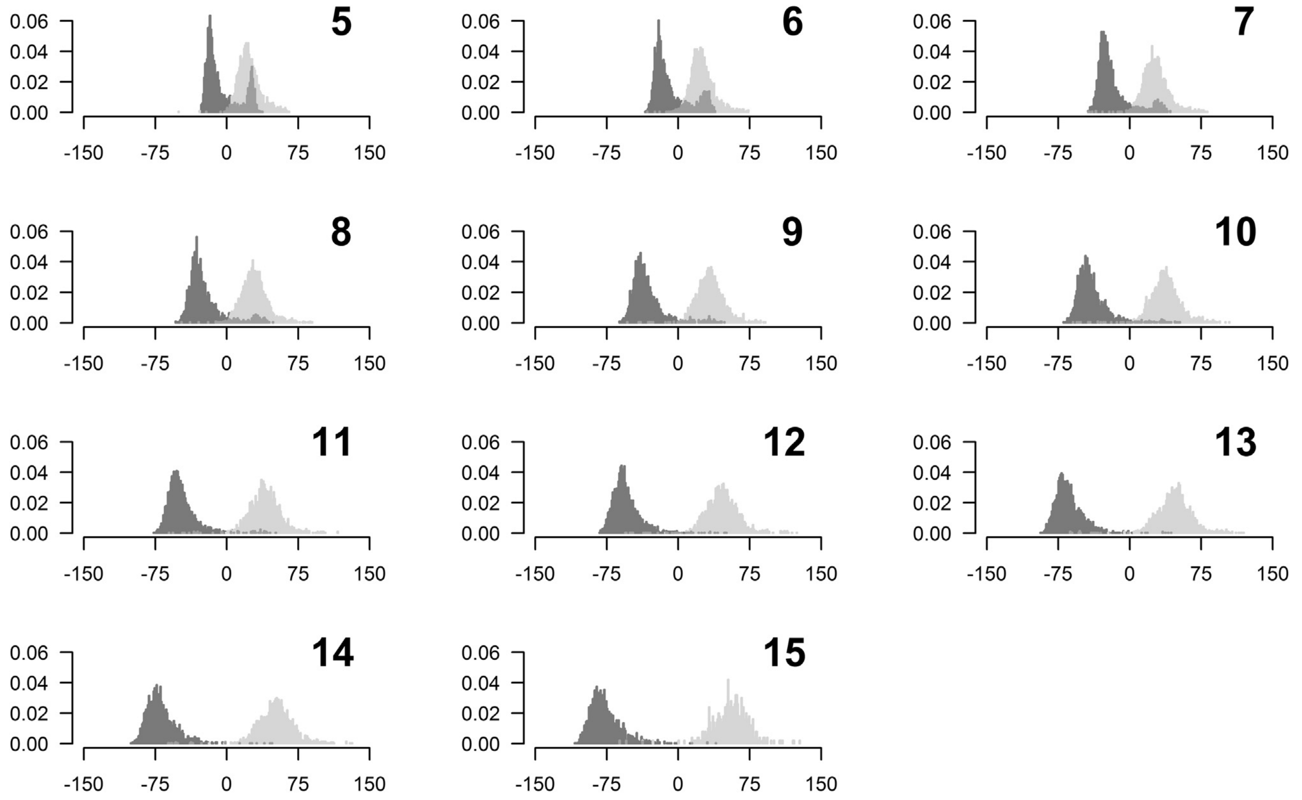

Fig. 5. Empirical distributions of similarity statistic values for both non-mated (dark grey) and mated (light grey) samples for feature quantities 5 through 15. The x-axis represents the global similarity statistic values. The y-axis represents the density.

foundation for the parameter estimations and modeling described in greater detail in Supplemental Appendix IV. Fig. 5 illustrates the empirical distributions in terms of density.

From Fig. 5, two important observations can be made. First, we see that the distributions appear to exhibit little overlap between the mated and non-mated datasets. Second, we see that the distributions appear to increase in separation as the feature quantities increase.

3.1. Sensitivity

The sensitivity was evaluated using the mated test dataset #1 (*known* to be mated). Mated test dataset #2 (*accepted* to be mated) and mated test dataset #3 (*believed* to be mated) were also utilized to evaluate the consistency between threshold ratio values and experts' interpretation of mated status ("consistency" is used here since it is not a true measure of sensitivity because mated status is not truly known). Each dataset was considered separately. Table 1 provides the sensitivity using mated test dataset #1. Table 2 provides the consistency between the method and experts' interpretation of mated status using mated test dataset #2. Table 3 provides the consistency between the method and experts' interpretation of mated status using mated test dataset #3.

With respect to the sensitivity calculations listed above, it is important to note that the values were generated *without* the examiners having direct feedback regarding their annotation precision. Without such feedback, examiners have become acclimated to a relaxed environment in which they were accustomed to annotating the mere presence of a feature and in which measurements were not taken directly from the annotations. In practice, where a fingerprint expert recognizes the importance of precise annotations and adjusts accordingly, it is a reasonable assumption that the sensitivity will be *higher* (and thus the false negative rate will be *lower*) than what is represented in this section; however, a quantitative measure of *how much* higher the sensitivity would be in practice is unknown at this time. Nevertheless, the sensitivity of the method is expected to increase as examiners gain more experience and become more precise in their feature annotations — similar to when examiners gain a better understanding of how feature annotations impact the performance of AFIS search results and adjust their annotation habits accordingly.

3.2. Specificity

The specificity was evaluated using the non-mated test dataset #1 (*known* to be non-mated) as well as the non-mated test datasets #2a and #2b (*known* to be non-mated, "close non-match" from AFIS database search). The use of both datasets provides two different perspectives of the specificity as a result of prints being paired with non-mated impressions selected arbitrarily (non-mated dataset #1) as well prints being paired with the most-similar non-mated impression selected from a database of approximately 100 million others. Table 4 provides the specificity using non-mated test dataset #1. Table 5a and 5b provides the specificity using non-mated test datasets #2a and #2b (Table 5a — "delta" region; Table 5b — "core" region).

With respect to the specificity calculations listed above, it is important to note that the values are limited to the output of the *FRStat* algorithm alone; thus, these values should not be confused with the overall specificity of the latent print examination method in general which is much improved by the input of the fingerprint expert. In practice, where a fingerprint expert's visual examination will precede the calculation of a similarity statistic value using *FRStat* and serve as an initial means of discrimination using details that *FRStat* is not designed to take into account, it is a reasonable assumption that the specificity will be much *higher* (and thus the false positive rate will be much *lower*) than what is represented in this section. However, because there are no publically available

Table 1
Sensitivity of the method using mated test dataset #1 (known to be mated) for each quantity of features (ranging from 5 to 15). Sensitivity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

Feature quantity	Number of configurations (Mated dataset #1)	Sensitivity (ratio >1)	Sensitivity (ratio >10)	Sensitivity (ratio >100)
5	2798	0.657	0.249	0.085
6	2703	0.708	0.381	0.145
7	2550	0.736	0.478	0.234
8	2367	0.823	0.593	0.402
9	2092	0.892	0.755	0.565
10	1898	0.928	0.824	0.645
11	1655	0.947	0.860	0.710
12	1432	0.970	0.925	0.799
13	1230	0.984	0.949	0.825
14	994	0.980	0.971	0.902
15	97	0.990	0.979	0.959

Table 2
Consistency between ratio values greater than 1, 10, and 100 and experts' interpretation of mated status using mated test dataset #2 (accepted to be mated) for each quantity of features (ranging from 5 to 15).

Feature quantity	Number of configurations (Mated dataset #2)	Consistency (ratio >1)	Consistency (ratio >10)	Consistency (ratio >100)
5	1772	0.730	0.201	0.052
6	1674	0.783	0.317	0.100
7	1512	0.830	0.446	0.163
8	1317	0.913	0.636	0.328
9	1166	0.959	0.852	0.595
10	988	0.966	0.899	0.721
11	781	0.968	0.948	0.827
12	706	0.965	0.965	0.905
13	583	0.971	0.971	0.949
14	480	0.973	0.960	0.960
15	47	0.979	0.957	0.957

Table 3
Consistency between ratio values greater than 1, 10, and 100 and experts' interpretation of mated status using mated test dataset #3 (believed to be mated) for each quantity of features (ranging from 5 to 15).

Feature quantity	Number of configurations (Mated dataset #3)	Consistency (ratio >1)	Consistency (ratio >10)	Consistency (ratio >100)
5	6050	0.794	0.287	0.088
6	6038	0.840	0.436	0.150
7	5982	0.870	0.530	0.239
8	5830	0.927	0.716	0.437
9	5526	0.955	0.889	0.690
10	5040	0.961	0.927	0.805
11	4441	0.965	0.934	0.868
12	3876	0.971	0.953	0.910
13	3226	0.970	0.958	0.920
14	2638	0.978	0.974	0.961
15	258	0.981	0.977	0.970

Table 4
Specificity of the method using non-mated test dataset #1 (known to be non-mated) for each quantity of features (ranging from 5 to 15). Specificity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

Feature quantity	Number of image pairs (Non-mated dataset #1)	Specificity (ratio <1)	Specificity (ratio <10)	Specificity (ratio <100)
5	500	0.818	1.000	1.000
6	500	0.850	0.992	1.000
7	500	0.900	0.994	1.000
8	500	0.912	0.986	1.000
9	500	0.940	0.952	0.990
10	500	0.970	0.976	0.992
11	500	0.978	0.982	0.990
12	500	0.988	0.992	0.998
13	500	0.988	0.994	0.996
14	500	0.988	0.992	0.994
15	500	0.996	1.000	1.000

Table 5a
Specificity of the method using non-mated test dataset #2a (known to be non-mated; "close non-match" from AFIS database searches of the delta region) for each quantity of features (ranging from 5 to 15). Specificity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

Feature quantity	Number of image pairs (Non-mated dataset #2a – "delta" region)	Specificity (ratio <1)	Specificity (ratio <10)	Specificity (ratio <100)
5	99	0.566	0.788	0.980
6	99	0.687	0.747	0.980
7	96	0.688	0.719	0.896
8	99	0.747	0.788	0.812
9	99	0.818	0.818	0.828
10	97	0.814	0.835	0.845
11	96	0.802	0.823	0.823
12	98	0.857	0.867	0.888
13	99	0.899	0.929	0.939
14	100	0.980	0.990	0.990
15	100	0.920	0.920	0.940

Table 5b
Specificity of the method using non-mated test dataset #2b (known to be non-mated; "close non-match" from AFIS database searches of the core region) for each quantity of features (ranging from 5 to 15). Specificity was evaluated using a ratio of 1, 10, and 100 as the thresholds.

Feature quantity	Number of image pairs (Non-mated dataset #2b – "core" region)	Specificity (ratio <1)	Specificity (ratio <10)	Specificity (ratio <100)
5	94	0.787	0.979	1.000
6	96	0.802	0.927	1.000
7	95	0.884	0.926	0.979
8	96	0.906	0.938	1.000
9	95	0.884	0.952	0.990
10	96	0.969	0.990	1.000
11	95	0.989	0.989	0.989
12	97	1.000	1.000	1.000
13	97	1.000	1.000	1.000
14	96	1.000	1.000	1.000
15	95	1.000	1.000	1.000

datasets to empirically measure how often non-mated impressions are falsely included by fingerprint experts *and* which result in sufficiently high similarity statistic values using this method, a quantitative measure of *how much* higher the specificity would be in practice cannot be determined at this time.

3.3. Receiver Operator Characteristic (ROC)

The Receiver Operator Characteristic (ROC) illustrates the performance of the method across the full range of potential threshold values. Fig. 6 illustrates the ROC curves for mated test dataset #1 (*known* to be mated) and non-mated test dataset #1 (*known* to be non-mated) as well as the non-mated test datasets #2a and #2b (*known* to be non-mated, "close non-match" from AFIS database search). The use of both non-mated datasets provides two different perspectives of the performance of the method as a result of prints being paired with non-mated impressions selected arbitrarily (non-mated dataset #1) as well prints being paired with the most-similar non-mated impression selected from a database of approximately 100 million others.

From Fig. 6 as well as Tables 4 and 5, we can make two important observations. First, the specificity rates from non-mated dataset #1 and non-mated dataset #2b are very similar to one another. Second, while the specificity rates from non-mated dataset #2a provides an indication of the "worst case-scenario" since it narrowly focuses on the #1 rank candidates out of approximately 100 million other non-mated prints as a result of AFIS searches *and* only considers the delta region of the fingerprint during the searches, the method still demonstrates the ability to accurately classify mated and non-mated impressions. Taking together, the performance characteristics discussed above may provide some general context to the results when non-mated samples are selected at random or whether they were selected on the basis of their similarity from large database searches. The samples comprising non-mated datasets #2a and #2b are limited in size due to operational constraints at the time of collection. A likely consequence of the small sample sizes is the subtle variability in the performance characteristics observed between the various feature quantities, particularly between 13, 14, and 15 features where the observed data suggests 14 features had better performance characteristics than 15 features. With a larger sample, the uncertainty associated with the performance characteristics will be reduced; therefore, further research into the impact of AFIS searches on the specificity rates is encouraged. Nevertheless, because the intent of the method is to estimate the relative prevalence of similarity statistic values among the broader population of non-mated impressions rather than focus only on "close non-mates" from large database searches, the low sample size of these datasets (#2a and #2b) is not considered a critical limitation – their selection as the #1 rank candidate means they were already distinguished from all other impressions in the system using the high performance AFIS algorithms.

3.4. Within-sample variability

The within-sample variability captures the variation of the similarity statistic values as a result of multiple measurements of the *same* features without re-annotations (due to the random resampling scheme discussed in greater detail in Supplemental Appendix I). Table 6 provides the within-sample variability of the method in terms of the combined standard deviation of similarity statistic values. These results demonstrate very low within sample variability and are insignificant compared to the between-sample variability.

3.5. Between-sample variability

The between-sample variability captures the variation of the similarity statistic values as a result of multiple (different) measurements of *different* features. Table 7 provides the

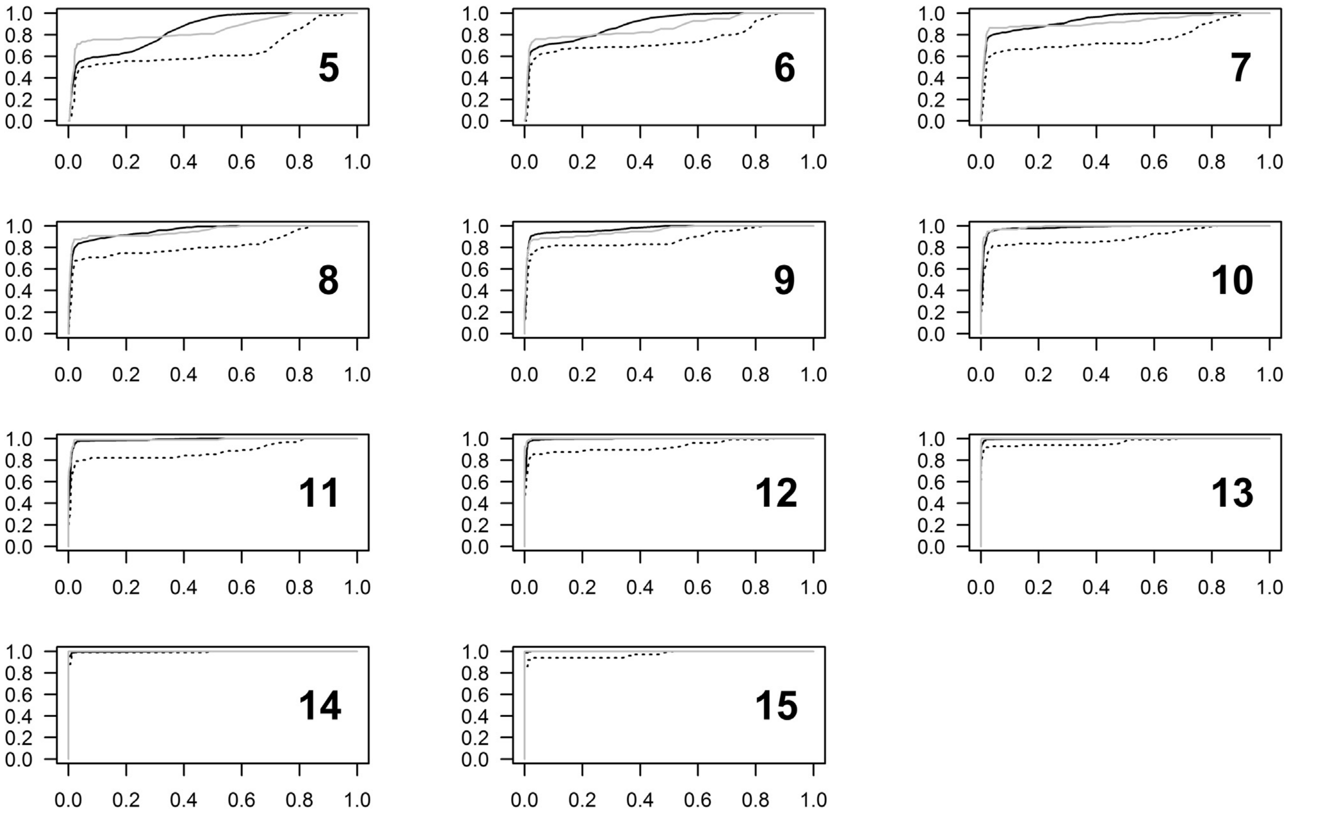

Fig. 6. ROC curves illustrating the performance of the method using mated test dataset #1 (known to be mated) and non-mated test datasets #1 (known to be non-mated) as well as non-mated test datasets #2a (known to be non-mated; "close non-match" from AFIS database searches of the delta region) and #2b (known to be non-mated; "close non-match" from AFIS database searches of the core region) for each quantity of features (ranging from 5 to 15). The solid black line represents the ROC using non-mated test dataset #1 (known to be non-mated). The dotted black line represents the ROC using non-mated test dataset #2a (known to be non-mated; "close non-match" from AFIS database searches of the delta region). The solid grey line represents the ROC using non-mated test dataset #2b (known to be non-mated; "close non-match" from AFIS database searches of the core region). The x-axis represents 1 − specificity. The y-axis represents the sensitivity.

between-sample variability of the method in terms of the similarity test statistic. These results demonstrate between-sample variabilities consistent with those represented by the estimated parameters of the population distributions discussed in further detail in Supplemental Appendix IV and are therefore consistent with expectations.

3.6. General discussion

3.6.1. Ratio values

The ratio values obtained with the method will vary depending on the measured similarity between the two impressions, reflected by the global similarity statistic, GSS(t), as well as the quantity of features. As the GSS(t) value and quantity of features increase, the ratio value will also increase indicating stronger significance of the association between the paired impressions. Theoretically, the ratio values can range from negative infinity to positive infinity; however, this provides little context to understanding the range of ratio values that one may plausibly observe in practice. Fig. 7 illustrates the range of ratio values based on the GSS(t) values corresponding to 95% of the theoretical distribution modeling the mated source dataset (ranging from a left tail probability of 0.025–0.975) for each quantity of features.

From Fig. 7, we observe a steady increase of ratio values as the quantity of features increases. This steady increase is a mathematical consequence of the algorithms for calculating the similarity statistic and consistent with the expected behavior of the method in terms of experience by forensic experts. Although the actual

Table 6
Within-sample variability (combined standard deviation from 25 repeat measurements each for 92 different images) of the similarity statistic value (GSS(t)) for each quantity of features (ranging from 5 to 15).

Feature quantity	Combined σ GSS(t)	Mean GSS(t)
5	0.593	20.742
6	0.648	20.202
7	0.651	24.736
8	0.692	25.104
9	0.831	25.869
10	0.903	32.910
11	0.916	33.371
12	0.969	37.555
13	1.067	39.275
14	1.196	42.979
15	1.244	47.464

Table 7
Between-sample variability (standard deviation) of the similarity statistic value (GSS(t)) for each quantity of features (ranging from 5 to 15).

Feature quantity	Number of configurations	Mean GSS(t)	σ GSS(t)
5	10,620	20.864	13.585
6	10,415	23.849	15.112
7	10,044	25.372	16.681
8	9514	29.557	18.41
9	8784	32.392	19.642
10	7926	36.602	21.666
11	6877	39.826	23.653
12	6014	44.864	25.133
13	5039	47.81	27.192
14	4112	52.908	27.698
15	402	56.952	29.233

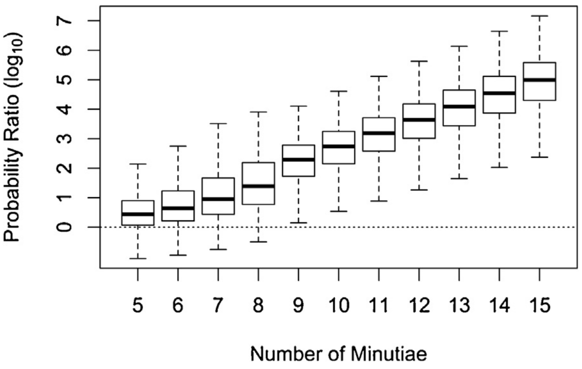

Fig. 7. Box plots illustrating the plausible range of ratio values that may be reasonably expected for each quantity of features based on GSS(t) values corresponding to 95% of the theoretical distribution modeling the mated source dataset (ranging from a probability of 0.025 to 0.975). The x-axis represents the number of features (ranging from 5 to 15). The y-axis represents the $\log_{10}$ ratio value.

ratio values are much lower than what experts might expect, these ratio values are highly conservative since: (1) the method does not take into account all aspects of the impression, such as pattern type, feature type, ridge counts, and other types of features considered by an expert, (2) the similarity statistic value provides a single dimensional summary of the similarity between two impressions and does not consider the prevalence of the specific arrangement of features under consideration within the population, (3) the empirical distributions of similarity statistic values were conditioned such that the non-mated distribution was biased towards higher similarity statistic values (in terms of randomly paired impressions) and the mated distribution was biased towards lower similarity statistic values, and (4) logistic mixture distributions were chosen to model the empirical distributions of similarity statistic values on the basis of their heavier tails thus providing more conservative estimates of probabilities in the extreme ends of the distributions compared to Gaussian mixture distributions.

Although the ratio values provide a measure of the significance (i.e. strength of an association) between two impressions, common practice by forensic experts is to conduct an experience-based judgment and classify an impression as originating from a specific individual (i.e. individualization decision) based on personal confidence and subjective observation. The accuracy of expert determinations of individualization has been evaluated by Ulery et al. [27] finding approximately 0.1% false individualization rate. In a subsequent study [28], Ulery et al. found that individualization determinations increase as the number of annotated features increase. Further, among all individualization decisions (n = 1,653), only 1% were based on mated comparisons containing less than 7 features and among all mated comparisons with 12 or more features, 98.4% resulted in individualization decision. Table 8 provides the percentage of individualization decisions for each number of features (ranging from 5 to 15) from Ref. [28]. Although a loose comparison, given the accuracy of individualization determinations from Ref. [27] and the breakdown of individualization decisions as it relates to the number of annotated features from Ref. [28], these data may provide some general context for understanding how the results from this method compare to performance metrics and individualization decision behaviors by experts in traditional practice. Interestingly, if we compare the inter-quartile range of ratio values for each quantity of features from Fig. 7 above to the individualization determinations in Table 8, we see that the inter-quartile ranges for 9 or more features exceeded a ratio of 10, which correspond to reasonably high specificity rates. Having discussed the comparisons between the ratio values of this method and experts' performance when making individualization decisions, caution should be exercised to ensure the probability estimates from this method are not incorrectly interpreted. The results provide the ratio of the estimated probabilities of a given similarity statistic value or more extreme among datasets of similarity statistic values from mated and non-mated comparisons. The results do not provide the probability of observing a *specific configuration* of features in the population or the probability that a *specific individual* is the source of an impression. Accordingly, although this method will provide an empirical foundation to the strength of an association between two impressions, determinations that specific individual is *the* source of an impression (i.e. individualization decisions) remain a subjective opinion by the expert.

Table 8
Percentage of individualization decisions by fingerprint experts on fingerprint images having different numbers of features (ranging from 5 to 15). Table values estimated from Fig. 3B in Ref. [28].

Feature quantity	% Individualization decisions
5	2
6	17
7	47
8	64
9	81
10	90
11	92
12	95
13	97
14	99
15	96

3.6.2. Method limitations

The major limitations of the method include: (1) the similarity statistic values are dependent upon the subjective detection and annotation of friction ridge skin features by the human expert. (2) The method is only able to consider what the expert annotates and is not able to evaluate the accuracy of feature annotations by the expert. (3) The method requires a minimum of five features and a maximum of fifteen features. The minimum of five features is due to the manner in which the similarity statistic is calculated. The maximum of fifteen features was a cutoff decision by the authors due to the computational impact of running the pairing algorithm on configurations containing higher numbers of features based on the current software implementation. For friction ridge skin impressions that contain more than fifteen features, only fifteen features can be encoded for statistical evaluation. This does not prevent the expert from making reference to the additional features available, but were not able to be encoded and evaluated by this version of the software application. (4) The weight functions are based on lateral distortions of friction ridge skin impressions on flat surfaces and may not capture all types of extreme distortions which may be encountered in practice, such as substrate, matrix, or photographic effects. (5) The method is not designed to evaluate all aspects of the impression, such as pattern type, feature type, ridge counts, and other types of features considered by an expert; thus, the quantitative results are artificially attenuated and conservative.

3.6.3. Considerations for policy & procedure

Taking into consideration the major limitations described above, general considerations for policy & procedure include: (1) the method should only be used *after* the expert has visually analyzed, detected, and annotated friction ridge skin features which are believed to correspond between two separate impressions of friction ridge skin. The method should not be used on impressions in which the analyst is able to visually exclude the two impressions as originating from the same source. (2) The method should be used in accordance with a set of strict policies and procedures to guard against potential cognitive biases in the analysis, detection, interpretation and annotation of friction ridge skin features as well as a quality assurance program to verify the accuracy of the annotated features. (3) The method should be used on digital images having a resolution of 500 pixels per inch or higher to ensure distance calculations are not impacted by lower resolution images.

Despite the limitations described above, this method provides several advantages which far outweigh the limitations. Most importantly, it provides fingerprint experts the capability to demonstrate the reliability of fingerprint evidence *for the case at hand* and ensure the evidence is reported with an empirically grounded basis. Further, having the ability to quantify the strength of fingerprint comparison, the evidence can be reported in a more transparent and standardized fashion with clearly defined criteria for conclusions and known error rate information. Supplemental Appendix V provides an example demonstrating the use of *FRStat*.

4. Conclusion

Over the years, the forensic science community has faced increasing amounts of criticism by scientific and legal commentators, challenging the validity and reliability of many forensic examination methods that rely on subjective interpretations by forensic practitioners. Among those concerns is the lack of an empirically demonstrable basis to evaluate and report the strength of the fingerprint evidence for a given case. In this paper, a method is presented which provides a statistical assessment of the strength of fingerprint evidence. The method measures the similarity between friction ridge skin impressions using details annotated by human experts to calculate a similarity statistic (i.e. score), which is then evaluated against databases of similarity statistic values derived from pairs of impressions made by mated (same) and non-mated (different) sources of friction ridge skin impressions relevant for forensic casework. The distributions of similarity statistic values were developed such that the non-mated data are biased to *higher* similarity statistic values and mated data are biased to *lower* similarity statistic values. For non-mated data, this was accomplished by conditioning on (1) the delta region of friction ridge skin which was determined to maximize the opportunities of observing higher similarity statistic values, and (2) any set of *n* features determined to be “optimally paired” from a larger set of *m* possible features with respect to a combinatorial optimization algorithm under any condition of rotation and translation such that the similarity statistic values are maximized. For mated data, the bias to lower values was accomplished by conditioning on lateral pressures and other distortions such that the similarity statistic values are minimized and ensuring that the distributions represent the full range of plausible similarity statistic values that could reasonably be observed in casework when impressions are subject to various distortions during deposition. The empirical distributions were statistically modeled and plausible estimates of population parameters were evaluated using the Kolmogorov–Smirnov (K–S) “goodness of fit” test. The K–S test was selected for this purpose on the basis of its ubiquitous use as a non-parametric test of the equality of continuous probability distributions. The strength of the fingerprint evidence is calculated as a ratio of the tail probabilities from the distributions of similarity statistic values of mated and non-mated impressions. The numerator is the left tail probability of a given similarity statistic value or *lower* among the distribution of values from mated sources. The denominator is the right tail probability of a given similarity statistic value or *higher* among the distribution of values from non-mated sources. Although similar in appearance, the ratio is not a true likelihood ratio or Bayes’ factor and therefore should not be used to estimate a posterior probability for a proposition.

The performance of the method was evaluated using a variety of different mated and non-mated datasets, including the most similar non-mated impressions from AFIS searches against a database of approximately 100 million other fingers. The results show strong performance characteristics. As the number of features increase, the magnitude of the ratio values increase as well as the ability to discriminate between mated and non-mated impressions, often with values supporting specificity rates greater than 99%. Despite the trend of increasing ratio values, there is still some overlap of the values between the different quantities of features. Consequently, similar to the findings in Refs. [17,22], these data demonstrate the importance of evaluating the strength of the fingerprint evidence based on the measurable attributes of the given comparison rather than relying on generalizations based solely on the number of features.

As with any method, there are limitations to consider. For example, this method relies on the features annotated by the expert but does not take into account all aspects of fingerprint evidence. As a result, the quantitative results for reported associations using this method (*FRStat*) will be artificially low. Despite the limitations, *FRStat* provides fingerprint experts the capability to demonstrate the reliability of fingerprint evidence *for the case at hand* and ensure the evidence is evaluated with an empirically grounded basis. Further, having the ability to quantify the strength of the fingerprint comparison, the evidence can be reported in a more transparent and standardized fashion with clearly defined criteria for conclusions and known error rate information.

Although various aspects of the method may be further optimized, the performance characteristics described are proposed as a sufficient basis to demonstrate the foundational validity of the method to perform within the scope of its intended purpose — as a means of providing a statistical measure of the strength of a given fingerprint comparison. Further optimizations which may improve upon the method’s performance are encouraged for future works.

Author contributions

Swofford: Lead in terms of overall concept, design and method development; software development and programming; data collection, analysis, and interpretation; article draft and revisions.

Koertner: Support in terms of suggestions and recommendations to concept, design, and method development; data collection; article review.

Zemp: Support in terms of evaluating and recommending optimizations to weight functions; input and recommendations to study design; article review.

Ausdemore: Support in terms of evaluating and recommending optimizations to weight functions; article review.

Liu: Support in terms of evaluating and recommending methods for data analysis and interpretation; article review.

Salyards: Support in terms of suggestions and recommendations to concept and design; article review.

Disclaimer

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the United States Department of the Army or United States Department of Defense.

Acknowledgments

The authors are deeply indebted to several individuals that have contributed to this effort. First and foremost, we are especially thankful to Dr. Hariharian Iyer, Dr. Karen Kafadar, and Dr. Hal Stern for their several iterations of review and helpful recommendations for improvements of the method. Additionally, we are thankful to Dr. Simone Gittelson, Dr. Matthew Bohn, Dr. Kate Schilling, Dr. McKay Allred, Dr. Tim Kalafut, Mr. Henry Maynard, and Ms. Jessica LeCroy for providing additional reviews and input following the development of the method. Further, we are thankful to the several latent print examiners at the Defense Forensic Science Center for contributing to the data collection, offering input from an end user's perspective during the development of the method, and being open and supportive of a new paradigm of friction ridge interpretation and reporting by the forensic fingerprint discipline. Last, but not least, we owe a sincere appreciation to Ms. Lauren Reed, Director of the U.S. Army Criminal Investigation Laboratory and rest of the leadership team at the Defense Forensic Science Center for encouraging an innovative and scientific working environment — a leadership culture that is integral to supporting the next generation of forensic scientists.

The authors express the gratitude to those individuals listed above for their willingness to provide input; however, these acknowledgements do not imply endorsement of these results or conclusions. Responsibility for any errors of fact or interpretation rests solely with the authors.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.forsciint.2018.03.043.

References

[1] S.A. Cole, The 'opinionization' of fingerprint evidence, BioSocieties 3 (1) (2008) 105–113.

[2] L. Haber, R.N. Haber, Scientific validation of fingerprint evidence under Daubert, Law Prob. Risk 7 (2) (2008) 87–109.

[3] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. Summary. Strengthening Forensic Science in the United States: A Path Forward; National Academy of Sciences, National Academies Press: Washington, DC, 2009.

[4] J. Koehler, M.J. Saks, Individualization claims in forensic science: still unwarranted, Brook. Law Rev. 75 (4) (2010) 1187–1208.

[5] M.J. Saks, Forensic identification from a faith-based science to a scientific science, Forensic Sci. Int. 201 (1–3) (2010) 14–17.

[6] S.A. Cole, Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the united states, Law Prob. Risk 13 (2) (2014) 117–150.

[7] REPORT TO THE PRESIDENT, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President President's Council of Advisors on Science and Technology. 2016

[8] Expert Working Group on Human Factors in Latent Print Analysis; The Latent Print Examination Process and Terminology. Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach, U.S. Department of Commerce, National Institute of Standards and Technology, 2012.

[9] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, D. Meuwly, et al., Computation of likelihood ratios in fingerprint identification for configurations of three minutiae, J. Forensic Sci. 51 (6) (2006) 1255–1266.

[10] Y. Zhu, S.C. Dass, A.K. Jain, Statistical Models for Assessing the Individuality of Fingerprints, MSU technical report MSU-CSE-06-25, Department of Computer Science, Michigan State University, 2006.

[11] N.M. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems: modelling within finger variability, Forensic Sci. Int. 167 (2–3) (2007) 189–195.

[12] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, J. Forensic Sci. 52 (1) (2007) 54–64.

[13] N.M. Egli, Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System, PhD thesis, University of Lausanne, 2009.

[14] C. Su, S.N. Srihari, Evaluation of rarity of fingerprints in forensics, Adv. Neural Inf. Process. Syst. 23 (2010) 1207–1215.

[15] C. Lim, S.C. Dass, Assessing fingerprint individuality using EPIC: a case study in the analysis of spatially dependent marked processes, Technometrics 53 (2) (2011) 112–124.

[16] H. Choi, A. Nagar, On the evidential value of fingerprints, International Joint Conference on Biometrics (IJCB) (2011) 1–8.

[17] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc. A 175 (2) (2012) 371–415.

[18] C. Neumann, I.W. Evett, J.E. Skerrett, I. Mateos-Garcia, Quantitative assessment of evidential weight for a fingerprint comparison. Part II: a generalisation to take account of the general pattern, Forensic Sci. Int. 214 (1–3) (2012) 195–199.

[19] J. Abraham, C. Champod, C. Lennard, C. Roux, Spatial analysis of corresponding fingerprint features from match and close non-match populations, Forensic Sci. Int. 230 (2013) 87–98.

[20] I. Alberink, A. de Jongh, C.M. Rodriguez, Fingermark evidence evaluation based on automated fingerprint identification system matching scores: the effect of different types of conditioning on likelihood ratios, J. Forensic Sci. 59 (1) (2014) 70–81.

[21] N.M. Egli Anthonioz, C. Champod, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems: modeling between finger variability, Forensic Sci. Int. 235 (2014) 86–101.

[22] C. Neumann, C. Champod, M. Yoo, T. Genessay, G. Langenburg, Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks, Forensic Sci. Int. 248 (2015) 154–171.

[23] A.J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, I. Alberink, Performance study of a score-based likelihood ratio system for forensic fingermark comparison, J. Forensic Sci. 62 (3) (2017) 626–640.

[24] H. Kuhn, The Hungarian method for the assignment problem, Naval Res. Logist. Q. 2 (1955) 83–97.

[25] M. Fagert, K. Morris, Quantifying the limits of fingerprint variability, Forensic Sci. Int. 254 (2015) 87–99.

[26] D. Garris, R.M. McCabe, NIST Special Database 27: Fingerprint Minutiae from Latent and Matching Tenprint Images. National Institute of Standards and Technology, Gaithersburg, MD. [CD-ROM] NISTIR 6534.

[27] B. Ulery, A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, PNAS 108 (19) (2011) 7733–7738.

[28] B. Ulery, A. Hicklin, M.A. Roberts, J. Buscaglia, Measuring what latent fingerprint examiners consider sufficient information for individualization determinations, PLoS One 9 (11) (2014) e110179.