Forensic Science International 320 (2021) 110703

Contents lists available at ScienceDirect

Forensic Science International

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

A method for measuring the quality of friction skin impression evidence: Method development and validation

H. Swofford[a,*], C. Champod[a], A. Koertner[b], H. Eldridge[a,c], M. Salyards[d]

[a] School of Criminal Justice, Forensic Science Institute, University of Lausanne, Switzerland
[b] U.S. Army Criminal Investigation Laboratory, Defense Forensic Science Center, USA
[c] RTI International, Inc., USA
[d] Compass Scientific, LLC, USA

ARTICLE INFO

Article history:
Received 25 September 2020
Received in revised form 13 January 2021
Accepted 19 January 2021
Available online 23 January 2021

Keywords:
Forensic science
Fingerprints
Quality metric
Probability

ABSTRACT

The forensic fingerprint community has faced increasing criticism by scientific and legal commentators, challenging the validity and reliability of fingerprint evidence due to the lack of an empirical basis to assess the quality of the friction ridge impressions. This paper presents a method, developed as a standalone software application, DFIQI ("Defense Fingerprint Image Quality Index"), which measures the clarity of friction ridge features (locally) and evaluates the quality of impressions (globally) across three different scales: value, complexity, and difficulty. Performance was evaluated using a variety of datasets, including datasets designed to simulate casework and a dataset derived directly from casework under operational conditions. The results show performance characteristics that are consistent with experts' subjective determinations. This method provides fingerprint experts: (1) a more rigorous approach by providing an empirical foundation to support their subjective determinations from the Analysis phase of the examination methodology, (2) a framework for organizations to establish transparent, measurable, and demonstrable criteria for Value determinations, (3) and a means of flagging impressions that are vulnerable to erroneous outcomes or inconsistency between experts (e.g., higher complexity and difficulty), and (4) a method for quantitatively summarizing the overall quality of impressions for ensuring representative distributions for samples used in research designs, proficiency testing and error rate testing, and other applications by forensic science stakeholders.

1. Introduction

Friction ridge examination is practiced by nearly every forensic laboratory throughout the world and is often relied upon as evidence that an individual touched an item or was present at the scene of a crime. The process for conducting friction ridge examination is described by the acronym ACE-V, which stands for "Analysis", "Comparison", "Evaluation", and "Verification". ACE-V has been described in the forensic literature as a means of comparative analysis of evidence since 1959 [1]. The process begins with the *analysis* of the latent print in which human analysts will visually observe and interpret friction ridge detail in a latent impression and determine if it is "suitable" or "of value" for comparison purposes. This determination is an experience-based judgment based on the quality and quantity of friction ridge detail discernible in the impression. If a latent print does not have "sufficient" detail to form a conclusion regarding the source of the impression, the impression is determined to be "not suitable" or "no value" and no comparison is made. If an impression is determined to be "of value", the analyst will perform a side-by-side *comparison* of the friction ridge detail between the latent print and the known prints from an individual. During comparison, and ultimately thereafter, the analyst will *evaluate* the similarities and differences of the friction ridge detail between the two impressions and form a conclusion regarding the source of the impression. *Verification* occurs when another qualified analyst repeats the observations and forms the same conclusion.

Within the ACE-V process, the "analysis" of the friction ridge skin detail is one of the most critical tasks of the examination as it establishes whether, and to what extent, the impression bears sufficiently discernible features that can be used for examination. More specifically, during the "analysis", the analyst is particularly concerned with identifying reproducible and discriminating attributes of the friction ridge detail which may be used for comparison and evaluation against a known source impression. The ability for the analyst to reliably detect these attributes

* Corresponding author.
E-mail address: henry.swofford@unil.ch (H. Swofford).

http://dx.doi.org/10.1016/j.forsciint.2021.110703

depends heavily on the clarity of the impression. Generally, as the clarity of an impression increases, analysts' have more confidence in their interpretation of the location, orientation, type, and spatial arrangement of features. Additionally, as the number of interpretable features increases, the discriminating strength of the impression as a whole is considered to increase as well. Once the features have been detected, the analyst will assess the overall quality of the impression and make a determination of the "suitability" or "value" for further comparison and evaluation [2]. This determination is not based on an empirical standard; rather, it is a subjective determination made by the analyst on a case-by-case basis and depends on whether the analyst believes the quality of the impression is sufficiently reproducible and selective to be compared to a known source and render a particular conclusion regarding the potential source of the impression. Consequently, assessments made during friction ridge examinations are susceptible to variation from one analyst to another (inter-analyst) as well as by the same analyst from one examination to another (intra-analyst). When considering borderline impressions which contain marginal quality or quantity of features, these variations often result in differences in the *analysis* conclusion. In the broad spectrum, however, while the lack of empirical standards and measurements do not necessarily imply the practice as a whole is unreliable or fraught with error, it does raise questions as to how reliable the evidence is for the case at hand. Thus, there is a critical need for the friction ridge community to move towards integrating tools to quantitatively assess the clarity and quality of friction ridge impression details to standardize and provide an empirical warrant for analysts' claims [1,3–5].

Over the years, there have been several notable efforts by researchers in which quantitative tools were introduced for assessing the quality of friction ridge impressions [6–19]. The majority of these efforts can be classified as suitability prediction models, which provide a predictive estimate of whether the impression is suitable for some intended purpose or utility, such as suitability for identification or exclusion purposes during manual comparisons or, more often, for predictions of search performance using automated fingerprint identification systems (AFIS). Early models are described by Alonso-Fernandez et al. (2005) and all focus on calculating quality as a means of predicting AFIS feature extraction or matcher performance. Most of the early methods entailed a variety of different image processing techniques, such as measuring ridge frequency, ridge thickness, and ridge to valley thickness ratio, using Gabor filters to increase contrast, measuring pixel intensity differences, two-dimensional Discrete Fourier Transform (DFT), and neural network classifiers to classify local regions as "good" or "bad" quality [6]. Alonso-Fernandez et al. note that all of the various methods tend to behave similarly to one another except for the method based on neural network classifiers, likely due to the low number of quality labels used for training, and propose the concept of integrating the various algorithms into a quality-based multimodal authentication system for future works.

In 2007, Nill developed Image Quality of Fingerprint (IQF) as a freeware software application designed to predict AFIS matching performance, alert operators to poor quality enrollment of known source standards or aid in performance assessments of capture devices [7]. The approach developed by Nill relies on the two-dimensional, spatial frequency power spectrum of the digital fingerprint image to produce a global assessment of quality [7]. In 2008, Fronthaler et al. studied the orientation tensor of fingerprint images to quantify signal impairments like noise, lack of structure, and blur with the help of symmetry descriptors when combining multiple AFIS matchers for improved matching performance [8].

In 2011, Hicklin et al. [9] attempted to understand how human latent fingerprint analysts assess fingerprint quality by surveying eighty-six latent print examiners from federal, state, local, international, and private sector laboratories using overlapping subsets of 1090 latent and exemplar fingerprint images to identify key features that will guide the development of automated quality metric algorithms in future works [9]. Up to this point, nearly every other method was focused entirely on optimizing AFIS matching performance or developing quality metrics to predict match performance rather than attempting to understand what was considered by human analysts during manual examinations. From the survey, Hicklin et al. note there is general concurrence of human assessments of local and overall image quality, but enough variation between examiners to result in differing conclusions and demonstrate the need to provide uniform definitions of quality and automated assessment tools to standardize the practice [9].

In 2012, two additional methods were proposed: both focused on optimizing or predicting AFIS match performance. While earlier methods tended to focus on biometric enrollments and known source impressions, these were geared more towards latent fingerprint impressions. Murch et al. (2012) proposed a method for automated feature extraction to improve the performance of AFIS searches of latent fingerprint impressions using image segmentation to differentiate the foreground impression from background noise [10]. Yoon et al. (work presented in 2012, but published in 2015) proposed a method for assessing latent fingerprint image quality using the product of the average ridge clarity bounded within the convex hull enclosing all annotated minutiae and total number of minutiae [11]. The calculation of average ridge clarity involved the application of two-dimensional Fourier analysis to a pre-processed contrast enhanced image. Although Yoon et al. was focused specifically on latent impressions, the quality algorithm was still geared towards predicting AFIS matcher performance and thus not necessarily tailored to attributes considered during human examinations [11,12].

In 2013, three additional approaches were introduced, which begin to steer focus towards latent fingerprint image clarity relevant during human examinations compared to prior methods. Hicklin et al. (2013) developed Latent Quality Assessment Software (LQAS), which applies a variety of image processing algorithms to assess the clarity of friction ridges in localized regions [13,14] (LQAS [13] was later enhanced and combined with Universal Latent Workstation (ULW). Within ULW, it is referred to as LQMetric. Details related to LQMetric development are provided by Kalka et al. 2020 [14]). Based on the clarity assessment, the software then applies a color-coded clarity map which corresponds to the color codes within the American National Standards Institute/National Institute of Standards and Technology (ANSI/NIST) 2011 standard "Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information" [15] for simple interpretation and a standardized framework for documentation [13,14]. Sankaran et al. (2013) propose a method which assesses ridge clarity and quality [16]. The former (Hicklin et al.) refers to the visual discernibility of the features irrespective of the presence or absence of features and the latter (Sankaran et al.) refers to the quantity and number of features present in a given local region (i.e. a predictor of AFIS matching performance). The local ridge clarity assessment is based on average eigenvalues from decomposed structure tensors following image smoothing using a Gaussian filter [16]. A local clarity map is generated as a result of the clarity assessment similar to that of Hicklin et al. (2013) [9]. The ridge quality assessment is calculated as the kurtosis of the weighted average histogram based on the local clarity map described previously along with the number of features present within a local region [16]. Pulsifer et al. (2013) propose a method for calculating overall quality based on a semi-automated assessment of the local clarity maps generated from LQAS developed by Hicklin et al. (2013) [13,14] to produce an alternative way of calculating the overall quality of the impression [17].

In 2014, Kellman et al. proposed a number of quantitative measures of image characteristics related to image quality metrics, such as intensity and contrast information, as well as measures of information quantity, such as total fingerprint area, to calculate image quality and predict analyst performance and perceived difficulty during comparisons by human analysts [18]. The work by Kellman et al. indicates a shift towards establishing quality metrics geared towards predicting human analyst performance rather than tailored specifically to predicting AFIS match performance. More recently in 2018, with a similar intent as Hicklin et al. [13,14] and Kellman et al. [18], Chugh et al. proposed a crowdsourcing framework to understand the underlying bases of suitability determinations by fingerprint analysts and use it to develop an automated means of predicting suitability determinations [19].

While there have been a number of different models proposed over the years, the majority of them are geared entirely towards optimizing or predicting AFIS match performance rather than focused on assessing local ridge clarity (discernibility of feature data) and predicting human performance using image quality attributes considered by human analysts during manual comparisons. Consequently, these types of predictive models are often based on the aggregate of qualitative and quantitative attributes of the entire impression to provide a single estimate of utility or quality. These approaches often lack transparency and often do not necessarily correspond to the same features considered by human analysts during traditional examinations. The motivation behind this focus is largely driven by industry desires to optimize the performance of AFIS in a "lights-out" environment. Indeed, this focus is important for the broader biometric industry; however, the narrow focus on AFIS platforms leaves a gap as it relates to manual examination and interpretation processes by human analysts in the traditional forensic setting. Thus, the need remains for the development and implementation of tools capable of quantitatively assessing the clarity of friction ridge detail in a transparent and objective manner within in a simple, accessible, and user-friendly software application that can be easily integrated into friction ridge examination practices. Such a tool would offer significant improvements to traditional practices and permit laboratories to establish standardized suitability criterion and provide empirical substantiation to analysts' opinions.

This paper presents a method, developed as a stand-alone software application, DFIQI ("Defense Fingerprint Image Quality Index"), designed to measure the clarity of friction ridge impression minutiae and provides a quantitative assessment of the quality of an impression for comparison and evaluation purposes. Although this method builds upon general approaches described earlier and considers well established means of assessing image clarity, it provides a simple and novel approach for quantifying the quality of friction ridge impressions. Further, having been developed as an automated stand-alone software application, this method is accessible to the forensic community[1] thereby providing the capability for laboratories to ensure the quality of friction ridge details are sufficient to permit reliable interpretations and move toward standardizing and improving traditional practices. In the sections that follow, this paper provides a brief overview of the 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 discussed as well as implications for future integrations with other tools to strengthen the foundations of friction ridge examination in general.

2. Materials & methods

2.1. Background

In general terms, the method assesses the clarity of friction ridges in localized "regions of interest" (ROIs) immediately surrounding the x,y location of features identified in the impression. Features can be identified by manual annotation or using automated feature extraction applications (followed by human-expert verification). Each region of interest is assessed using five variables (described below) consisting of various measures of friction ridge image clarity and quality. The five variables were selected by the authors based on domain expertise, reduced mathematical complexity, and algorithmic transparency. The output of each variable measured is normalized by a scoring function and combined to create a single quantitative value representing the clarity and quality of the friction ridges within the localized ROI. Each local ROI score is then combined to a single quantitative value representing the quality of the ROIs combined across the entire impression, which accounts for both the quality and quantity of detail in the impression.

Once the x,y coordinates are identified for the features in the impression (e.g. by an analyst marking the location), the application creates an inverted 8-bit digital grey-scale copy of the image on which all subsequent digital processing is performed. For each feature, a 2.54 mm × 2.54 mm (i.e. 0.1-inch × 0.1-inch) square ROI, centered on the location of the feature, is applied to the image and cropped (as a copy). The size of the ROI was selected to ensure it is small enough to represent a local region of the impression immediately surrounding a feature, but large enough to cover multiple ridges and enable a meaningful discrete Fourier transform related to the spatial frequency variable (described below); however, it was not subject to formal parameter optimization methods Each ROI is large enough to generally contain between four and seven ridges, depending on the width and orientation of the ridges. The five variable measures are taken from the cropped ROI to calculate the clarity and quality of the ridge detail immediately surrounding each individual feature in the impression.

Before the variable values are calculated, each ROI is split into two separate images to separate the "ridges" from the "furrows" (or more appropriately referred to as "signal" from "background") by applying adaptive mean thresholding to the pixel intensity values with a local neighborhood radius of 0.38 mm. The 0.38 mm radius was selected based on ad hoc testing and not subject to formal parameter optimization methods. Unlike simple thresholding methods, adaptive thresholding determines the threshold for a pixel based on a small region around it resulting in different thresholds for different regions of the same image. This generally provides greater segmentation accuracy as illumination conditions may vary throughout an image. Fig. 1 illustrates the results of applying adaptive thresholding to a cropped ROI.

2.2. Variables

Using the cropped and segmented ROIs, the following measures of clarity and quality are calculated:

- Signal Percent Pixels Per Grid (S3PG): This variable calculates the percentage of pixels that have been segmented as "signal" compared to the total number of pixels available in the ROI. For a high-quality impression of friction ridges, an approximate value of 50, accounting for approximately 50% of total pixels segmented as "signal," is expected. As S3PG values deviate from the expected output of 50 in one direction or another, it suggests

[1] The software application can be accessed at: https://doi.org/10.5281/zenodo.4426344.

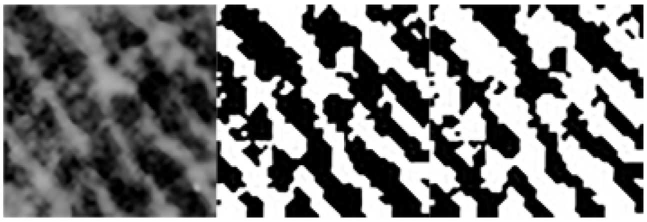

Fig. 1. The image on the left represents the original ROI (the darker color pixels correspond to friction ridges). The image in the center represents the binary mask of the segmented ROI for which the black areas correspond to pixels thresholded as “signal”. The image on the right represents the binary mask of the segmented ROI for which the black areas correspond to pixels thresholded as “background” (i.e. the image on the right is the inverse of the image in the center). NOTE: Actual size of images are 2.54 mm × 2.54 mm. Images are enlarged and pixels interpolated for illustration.

there are distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail.

- Bimodal Separation (BS): This variable calculates an index value summarizing the extent to which two histograms of pixel intensity values are separated from one another. Using the pixel intensity values of those segmented as “signal” and those segmented as “background”, the index is calculated using the formula below. As the difference between the mean values increase and the standard deviations decrease between the segmented images, the value of the bimodal separation index increases, which indicates greater contrast between pixels classified as “signal” versus “background”. On the other hand, as the difference between the mean values decrease and the standard deviations increase between the segmented images, the value of the bimodal separation index decreases, which indicates lower contrast and may interfere with accurate detection of friction ridge detail. The bimodal separation variable is calculated using the formula in equation 1.

$$\frac{x = \bar{S} - \bar{B}}{2(\sigma_S + \sigma_B)}$$

Equation 1: The formula for which the bimodal separation variable is calculated for each ROI.

- Acutance (ACUT): This variable calculates an index value summarizing the natural log of the mean acutance across the entire ROI and is applied to the non-segmented copy of the image. Acutance is described as the physical characteristics that underlay the subjective perception of “sharpness” in an image. In general terms, the acutance is calculated as the mean squared difference between a center pixel and its eight neighboring pixels in a 3 × 3 window iteratively calculated across an entire image. As the difference of pixel intensities increase, the perceived sharpness of the objects represented in the image also increase. This perceived increase of sharpness is represented by a higher acutance index value. As the acutance index value decreases, the perceived sharpness of the image decreases resulting in lower contrast which may interfere with accurate detection of friction ridge detail. The acutance variable calculation routine is illustrated in Figs. 2a and b and stated in equation 2 (adapted from Choong et al. [20]).

$$x = ln\left(\frac{\sum\left(\sum_{n=1}^{8}(I_c - I_n)^2\right)}{8(p-2)^2}\right)$$

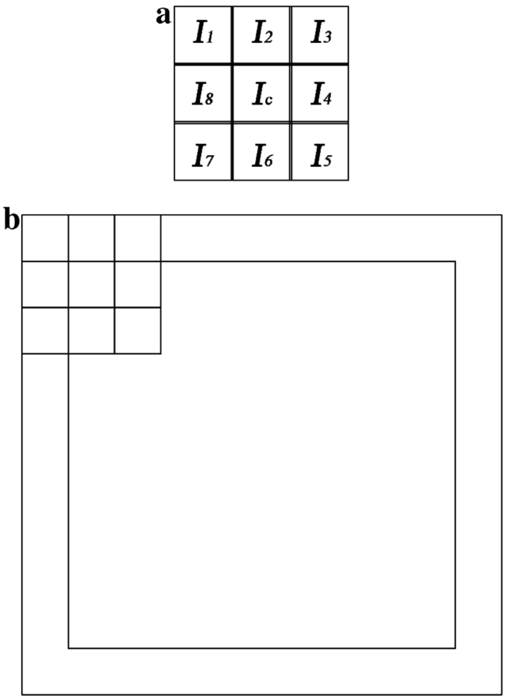

Fig. 2. a The 3 × 3 window representing a neighborhood of pixel values (the center pixel surrounding by its 8 contiguous neighbors). 2b The external box is a simplistic illustration representing the entire ROI containing p × p pixels (e.g. for an image resolution of 500 pixels per inch, p = 50 pixels. The inner box is a simplistic illustration representing the inner window of p-1 × p-1 pixels for the ROI in which every pixel serves as the center pixel of the scrolling 3 × 3 pixel window. The 3 × 3 window at the top left is a simplistic illustration of the 3 × 3 window represented in Fig. 2a.

Equation 2: The formula for which acutance is calculated for each ROI.

- Mean Object Width (MOW): This variable calculates the mean width of objects segmented as “signal” in the ROI. The term “objects” refers to a set of contiguously thresholded pixels within the “signal”. The width of each object is calculated by fitting an ellipse and measuring the width of the minor axis. In the context of friction ridge impressions having perfect quality, those pixels thresholded as “signal” would correspond to separate and distinct “objects” in the image, representing separate friction ridges having nearly uniform and predictable widths. As the values for the mean object width deviate from the expected width of friction ridges in one direction or another, it suggests there are distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail. The manner in which the mean object width variable is calculated is illustrated in Figs. 3a and b.
- Spatial Frequency (SF): This variable calculates the spatial frequency of the ridges in the non-thresholded ROI using the two-dimensional discrete Fourier transform. For high-quality impressions of friction ridges, the ridges have been shown to have a predictable spatial frequency of approximately 2.1 ridges per millimeter for males and 2.4 ridges per millimeter for females [21] (combined mean of approximately 2.25 ridges per millimeter). As the spatial frequency values deviate from the expected output of approximately 2.25 ridges per millimeter in one direction or another, it suggests there are distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail. The two-dimensional discrete Fourier transform for a sample ROI is shown in Figs. 4a and b to illustrate how the system calculates this variable.

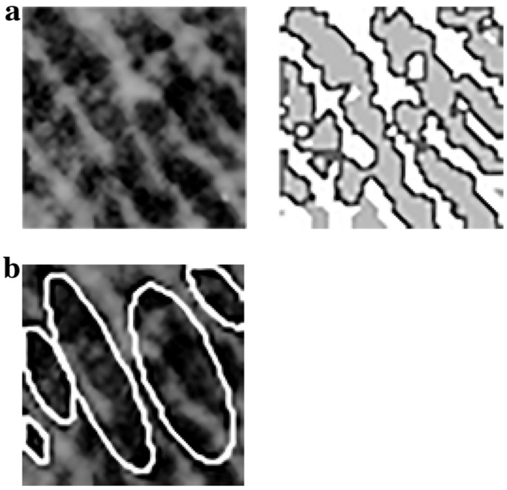

Fig. 3. a The image on the left represents the original ROI (the darker color pixels correspond to friction ridges). The image on the right represents the mask of the segmented ROI for which the light grey areas correspond to pixels thresholded as “signal”. The dark grey borders represent the borders around groups of contigous pixels represenging the various “objects” in the impression. NOTE: Actual size of images are 2.54 mm × 2.54 mm. Images are enlarged and pixels interpolated for illustration. 3b An ellipse is fit to each distinct “object” in the image (ellipses overlaid on the origial image of friction ridges). The object width is calculated by measuing the width of the minor axis of each ellipse. In this example, two ridges appear connected together due to smudging in the impression resulting in a larger mean object width for the ROI; thus indicating the presence of distorting factors which may interfere with accurate interpretation of friction ridge detail. NOTE: Actual size of image is 2.54 mm × 2.54 mm. Images are enlarged and pixels interpolated for illustration.

2.3. Local quality score

As described earlier, the five variable values are calculated for each ROI in an image. Let x_*i* denote the i^{th} variable, with *i* = 1 . . . 4 corresponding to S3PG, BS, MOW, and SF, respectively, and x_5 denote ACUT. The raw variable values for S3PG, BS, MOW, and SF are each normalized and scored using a symmetrical distribution scaled between 0 and 1 as provided by *f*(x) in equation 3 below. A symmetrical distribution is used for these variables since a value that deviates too far on either side of the expected value indicates the presence of distorting artifacts in the ROI that may interfere with accurate detection of friction ridge detail.

$$f(x) = e^{\frac{-(x-\hat{\mu})^2}{2\hat{\sigma}^2}}$$

Equation 3: Scoring function for raw variable values S3PG, Bimodal Separation, Mean Object Width, and Spatial Frequency. The scoring function provides a maximum value of 1 if the raw variable value is equal to the expected value ($\hat{\mu}$) (i.e. location parameter). As the raw variable value deviates from the expected value on either side, the score is reduced and trends toward 0 at a rate determined by the scale parameter ($\hat{\sigma}$).

The raw variable value for ACUT is provided by a simple logistic cumulative distribution (scaled between 0 and 1) as provided by g (x) in equation 4 below. A cumulative distribution is used for this variable since only values that are less than the expected value indicates the presence of lower sharpness and contrast of ridges in the ROI that may interfere with accurate detection of friction ridge detail.

$$g(x) = \frac{1}{1 + e^{-\frac{x-\hat{\mu}}{\hat{s}}}}$$

Equation 4: Scoring function for the raw variable value Acutance. The scoring function provides a maximum value of 1 if the raw variable value is equal to the expected value ($\hat{\mu}$) (i.e. location parameter). As the raw variable value deviates from the expected value (lower acutance values), the score is reduced and trends toward 0 at a rate determined by the scale parameter ($\hat{s}$). The Acutance is scored on a cumulative distribution since lower quality is only manifest with lower acutance values.

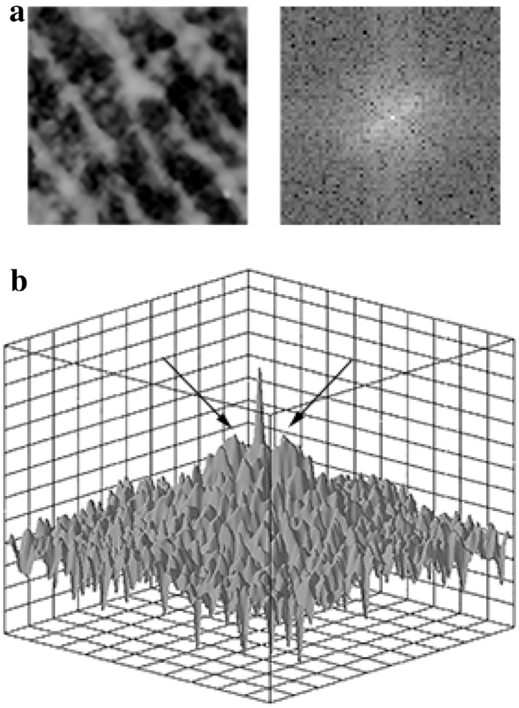

Fig. 4. a The image on the left represents the original ROI (the darker color pixels correspond to friction ridges). The image on the right represents the discrete two dimensional Fourer transform of the image on the left. NOTE: Actual size of images are 2.54 mm × 2.54 mm. Images are enlarged and pixels interpolated for illustration. 4b A three-dimensional representation of the pixel intensity values of the discrete two-dimensional Fourier transform image in Fig. 4a. The vertical axis represents the pixel intensity values corresponding to lighter colored pixels in the Fourier transform image in Fig. 4a. The tallest point on the vertical axis in the middle represents the DC-value for the image. The second two tallest points on each side of the DC-value represent the spatial frequency of the ridges in the image (indicated by the arrows).

The input parameters for the scoring functions for each variable consist of the location parameter (i.e. mean raw variable value) and scale parameter (e.g. standard deviation of the raw variable value) empirically estimated from a reference dataset. The reference dataset consisted of 1373 ROIs selected from pristine quality exemplar impressions. The impressions in this dataset were deposited under controlled conditions using a mixture of traditional ink and Livescan device. Table 1 provides the input parameters for the scoring function related to each variable.

The five normalized variable values are then combined to create a mean univariate quantitative score summarizing the clarity and quality of the feature represented in the ROI on a scale from 0 to 1 (higher values indicate higher clarity and quality of the friction ridges in the ROI). This ROI score (i.e. Local Quality Score, or “LQS”) provides a proxy estimate of the quality of the feature contained within the ROI on the basis of the clarity of the friction ridge detail immediately surrounding it. The LQS is calculated using the formula below:

$$LQS = \frac{\sum_{i=1}^{5} f(x)_i}{5}$$

Equation 5: Local Quality Score (LQS) function – calculated for each ROI as the mean of the normalized variable scores, where *f(x)*$_i$

Table 1
Input parameters for the scoring functions for each variable.

Variable	Location Parameter ($\hat{\mu}$)	Scale Parameter ($\hat{\sigma}$ or $\hat{s}$)
S3PG	51.408	4.134
Bimodal Separation	0.843	0.147
Acutance	6.869	0.532
Mean Object Width	1.383	0.391
Spatial Frequency	2.078	0.397

is the normalized variable score for *i*-th function in the set containing the normalized variable scores for all 5 variables.

The LQS value is then used as a basis to categorize and color-code the quality of the feature as a graphical output to the user (e.g. high, medium, and low) in terms that align with subjective determinations by human analysts, such as that proposed by Langenburg & Champod (2011) [22]. Features color-coded as green generally indicate areas of high quality, features color-coded as yellow generally indicate areas of medium quality, and features color-coded as red generally indicate areas of low quality.

2.4. Global quality scores

Three different Global Quality Score (GQS) values are calculated, each of which represent a summary of the overall quality of the impression for different purposes: to predict analysts' determinations of "value", "complexity", and "difficulty" as proposed by Eldridge et al. (2020) [23] and as part of the Analysis phase of the examination methodology. For all three prediction categories (value, complexity, and difficulty), the GQS is calculated as a multinomial combination of two variables: (a) $\mathrm{LQS}_{\mathrm{sum}}$ – the sum of all LQS values, and (b) *n*FEAT – the total quantity of features identified in the impression. Taken together, these provide explainable quantitative representations and variables of the overall quality of the impression for manual comparison purposes.

The multinomial coefficients for each outcome class (value, complexity, and difficulty) were derived using a multinomial regression model provided by the *nnet* package in R [24] against a training/test-dataset of feature measurements from impressions for which latent print examiners previously analyzed and categorized based on their "value", "complexity", and "difficulty" for comparison. The multinomial model was selected after testing a range of machine learning techniques with the variables $\mathrm{LQS}_{\mathrm{sum}}$ and *n*FEAT (naïve based classifier, tree-based classifiers, discriminant analysis techniques, neural networks and support vector machines). Overall, the multinomial regression offers a competitive accuracy while maintaining easy explainability (see Supplemental Appendix for raw model diagnostics and uncertainty values). The training-dataset was derived as a random 50/50 training-test split obtained from the full dataset provided by Eldridge et al. (2020) [23]. The full dataset consisted of a total of 3241 determinations made by 116 analysts rendering "value", "complexity", and "difficulty" decisions for each image they viewed from a set of 100 different latent print impressions – each participant was provided a set of approximately 30 impressions to analyze, resulting in each impression being analyzed by between 26 and 41 different analysts. The impressions were generated during the course of normal casework at a large metropolitan police laboratory using standard powder processing and lifting techniques. All participants were practicing latent print examiners recruited by several outreach methods, such as email distribution lists, presentations given at professional educational meetings, and professional contacts. Half of this dataset was used to train the models (1621 responses) and the other half of this dataset was used to test the models (1620 responses) described by GQS Test-Dataset 1 below. It should be noted that the model was trained and tested using the results of each examiner's individual observations and judgments of the impressions rather than artificially combining them. Ground truth for these types of judgments is non-existent. Although consensus judgments could be declared as a surrogate to ground truth for each *image*, the examiners' observations for which their subjective judgments are based are variable which would require artificially aggregating examiners' judgments and disconnecting their individual observations from their individual judgments. As a result, the authors believe a model that is trained using individual examiners' observations and resulting judgments is appropriate in this context. The output of the model, effectively, then reflects a proxy consensus of examiners' judgments for a given input in a specific case impression. Table 2a, Table 2b, and Table 2c provide the coefficients related to the multinomial models from the training partition (see Supplemental Appendix for raw model diagnostics and uncertainty values on the coefficients). Each multinomial model provides a probability of class inclusion (ranging from 0.00 to 1.00) for each outcome class (e.g., for the Value determination the three outcome classes are no-value, value for exclusion only, and value for identification).

Table 2a
Multinomial coefficients for each outcome class probability (no-value, value for exclusion only, value for identification) of the "value" determination. Note: In Eldridge et al. [23], participants were given the following response choices: "no value", "some probative or investigative value but insufficient for identification or exclusion", "value for exclusion only", "value for identification only", "value for both identification and exclusion". Responses of "some probative or investigative value but insufficient for identification or exclusion" were categorized as "value for exclusion" to represent the middle bin of the value spectrum. Responses of "value for both identification and exclusion" and "value for identification only" were categorized as "value for identification".

"Value" coefficients	Intercept	$\mathrm{LQS}_{\mathrm{sum}}$	*n*FEAT
No Value	0.000	0.000	0.000
Value for Exclusion	−1.736	−0.051	0.277
Value for Identification	−6.042	0.495	0.726

Table 2b
Multinomial coefficients for each outcome class probability (highly complex, complex, non-complex) of the "complexity" determination. Note: In Eldridge et al. [23], participants were given the following response choices: "no value", "of value, complex", "of value, non-complex; requiring documentation", and "of value, non-complex; self-evident". Responses of "of value, non-complex; requiring documentation" and "of value, non-complex; self evident" were both categorized as "non-complex". Responses of "no value" were re-labeled "highly complex" to represent the extreme end of the complexity spectrum.

"Complexity" coefficients	Intercept	$\mathrm{LQS}_{\mathrm{sum}}$	*n*FEAT
Highly Complex	3.325	−0.100	−0.459
Complex	0.000	0.000	0.000
Non-Complex	−1.781	0.741	−0.025

Table 2c
Multinomial coefficients for each outcome class probability (high, medium, low) of the "difficulty" determination.

"Difficulty" coefficients	Intercept	$\mathrm{LQS}_{\mathrm{sum}}$	*n*FEAT
High	0.000	0.000	0.000
Medium	−1.896	0.289	0.125
Low	−3.071	0.965	−0.004

Recognizing each class represents an outcome along a spectrum (e.g. for the "value" determination: No Value represents the left-most extreme and Value for Identification represents the right-most extreme) and the sum across all classes equals 1.00, we can combine to create single values representing the GQS for each determination (value, complexity, difficulty) by subtracting the probability of class inclusion representing the left-most extreme from the probability of class inclusion representing the right-most extreme to produce a number ranging from −1.00 to 1.00, where higher values indicate stronger support for "value for identification", "non-complex", and "low difficulty" and lower values indicate stronger support for "no value", "highly complex", and "high difficulty". The GQS values for each determination are calculated using the formulae below:

$$Value_{GQS} = p(VID) - p(NV)$$

Equation 6: GQS function for Value determination – calculated by subtracting the probability of class inclusion for No Value outcome (NV) from the probability of class inclusion for Value for Identification outcome (VID). Values near -1.00 indicate no-value determinations, values near 1.00 indicate value for identification determinations, and values near 0 indicate value for exclusion only determinations (or inconclusive determinations in lieu of value for exclusion only).

$$Complexity_{GQS} = p(NC) - p(HC)$$

Equation 7: GQS function for Complexity determination – calculated by subtracting the probability of class inclusion for Highly Complex outcome (HC) from the probability of class inclusion for Non-Complex outcome (NC). Values near -1.00 indicate no-value determinations, values near 1.00 indicate non-complex determinations, and values near 0 indicate complex determinations.

$$Difficulty_{GQS} = p(L) - p(H)$$

Equation 8: GQS function for Difficulty determination – calculated by subtracting the probability of class inclusion for High difficulty outcome (H) from the probability of class inclusion for Low difficulty outcome (L). Values near -1.00 indicate high difficulty determinations, values near 1.00 indicate low difficulty determinations, and values near 0 indicate medium difficulty determinations.

ROC curves will be used to illustrate model performance. The associated areas under the curve (AUC), and confidence intervals have been computed taking advantage of the pROC package [25].

2.5. Method performance

The performance of the method was evaluated in different conditions capturing performance characteristics both locally and globally. The local performance characteristics were evaluated in terms of (i) the ability of the LQS value to accurately distinguish between the extreme conditions of "good" and "bad" quality ROIs and (ii) the ability of the LQS value to predict analysts' subjective determinations of feature quality according to the GYRO annotation scheme proposed by Langenburg & Champod (2011) [22]. The global performance characteristics were evaluated in terms of the ability of the GQS values to distinguish between analysts' subjective determinations of "value", "complexity", and "difficulty" from test-datasets of feature measurements from impressions for which latent print examiners previously analyzed and categorized.

2.5.1. Local performance characteristics

The local performance characteristics were evaluated to understand the behavior of the system as the clarity of friction ridge detail within the ROIs change. This was evaluated using measurements from two different test-datasets:

(1) LQS-Test-Dataset-1: This dataset consists of 867 "good" quality ROIs selected from high quality regions of exemplar friction ridge impressions and a dataset of 3816 "bad" quality ROIs selected from low quality regions of latent lift cards submitted under operational conditions as attempts to lift latent images from a variety of different surfaces during normal forensic casework. The "bad" quality ROIs represented impressions with excessive smudging, indiscernible ridge detail, background interference and artifacts, and related factors impacting reliable interpretation of friction ridges, yet still having artifacts present bearing reasonable contrast and clarity but lacking morphological representations of friction ridge detail. The purpose of this dataset is to evaluate how well the LQS values distinguish between the extremes of "good" and "bad" quality ROIs collected under operational conditions.

(2) LQS-Test-Dataset-2: This dataset consists of 4480 ROIs containing features annotated as "high confidence" (i.e. green) and 920 ROIs containing features annotated as "medium confidence" (i.e. yellow) by practicing latent print examiners according to the GYRO annotation scheme proposed by Langenburg & Champod (2011) [22] across 300 different impressions deposited using normal handling of objects and developed using common latent print processing methods representative of typical casework. This dataset was obtained from John & Swofford (2020) [26]. The purpose of this dataset is to evaluate how well the LQS color-coded quality categories correspond to fingerprint experts' subjective assessment of feature confidence ("high" confidence vs. "medium" confidence).

2.5.2. Global performance characteristics

The global performance characteristics were evaluated to understand the ability of the method to predict human analysts' subjective assessments of whether impressions are considered "suitable" or "of value" as well as assessments of "complexity" and "difficulty" for comparison purposes. These were evaluated using measurements from two different test-datasets:

(1) GQS-Test-Dataset-1: This dataset represents the test fraction derived as a random 50/50 training-test split of the full dataset obtained from Eldridge et al. (2020) [23]. The full dataset consisted of a total of 3241 analysts' determinations of "value", "complexity", and "difficulty" and documentation of features across a set of 100 different latent print impressions by approximately 116 different participants – each participant was provided a set of approximately 30 impressions to analyze resulting in each impression being analyzed by between 26 and 41 different analysts. The impressions were generated during the course of normal casework at a large metropolitan police laboratory using a variety of standard processing techniques. All participants were practicing latent print examiners recruited by several outreach methods, such as email distribution lists, presentations given at professional educational meetings, and professional contacts. Half of this dataset was used to train the models (1621 responses) and the other half of this dataset was used to test the models (1620 responses). The purpose of this dataset is to evaluate how well the GQS values correspond to subjective determinations of value, complexity, and difficulty when examined under pseudo-operational conditions.

(2) GQS-Dataset-2: This dataset consists of 605 latent impressions collected from casework during the course of routine

operations by fingerprint experts in a federal crime laboratory in the United States for which fingerprint experts conducted examinations and identified the impressions to corresponding reference standards. All impressions in this dataset were determined to be "suitable" or "of value" for identification purposes. The purpose of this dataset is to evaluate the distribution of GQS values and implications thereof when applied to impressions derived from actual casework and assessed under normal operational conditions.

3. Results & discussion

3.1. Local performance characteristics

The local performance characteristics were evaluated on the basis of how well the LQS values were able to distinguish between the extremes of "good" and "bad" quality ROIs collected under operational conditions using LQS-Test-Dataset-1 and how well the LQS color-coded quality categories correspond to fingerprint experts' subjective assessment of feature confidence ("high" confidence vs. "medium" confidence) using LQS-Test-Dataset-2. Figs. 5a and b illustrates the degree of separation observed between the extremes of "Good" and "Bad" quality ROIs using the LQS value.

From Figs. 5a and b, we see remarkable separation between the two extremes of "Good" and "Bad" quality ROIs. Although these results may be expected for this dataset since they represent the extreme ends of the spectrum, it establishes an important baseline which validates the relevance of the input variables which comprise the LQS value and its ability to distinguish between high-quality friction ridge impressions and low-quality non-friction ridge artifacts. Further, from these data, we can establish thresholds for distinguishing between "high", "medium", and "low" color-coded bins categorizing ROI quality as an overlay output to the user. For this purpose, LQS values between 0.35 and 1.00 are color-coded green (high quality), LQS values between 0.20 and 0.35 are color coded yellow (medium quality), and LQS values between 0.00 and 0.20 are color-coded red (low quality). Using this color-coding scheme, Table 3 provides the distribution of "Good" and "Bad" quality ROIs categorized as green, yellow, and red.

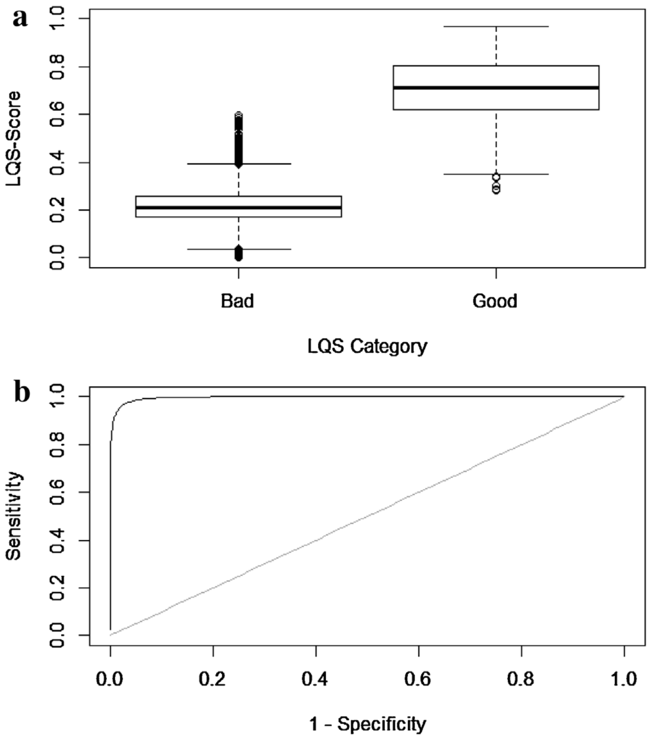

Fig. 5. a Boxplot of LQS values for "Bad" ($n = 3816$) and "Good" ($n = 867$) quality ROIs from LQS-Test-Dataset-1. 5b Receiving Operating Characteristic (ROC) curve of LQS values for "Bad" ($n = 3816$) and "Good" ($n = 867$) quality ROIs from LQS-Test-Dataset-1. The area under the curve (AUC) is 99.7% with a 95% confidence interval of (99.6% - 99.8%).

Table 3
Number of LQS values color-coded as green, yellow, and red compared for "Good" and "Bad" quality ROIs using LQS-Test-Dataset-1. LQS values between 0.35 and 1.00 are color-coded green (high quality), LQS values between 0.20 and 0.35 are color coded yellow (medium quality), and LQS values between 0.00 and 0.20 are color-coded red (low quality).

ROI Quality LQS Color Code	Good	Bad	Total
Green	862	318	1180
Yellow	5	1892	1897
Red	0	1606	1606
Total	867	3816	4683

Having established the baseline performance of the LQS values to distinguish between "Good" and "Bad" quality ROIs and a threshold for categorizing as "high", "medium", or "low" quality (i.e. green, yellow, red), we can use LQS-Test-Dataset-2 to evaluate how well the color-coding output correspond to fingerprint experts' subjective assessment of feature quality ("high" quality vs. "medium" quality due to insufficient annotations of "low" quality features in the dataset). Table 4 demonstrates the consistency between automated predictions of quality using the LQS color-code scheme and experts' subjective judgments.

From Table 4, we see that approximately 94% of the features annotated by experts as green (high quality) were categorized by the LQS color-code scheme as either green (69%) or yellow (25%). Approximately 6% of the features annotated by experts as green were categorized by the LQS color-code scheme as red. Of the features annotated by experts as yellow (medium quality), approximately 87% were categorized by the LQS color-code scheme as green (49%) or yellow (38%). Approximately 13% of the features annotated by experts as yellow were categorized by the LQS color-code scheme as red. Although not perfect correspondence between green vs. green and yellow vs. yellow (which may be expected given the variable nature of experts' judgements), these results indicate reasonable agreement between experts' subjective assessments of feature quality and LQS color-coded classifications as it relates to general groupings of medium or high-quality features. Taken together, among the 5400 total features annotated as either green or yellow by experts' subjective judgments, approximately 93% were categorized as either green or yellow by the LQS color-code scheme. Recognizing the variability in subjective judgments

Table 4
Number of LQS values color-coded as green, yellow, and red compared to experts' subjective judgments of feature quality / confidence using GYRO [22] using LQS-Test-Dataset-2. LQS values between 0.35 and 1.00 are color-coded green (high quality), LQS values between 0.20 and 0.35 are color coded yellow (medium quality), and LQS values between 0.00 and 0.20 are color-coded red (low quality). NOTE: As discussed by John & Swofford (2020) [26] from which this dataset was obtained, experts mostly only annotated features as green and yellow. Experts rarely annotated features as low quality (red), thus those data were insufficient for this assessment.

Expert Judgement LQS Color Code	Green	Yellow	Total
Green	3077	450	3527
Yellow	1119	348	1467
Red	284	122	406
Total	4480	920	5400

of feature quality (e.g. green-yellow or yellow-red), the most significant contribution of the LQS color-code scheme is the ability for it to provide a standardized framework for establishing consistency between examiners related to the relative contribution of features for comparison and flag conditions warranting additional quality assurance review such as those situations where examiners' judgments and the LQS color-code scheme contradict each other on the extreme ends of the spectrum (e.g. green vs. red). While the local performance characteristics are important, the global performance characteristics have the most significant impact on the ultimate outcome of the examination.

3.2. Global performance characteristics

The global performance characteristics were evaluated on the basis of how well the GQS values correspond to analysts' subjective assessments of "value", "complexity", and "difficulty" using a dataset representing casework-like conditions (GQS-Test-Dataset-1). The implications of applying GQS values to impressions under operational conditions is further explored using a dataset derived directly from casework (GQS-Test-Dataset-2). Each dataset is evaluated separately so that the results can be considered within context of the conditions from which the datasets were obtained (e.g. casework-like conditions vs. casework conditions).

3.2.1. "Value" determinations

The Value$_{GQS}$ score is calculated by equation 6 and can range from -1.0 to 1.0. Values near -1.0 indicate the impression is "not suitable" or "no value" and thus should not proceed for further comparison or should do so with caution and additional quality assurance safeguards in place. Values near 1.0 indicate the impression is "suitable" or "of value for identification" and may proceed for further comparison in accordance with normal operating protocols. Fig. 6 illustrates how well the Value$_{GQS}$ score correspond to experts' subjective judgments of impressions deemed to be "no value" (n = 252), "value for exclusion only" (n = 227), or "value for identification" (n = 1141).

From Fig. 6, we see the Value$_{GQS}$ score is able to reasonably distinguish between impressions determined to be "no value" and "value for identification", which represent the ends of the value spectrum. There is overlap between the classes; however, the results demonstrate a trend consistent with expectations—the majority of impressions judged as "VID" have higher values compared to those judged as "NV." The impressions deemed "value for exclusion only" represent a broad span of Value$_{GQS}$ scores and are more difficult to predict. This is understandable, however, since the impressions deemed "value for exclusion only" represent the broad category of impressions in the middle of the value spectrum for which disagreement between examiners was most significant. Looking closer at the inter-rater reliability across the full dataset (train and test partitions combined), *none* of the impressions resulted in consensus determination (defined as two-thirds agreement among participants) for this decision outcome. Consequently, and more practically in an operational setting, the Value$_{GQS}$ score has greater applicability to predicting whether an impression should be categorized as "no value" or "value for identification" and the lack of support for one of those categories should be indicative of the potential for disagreements between experts' interpretations in the middle of the spectrum, thus triggering the impression to be raised for further quality assurance review. Fig. 7 illustrates the performance of the Value$_{GQS}$ score when distinguishing against those impressions determined to be "no value" and "value for identification" using the receiver operator characteristic (ROC). Table 5 demonstrates the performance tradeoff when different threshold values are applied.

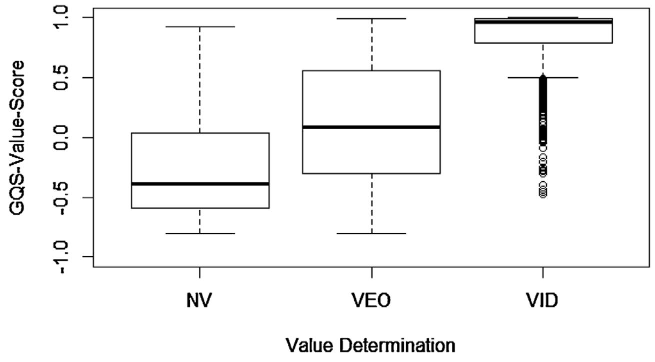

Fig. 6. Boxplot of Value$_{GQS}$ scores for impressions subjectively judged by experts to be "no value" (NV) (n = 252), "value for exclusion only" (VEO) (n = 227), or "value for identification" (VID) (n = 1141) from GQS-Test-Dataset-1.

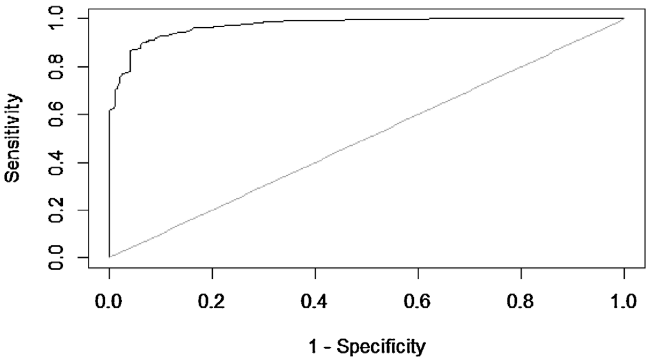

Fig. 7. Receiving Operating Characteristic (ROC) curve of Value$_{GQS}$ scores for impressions subjectively judged by experts to be "no-value" (n = 252) and "value for identification" (n = 1141) from GQS-Test-Dataset-1. The area under the curve (AUC) is 97.3% with a 95% confidence interval of (96.5% - 98.2%).

3.2.2. "Complexity" determinations

The Complexity$_{GQS}$ score is calculated by equation 7 and can range from -1.0 to 1.0. Values near -1.0 indicate the impression is "not suitable" or "highly complex" and thus should only proceed to comparison with caution and additional quality assurance safeguards in place. Values near 1.0 indicate the impression is "non-complex" and may proceed for further comparison in accordance with normal operating protocols. Fig. 8 illustrates how well the Complexity$_{GQS}$ score corresponds to experts' subjective judgments of impressions deemed to be "no value" or "highly complex" (n = 291), "complex" (n = 452), or "non-complex" (n = 877).

It transpires from Fig. 8, that the Complexity$_{GQS}$ score is able to reasonably distinguish between impressions determined to be "highly complex" and "non-complex", which represent the ends of the complexity spectrum. There is overlap between the classes;

Table 5
Proportion of responses resulting in Value$_{GQS}$ score greater than threshold values (-0.50, -0.33, -0.25, 0.00, 0.25, 0.33, 0.50) and assessed as "no-value" (n = 252) and "value for identification" (n = 1141) by experts from GQS-Test-Dataset-1. Confidence intervals are indicated (lower CI - upper CI).

Threshold Value$_{GQS}$	"No Value"	"Value for Identification"
−0.50	0.623 (0.563–0.683)	1.00 (1.00–1.00)
−0.33	0.484 (0.425–0.548)	0.996 (0.991–0.999)
−0.25	0.405 (0.345–0.464)	0.992 (0.987–0.996)
0.00	0.274 (0.218–0.329)	0.979 (0.97–0.987)
0.25	0.159 (0.115–0.206)	0.954 (0.942–0.966)
0.33	0.127 (0.087–0.171)	0.938 (0.924–0.952)
0.50	0.063 (0.036–0.095)	0.895 (0.876–0.912)

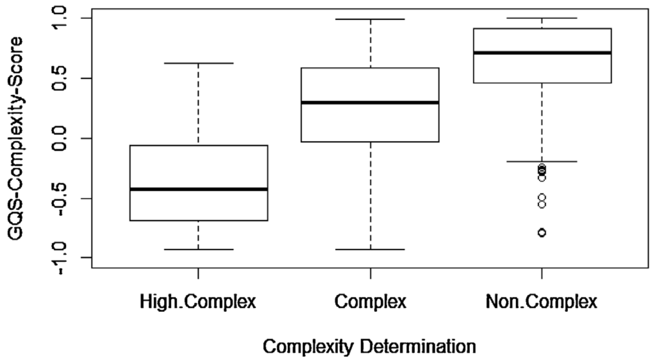

Fig. 8. Boxplot of Complexity$_{GQS}$ scores for impressions subjectively judged by experts to be "highly complex" (n = 291), "complex" (n = 452), or "non-complex" (n = 877) from GQS-Test-Dataset-1.

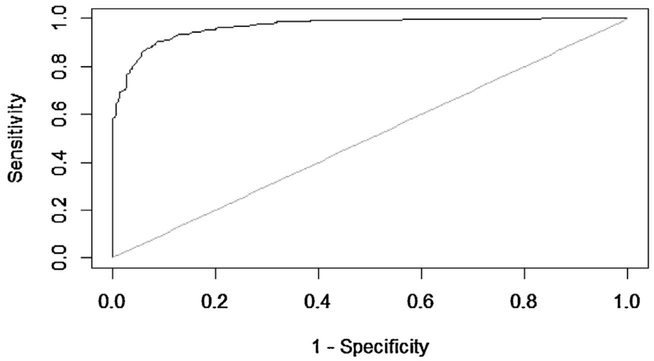

Fig. 9. Receiving Operating Characteristic (ROC) curve of Complexity$_{GQS}$ scores for impressions subjectively judged by experts to be "highly complex" (n = 291) and "non-complex" (n = 877) from GQS-Test-Dataset-1. The area under the curve (AUC) is 96.8% with a 95% confidence interval of (95.9% - 97.7%).

however, the results demonstrate a trend consistent with expectations—the majority of impressions judged as "non-complex" have higher values compared to those judged as "highly complex." The impressions deemed "complex" represent a broad span of Complexity$_{GQS}$ scores and are more difficult to predict. Similar to the "value" spectrum, this is understandable since impressions deemed "complex" represent the broad category of impressions in the middle of the complexity spectrum for which disagreement between examiners was most significant. Consequently, and more practically in an operational setting, the Complexity$_{GQS}$ score has greater applicability to predicting whether an impression should be categorized as "highly complex" or "non-complex" and the lack of support for one of those categories should be indicative of the potential for disagreements between experts' interpretations in the middle of the spectrum, thus triggering the impression to be raised for further quality assurance review. Fig. 9 illustrates the performance of the Complexity$_{GQS}$ score when distinguishing against those impressions determined to be "highly complex" and "non-complex" using the receiver operator characteristic (ROC). Table 6 demonstrates the performance tradeoff when different threshold values are applied.

3.2.3. "Difficulty" determinations

The Difficulty$_{GQS}$ score is calculated by equation 8 and can range from −1.0 to 1.0. Values near -1.0 indicate the impression is "high difficulty" and thus should only proceed to comparison with caution and additional quality assurance safeguards in place. Values near 1.0 indicate the impression is "low difficulty" and may proceed for further comparison in accordance with normal operating protocols. Fig. 10 illustrates how well the Difficulty$_{GQS}$ score corresponds to experts' subjective judgments of impressions deemed to be "high difficulty" (n = 487), "medium difficulty" (n = 556), or "low difficulty" (n = 577).

Table 6
Proportion of responses resulting in Complexity$_{GQS}$ score greater than threshold values (−0.50, −0.33, −0.25, 0.00, 0.25, 0.33, 0.50) and assessed as "highly complex" (n = 291) and "non-complex" (n = 877) by experts from GQS-Test-Dataset-1. Confidence intervals are indicated (lower CI - upper CI).

Threshold Complexity$_{GQS}$	"Highly Complex"	"Non-Complex"
−0.50	0.570 (0.512–0.625)	0.997 (0.992–1.00)
−0.33	0.419 (0.364–0.478)	0.994 (0.989–0.999)
−0.25	0.378 (0.323–0.433)	0.989 (0.981–0.995)
0.00	0.206 (0.162–0.254)	0.962 (0.950–0.975)
0.25	0.076 (0.048–0.107)	0.886 (0.864–0.906)
0.33	0.055 (0.031–0.082)	0.854 (0.830–0.877)
0.50	0.027 (0.010–0.048)	0.717 (0.688–0.747)

From Fig. 10, we see the Difficulty$_{GQS}$ score is able to generally distinguish between impressions determined to be "high difficulty" and "low difficulty", which represent the ends of the difficulty spectrum. There is overlap between the classes; however, the results demonstrate a trend consistent with expectations—the majority of impressions judged as "low" difficulty have higher values compared to those judged as "high" difficulty. The impressions deemed "medium difficulty" represent a broad span of Difficulty$_{GQS}$ scores and are more difficult to predict. Similar to the "value" and "complexity" spectrums, this is understandable since impressions deemed "medium difficulty" represent the broad category of impressions in the middle of the spectrum for which disagreement between examiners was most significant. Consequently, and more practically in an operational setting, the Difficulty$_{GQS}$ score has greater applicability to predicting whether an impression should be categorized as "high difficulty" or "low difficulty" and the lack of support for one of those categories should be indicative of the potential for disagreements between experts' interpretations in the middle of the spectrum, thus triggering the impression to be raised for further quality assurance review. Fig. 11 illustrates the performance of the Difficulty$_{GQS}$ score when distinguishing against those impressions determined to be "high difficulty" and "low difficulty" using the receiver operator characteristic (ROC). Table 7 demonstrates the performance tradeoff when different threshold values are applied.

3.2.4. Casework evaluation

From GQS-Test-Dataset-1, we see that the GQS values are able to reasonably distinguish between impressions on the ends of the

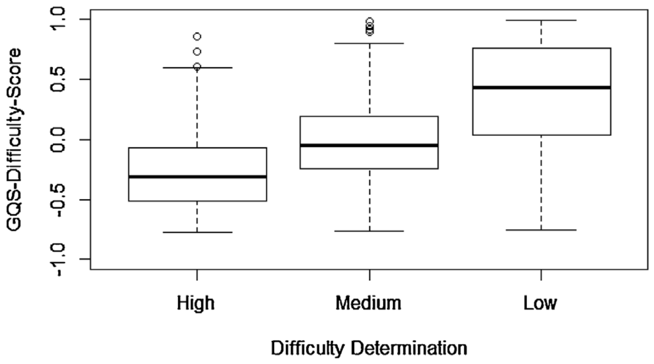

Fig. 10. Boxplot of Difficulty$_{GQS}$ scores for impressions subjectively judged by experts to be "high difficulty" (n = 487), "medium difficulty" (n = 556), or "low difficulty" (n = 577) from GQS-Test-Dataset-1.

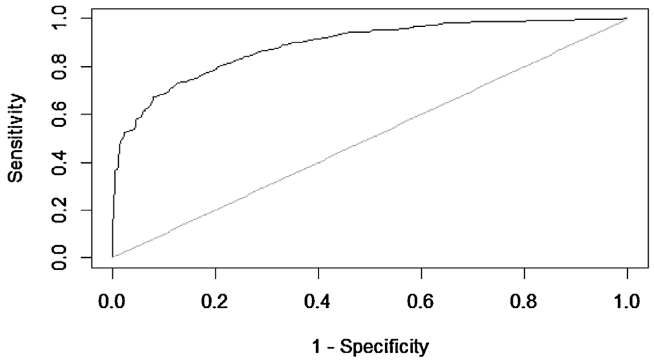

Fig. 11. Receiving Operating Characteristic (ROC) curve of Difficulty$_{GQS}$ scores for impressions subjectively judged by experts to be “high difficulty” (n = 487) and “low difficulty” (n = 577) from GQS-Test-Dataset-1. The area under the curve (AUC) is 88.8% with a 95% confidence interval of (86.9% - 90.7%).

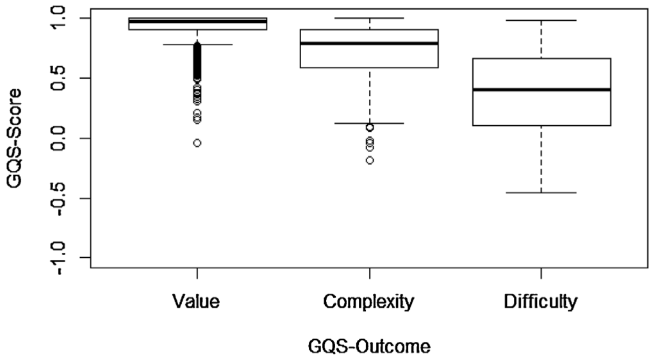

Fig. 12. Boxplot of Value$_{GQS}$, Complexity$_{GQS}$, and Difficulty$_{GQS}$ scores for 605 impressions subjectively judged by experts to be “value for identification” and subsequently identified to exemplar impressions during normal casework conditions from GQS-Test-Dataset-2.

Table 7
Proportion of responses resulting in Difficulty$_{GQS}$ score greater than threshold values (−0.50, −0.33, −0.25, 0.00, 0.25, 0.33, 0.50) and assessed as “high difficulty” (n = 487) and “low difficulty” (n = 577) by experts from GQS-Test-Dataset-1. Confidence intervals are indicated (lower CI - upper CI).

Threshold Difficulty$_{GQS}$	“High Difficulty”	“Low Difficulty”
−0.50	0.729 (0.690–0.768)	0.986 (0.976–0.995)
−0.33	0.515 (0.470–0.561)	0.953 (0.936–0.969)
−0.25	0.415 (0.372–0.458)	0.922 (0.899–0.943)
0.00	0.193 (0.158–0.228)	0.782 (0.747–0.815)
0.25	0.057 (0.037–0.080)	0.610 (0.570–0.650)
0.33	0.045 (0.029–0.064)	0.555 (0.515–0.594)
0.50	0.012 (0.004–0.023)	0.449 (0.409–0.490)

value, complexity, and difficulty spectra thus indicating those impressions which may proceed to further comparison in accordance with normal operational protocols versus those impressions which may be flagged for further quality assurance review and additional safeguards. Having established the baseline performance characteristics under case-work like conditions, we can consider the implications if this quality metric were to be applied in an operational setting on actual casework to demonstrate the distribution of GQS values and potentially indicate the need for intervention from a quality assurance perspective when GQS values fall below an established threshold. To evaluate this, we use GQS-Test-Dataset-2, which consists of 605 impressions that were deemed “value for identification” by experts’ subjective judgements (and subsequently identified to exemplar impressions). Although this dataset does not include those impressions deemed to be “no value” since operational procedures did not require retention of annotated images for that outcome category, we can consider the proportion of impressions for which the determination of “value for identification” was supported. Similarly, despite the impressions not being pre-categorized against the complexity spectrum or difficulty spectrum, we can visualize the distribution of the impressions against each metric for general context.

Fig. 12 illustrates the distribution of Value$_{GQS}$ scores, Complexity$_{GQS}$ scores, and Difficulty$_{GQS}$ scores for the GQS-Test-Dataset-2.

If we were to apply threshold values to the GQS metrics to evaluate how often the experts’ assessment of “value for identification” was supported or to indicate circumstances in which the impressions may be flagged for additional quality assurance review, we can consider the implications to practice more clearly. For the Value determination, Table 5 suggests a Value$_{GQS}$ score of 0.50 is a reasonable threshold. For the Complexity determination, Table 6 suggests a Complexity$_{GQS}$ score of 0.33 is a reasonable threshold. For the Difficulty determination, Table 7 suggests a Difficulty$_{GQS}$ score of 0.00 is a reasonable threshold. Table 8 provides the proportion of impressions for which normal procedures are sufficient and those for which additional quality assurance review may be considered based on the results of the GQS metrics. From these data, we see reasonably strong support for experts’ subjective judgement on the casework sample (GQS-Test-Dataset-2) and only a small percentage of impressions for which additional quality assurance review might be considered (~2% lacking support for value, ~6% categorized as complex, and ~16% categorized as difficult).

3.3. General discussion

The method proposed provides three different quality metrics which can be used as a means to provide empirical support to experts’ subjective assessments and a framework for establishing policies and procedures to flag impressions warranting further quality assurance review. Determinations of “value” have been considered by the friction ridge discipline for decades and are

Table 8
Proportion of impressions for which normal procedures are warranted and those for which additional quality assurance review may be considered based on the results of the GQS metrics from GQS-Test-Dataset-2 (n = 605) and the following thresholds: Value$_{GQS}$ scores less than 0.50, Complexity$_{GQS}$ scores less than 0.33, and Difficulty$_{GQS}$ scores less than 0.00. Note: GQS-Test-Dataset-2 is a dataset of impressions taken from a single federal laboratory in the United States which were considered “value for identification” and subsequently identified to exemplar impressions. Given the lack of quantifiable standards for “value for identification” at the time these impressions were examined, the extent to which these results are generalizable is unclear.

GQS Metric	Proportion of Cases with Normal Procedures Warranted	Proportion of Cases to Consider Additional Quality Assurance Review
Value	0.977	0.023
Complexity	0.942	0.058
Difficulty	0.843	0.157

familiar to all practicing examiners. Determinations of “complexity” and “difficulty”, however, are more recent terms to categorize impressions which tend to have lower quality and quantity of features and are therefore more susceptible to erroneous outcomes. With limited time and resources due to growing backlogs and operational demands, it is critical to have a means of focusing efforts on those impressions most vulnerable to errors or may require additional quality control measures. This method provides a means of accomplishing this goal in a more objective, transparent, and consistent fashion grounded by empirical validation. Although ground truth is non-existent for determinations of “value”, “complexity”, and “difficulty”, the results demonstrate reasonable agreement to experts’ subjective assessments and illustrate a consistent general trend of increasing GQS values across the ordinal scale of “value”, “complexity”, and “difficulty” determinations. Having these quantitative outputs along ordinal scales, further work could enable a visual illustration and representation of the overall quality of an impression in three-dimensional space based on axes of “value”, “complexity”, and “difficulty”.

Two important limitations for this method remain. First, the LQS and GQS values are dependent upon the subjective detection and annotation of friction ridge skin features (minutiae) by the human expert. Second, the method relies on clarity attributes of friction ridge minutiae and does not consider all of the attributes that experts may consider when making subjective determinations, such as pattern type, feature type, rarity of features and their configurations, continuity of ridge detail between features, and other types of features (non-minutiae) available.

To attenuate these limitations, two general recommendations for policy and procedure could be considered. First, the method should be used *after* the expert has visually analyzed, detected, and annotated friction ridge skin features for which the expert has reasonably high confidence of their presence. Second, the method should be used as a framework for flagging impressions which may require additional quality assurance review. Although the method demonstrates reasonable consistency with experts’ judgements, it should not be considered a replacement for the experts’ interpretation. This method is a step toward greater transparency and objectivity, but is not designed or intended to supplant the careful interpretation of experts.

This method provides fingerprint experts the capability to provide an empirical foundation to support their subjective interpretations following *Analysis*. It also offers a framework for organizations to establish transparent, measurable, and demonstrable criteria for Value determinations and a means of flagging impressions that are vulnerable to erroneous outcomes or inconsistency between experts (e.g. higher Complexity and/or Difficulty). Finally, it provides a means for quantitatively summarizing the overall quality of the impression in terms of Value, Complexity, and Difficulty for ensuring representative distributions in samples used for research designs, proficiency testing, error rate testing, and other applications by forensic science stakeholders. As a stand-alone application, this method enables the forensic science community to take a step toward greater transparency and empiricism – particularly as it relates to Value and Complexity determinations during casework examinations and assessments of Difficulty for research, training, and testing purposes. Further, because this method provides quality assessments at both the local and global levels (LQS and GQS), its development lends the possibility of integrating with other quality assessment and statistical evaluation software applications, such as *FRStat* [27], to provide a complete tool-pack to ensure experts’ interpretations are empirically supported for all major decisions throughout the entire examination methodology.

4. Conclusion

Over the years, the forensic science community has faced increasing 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 assess the quality of fingerprint evidence for a given case. In this paper, a method is presented which measures the clarity of friction ridge features and evaluates the quality of impression across three different scales: Value, Complexity, and Difficulty. The local quality scores (LQS) provide a quantitative assessment of the quality of individual features based on the clarity of the local region of friction ridge detail immediately surrounding each feature. Individual features are then color-coded green, yellow, or red indicating high, medium, or low quality. The results demonstrate remarkable separation between regions representing the extreme ends of “good” and “bad” quality of friction ridge detail and general agreement with experts’ subjective assessments of feature quality based on features categorized as “high” or “medium” quality. While quality assessments at localized regions are important, quality assessments for the overall impression have the most significant impact on the ultimate outcome of the examination. The global quality scores (GQS) provide quantitative assessments of the quality of the entire impression against different outcome scales (value, complexity, and difficulty) based on the quality and quantity of individual features. The results demonstrate reasonable consistency between automated predictions and experts’ subjective assessments. In an operational environment, the tool is intended to provide an empirical foundation to support experts’ subjective judgments, provide transparency to the overall quality of the impression for a given outcome (e.g. determination of value, complexity, or difficulty), and provide a framework to establish policies and procedures for examination decisions geared toward flagging impressions that are generally lower quality and more vulnerable to disagreements between experts or potentially erroneous interpretations.

As with any method, there are limitations to consider. The most significant is that this method relies on the features annotated by the expert and does not take into account all aspects of the friction ridge detail. Consequently, the system should not be considered as a means of supplanting expert interpretation and judgement when analyzing friction ridge detail. Rather, the method should be considered a tool to support experts’ judgements or detect potentially problematic impressions necessitating further quality assurance review.

Although various aspects of this 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 quantitative measure of the quality of a fingerprint. Further optimizations which may improve upon the method’s performance are encouraged for future works.

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.

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.

Champod: Support in terms of data collection, analysis, and interpretation; suggestions and recommendations to concept and design; article review.

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

Eldridge: Support in terms of data collection, analysis, and interpretation; article review.

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

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

The authors are grateful to several individuals that have contributed to this effort. Dr. Hariharian Iyer, Dr. Steven Lund, Dr. John Buckleton, Dr. Simone Gittelson, Dr. Hal Stern, and Dr. Karen Kafadar for their early guidance and suggestions for improvements; Marco De Donno and Tommy Jones for their assistance with R programming for data analysis; Rick Tontarski and Lauren Reed for their encouragement and support; and the several latent print examiners that participated in the studies for which the method was developed and tested against and contributed to the data collection.

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 material related to this article can be found, in the online version, at https://doi.org/10.1016/j.forsciint.2021.110703.

References

[1] 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.

[2] Standards for Examining Friction Ridge Impressions and Resulting Conclusions, Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST), Ver 2.0, (2013) Available online https://www.nist.gov/sites/default/files/documents/2016/10/26/swgfast_examinations-conclusions_2.0_130427.pdf. Accessed 6/7/2020.

[3] 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 Available online https://nvlpubs.nist.gov/nistpubs/ir/2012/NIST.IR.7842.pdf. Accessed 6/7/2020.

[4] 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) Available online https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf. Accessed 6/7/2020.

[5] American Academy for the Advancement of Science (AAAS), Forensic Science Assessments: A Quality and Gap Analysis, Latent Fingerprint Examination, AAAS, 2017 Available online https://www.aaas.org/report/latent-fingerprint-examination. Accessed 6/7/2020.

[6] F. Alonso-Fernandez, J. Fierrez-Aguilar, J. Ortega-Garcia, A review of schemes for fingerprint image quality computation, Proc. COST-275 Workshop on Biometrics on the Internet (2005) 3–6.

[7] N. Nill, IQF (Image Quality of Fingerprint) Software Application, MTR 070053, MITRE Technical Report, (2007) .

[8] H. Fronthaler, K. Kollreider, J. Bigun, J. Fierrez, F. Alonso-Fernandez, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Fingerprint image quality estimation and its application to multi-algorithm verification, Ieee Trans. Inf. Forensics Secur. 3 (2) (2008) 331–338.

[9] R.A. Hicklin, J. Buscaglia, M.A. Roberts, S. Meagher, W. Fellner, M. Burge, M. Monaco, D. Vera, L. Pantzer, C. Yeung, T. Unnikumaran, Latent fingerprint quality: a survey of examiners, J. For. Ident 61 (4) (2011) 385–418.

[10] R. Murch, A.L. Abbott, E. Fox, M. Hsiao, B. Budowle, Establishing the Quantitative Basis for Sufficiency Thresholds and Metrics for Friction Ridge Pattern Detail and the Foundation for a Standard, Technical Report, Technical Report, National Institute of Justice, U.S. Department of Justice, 2012 Available online https://www.ncjrs.gov/pdffiles1/nij/grants/239049.pdf. Accessed 6/7/2020.

[11] S. Yoon, E. Liu, A. Jain, On latent fingerprint image quality, Computational Forensics: Proc. 5th International Workshop on Computational Forensics, Tsukuba, Japan, 2015, pp. 67–82 November 11, 2012 and 6th International Workshop, IWCF 2014, Stockholm, Sweden, August 24, 2014, Revised Selected Papers. Garain U. and Shafait F. (Eds), Springer.

[12] S. Yoon, K. Cao, E. Liu, A.K. Jain, LFIQ: latent fingerprint image quality. Biometrics: theory, applications and systems (BTAS), 2013 IEEE Sixth International Conference on (2013) 1–8.

[13] R.A. Hicklin, J. Buscaglia, M.A. Roberts, Assessing the clarity of friction ridge impressions, Forensic Sci. Int. 226 (2013) 106–117.

[14] N. Kalka, M. Beachler, A. Hicklin, LQ Metric: A Latent Fingerprint Quality Metric for Predicting AFIS Performance and Assessing the Value of Latent Fingerprints, J. For. Ident. 70 (4) (2020) 443–463.

[15] National Institute of Standards and Technology, ANSI/NIST-ITL 1-2011, American National Standard for Information Systems: Data Format for the Interchange of Fingerprint Facial & Other Biometric Information, (2011) .

[16] A. Sankaran, M. Vatsa, R. Singh, Automated clarity and quality assessment for latent fingerprints, IEEE International Conference on Biometrics: Theory, Applications and Systems, 2013, pp. 1–6.

[17] D. Pulsifer, S. Muhlberger, S. Williams, R. Shaler, A. Lakhtakia, An Objective Fingerprint quality-grading system, Forensic Sci. Int. 231 (2013) 204–207.

[18] P. Kellman, J. Mnookin, G. Erlikhman, P. Garrigan, T. Ghose, E. Mettler, D. Charlton, I. Dror, Forensic comparison and matching of fingerprints: using quantitative image measures for estimating error rates through understanding and predicting difficulty, PLoS One 9 (5) (2014) 1–14.

[19] T. Chugh, K. Cao, J. Zhou, E. Tabassi, A. Jain, Latent fingerprint value prediction: crowd-based learning, Ieee Trans. Inf. Forensics Secur. 13 (1) (2018) 20–34.

[20] Choong, et al., Acutance, an objective measure of retinal nerve fibre image clarity, Br. J. Ophthalmol. 87 (2003) 322–326.

[21] R.T. Moore, Analysis of Ridge –to-Ridge distance in fingerprints, J. For. Ident. 39 (4) (1989) 231–238.

[22] G. Langenburg, C. Champod, The GYRO system – a recommended approach to more transparent documentation, J. For. Ident. 61 (4) (2011) 373–384.

[23] H. Eldridge, J. Furrer, M. De Donno, C. Champod, Examining and expanding the friction ridge value decision, For. Sci. Int. 314 (2020)110408.

[24] R Core Team, R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2019) . URL: https://www.R-project.org/.

[25] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformatics 12 (1) (2011) 77.

[26] J. John, H. Swofford, Evaluating the accuracy and weight of confidence in examiner minutiae annotations, J. For. Ident. 70 (3) (2020) 289–309.

[27] H. Swofford, A. Koertner, F. Zemp, M. Ausdemore, A. Liu, J. Salyards, A method for the statistical interpretation of friction ridge skin impression evidence: method development and validation, For. Sci. Int. 287 (2018) 113–126.