Forensic Science International 284 (2018) 15–32

Contents lists available at ScienceDirect

Forensic Science International

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

Estimating error rates for firearm evidence identifications in forensic science

John Song[a], Theodore V. Vorburger[a,*], Wei Chu[a], James Yen[b], Johannes A. Soons[a], Daniel B. Ott[a], Nien Fan Zhang[b]

[a] Engineering Physics Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA
[b] Statistical Engineering Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA

ARTICLE INFO

Article history:
Received 21 July 2017
Received in revised form 6 November 2017
Accepted 6 December 2017
Available online 13 December 2017

Keywords:
Forensics
Firearm
Ballistics identification
Error rate
Congruent matching cell
CMC

ABSTRACT

Estimating error rates for firearm evidence identification is a fundamental challenge in forensic science. This paper describes the recently developed congruent matching cells (CMC) method for image comparisons, its application to firearm evidence identification, and its usage and initial tests for error rate estimation. The CMC method divides compared topography images into correlation cells. Four identification parameters are defined for quantifying both the topography similarity of the correlated cell pairs and the pattern congruency of the registered cell locations. A declared match requires a significant number of CMCs, i.e., cell pairs that meet all similarity and congruency requirements. Initial testing on breech face impressions of a set of 40 cartridge cases fired with consecutively manufactured pistol slides showed wide separation between the distributions of CMC numbers observed for known matching and known non-matching image pairs. Another test on 95 cartridge cases from a different set of slides manufactured by the same process also yielded widely separated distributions. The test results were used to develop two statistical models for the probability mass function of CMC correlation scores. The models were applied to develop a framework for estimating cumulative false positive and false negative error rates and individual error rates of declared matches and non-matches for this population of breech face impressions. The prospect for applying the models to large populations and realistic case work is also discussed. The CMC method can provide a statistical foundation for estimating error rates in firearm evidence identifications, thus emulating methods used for forensic identification of DNA evidence.

1. Introduction

Tool marks are permanent changes in the topography of a surface created by forced contact with a harder object (the tool). When bullets and cartridge cases are fired or ejected from a firearm, the parts of the firearm that make forcible contact with them create characteristic tool marks called “ballistic signatures” [1]. By examining these ballistic signatures side-by-side in a comparison microscope, firearm examiners can determine whether a pair of bullets or cartridge cases was fired or ejected from the same firearm. Firearm examiners can then connect a recovered firearm or other firearm evidence to criminal acts.

Successful identification requires that the relevant firearm surfaces have individuality and that the tool marks are reproducible [1]. In general, tool marks have so-called “class characteristics” that are common to certain firearm designs and manufacturing methods, and “individual characteristics” arising from random variations in firearm manufacturing and wear [1]. While class characteristics can be used to exclude a firearm as a source of a recovered cartridge case or bullet, the patterns of individual characteristics are often unique to individual firearms and can therefore form the basis for identification [1]. These individual characteristics are marks produced by the random imperfections or irregularities of the firearm surfaces, which may arise during manufacture or by corrosion or damage during use [2]. In mechanical engineering terms, individual characteristics are approximately equivalent in scale to surface roughness irregularities [3].

Side-by-side tool mark image comparisons for firearm identification have a history of more than a hundred-years [1]. However, the scientific foundation of firearm and tool mark identification has been challenged by recent reports and court decisions. As stated in a 2008 National Academies Report [4], “*The validity of the fundamental assumptions of uniqueness and reproducibility of*

* Corresponding author at: NIST, 100 Bureau Drive, Gaithersburg, MD 20899, USA.
E-mail address: tvtv@nist.gov (T.V. Vorburger).

https://doi.org/10.1016/j.forsciint.2017.12.013

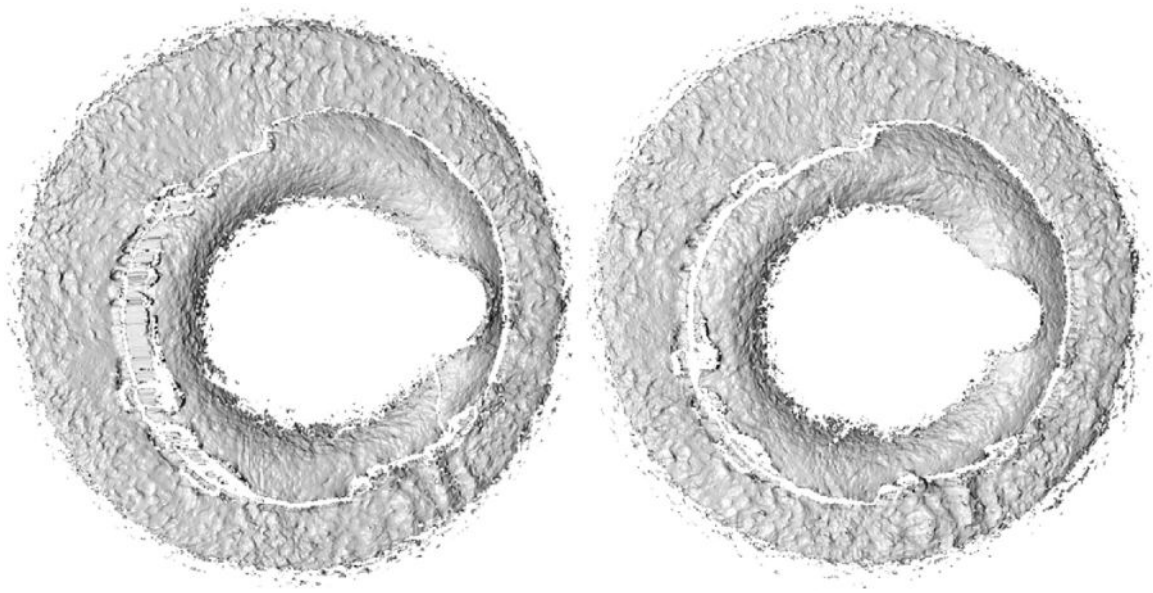

Fig. 1. Topography images of breech face impressions obtained from a pair of cartridge cases ejected from slide 3 in the Fadul data set [19] discussed here. The data set consisted of test fires of Federal[1] cartridges from consecutively manufactured Ruger 9 mm slides. The images have several features in common. The diameter of each image is about 3.5 mm. The topography contrast is rendered with a virtual light source from the left.

firearms-related tool marks has not yet been fully demonstrated . . . " and "*Since the basis of all forensic identification is probability theory, examiners can never really assert a conclusion of an 'identification to exclusion of all others in the world,' but at best can only assert a very small (objective or subjective) probability of a coincidental match.*"

The legal standard for the acceptance of scientific evidence contained in the U.S. Supreme Court decision, called the Daubert standard [4], "*places high probative weight on quantifiable evidence that can be tested empirically and for which known or potential error rates may be estimated, such as identification using DNA markers*" [4]. However, as stated in a 2009 National Academies Report [5], "*But even with more training and experience using newer techniques, the decision of the toolmark examiner remains a subjective decision based on unarticulated standards and no statistical foundation for estimation of error rates.*"

Since the 1980's, estimates of coincidental match probability (CMP) have been used for specifying uncertainty of DNA identifications: "*The courts already have proven their ability to deal with some degree of uncertainty in individualizations, as demonstrated by the successful use of DNA analysis (with its small, but nonzero, error rate)*" [5]. It is therefore a fundamental challenge in forensic science to establish a scientific foundation and statistical procedures providing quantitative error rate reports to support firearm identifications, in the same way that reporting procedures have been established for forensic identification of DNA evidence [5]. Several experimental and theoretical efforts have been pursued along this line including the computer learning approach of Petraco et al. [6,7], the work on likelihood ratio by Riva and Champod [8], the study of examiner error rates by Baldwin et al. [9], the feature-based matching algorithm of Lilien [10,11], the work on image cross correlation and congruent matching cells (CMC) of Song et al. [12–17], and the random forest approach of Hare et al. [18].

In this paper, we apply the CMC method [14–17] to estimations of error rates for false identifications and exclusions for two sets of topography image data of breech face impressions from fired cartridge cases. We discuss the CMC method in Section 2, then describe validation tests, error rate estimation procedures and initial results in Sections 3–5, and provide observations about future directions and the prospect for application to case work in Section 6.

2. Congruent matching cells (CMC) method

We begin with pairs of measured 3D topography images of breech face impressions whose similarity we wish to quantify (see Fig. 1). A common approach would be to calculate the value of the normalized cross-correlation function (Pearson's correlation coefficient) for the pair of images as a whole [12,13], when they are registered at a position of maximum correlation. Instead, the CMC method divides the reference image into a rectangular array of cells as shown in Fig. 2. For each cell on the reference image, an automated search is made on a compared image for a highly similar region. The cell-by-cell analysis is done because a firearm often produces characteristic marks, or individual characteristics [1], on only a portion of the bullet or cartridge case surface, depending on its degree of contact with the firearm during firing. Carrying over the terminology from previous research in firearms identification [14,15], a region of the surface topography is termed a "valid correlation region" if it contains individual characteristics of the ballistic signature that can be used effectively for firearm identification. Conversely, a region of the surface topography that does not contain individual characteristics of the firearm's ballistic signature is termed an "invalid correlation region" that should be eliminated from consideration for firearm identification. Invalid correlation areas can occur, for example, due to insufficient contact between the firearm's surface and the bullet or cartridge case during firing.

If two ballistic topographies A and B originate from the same firearm, both will likely contain valid and invalid correlation regions. When A and B are compared with each other, their common valid correlation region is the overlap of the individual valid correlation regions of A and B, which comprise only part,

[1] Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

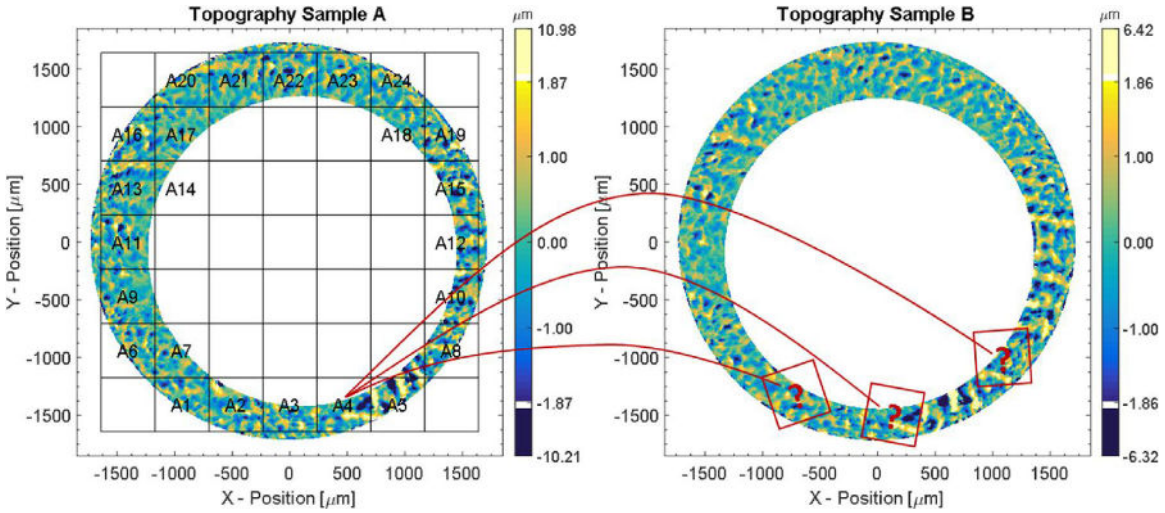

Fig. 2. Conceptual diagram of a topography image from Fig. 1 overlaid by a 7 × 7 grid, dividing the reference image (left) into cells. The drag mark at the 3 o'clock position in Fig. 1 and the central hole and surrounding bulge from the firing pin impression are masked out. Only cells with a sufficient fraction of measured pixels are used for the correlation analysis. Also shown is an illustration of the automated search procedure to find an area in the compared image (right) that has a strong correlation with one of the cells in the reference image (left). Here the topography is represented by a color scale.

sometimes even a small part, of the entire areas of A and B. If a quantitative measure of correlation is obtained from the entire images of A and B, the correlation accuracy may be relatively low because large invalid regions may be included in the correlation. If the correlation areas are divided into cells, the valid correlation regions may be analyzed without being combined with invalid regions. The CMC procedure to identify cells containing valid regions can significantly increase the correlation effectiveness and accuracy. Furthermore, the use of a statistically large number of congruently matched cells identified by multiple parameters can facilitate the estimation of an error rate [15] from a well characterized population.

A correlation cell is a rectangular sub-region of the surface topography image that contains a sufficient quantity of distinguishing peaks, valleys, and other topographic features so that an assessment of topography similarity can be made. If topographies A and B originating from the same firearm are registered at their position of maximum correlation (Fig. 3), the cell pairs located in their common valid correlation regions can be identified, as shown by the solid cell pairs located in (A_1, B_1), (A_2, B_2), and (A_3, B_3). These cell pairs are necessarily characterized by [14,15]:

1) High pairwise topography similarity as quantified by a high value of the normalized cross correlation function maximum CCF_{max};
2) Similar registration angles θ for all correlated cell pairs in valid regions A and B; and
3) "Congruent" x–y spatial distribution patterns for the correlated cell arrays (A_1, A_2, A_3 . . .) and (B_1, B_2, B_3 . . .) or nearly so.

On the other hand, if the registered cell pairs are located in the invalid correlation regions of A and B, such as the dotted cells (a', a'', a''') and (b', b'', b''') in Fig. 3, or if they originate from different firearms, their maximum cross correlation value CCF_{max} would be relatively low, and their cell arrays would show significant variation in their x–y distribution patterns and registration angles θ.

Congruent matching cell pairs, or CMCs, are therefore determined by four dentification parameters for quantifying both the topography similarity of the correlated cell pairs and the pattern congruency of the cell distributions. The former is quantified by the normalized cross correlation function maximum CCF_{max} with threshold T_{CCF}; the latter is quantified by the registration angle θ and translation distances in x and y with corresponding thresholds

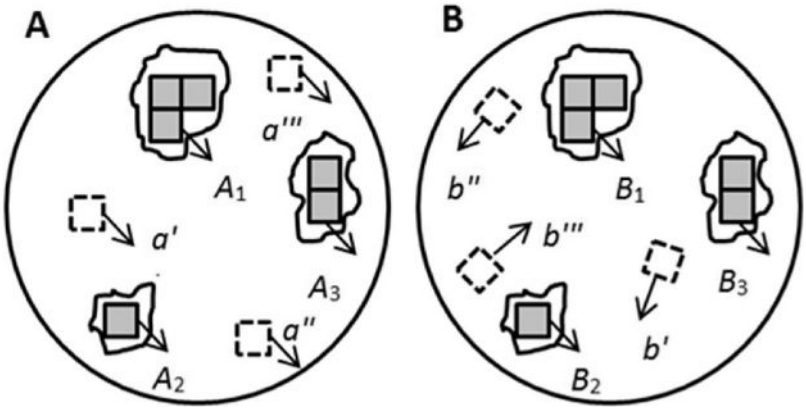

Fig. 3. Schematic diagram of topographies A and B originating from the same firearm and registered at the position of maximum correlation. The six solid cell pairs in each image are in three valid correlated regions (A_1,B_1), (A_2, B_2), and (A_3, B_3). The dotted cell pairs (a', b'), (a'', b''), and (a''', b''') are in the invalid correlation region.

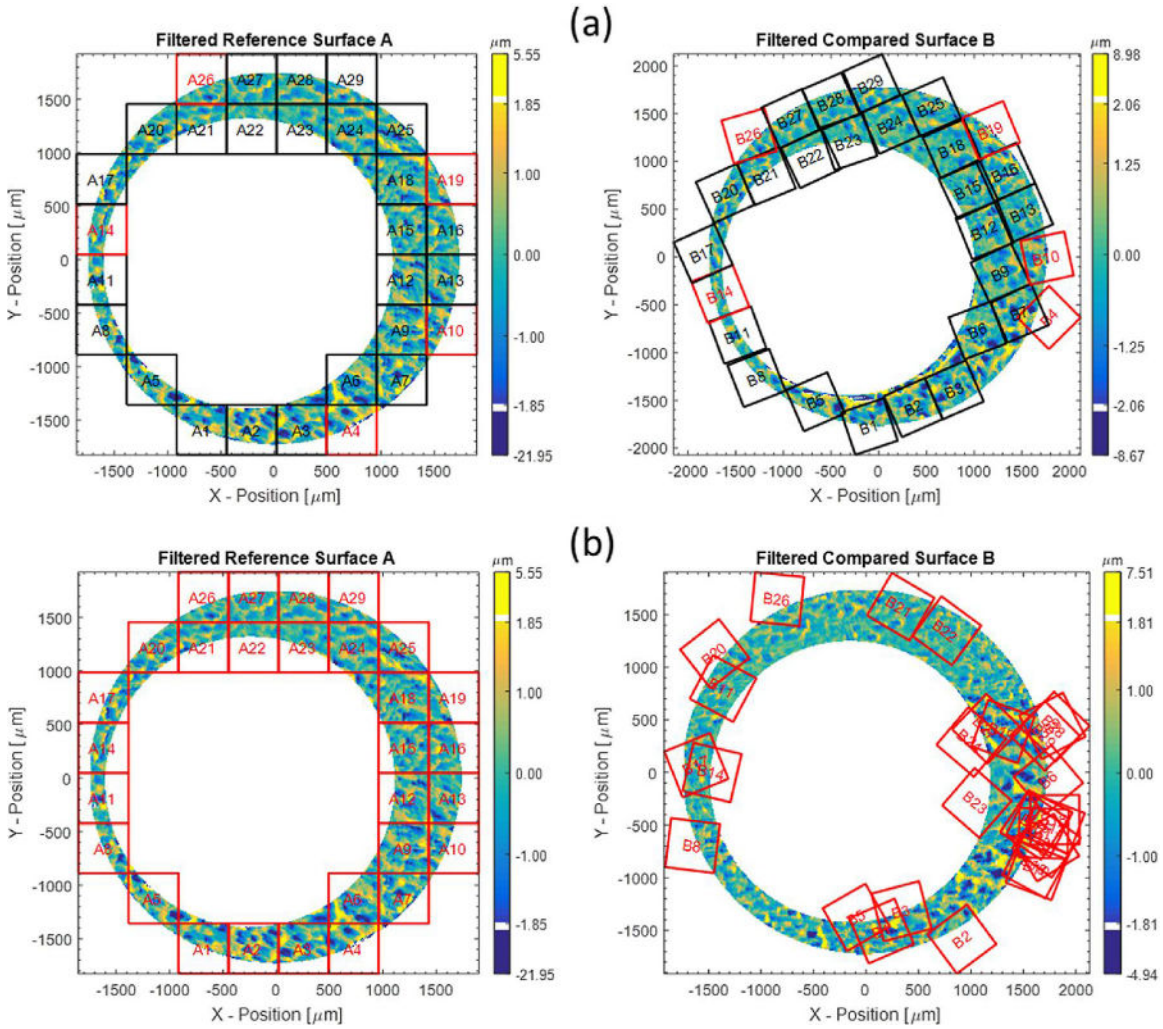

Fig. 4. Typical results for a CMC comparison of (a) breech face impressions from the same firearm and (b) breech face impressions from different firearms.

T_θ, T_x, and T_y. A correlated cell pair is considered a CMC — that is, part of a congruent matching cell pattern — when its correlation value CCF_{max} is greater than a chosen T_{CCF}, and the registration angle θ and x, y registration positions are within chosen thresholds T_θ, T_x and T_y. The automated search and registration procedure is performed for each individual cell in the reference image A, shown for example in Fig. 2 (left), by scanning through compared image B (right) for a suitable matching area that yields the highest CCF value.

Fig. 4 shows typical results for CMC comparisons. The upper diagrams (Fig. 4a) show a CMC comparison of two breech face impressions from the same firearm. 24 out of 29 cells, outlined in black, satisfy all the criteria discussed above and are counted as CMC cells. That is, the cross correlation values between comparable cells are above a chosen threshold and the 24-cell pattern on the left is congruent with that on the right. Only five of the cell pairs, outlined in red, do not satisfy all the CMC critera. The lower diagrams (Fig. 4b) show a CMC comparison of breech face impressions from different firearms. The cells in the right hand image having the largest CCF, when compared with each cell in the left hand image, do not form a pattern that is congruent with the cell pattern on the left.

How many CMC pairs are required so that the two surface topographies can be identified as matching? Ideally, this would be determined after carefully designed experiments and error rate estimations. Threshold values for identification of matching topographies based on breech face impressions will doubtless depend on many aspects of the firearms and the ammunition, including the area of the impressed surface, the quality of the impression marks left by the firing process, the manufacturing method for the breech face resulting in roughness features that form the impression, and the manufacturing method of the cartridge case primer resulting in pre-fire roughness features that can obscure the impression from firing. As a starting point for the current population, we use a single identification criterion C, about midway between distributions of matching and non-matching pairs of images (see Fig. 6). We demonstrate that this criterion works well for the tests that we present in the following sections. After further studies, depending on target error rates for declared matches and non-matches, the single criterion C may evolve into two separated criteria. When applying similar algorithms to other types of tool marks, such as firing pin impressions, different criteria will likely be required [20]. Even for other types of breech face impressions, C would be determined from the population statistics and from estimated targets for error rates. Estimation of error rates for a specific data set are discussed in Section 5.

3. Validation tests: materials and methods

Validation tests of the CMC method have been conducted previously [15–17] using a set of cartridge cases originally created by Fadul et al. [19] for a study of visual firearm identifications by

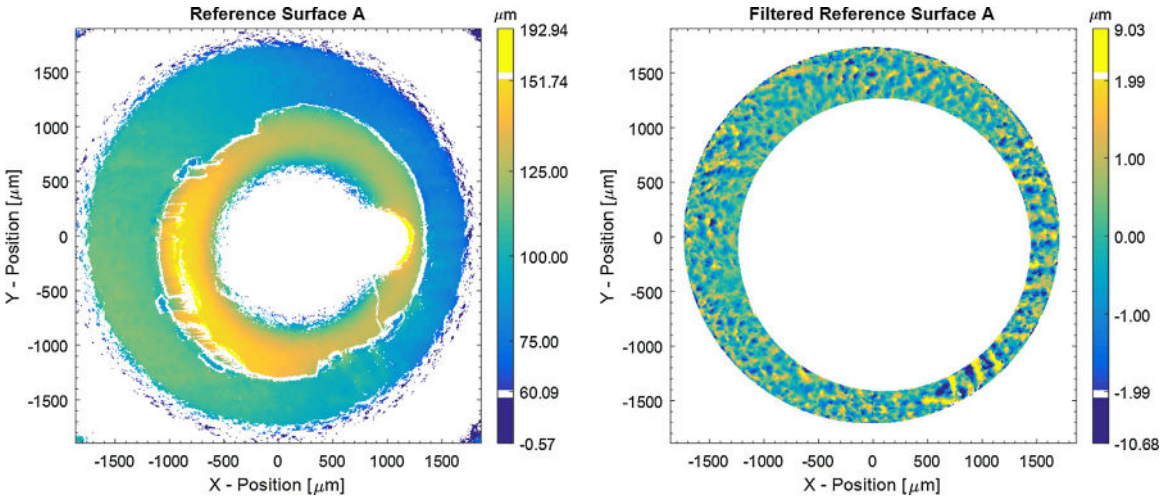

Fig. 5. Color coded topography image of one of the breech face impressions before (left) and after (right) trimming, leveling, and filtering. The prominent annular ridge on the left-hand image is due to flow back into the firing pin aperture. This feature is trimmed away in the right-hand image.

ballistics examiners. The current test is intended mainly to demonstrate the error rate procedure rather than to show application to a real result from case work. The set contains 40 cartridge cases ejected from handguns with ten consecutively manufactured Ruger 9 mm pistol slides. Three slides were used to fire three cartridge cases each, four slides were used to fire four cartridge cases each, and three slides were used to fire five cartridge cases each. The slide is a component of a semi-automatic pistol firing mechanism that absorbs the recoil impact of the cartridge case on its breech face. Thus, the surface topography of the slide's breech face is impressed on the soft primer of the cartridge case upon impact.

Comparisons involving a population of consecutively manufactured firearm parts represent a challenging scenario for accurately identifying bullets or cartridge cases as being fired or ejected from the same firearm. Consecutively manufactured parts can have similar topographic features arising from temporary imperfections in the manufacturing process, such as a worn tool. The presence of these *sub-class characteristics* can lead to false identifications [1]. For this studied set, the breech face was machined using a straight pull step broach [19]. However, the manufacturer finished the surfaces of the slides by sand and bead blasting, a process that should produce random surface topographies [21] with clear individual characteristics and mitigate the effect of sub-class characteristics. The task then for topography measurement and analysis is to distinguish the individual characteristics of the surface impressions from any underlying similarities in consecutively manufactured slides resulting from earlier phases of the manufacturing process. The objective for this set of materials is to draw a correct conclusion of match or non-match with error rate estimation for any pair of topography images drawn from the 40 cartridge cases that were measured. In Section 6, we will discuss the ultimate objective of extrapolating to larger databases and real casework.

The breech face impression topographies on the cartridge cases were measured by a disk scanning confocal microscope described elsewhere [22]. Briefly, illumination from a white light source is reflected from the surface under investigation and is focused onto a pinhole aperture in the disk. If the surface is at the correct height, the reflected light will be focused through the pinhole and a strong optical signal will pass onto the detector. If the surface is not at the correct height, the light arriving at the aperture will be out of focus and little or no signal arrives at the detector. Scanning the surface vertically enables one to determine the surface height at a single lateral location by looking for a maximum in the light transmitted to the detector. The disk contains a large number of pinholes, and spinning the disk serves to provide a lateral scan over the surface.

The confocal microscope was operated with a 10× objective having a numerical aperture of 0.3, a nominal working distance of approximately 10.1 mm, and a field of view of approximately 1.6 mm × 1.6 mm, comprising 512 × 512 pixels. The topography images of the entire breech face impressions were achieved by stitching 3 × 3 fields of view and were approximately 3.9 mm × 3.9 mm with approximately 1240 × 1240 pixels and a nominal pixel spacing of 3.125 μm. The images were down sampled to a pixel spacing of 6.25 μm to improve the speed of the subsequent image correlations. The sample spacing in the vertical scan was 0.2 μm, but the vertical resolution limit of confocal microscopes is significantly smaller than the vertical sample spacing because the signal is interpolated to find a maximum. The root mean square instrument noise was approximately 13 nm, tested by measuring an optical flat at 10× with a long wavelength cutoff of 250 μm.

Before correlating, the images were manually trimmed to extract the breech face impression of interest, yielding, on average, an image size of 3.5 mm × 3.5 mm. Specifically, drag marks and central firing pin impressions with any surrounding flow back ridges are not considered as part of the breech face impression (Fig. 1 vs. Fig. 2). The images were then bandpass filtered to attenuate noise with short spatial wavelengths and attenuate surface form and waviness with long wavelengths thus highlighting individual characteristics. The short wavelength cutoff of the Gaussian filter was 16 μm, and the long wavelength cutoff was 250 μm. Fig. 5 shows a topography image of a breech face impression before and after trimming and filtering.

The topography images were correlated using the CMC method. A total of 780 (=40 × 39/2) image correlations were performed, comprising 63 (=3 × 3 + 4 × 6 + 3 × 10) known matching (KM) and 717 (=780 − 63) known non-matching (KNM) image pair comparisons. The images were divided into cell arrays. There is a trade-off on the chosen cell size. Each cell should be large enough to include a statistically large number of pixels, but there should also be enough cells in the image to distinguish valid and invalid regions. For these tests, the images were divided into arrays of 49 (=7 × 7) cells. Each cell size for the set of correlation tests was chosen to be

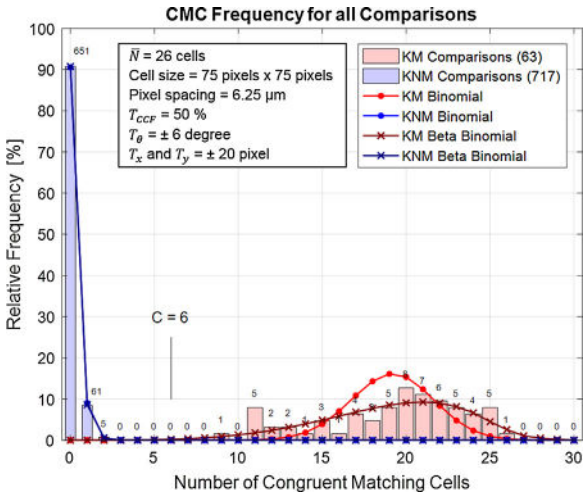

Fig. 6. Relative frequency distribution of image pairs vs. CMC number for 63 KM and 717 KNM image pairs. The KM and KNM distributions are each scaled to their sample size. The red and brown curves represent binomial and beta binomial distribution models, respectively, for the KM data, estimated by Eqs. (8) and (9), respectively. The overlapping blue curves represent the two models for the KNM data (see Section 4). Note that the distribution models are discrete, with the connecting lines drawn for visualization. The number of image pairs having a particular CMC value is shown just above each bar in the histograms.

75 × 75 pixels (nominally 468.75 μm × 468.75 μm), and the range of cell registration angles was restricted to ±30° with respect to their initial orientation.

Although the nominal number, N_{nom}, of compared cell pairs for each topography correlation equals 49, the actual number N of effective cell pairs for each correlation depends on the number of cells in the reference image that contain enough measured pixels for effective correlation. For example, the empty center portion of the surface shown in Fig. 5, corresponding to the firing pin impression, leads to fewer effective correlation cells than N_{nom}. A cell was not used unless at least 10% (approximately 563) of its pixels represented measured data. For this study, the number of evaluated cell pairs in a comparison ranged from 24 to 30, with an average of 26.

One set of test results is shown in Fig. 6 [16]. The cell size a, the pixel spacing, and the thresholds T_{CCF}, T_θ, T_x and T_y are shown on the upper left side. The number of congruent matching cell pairs (CMCs) for the 63 KM topography pairs ranges from 9 to 26; while the number of CMCs for the 717 KNM topography pairs ranges from 0 to 2.

Of the 717 KNM topography pairs, 651 pairs have CMC = 0 (no congruent matching cells). There are only five non-matching topography pairs that have as many as two congruent matching cells, i.e. CMC = 2 (Fig. 6); one of them is shown in Fig. 7A. For the 63 KM topography pairs, only one topography pair has a CMC number as low as 9. This topography pair is shown in Fig. 7B. All the other KM topography pairs have a CMC number ranging from 11 to 26 (Fig. 6). A close-up of one pair of matching cells from Fig. 7B, cell A1 vs. cell B1, is shown in Fig. 8. Their topography similarity is quantified by the maximum value of the normalized cross-correlation function CCF_{max} = 67.6%.

The KM and KNM distributions of Fig. 6 show a significant separation. Additional tests using slightly different versions of the correlation software and different parameter values show similar results without an overlap [16]. Tests performed with optical intensity images of the breech face impressions, instead of 3D topography images, also show similar results without any overlap [17]. In standard binary classifier terms, these results indicate both high sensitivity and specificity [23] for this data set.

The separation between matching and non-matching image pairs shown in Fig. 6 can likely be improved further by designed experiments to optimize the image processing, cell size, parameter threshold values, and registration intervals. The focus here, however, is on reporting an error rate from such results.

4. Error rate analysis and results

4.1. A statistical framework

We seek to develop an approach for estimating the expected error rates of ballistic identifications based on the CMC method. Error rates can be considered from two points of view [24,25]. The first point of view addresses the reliability of the identification procedure. This reliability can be expressed by the false positive and false negative error rates for a given set of KM and KNM samples. The false positive error rate (Fig. 9a) represents the expected frequency or probability of obtaining an erroneous result of identification (declared match) when comparing samples from different sources (KNM). The false negative error rate represents the probability of obtaining an erroneous result of exclusion (declared non-match) when comparing samples from the same source (KM). The false positive and false negative error rates can be used as a measure of the reliability of the identification procedure. In this paper, the false positive and false negative error rates are represented by the "cumulative error rates" E_1 and E_2 (see Eqs. (11) and (13)).

The second point of view addresses the probability of an incorrect conclusion for an identification (declared match) or exclusion (declared non-match). It represents the expected frequency or error rate that a result of either identification or exclusion is false (Fig. 9b). In this paper, false identification and false exclusion error rates are represented by the "individual error rates" R_1 and R_2, respectively (see Eqs. (14) and (15)). This way of describing error rate is of interest during legal proceedings. For example, when a firearms examiner concludes that the evidence and reference items are from the same source, an attorney may ask: "What is the probability that these two items are actually from different sources?" However, error rates in this class depend not only on the reliability of the identification procedure, but also on the ratio of same-source image pairs to different-source image pairs in the population of comparisons relevant to the case [8], or (for this paper) relevant to the validation test (see Section 4.6).

Another way to describe this second point of view is with a Bayesian approach, where the ratio of same-source to different-source populations is cast as prior odds. Multiplying this factor by the likelihood ratio [26,27,28] yields posterior odds, say, for a declared match being correct. The likelihood ratio is the ratio of the probabilities of obtaining a specific comparison result under the competing hypotheses of same-source and different-source samples. Thus, the likelihood ratio expresses the strength of the obtained evidence irrespective of the prior odds. It can be calculated from data and models such as those in Fig. 6.

In this paper, we calculate both the cumulative (false positive and false negative) error rates E_1 and E_2, and the individual (false identification and false exclusion) error rates R_1 and R_2 from the distributions obtained with the CMC method. Thus, the cumulative false positive error rate E_1 (Eq. (11)) represents the probability of obtaining a CMC score larger than or equal to the identification criterion C, when comparing samples from different sources (KNM). Alternatively, for a specific CMC comparison score, we calculate individual error rates of identifications R_1 (Eq. (14)) and exclusions R_2 (Eq. (15)). For example, when CMC = 15, the individual identification error rate R_1 represents the probability that an identification based on a CMC score of 15 is a falsely declared match.

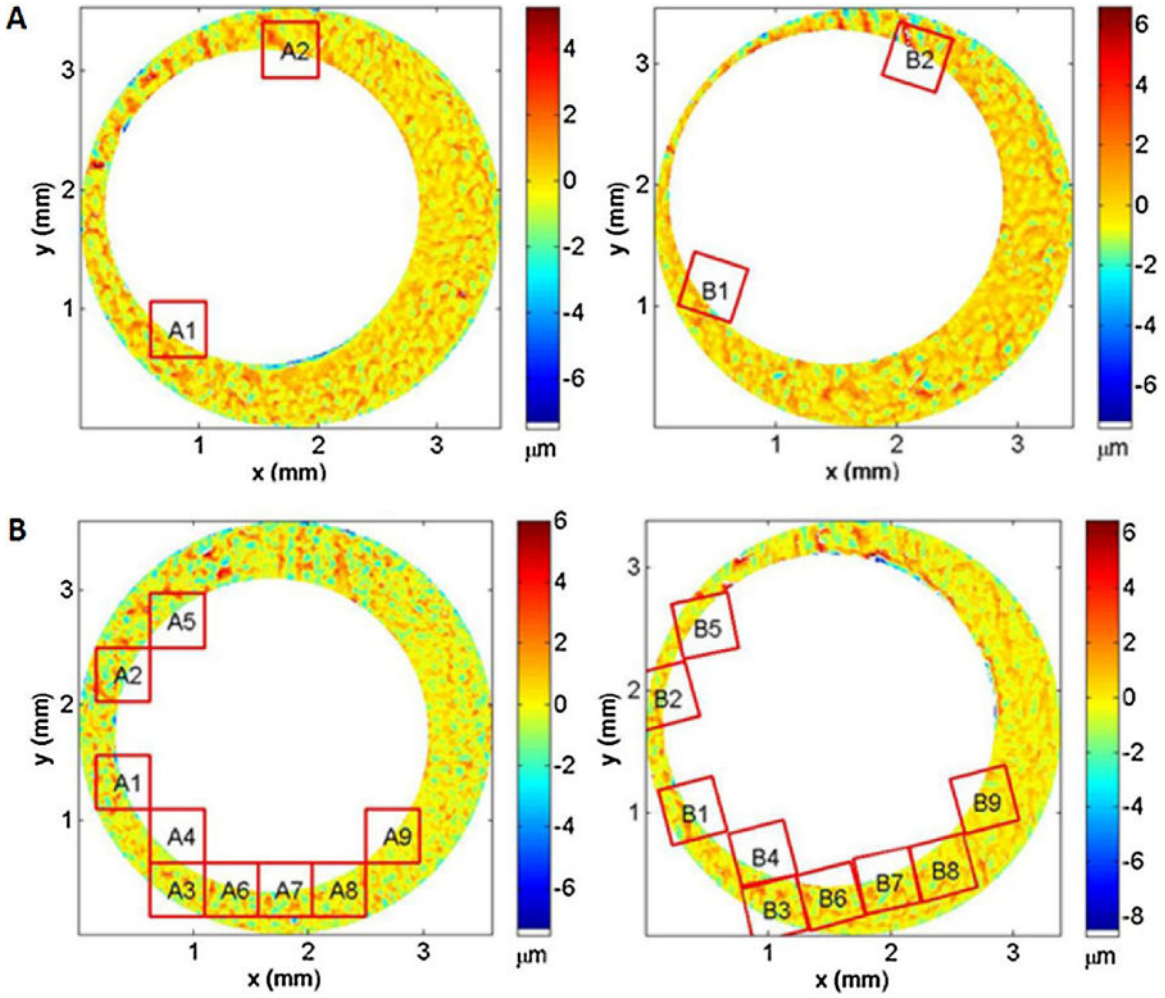

Fig. 7. Depiction of congruent matching cells for two correlated topography pairs. For the 717 KNM topography pairs, only five pairs have a CMC value as high as 2; one of these image pairs is shown in (A). For the 63 KM topography pairs, only one has a CMC value as low as 9; that pair is shown in (B). The cell pattern A1–A9 on the left of Fig. 7B is congruent with the cell pattern B1–B9 on the right. The filtered surface topographies of the breech face impressions are depicted by the color scale of the diagram.

The large number of cell correlations associated with the CMC method using multiple identification parameters facilitates a statistical approach to modeling error rates. The CMC method is based on pass-or-fail tests of individual cell pairs comprising an image pair of breech face impressions. In this section, we develop statistical models for the probability distribution of the number of successful tests in a comparison, i.e., the CMC numbers of KM and KNM comparisons. After estimating model parameters from experimental results for KM and KNM comparisons, the models are applied to estimate potential error rates.

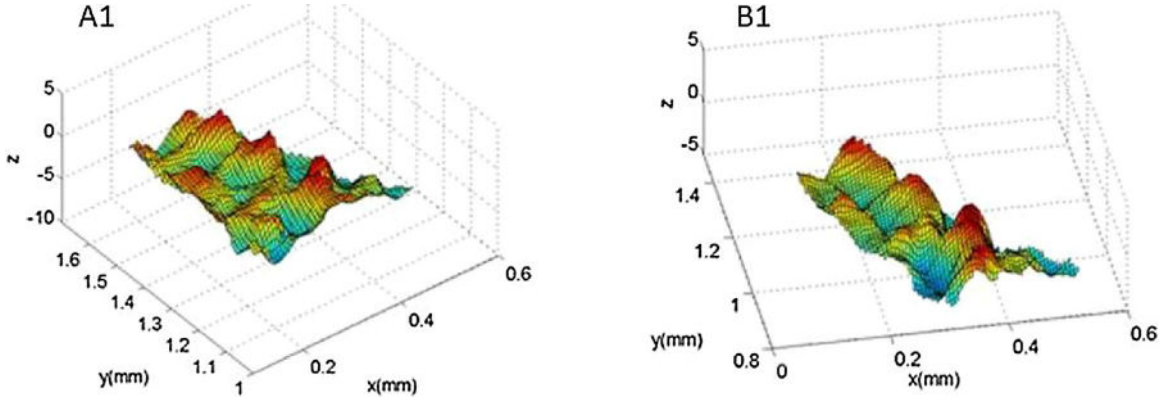

Fig. 8. Topography comparison of KM cell pair A1 and B1 from the KM image pair of Fig. 7B. Common topography features are apparent. The normalized correlation value, CCF_{max}, is 67.6 %.

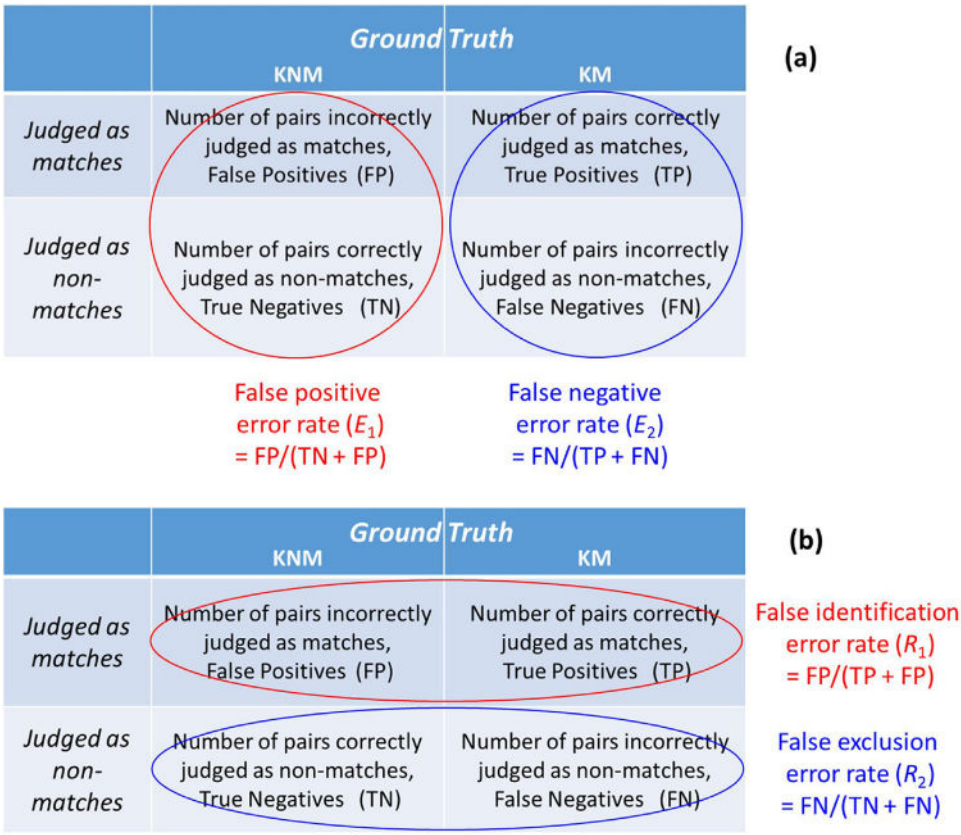

Fig. 9. Two points of view for describing error rates for firearms identifications.

4.2. A binomial probability model for the distribution of CMCs

For a pair of images, N represents the number of correlated cell pairs. If, for example, there are 49 cells in the array of the reference image ($N_{nom} = 49$) but nine of those have an insufficient fraction of pixels with measurement values, then N is reduced to 40. For a given correlated cell pair, a random variable X represents the outcome of the CMC method for that cell pair. When the CMC method determines that the cell pair is part of the set of congruent matching cells, i.e. when its correlation value CCF_{max} is greater than a chosen threshold T_{CCF} and the registration angle θ and x, y registration positions are within the chosen threshold limits T_θ, T_x and T_y, then $X = 1$; otherwise $X = 0$. We use the symbol P to represent probability in general and the symbol p to represent the probability that $X = 1$. That is, $P(X = 1) = p$, and $P(X = 0) = 1 - p$.

We now make two key approximations that will be revisited in later sections: (1) the comparisons between cell pairs are independent from each other, and (2) each cell pair comparison for the KNM images has the same probability $p = p_{KNM}$ to qualify as a CMC and each cell pair comparison for the KM images has the same probability $p = p_{KM}$ to qualify as a CMC. Thus, for the first image pair with N_1 correlated cell pairs, we have a sequence of Bernoulli trials [29], $X_{11}, \ldots, X_{1N_1}$, which are independent from each other but have a common probability, p_{KNM} or p_{KM}. We denote the number of successful trials, the CMC number, for the first image pair by Y_1. That is, $Y_1 = \sum_{i=1}^{N_1} X_{1i}$. Under the stated assumptions, Y_1 is a binomially distributed random variable [29], namely, $Y_1 \sim Bin\ (N_1, p)$. The functional form of *Bin* is shown later in Eq. (7). Similarly, for M KNM or KM image pairs, we have $Y_1, \ldots, Y_M$. Assuming that $\{Y_j, j = 1, \ldots, M\}$ are independent from each other, we have a sequence of binomially distributed random variables, $Y_j = \sum_{i=1}^{N_j} X_{ji}$, for $j = 1,\ldots,M$ and $Y_j \sim Bin\ (N_j, p)$. In addition, we can state

$$\sum_{j=1}^{M} Y_j \sim Bin\left(\sum_{j=1}^{M} N_j, p\right). \tag{1}$$

For observed values of $\{Y_j, j = 1, \ldots, M\}$, the maximum likelihood estimator of p is given by [30]:

$$\hat{p} = \frac{\sum_{j=1}^{M} Y_j}{\sum_{j=1}^{M} N_j} = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j} X_{ji}}{\sum_{j=1}^{M} N_j} \tag{2}$$

Therefore, for the sub-population of KNM image comparisons, the false positive cell probability, denoted by p_{KNM}, is the probability that a KNM cell pair comparison results in a CMC. Likewise, for the sub-population consisting of KM image pairs, the false negative cell probability is denoted by $(1 - p_{KM})$.

To estimate p_{KNM} and p_{KM} from the data, we apply Eq. (2) to the sub-population of 717 KNM image pairs and the sub-population of 63 KM image pairs. For each sub-population, we estimate p by counting all the CMC cell pairs that pass the four threshold criteria for a match:

$$\hat{p}_{KNM} = \frac{\text{Number of KNM CMC cell pairs}}{\text{Total number of evaluated KNM cell pairs}}, \tag{3a}$$

$$\hat{p}_{KM} = \frac{\text{Number of KM CMC cell pairs}}{\text{Total number of evaluated KM cell pairs}}. \tag{3b}$$

For the test results depicted in Fig. 6, the estimates obtained from Eq. (3) are:

$$\hat{p}_{KNM} = 71/18859 = \mathbf{0.003765},$$
$$\hat{p}_{KM} = 1207/1628 = \mathbf{0.7414}.$$

It is also instructive to plot the experimental frequency distributions of registered KM and KNM cell pairs with respect to each CMC identification parameter to observe how the overlap between KM and KNM cell pairs is eliminated when all four identification parameters are combined. By this method, the value of $\hat{p}_{KNM}$ and $\hat{p}_{KM}$ can also be calculated with the same results (see Appendix A).

To evaluate the model, we compare the observed frequency of correlations as a function of CMC number to the respective modeled frequency. For KNM correlations, the observed frequency distribution is obtained as:

$$f_{KNM}(CMC = h) = \frac{\text{Number of KNM image pair correlations with } CMC=h}{\text{Total number of KNM image pair comparisons}}.$$

In Fig. 6, the observed frequency distribution is depicted by the blue histogram. The respective modeled frequency distribution, depicted by the blue curve, is obtained as:

$$\hat{f}_{KNM}(CMC = h) = \sum_{j=1}^{M} Bin(h|N_j, \hat{p}_{KNM})/(\text{Total Number of KNM correlations}),$$

where the summation of the binomial probability values is performed over all KNM comparisons. If all correlations have the same number of evaluated cells N, Eq. (6) would be simplified to:

$$\hat{f}_{KNM}(CMC = h) = Bin(h|N, \hat{p}_{KNM}) = C_N^h \cdot \hat{p}_{KNM}^h \cdot \left(1 - \hat{p}_{KNM}\right)^{N-h}, \quad (7)$$

where the binomial coefficient C_N^h is the number of possible combinations of h out of N elements.

Likewise, for KM image correlations, the modeled distribution, depicted by the red curve in Fig. 6, is:

$$\hat{f}_{KM}(CMC = g) = Bin(g|N, \hat{p}_{KM}) = C_N^g \cdot \hat{p}_{KM}^g \cdot \left(1 - \hat{p}_{KM}\right)^{N-g}. \quad (8)$$

4.3. Re-assessing the binomial model

The binomial model described above contains the assumption that a single value (p_{KNM}) characterizes the probability that a pair of cells from KNM images will pass all criteria and qualify as a false positive CMC cell pair. The resulting model fits the KNM data quite well (see Fig. 6, blue line), and theoretically, we expect the use of a single false positive cell probability p_{KNM} to be a good approximation for KNM data. If two cells were from images of breech face impressions from different firearms, the fact that they appear to qualify as a CMC cell pair is likely driven by random, non-selective factors as long as subclass characteristics, the carry-over of pre-fire tool marks, and systematic measurement errors are not significant factors in the evaluation.

The situation is more complicated for cell pairs of KM images. Variations in firing conditions, firearm wear, and contaminants cause variations in the tool marks imparted on the cartridge case and the domain of the breech face impression area. These effects and others cause variations in the size and quality of the common valid correlation areas of a KM image pair comparison, which may cause variations in the probability p_{KM} of the cell pairs to qualify as CMCs. For comparisons of KNM samples, these effects simply add additional random factors to a comparison result, which is already largely driven by random factors and which is unlikely to cause major variations in the false positive cell probability p_{KNM}. Variation in the false negative cell probability ($1 - p_{KM}$) is consistent with the higher dispersion of the observed CMC numbers for KM comparisons than predicted by the binomial model (red curve in Fig. 6), which is based on the assumption of a single value of p_{KM} for all KM comparisons. To account to some extent for these observations, we relax the assumption of the same cell trial success probability p_{KM} for all KM comparisons as described below.

4.4. A beta-binomial probability model for the distribution of CMCs

In this approach, we still assume that a CMC image comparison can be modeled as a set of independent Bernoulli trials characterized by the same cell trial success probability. However, we now allow the cell trial success probability p to vary between image comparisons. Here we assume that the parameter p can be modeled as a random variable with a beta distribution. The choice

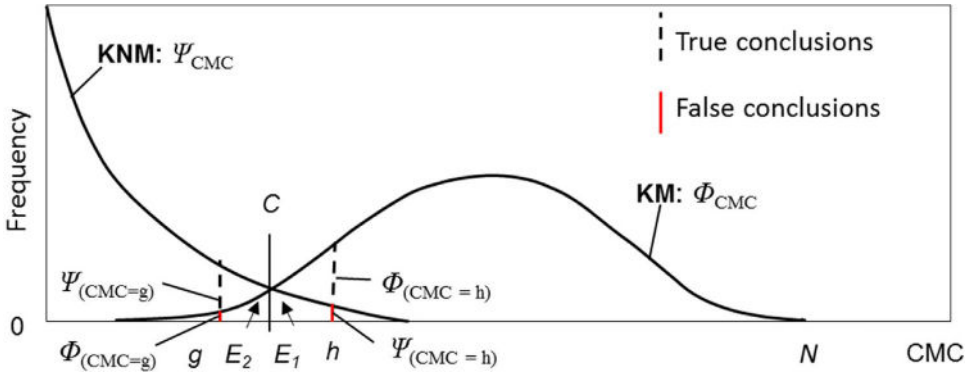

Fig. 10. Conceptual diagram of the CMC probability mass functions for KM and KNM comparisons, Φ_{CMC} and Ψ_{CMC}. To illustrate clearly the listed quantities, the schematic depicts the discrete probability distributions as continuous density functions that overlap much more than they would be expected to in practice. The regions E_1 and E_2 under the curves represent cumulative false positive and false negative error rates. For each "matching" conclusion, $h \geq C$, and "non-matching" conclusion, $g < C$, there are probabilities for both "True" and "False" conclusions as demonstrated by the black dashed bars (which extend down to the x-axis) and the solid red bars, respectively.

of the beta distribution has several advantages. The beta distribution is defined by two parameters, α and β, which allow for a wide range of distribution shapes. The domain of the beta distribution is restricted to the interval [0,1], which makes it a convenient distribution to model probabilities. In a Bayesian framework, the beta distribution is a conjugate distribution of the binomial distribution, yielding an analytical expression for the resulting compound beta-binomial distribution [31]. Finally, the resulting beta-binomial distribution can approximate the binomial distribution to arbitrary precision when needed [32].

Like Section 4.2, we model an image correlation j with N_j evaluated cell pairs as a sequence of Bernoulli trials, $X_{j1}, \ldots, X_{jNj}$, which are independent from each other and have a common success probability $p = p_j$. The CMC number of the comparison, i.e., the sum of the trial outcomes X_{ji}, is Y_j, which for a given $p = p_j$ has a binomial distribution $Y_j|p_j \sim Bin\left(N_j, p_j\right)$. For M image comparisons, we have $Y_j|p_j \sim Bin\left(N_j, p_j\right)$, for $j = 1$ to M, where p_j now has a beta distribution, i.e., $p_j \sim Beta(\alpha, \beta)$ with positive α and β. The probability mass function of the resulting beta-binomial random variable Y for given values of N, α, and β is given by Ref. [32]:

$$P(Y = k|N, \alpha, \beta) = C_N^k \frac{B(k+\alpha, N-k+\beta)}{B(\alpha, \beta)}, \tag{9}$$

where $B(\alpha, \beta)$ is a beta function with parameters α and β, and k is a CMC value.

For the KM and KNM correlation results discussed in Section 3, we obtained maximum likelihood estimates of the parameters α and β using the algorithm described by Smith [32]. The respective values are: $\hat{\alpha}_{KNM} = 2.15$ and $\hat{\beta}_{KNM} = 569.1$ for the KNM comparisons and $\hat{\alpha}_{KM} = 6.55$ and $\hat{\beta}_{KM} = 2.29$ for the KM comparisons. The modeled frequency distributions for the KM and KNM CMC results are depicted by the curves in Fig. 6. The beta-binomial model for the KNM comparisons is nearly indistinguishable from the respective binomial model. For the KM comparisons, on the other hand, the beta-binomial model shows a significant improvement in the ability to model the dispersion of the experimental results.

4.5. Error rate estimation

The estimated false positive cell probability, $\hat{p}_{KNM}$ for the binomial model of the KNM cells and the parameters, $\hat{\alpha}_{KM}$ and $\hat{\beta}_{KM}$, of the beta binomial distribution for the KM cells are inserted into the respective models to estimate potential error rates for cartridge cases fired from different firearms (KNM) and the same firearm (KM) under similar conditions. Fig. 10 shows a conceptual diagram for two CMC probability mass functions, Φ_{CMC} and Ψ_{CMC}, for KM and KNM topography pairs, respectively. As discussed in Section 4.2, the probability mass function Ψ_{CMC} for KNM comparisons is modeled as:

$$\Psi\left(CMC = h|N, \hat{p}_{KNM}\right) = Bin(h|N, \hat{p}_{KNM}) = C_N^h \cdot \hat{p}_{KNM}^h \cdot \left(1 - \hat{p}_{KNM}\right)^{N-h}. \tag{10}$$

The cumulative false positive error rate E_1 is given by the sum of the probability mass function values Ψ_{CMC} for CMC values between C and N:

$$\begin{aligned} E_1 &= \sum_{CMC=C}^{CMC=N} \Psi_{(CMC)} = \Psi_{(CMC=C)} + \Psi_{(CMC=C+1)} + \cdots + \Psi_{(CMC=N)} \\ &= 1 - \left(\Psi_{(CMC=0)} + \Psi_{(CMC=1)} + \cdots + \Psi_{(CMC=C-1)}\right). \end{aligned} \tag{11}$$

The cumulative false positive error rate E_1 is determined by three factors: the number of correlation cell pairs N in a comparison, the numerical identification criterion C of the CMC method, and the false positive cell probability $\hat{p}_{KNM}$ of each correlated cell pair estimated from Eq. (3a).

Similarly, the probability mass function Φ_{CMC} for KM correlations (Fig. 10) is modeled as:

$$\Phi(CMC = g|N, \alpha, \beta) = C_N^k \frac{B(g+\alpha, N-g+\beta)}{B(\alpha, \beta)}. \tag{12}$$

The cumulative false negative error rate E_2 is given by the sum of the probability mass function values Φ_{CMC} for CMC values between 0 and $(C-1)$:

$$E_2 = \sum_{CMC=0}^{CMC=C-1} \Phi_{(CMC)} = \Phi_{(CMC=0)} + \Phi_{(CMC=1)} + \cdots \Phi_{(CMC=C-1)}. \tag{13}$$

We note again the approximations underlying the binomial distribution model for the KNM image pairs:

1) the comparisons between cell pairs are independent from each other, and
2) each cell pair comparison for the KNM images has the same probability $p = p_{KNM}$ to qualify as a CMC.

The second condition is partially relaxed for the KM image pairs with the introduction of the beta binomial distribution. Each cell pair comparison within a KM image pair is still assumed to have the same p_{KM} value, but the beta binomial distribution, with parameters α and β, is introduced to model the distribution of p_{KM} values for different KM image comparisons. The error rates estimated from the values of $\hat{p}_{KNM}$, $\hat{a}$, and $\hat{\beta}$ are random variables themselves and their uncertainties should also be assessed [6,7].

The cumulative false positive and false negative error rates E_1 and E_2 associated with the data of Fig. 6 may be estimated from Eqs. (11) and (13) using the known number of effective cells N (the average number of evaluated cells in the image comparisons is $N = 26$), the CMC identification criterion C (=6 here) and the estimated parameters $\hat{p}_{KNM}$ (Eq. (4)), $\hat{a}$, and $\hat{\beta}$. For the 717 KNM comparisons, the cumulative false positive error rate is $E_1 = 6.1 \times 10^{-10}$, which represents the sum of the Ψ_{CMC} probabilities between 6 and N when using the binomial model. The cumulative false negative error rate is $E_2 = 2.1 \times 10^{-3}$, which represents the sum of the Φ_{CMC} probabilities between 0 and 5 when using the beta binomial model. These error rates will vary depending on the specific data population, the distribution models, and all the parameters chosen for the correlation, such as the cell size. Error rates for different models are discussed below.

4.6. Individual error rates for identifications and exclusions

Fig. 10 also illustrates the probabilities of true and false conclusions by the dashed black bars (which extend down to the x-axis but are partially hidden) and the solid red bars for specific CMC values h and g. These probabilities can be used to calculate likelihood ratios for various scores, that is, the ratio of the likelihoods of obtaining the score under two competing hypotheses (matching or non-matching samples) [25–28]. We define here the individual false identification probability R_1 that an identification is false as the probability that an image pair is non-matching when its CMC value appears in the matching region with a specific score h $(h \geq C)$, and conversely the individual false exclusion probability R_2 that an exclusion is false as the probability that an image pair is matching when its CMC value appears in the non-matching region with a specific score g $(g < C)$. For our experiment,

R_1 can be estimated as:

$$R_{1(CMC=h)} = \frac{K \times \Psi_{(CMC=h)}}{K \times \Psi_{(CMC=h)} + \Phi_{(CMC=h)}}, \quad (h \geq C), \tag{14}$$

where K is the ratio of the sample sizes of KNM and KM topography image pairs. In a Bayesian approach, K represents the prior odds against obtaining a match in the current population of 40 breech face images before conducting the forensic test. For this study, K is equal to 717/63 (=11.38) under the condition that we randomly select two cartridge cases from our materials set before comparing their topographies to determine whether they are matching.

Conversely R_2 can be estimated by

$$R_{2(CMC=g)} = \frac{\Phi_{(CMC=g)}}{\Phi_{(CMC=g)} + K \times \Psi_{(CMC=g)}}, (g < C). \tag{15}$$

The parameters, R_1 and R_2, could be useful when addressing questions, such as "given the conclusion of identification based on a CMC score h ($h \geq C$), what is the probability that the cartridge cases were actually ejected from different firearms (individual false identification error rate R_1)?", or "given the conclusion of exclusion based on a CMC comparison score g ($g < C$), what is the probability that the two cartridge cases were actually ejected from the same firearm (individual false exclusion error rate R_2)?" In Bayesian terms, R_1 and R_2 represent posterior probabilities of erroneous identifications and exclusions, respectively. The models produce very little overlap between the KNM and KM distributions. Even at the extrema of the experimental distributions, the modeled values are small. The value of R_1 for CMC = 9, computed for actual cell number $N = 25$, is 3.3×10^{-13}, and the R_2 value for CMC = 2, for $N = 26$, is 2.0×10^{-3}. The small estimated value for the individual false identification probability is largely due to the rapid decline in the modeled probability mass function curve Ψ_{CMC} for KNM comparisons, which matches well with the experimental data (Fig. 6). The larger values for the individual false exclusion probability follow from the fact that the distribution is expected to be wider for the reasons discussed in Section 4.3. For realistic databases with many entries of firearms and ammunition, even when classified according to model and manufacturer, the overlap of KM and KNM distributions can become significant and the error rates will likely increase significantly.

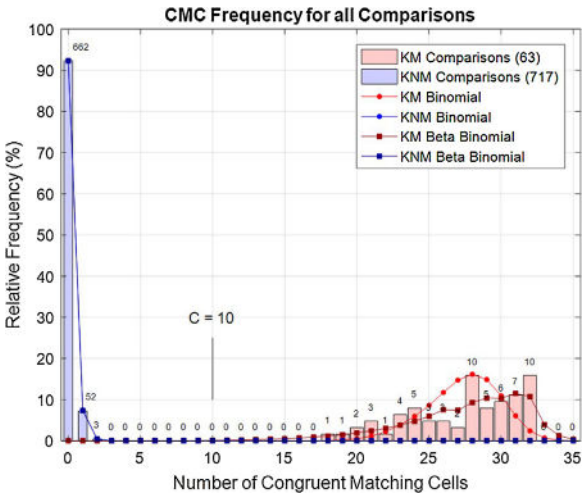

Fig. 11. Relative frequency distribution of CMC numbers for KM and KNM image pairs obtained with modified software and correlation parameters. The red and brown curves represent the binomial and beta-binomial distribution models for the KM data. The overlapping blue curves represent the two respective models for the KNM data. The models are estimated from the histogram data. Note that the distribution models are discrete, with the connecting lines drawn for visualization.

5. Further analysis and results

5.1. Software development

We repeated the analysis with several changes to the algorithms and correlation parameters. First, the images are no longer down sampled, resulting in an average pixel spacing of 3.125 μm instead of 6.25 μm. Second, the low-pass and high-pass filters are now, respectively, zeroth order and second order Gaussian regression filters [33] to attenuate filtering edge effects at the image domain boundaries. Their cutoffs are now, respectively, 25 μm and 250 μm. Third, cell registration was improved through a combination of Fourier-based and direct optimization of the normalized cross correlation value at overlapping image areas as a function of sample translation and rotation. Fourth, the effect of spurious local registration optima was reduced by increasing requirements for the minimum percentage of measured pixels in a cell from 10% to 25% and by decreasing the size of the search domain to ±0.75 mm for sample translations. Finally, we optimized the initial placement of the cells on the donut shape of the reference image to ensure maximum coverage of the respective sample domain, resulting in an increase of the average number of evaluated cells.

Fig. 11 shows the results of the revised analysis. The chosen cell size remains the same, but now comprises 150 × 150 pixels. The x–y registration thresholds remain at ±20 pixels (±62.5 μm considering the pixel spacing is 3.125 μm). The search range of registration angles and x-y displacements was limited to ±45° and ±750 μm, respectively. The number of evaluated cells per comparison varies between 26 and 35 cells, with an average of 31 cells. For the 63 KM cartridge pairs, the number of CMCs ranges from 18 to 32, and for the 717 KNM cartridge pairs, the number of CMCs remains in the same range (0–2) as that shown in Fig. 6.

A cutoff between declared matches and non-matches is chosen at $C = 10$, midway between the extremes of the two distributions. Using the binomial model with 31 cells, the estimated cumulative false positive error rate for a CMC cutoff of 10 decreases to $E_1 = 5.6 \times 10^{-19}$, and using the beta-binomial model, the cumulative false negative error rate decreases to $E_2 = 1.2 \times 10^{-3}$ with respect to the analysis for Fig. 6 described in Section 4.5. The value of R_1 for $h = 18$, the lower extreme of the KM data, is 1.5×10^{-35} (N = 32), and the value of R_2 for $g = 2$, the higher extreme of the KNM data, is 1.7×10^{-4} (N = 30). The KNM binomial distribution is based on an estimated cell success probability of $\hat{p}_{KNM} = 2.58 \times 10^{-3}$. The KM

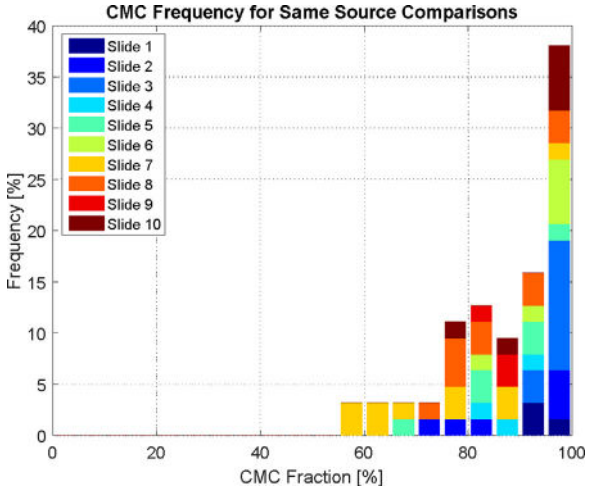

Fig. 12. Frequency distribution of KM image comparisons (Fig. 11) for the fraction of evaluated cells that were classified as CMC cells.

beta binomial distribution used is based on: $\hat{\alpha}_{KM} = 6.24$ and $\hat{\beta}_{KM} = 0.844$.

5.2. Clustering

It is important to note that the image comparisons in our experiments are not independent because each cartridge case and slide was used in more than one comparison. This lack of independence can lead to clustering effects in the experimental frequency distributions. For example, if one cartridge case is poorly marked, comparisons of this cartridge case with others fired from the same slide are affected in a similar manner. Of the five comparisons yielding a CMC number of 11 in Fig. 6, three involved the same cartridge case and four involved the same firearm slide. This clustering is currently not addressed by our models, which assume independence of the various comparisons in our set. Note that we do not expect clustering to be apparent with non-matching distributions.

The KM data of Fig. 11 are plotted again in Fig. 12 with the identity of the ten slides indicated by different colors. In Fig. 12, the histogram abscissa represents the percentage of evaluated cells in a correlation that were classified as CMC cells. There is clearly a difference between images from different slides. For example, all KM image comparisons involving samples from slides 1 or 3 yield CMC numbers exceeding 90% of the number of evaluated cells. For slide 7, on the other hand, half the comparisons yielded CMC numbers between 55% and 70% of evaluated cells.

The differences are illustrated further in Fig. 13 by a scatter plot of the KM images' CMC fractions for each slide. The differences in the spread of the CMC fractions from one slide to another suggest differences in the capacity of the slides to impress similar topographies on the breech faces. Clearly, the five images of slide 3 when compared among themselves to yield the ten KM CMC fractions on the right-hand side of Fig. 13 have consistently higher similarity than the five images of slide 7 whose ten KM CMC fractions are shown on the left-hand side.

Fig. 14 shows one breech face impression topography, image A, correlated with two other cartridge case topographies, B and C, all fired from a firearm using slide 7. In the first correlation, A vs. B, all 32 evaluated cells were classified as CMC cells. In the second correlation, A vs. C, only 19 cells were classified as CMCs. Some of the failed cells can be explained by insufficient trimming of the firing pin impression area on the inside edge in the upper right quadrant of image C. However, overall, stronger matching features are present in correlation A–B than A–C, as reflected by the difference in the average normalized CCF values of the CMC cells of 86 % vs. 66 %.

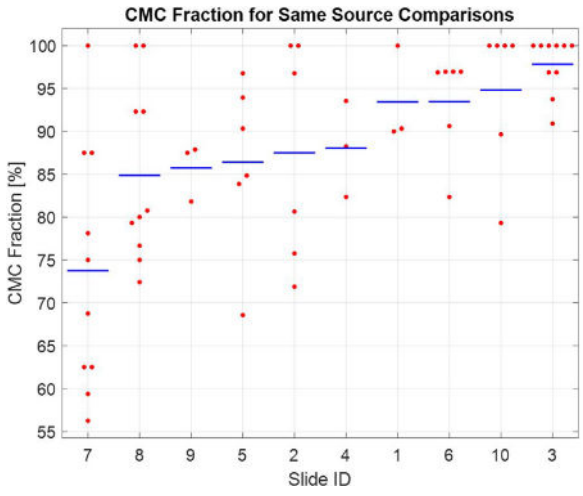

Fig. 13. Scatter plot of CMC fractions for different slides using the data shown in Fig. 12. The blue line represents the mean value for each slide. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

The lack of independence is investigated further in Appendices B and C. Appendix B shows a preliminary calculation that takes into account potential correlations between cell pairs. Appendix C shows results for independent image pairs.

5.3. Testing the models

We evaluated the models derived in Section 4 on a different set of cartridge cases created by Weller et al. [34]. The cartridge cases were obtained from another set of eleven firearm slides produced by the same manufacturer using the same process as that of the Fadul set described in Section 3. The Weller set consists of 95 Winchester cartridge cases, 9 cartridge cases each for 10 consecutively manufactured slides and 5 cartridge cases for one extra slide that was manufactured using the same process but not consecutively with the others. This resulted in 370 KM image pairs and 4095 KNM pairs, a data set that is significantly larger than the Fadul data of Figs. 6 and 11. The measurement procedure, image processing parameters, and the CMC threshold parameters were the same as those discussed in Section 5.1. For this dataset, the domain of the trimmed breech face impressions typically consists of thicker "donut" areas. This resulted in a larger number of evaluated cells per comparison, ranging from 28 to 49 cells with an average of 42 cells.

Fig. 15 shows the relative frequency distribution of the observed CMC numbers for the 370 KM and 4095 KNM image correlations. Once again, the KNM and KM data are widely separated, and with an identification criterion (C) equal to 10, there are no false positive or false negative results. For the 370 KM cartridge pairs, the number of CMCs ranges from 21 to 47; for the 4095 KNM cartridge pairs, the number of CMCs again ranges from 0 to 2. Also shown in Fig. 15 are the modeled frequency distributions for the CMC results. For the average number of evaluated cells, 42, the cumulative false positive and false negative error rates are, respectively, E_1 $(h \geq 10) = 4.5 \times 10^{-21}$ and E_2 $(g < 10) = 7.5 \times 10^{-09}$.

We have also compared the Weller data [34] with models using the same $\hat{p}_{KNM}$, $\hat{a}$, and $\hat{\beta}$ parameter values that were estimated from the Fadul data set discussed in Section 4.4. Fig. 16 shows good agreement with the KNM data and reasonable agreement with the KM data. For the average number of evaluated cells, 42, the cumulative false positive and false negative error rates are $E_1 = 1.8 \times 10^{-17}$ and $E_2 = 2.2 \times 10^{-4}$. This evaluation shows the consistency of the results and suggests the general applicability of the binomial model for describing KNM data and the applicability of the beta binomial model for describing KM data from these types of manufactured slides.

6. Summary and discussion

6.1. Summary observations on the test results

Reporting error rates for firearm identification is a fundamental challenge in forensic science. We have developed a practical statistical approach to estimating error rates based on the CMC method. Initial results for correlations of the breech face impressions on cartridge cases ejected from two different sets of pistol slides of the same brand show wide separation between the CMC scores of KM and KNM samples. The slides in each set were consecutively manufactured using the same process. Models

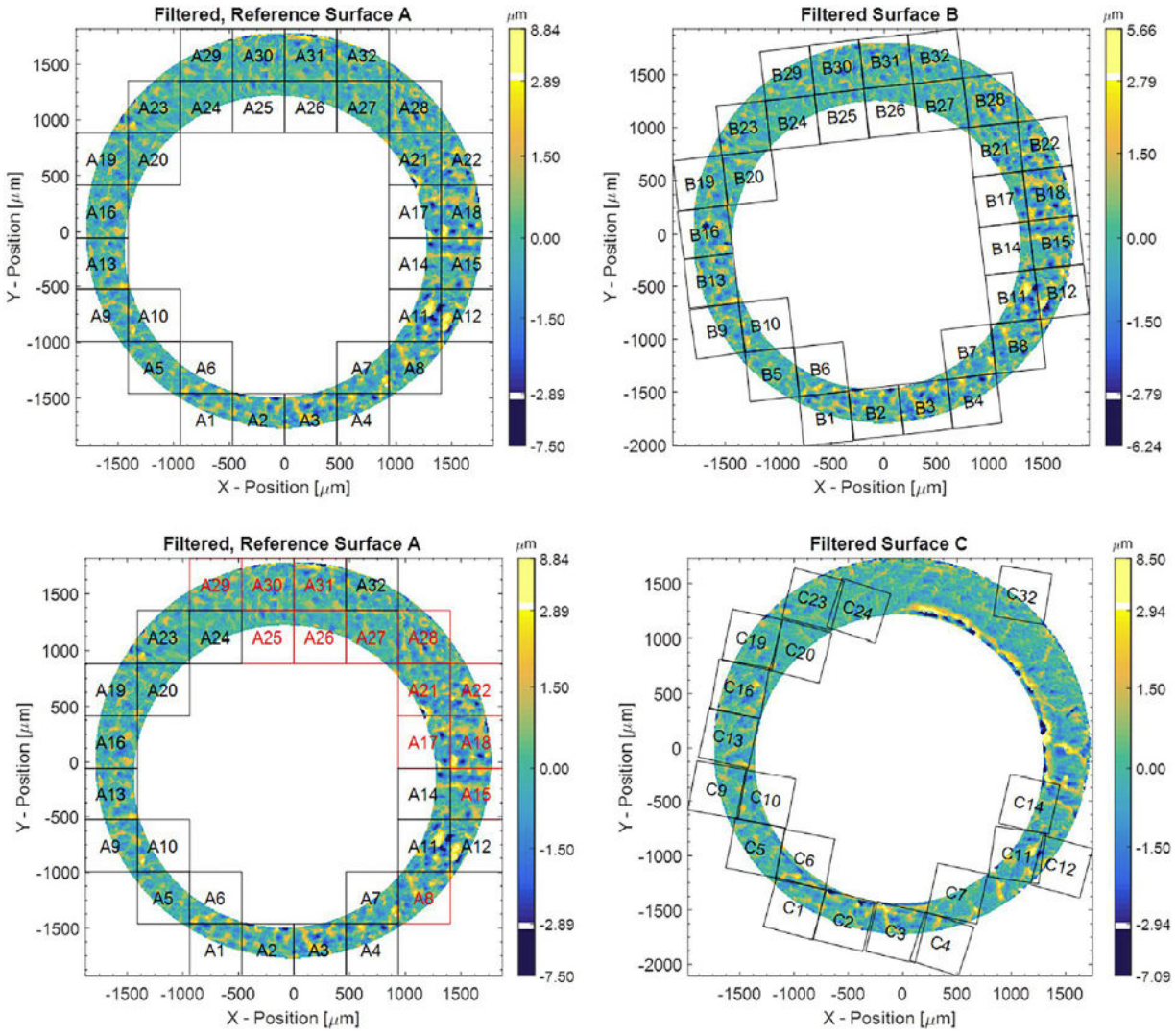

Fig. 14. Top-comparison of the breech face impression of cartridge cases A and B fired from a firearm using slide 7. Bottom-comparison of the breech face impression of cartridge cases A and C fired also from a firearm using slide 7.

for the frequency distribution of the CMC scores show good agreement with the experimental data and yield very small error rates for erroneous classifications, in particular for false positive identifications.

The error rate estimates are derived from models for the probability distribution of the similarity metric, the CMC score, for KNM and KM image correlations. The models were developed for data from the smaller population, then applied successfully to the larger population. The initial binomial model was based on two key approximations: (1) the CMC cell trials in a comparison are statistically independent, and (2) all the KNM cell pairs in our population have, the same false positive probability p_{KNM}. Barring the presence of sub-class characteristics, which is unlikely for sand-blasted breech face surfaces, these approximations seem reasonable. The resulting estimated binomial distribution Ψ_{CMC} for the respective CMC scores matches the experimental KNM data quite well for both the Fadul and Weller datasets (blue lines in Figs. 6,11 and 15). From a legal perspective, the KNM distribution is critical for ballistics identifications as it can be used to yield a probability of false positives (false identifications), which are to be avoided at almost any cost.

On the other hand, the binomial distribution for KM image comparisons (red line in Figs. 6 and 11) shows a lower dispersion than the experimental data, resulting in error rate estimates that are too low. This is not surprising, as variations in firing conditions, wear, and contaminants commonly cause variations in the tool marks imparted on the cartridge cases. These effects cause variations in the size and quality of the valid regions on matching pairs, which, in turn, are likely to cause variations in the probability p_{KM} of a matching cell pair to be qualified as a CMC. The beta-binomial model described in Section 4 allows the average cell trial success probability p to vary from image comparison to image comparison. This approach improves agreement between the modeled and experimental KM CMC distributions (brown lines in Figs. 6 and 11).

We emphasize that the estimated error rates in this report are specific to the sets of firearms studied here and are not applicable to other firearm scenarios. Furthermore, the presented models do not address correlations between the experimental comparison results due to the use of a sample image or firearm slide in more than one comparison. Some differences between the modeled and experimental results, such as the CMC = 11 values in Fig. 6, might be

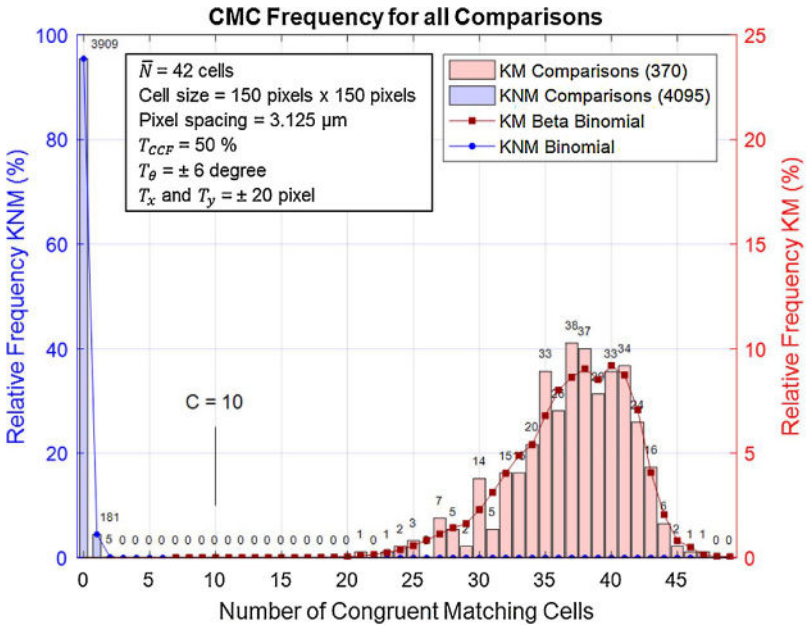

Fig. 15. Relative frequency distribution of CMC numbers for KM and KNM image pairs of the Weller dataset [34]. The red curve represents the beta-binomial distribution model for the KM data. The brown curve represents the binomial model for the KNM data. Note that the distribution models are discrete, with the connecting lines drawn for visualization. The right-hand scale for the KM data is magnified by a factor of four to show differences more clearly.

attributable to this effect. We discuss this issue further in Appendices B and C.

6.2. A scenario for case work

The results reported here were derived from a small data set of 780 image pairs and tested successfully on a larger data set of 4465 image pairs. These are far smaller populations than those anticipated in real forensic science case work. In addition, the question we posed is necessarily different from the issues associated with the traditional prosecution and defense hypotheses in forensic science case work. The question we posed was: If two cartridge cases are selected randomly from a set of cartridge cases that have been characterized for their breech face topography, can we compare the topographies and accurately identify whether the two cartridge cases were fired by the same firearm and estimate the error rate?

We believe yes, this has been demonstrated for the specific population here. However, can such a method ever be applied to real case work where potentially hundreds of thousands of firearms of different manufacture could be considered as possible sources of a piece of evidence in a crime? Because of the inherent

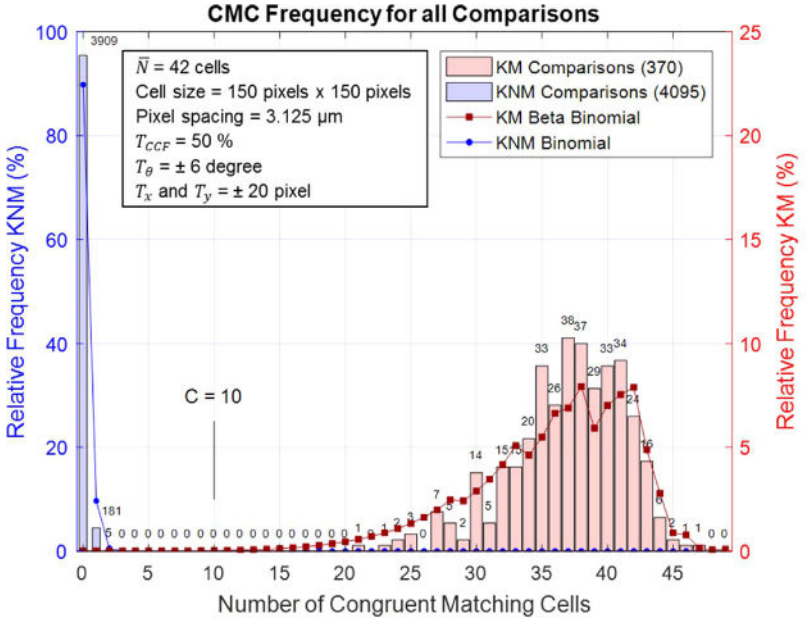

Fig. 16. Relative frequency distribution of CMC numbers for KM and KNM image pairs of the Weller dataset. The brown curve represents the beta-binomial distribution model for the KM data, using parameters also estimated from the Fadul data. The blue curve represents the binomial model for the KNM data. Note that the distribution models are discrete, with the connecting lines drawn for visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

variability of the firing process, we do not expect evidence from firearms to exhibit the extremely low error rates that are characteristic of DNA evidence [35]. However, the wide separation of the KM and KNM image correlations and the extremely small false positive error rates calculated from the models suggest the feasibility of applying the CMC method to a large number of firearms manufactured under similar conditions and producing correlation results to support identification, if not exclusion, decisions. The probability models for the results estimated from one firearm set were consistent with the distributions observed for the second set, indicating consistency of the statistical models and estimated error rates. In general, the shape of CMC distribution curves for KNM image pairs is expected to be narrow and stable.

The extremely small false identification error rates calculated from the models and population sizes discussed here suggest that it would be feasible to scale up the statistical procedure to case work with large population sizes and still arrive at reasonable and usefully small false identification error rates. However, practical case work would require (1) a database with accurate counts of firearms manufactured by different methods with different class characteristics, (2) data like Figs. 11 and 15 for different types of firearms, and (3) a statistical procedure to combine data sets from different types of firearms from different manufacturers – one could not generalize the results seen here to other types of manufactured firearms.

6.3. Future work

The work reported here is a demonstration of concept for the objective CMC method. Studies with larger databases, including direct comparisons with manual evaluations, will be required to demonstrate feasibility of the CMC method for crime lab casework. We are working to participate in black-box studies of firearms experts like that of Baldwin et al. [9], which was favorably reviewed by a recent report of the President's Council of Advisors on Science and Technology [36]. This could yield a comparison of error rates of the CMC method and of subjective methods.

We are also working to test the CMC method and error rate procedure on different sets of consecutively manufactured firearms, where the fabrication process leaves stronger common tool marks than the sand blasting process studied here. We have begun to adapt congruent matching methods to firing pin image correlations [20] and to 2D bullet image correlations. We are also working on a procedure based on international standards [37] to incorporate uncertainty statements into the reported error rates, and we aim to scale the approach to be usable with large databases of forensic samples.

We envision a time when ballistic examiners can input either topographies or optical intensity images into a program that automatically conducts correlations, and generates objective conclusions (declared match, for example) and error rate

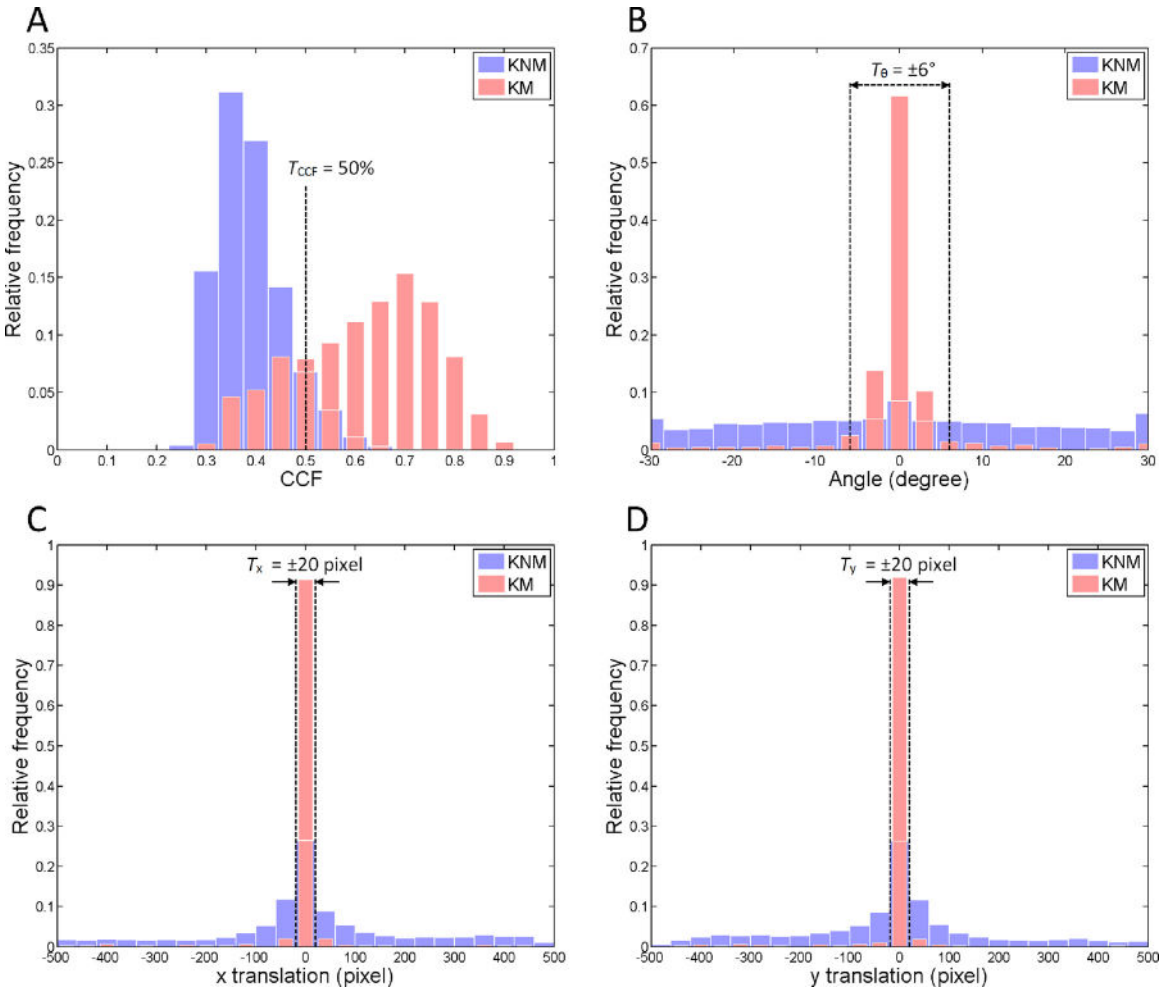

Fig. 17. Experimental relative frequency distributions of registered cell pairs for the KM (red) and KNM (blue) correlations with respect to the identification parameters: (A) CCF_{max} with a threshold T_{CCF} = 50%; (B) θ with $T_\theta = \pm 6°$; (C) x with $T_x = \pm 20$ pixels (or ± 0.125 mm); and (D) y with $T_y = \pm 20$ pixels (or ±0.125 mm). The KM (63 pairs) and KNM (717 pairs) distributions are each scaled to their sample size. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

estimates. The CMC method and statistical procedure can provide a scientific foundation and practical methods to do this.

Funding

This work was supported by the Forensic Measurement Challenge Program (FMC2012) and the Special Programs Office (SPO) of NIST.

Acknowledgments

The authors are grateful to T. Fadul of the Miami-Dade Crime Lab and to T. Weller for providing test samples, to X. Zheng of NIST for providing the topography images, to R. Silver, R.M. Thompson, and S. Ballou of NIST for their support of this work and their suggestions, and to J. Lu, J. Filliben, J. Butler, and J.M. Libert of NIST for their careful manuscript reviews.

Appendix A. Distributions for individual identification parameters

It is instructive to plot the experimental frequency distributions of registered KM and KNM cell pairs with respect to each CMC identification parameter and to estimate the respective cell trial success probabilities. Fig. 17 shows the experimental frequency distributions of registered cell pairs for KM (red) and KNM (blue) correlations with respect to each of the four identification parameters: CCF_{max} (Fig. 17A), registration angle θ (Fig. 17B), and x-, y-registration distances (Fig. 17C and D). The thresholds T_{CCF}, T_{θ}, T_x, and T_y are also shown.

Although there are large overlaps between the KM and KNM cell distributions for each parameter, combining all four parameters yields the significant separation between the CMC distributions of KM and KNM image pairs shown in Fig. 6. The combined false positive and false negative probabilities, p_{KNM} and $(1-p_{KM})$, for each correlated cell pair can likewise be estimated by combining the estimated false positive and false negative frequencies associated with each of the four identification parameters (CCF_{max}, θ, x and y) considering any correlation between them. First, the number of KNM cell pairs that pass the T_{CCF} threshold ($CCF_{max} \geq 50\%$ for the test case) are counted and compared with the total number of cell pairs to derive an estimation for the individual probability $p_{KNM(CCF)}$. Then only the cell pairs passing the T_{CCF} test are included in the conditional frequency distribution for the next parameter θ, from which the conditional probability $p_{KNM(\theta|CCF)}$ is estimated, and so on [15]. The false positive and true positive cell trial probabilities associated with all thresholds estimated in this way are the same as those calculated by Eqs. (2) and (3).

Appendix B. Relaxing the assumption of independence of cell pair comparisons

In Section 4.2, it was assumed that the cell pairs in an image pair are independent of each other. That is, the random variable X, which represents the outcome of the CMC method for a cell pair comparison is independent of the random variables for other cell pair comparisons in the image pair. Therefore, for an image with N cell pairs, we assumed that the sequence of $X_1, \ldots, X_N$ is a sequence of Bernoulli trials. However, in practice, cell pairs may not be independent in general. To address this we use a model for dependent Bernouilli trials proposed by Bahadur [38], which is sometimes called the Bahadur–Lazarsfed model [39]. The model allows the Bernoulli trials, $X_1, \ldots, X_N$ to be correlated while $P(X=1)=p$ and $P(X=0)=1-p$ for each of $\{X_1, \ldots, X_N\}$. For simplicity, we only consider the second order correlation and assume that the correlations are symmetric [38]. In this case, the probability mass function of the sum of the sequence of the correlated X's denoted by Y is expressed by

$$P(Y) = P_{[1]}(Y)\left[1 + r_{(2)}g_2(Y,p)\right], \tag{16}$$

where $P_{[1]}(Y)$ is the probability mass function of a binomially distributed random variable discussed in Section 4.2 with the parameters p and N, and $r_{(2)}$ is a parameter characterizing the second order correlation of the correlated X's. The function $g_2(Y,p)$ is a second order polynomial in Y. In this case, Y has a correlated binomial distribution.

For the KNM comparison results of Fig. 11, for example, we obtain maximum likelihood estimates (MLE) of the parameters $\hat{p}_{KNM} = 0.00258$ and $\hat{r}_{(2)} = 0.000972$. The inclusion of correlations does not significantly change the result for $\hat{p}$ or any conclusions drawn about false positives. This is illustrated by Fig. 18 where the model that includes the effect of correlations is indistinguishable from the model for the original binomial distribution.

For the KM distribution, the MLEs are $\hat{p}_{KM} = 0.8786$ and $\hat{r}_{(2)} = 0.0470$. The effect of correlations produces a significant change in the model for the KM distribution, as is illustrated in Fig. 18.

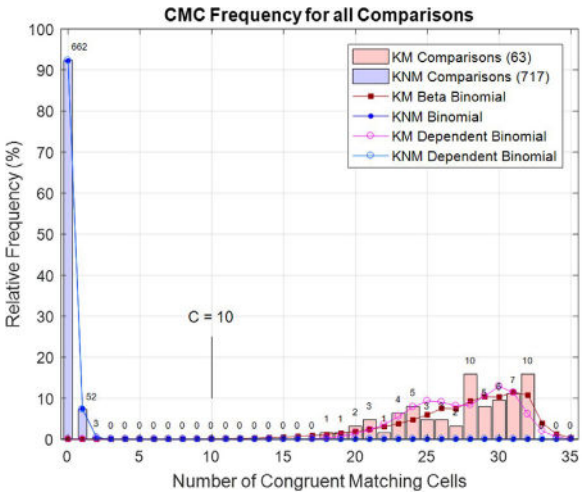

Fig. 18. CMC distribution data of Fig. 11 and comparison of distribution models with and without dependence. The brown curve is the KM beta binomial plot of Fig. 11; the red curve is the KM dependent binomial model; the indistinguishable blue curves are the KNM binomial and dependent binomial models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix C. Exploring the assumption of independence of image pair comparisons

In Section 4.2, we make the assumption that all the image pairs are independent even though each image is used more than once. Image A, for example, is compared with images B, C, etc. We provide here an alternative to this assumption by considering subsets of the image-pair population in which each image is used only once, that is, A is compared with B, C is compared with D, etc. This results in a much smaller sample size. For the Fadul set [19], there are 18–20 independent KNM pairs and 17 independent KM pairs depending on how the images are paired up. Note that such a sample is consistent with the original sample because it is a subset of those data; however, the value of probability p calculated from a small sample of KNM pairs varies depending on how the images are paired up. Fig. 19 is a semilog plot of calculated values for p_{KNM} for 100,000 reshuffled samples of the independent image pairs. The mean value for $\hat{p}_{KNM}$ is 2.57×10^{-3}, which is consistent with

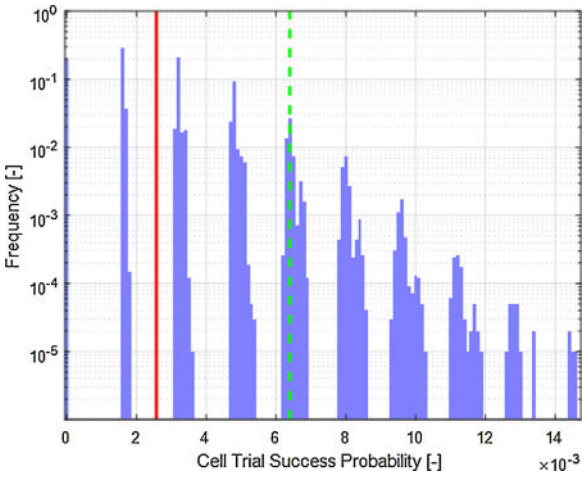

Fig. 19. Semilog plot of the frequency distribution of $\hat{p}_{KNM}$ values calculated for randomly chosen samples of independent image pair comparisons using the experimental data of Fig. 11. The solid red line shows the mean value, and the dashed green line shows the value below which 95 % of the $\hat{p}_{KNM}$ values lie.

the calculated value for $\hat{p}_{KNM}$ of 2.58×10^{-3} for the whole data set calculated in Section 5.1 for Fig. 11. An upper limit, given by the value, below which 95% of the p values lie is equal to 6.39×10^{-3} and is shown in green. The E_1 value corresponding to that p value is 4.45×10^{-15}, several orders of magnitude larger than the value of 5.6×10^{-19} obtained in connection with Fig. 11, but still extremely small.

Appendix D. Sampling issues

The error rate estimates discussed in Section 5.1 have an uncertainty due to many sources of variation. These may be associated with the accuracy of the topography images, the choice of the CMC procedure and the choices of its parameters, the assumptions underlying the distribution models, and the limited sample of comparison results used to calculate error rates. Development of an uncertainty budget encompassing all the significant sources of error is beyond the scope of the present paper, but is an important topic that we and others will be researching.

We emphasize here only the last factor related to the sample population, which was previously discussed by Petraco et al. [7]. An estimate of the sample size limitations can be obtained using bootstrapping. Bootstrapping [40] is a statistical technique where either the data or a model of the data is resampled with replacement to generate a new dataset. This resampled dataset is then used to obtain a new estimate for the parameter of interest, after which the process is repeated. The variability of the thus obtained parameter values is then used as an estimate for the variability that would be observed if new datasets were collected. Fig. 20 shows the result of a bootstrapping analysis for the false positive cumulative error rate E_1, estimated from the Fadul KNM dataset results of Fig. 11 using the binomial model for the CMC distribution, $N = 31$, and a CMC criterion of 10. The figure shows a histogram of 10,000 error rate values E_1, each calculated from a resampled dataset obtained by randomly drawing, with replacement, 717 KNM comparisons from the original dataset. Moreover, this is a blocked bootstrapping analysis to estimate firearm clustering effects. Here, the resampling is done by randomly selecting from the ten firearms and including all the results from each firearm selected until the population of 717 KNM image comparisons is reached. This provides an estimate of the potential uncertainty due to sampling effects. Fig. 20 indicates that 95% of the calculated error rate values are less than or equal to 3.00×10^{-18}.

Although bootstrapping yields insights into the variability of the error rate estimate, which in turn can be used to select a more conservative error rate, it does not replace the need for additional data. A key assumption underlying the bootstrapping approach is that the variability due to resampling the dataset is, in approximation, similar to the variability that would be observed when selecting new datasets from the population of all comparisons of interest. This, in turn, requires that the original dataset represents a reasonable approximation of this population. Our conclusion is that sampling effects increase the false positive error rate calculated here by amounts that are insignificant on an absolute scale.

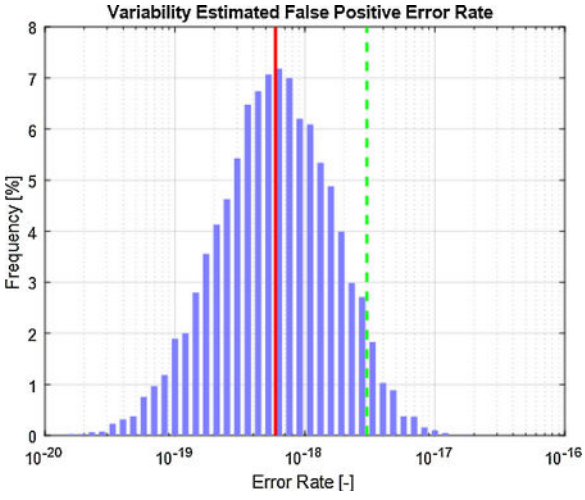

Fig. 20. Semilog plot of bootstrapping results showing the frequency distribution of 10,000 values for the false positive error rate E_1, assuming a binomial CMC distribution, obtained by resampling the Fadul KM dataset results of Fig. 11. The distribution of E_1 values was obtained by randomly selecting firearms. The solid red line indicates the median value of 5.94×10^{-19} for E_1, which is consistent with the value of 5.56×10^{-19} calculated in connection with Fig. 11; the dashed green line indicates the value (3.00×10^{-18}) at the 95th percentile of the sampled E_1 values.

References

[1] Scientific Working Group for Firearms and Toolmarks (SWGGUN), The Foundations of Firearm and Toolmark Identification, https://www.nist.gov/sites/default/files/documents/2016/11/28/swggun_foundational_report.pdf. (Accessed 19 July 2017).
[2] Firearm Examiner Training, Glossary, http://projects.nfstc.org/firearms/glossary.htm. (Accessed 5 April 2017).
[3] American National Standard ASME B46.1-2009, Surface Roughness, Waviness, and Lay, Am. Soc. Mech. Eng., New York, 2009.
[4] The National Research Council, Ballistic Imaging, NRC, Washington, DC, 2008 pp. 3, 82, 20.
[5] The National Research Council, Strengthening Forensic Science in the United States — A Path Forward, NRC, Washington, DC, 2009 pp. 153–154, 184, 155.
[6] N.D.K. Petraco, et al., Application of Machine Learning to Tool Marks: Statistically Based Methods for Impression Pattern Comparisons, NIJ Report 239048, National Institute of Justice, Washington, DC, 2012.
[7] N.D.K. Petraco, L. Kuo, H. Chan, E. Phelps, C. Gambino, P. McLaughlin, F. Kammerman, P. Diaczuk, P. Shenkin, N. Petraco, J. Hamby, Estimates of striation pattern identification error rates by algorithmic methods, AFTE J. 45 (3) (2013) 235–244.
[8] F. Riva, C. Champod, Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases, J. Forensic Sci. 59 (3) (2014) 637–647, doi: http://dx.doi.org/10.1111/1556-4029.12382.
[9] D.P. Baldwin, S.J. Bajic, M. Morris, D. Zamzow, A Study of False-positive and False-negative Error Rates in Cartridge Case Comparisons, USDOE Technical Report # IS-5207, Defense Forensics Science Center, Forest Park, Georgia, 2014 April.
[10] R. Lilien, Applied Research and Development of a Three-dimensional Topography System for Firearm Identification Using GelSight, NIJ Report 248639, National Institute of Justice, Washington, DC, 2016. https://www.ncjrs.gov/pdffiles1/nij/grants/248639.pdf.

[11] T. Weller, N. Brubaker, P. Duez, R. Lilien, Introduction and initial evaluation of a novel three-dimensional imaging and analysis system for firearms forensics, AFTE J. 47 (2015) 198–208.
[12] J. Song, E. Whitenton, D. Kelley, R. Clary, L. Ma, S. Ballou, SRM 2460/2461 standard bullets and cartridge cases project, J. Res. Natl. Inst. Stand. Technol. 109 (6) (2004) 533–542.
[13] NIST SRM 2460/2461 Standard Bullet and Cartridge Cases, Available at http://www.nist.gov/pml/div683/grp02/sbc.cfm (updated, May 25, 2017).
[14] J. Song, Proposed NIST ballistics identification system (NBIS) using 3D topography measurements on correlation cells, AFTE J. 45 (2) (2013) 184–189.
[15] J. Song, Proposed "congruent matching cells (CMC)" method for ballistic identification and error rate estimation, AFTE J. 47 (3) (2015) 177–185.
[16] W. Chu, M. Tong, J. Song, Validation tests for the congruent matching cells (CMC) method using cartridge cases fired with consecutively manufactured pistol slides, AFTE J. 45 (4) (2013) 361–366.
[17] M. Tong, J. Song, W. Chu, R.M. Thompson, Fired cartridge case identification using optical images and the congruent matching cells (CMC) method, J. Res. Natl. Inst. Stand. Technol. 119 (2014) 575–582, doi:http://dx.doi.org/10.6028/jres.119.023.
[18] E. Hare, H. Hofmann, A. Carriquiry, Automatic matching of bullet land impressions, Ann. Appl. Stat. 11 (4) (2017) 2332–2356.
[19] T.G. Fadul Jr., G.A. Hernandez, S. Stoiloff, S. Gulati, An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides. NIJ Report No. 237960, National Institute of Justice, 2012.
[20] H. Zhang, J. Song, M. Tong, W. Chu, Correlation of firing pin impressions based on the congruent matching cross-sections (CMX) method, Forensic Sci. Int. 263 (2016) 186–193.
[21] J.B.P. Williamson, The shape of solid surfaces, in: T.R. Thomas (Ed.), Rough Surfaces, 1st ed., Longman, Harlow, Essex, UK, 1982.
[22] T.V. Vorburger, J. Song, N. Petraco, Topography measurements and applications in ballistics and tool mark identifications, Surf. Topogr.: Metrol. Prop. 4 (2016) 013002, doi:http://dx.doi.org/10.1088/2051-672X/4/1/013002.
[23] L.A. Thibodeau, Sensitivity and specificity, in: S. Kotz, N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, 1st ed., Wiley, New York, 1988, pp. 370–372.
[24] J.J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter, Hastings Law J. 59 (5) (2008) 1077–1100.
[25] C. Aitken, P. Roberts, G. Jackson, Fundamentals of Probability and Statistical Evidence in Criminal Proceedings, Royal Statistical Society, 2010.
[26] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, R. Hermsen, The likelihood ratio approach in cartridge case and bullet comparison, AFTE J. 45 (3) (2013) 284–289.
[27] S. Bunch, G. Wevers, Application of likelihood ratios for firearm and toolmark analysis, Sci. Justice 53 (2013) 223–229.
[28] S.G. Bunch, Consecutive matching striation criteria: a general critique, J. Forensic Sci. 45 (2000) 955–963.
[29] A. Papoulis, Probability, Random Variables and Stochastic Processes, 2nd ed., McGraw-Hill, New York, 1984 pp. 42–43.
[30] Ref. 29, p. 75.
[31] R.R. Wilcox, A review of the beta-binomial model and its extensions, J. Educ. Stat. 6 (1) (1981) 3–32.
[32] D.M. Smith, A.S. Algorithm, 189: maximum likelihood estimation of the parameters of the beta-binomial distribution, J. R. Stat. Soc., C (Appl. Stat.) 32 (2) (1983) 196–204.
[33] S. Brinkman, H. Bodschwinna, Advanced Gaussian filters, in: L. Blunt, X. Jiang (Eds.), Advanced Techniques for Assessment Surface Topography, Elsevier, London, 2003 Chapter 4.
[34] T.J. Weller, A. Zheng, R. Thompson, F. Tulleners, Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides, J. Forensic Sci 57 (4) (2012) 912–917, doi:http://dx.doi.org/10.1111/j.1556-4029.2012.02072.x.
[35] Ref. 5, pp. 128-133.
[36] Executive Office of the President, Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, President's Council of Advisors on Science and Technology, 2016. September https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
[37] JCGM 100:2008, Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement (GUM), Joint Committee on Guides in Metrology, 2008. http://www.bipm.org/en/publications/guides/gum.html.
[38] R.R. Bahadur, A representation of the joint distribution of response to n dichotomous items, in: H. Solomon (Ed.), Studies in Item Analysis and Prediction, Stanford University Press, Stanford, 1961, pp. 158–168.
[39] P.A.G. Van Der Geest, The binomial distribution with dependent Bernoulli trials, J. Stat. Comput. Simul. 75 (2) (2005) 141–154.
[40] B.B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, London, 1993.