Identifying persistent and characteristic features in firearm tool marks on cartridge cases

Daniel Ott, Johannes Soons, Robert Thompson and John Song[1]

Engineering Physics Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, United States of America
[1] Author to whom any correspondence should be addressed.

E-mail: jun-feng.song@nist.gov

Keywords: forensics, feature identification, toolmark comparisons, firearm and tool mark identification, surface topography

Abstract

Recent concerns about subjectivity in forensic firearm identification have motivated the development of algorithms to compare firearm tool marks that are imparted on ammunition and to generate quantitative measures of similarity. In this paper, we describe an algorithm that identifies impressed tool marks on a cartridge case that are both consistent between firings and contribute strongly to a surface similarity metric. The result is a representation of the tool mark topography that emphasizes both significant and persistent features across firings. This characteristic surface map is useful for understanding the variability and persistence of the tool marks created by a firearm and can provide improved discrimination between the comparison scores of samples fired from the same firearm and the scores of samples fired from different firearms. The algorithm also provides a convenient method for visualizing areas of similarity that may be useful in providing quantitative support for visual comparisons by trained examiners.

1. Introduction

In the field of forensic firearm identification, tool marks are used to determine if two pieces of evidence were fired from the same firearm. Any marks that are imparted onto the ammunition during loading, firing, or ejection can be used. Typically, these marks are found as striations on the bullet imparted by the rifling of the barrel and as striations and impressions left on the surface of the cartridge case during the firing sequence. Among the most common markings for analysis are the firing pin impression and the breech face impression left on the cartridge case. The firing pin impression is created when the firing pin impacts the primer cup on a cartridge to initiate the burning of propellant which forces the bullet through the barrel. In response to this, the cartridge case is thrust backwards and another impression on the primer cup is made by the firearm's breech face, i.e. the surface from which the firing pin protrudes during firing.

The basis of forensic firearms identification is that random components in the microscopic surface texture of a tool (e.g. the firearm breech face) produced during the manufacturing process and subsequent use can be imparted on a softer object (e.g. the primer cup) that contacts the tool with sufficient force. Expert examiners compare the marks on two different pieces of evidence using a comparison microscope. The comparison microscope has an optical bridge consisting of a lens and prism system that combines images from two different objectives into a single eyepiece. The examiner is able to manipulate and align the evidence such that a side by side comparison of tool marks is possible. The current practice is for expert examiners to compare the pattern and shape of the markings on both samples and use their skill, experience, and training to determine whether the markings were generated by the same source (tool). The conclusions generated by this method are subjective in nature [1].

This methodology has been criticized in a report by the National Academies [2], which also expressed concerns over visual comparisons of pattern evidence in other forensic areas. The report recommended the development of objective comparison metrics and science-based estimates for the identification uncertainty or error rate. These goals have stimulated research into quantitative similarity metrics for pattern evidence comparisons, primarily in the form of computer comparison algorithms. In conjunction with appropriate reference populations, these algorithms can then be used to address concerns raised about applicable expressions for the weight of evidence that ultimately is presented in court.

The development of quantitative similarity metrics can, in turn, address the Academies' other primary concern which is the validity of the fundamental assumptions of uniqueness and reproducibility of firearm tool marks [3]. Through experience and training, practitioners hold these assumptions to be generally true, but to confirm this, it is necessary to establish well defined protocols and metrics by which the reliability of pattern matching can be validated. Certain factors can cause similarities in the tool marks generated by two different tools and thus confound the uniqueness of a particular tool mark. These factors must be considered carefully in developing a comparison algorithm and similarity metric. Namely, a tool mark can have class, subclass, and individual characteristics [4–6]. Class characteristics are common to tools from a large group and typically result from the tool's design. Class characteristics usually consist of larger features. Therefore, high pass filters are often employed to attenuate or remove these features along with surface form which is not unique or reproducible. Subclass characteristics are details incidentally imparted during manufacturing that typically persist across a limited set of firearm components sequentially produced over a particular time frame. An example of a subclass characteristic are the striations imparted on a firearm component by a worn or chipped manufacturing tool prior to its removal by the manufacturer. Class and subclass characteristics can be used for a determination of exclusion (non-match) but they cannot be used for a determination of identification (match) since they are not unique to a particular firearm. The individual characteristics are usually microscopic in scale and can be used for identification. However, expert care must be taken to differentiate individual and sub-class characteristics as they often have dimensions that are similar in scale. Therefore, it is critical that expert examiners are intimately aware of the firing process as well as the manufacturing process associated with a wide range of firearms to identify possible subclass features. This subtlety is difficult to capture with a computer algorithm and therefore necessitates some interaction or final judgement from a trained expert. To assist in this, the algorithm presented below places an emphasis on producing a quantitative tool mark similarity visualization for breech face comparisons.

Another confounding factor is the variability between marks made by the same tool. This issue is not unique to tool mark identification but must be considered in many forensic areas that rely on measures of similarity. In firearms identification, variability occurs because the cartridge case is not always registered and aligned rigidly relative to the firearm components before the firing sequence. Other sources of variability include the buildup of surface residues within the firearm, wear which causes the tool surface to change, cartridge-firearm physical tolerances, differences in firing pressure, and the surface texture and hardness of the primer cup before firing; all of which can affect the final impression evidence. Our algorithm aims to address the variability in impressions between firings and determine the underlying, reproducible patterns for a given firearm. This work may find application in other fields where there is some degree of variability in the relevant structure of the images that are compared.

The focus of many computer algorithms that attempt to automate firearm tool mark identification is on the cartridge case markings. This is because a fired bullet will often fragment and deform upon impact making it difficult to observe the remaining striated tool mark pattern on the bullet, whereas the markings on the cartridge case are typically preserved in their entirety. Furthermore, cartridge cases are often easier to obtain at crime scenes. For this reason, this study will focus on cartridge cases, specifically the breech face impression generated by handguns. However, the analysis technique applies to the more general case of comparing tool marks, including firing pin impressions, ejector marks, the cross-section profile of striated marks, and marks from other types of firearms.

The breech face impressions that are input into the proposed algorithm are surface topographies, that is, digital representations of a surface obtained using metrology instruments which map the surface topography either as a three-dimensional (3D) point cloud or as a two-dimensional (2D) matrix of surface heights. In this paper, we will use the term map or surface map for the measurement data of the sample surface topography. All measurements discussed here were obtained using a spinning-disk confocal microscope which generates a 2D matrix of surface heights. Other measurement technologies exist which produce comparable results for firearm markings [7–9]. Direct measurement of 3D surface topography differs from the traditional method of comparison which relies on reflectance microscopy to capture 2D images of surfaces. The benefit of using 3D technology is that there is no dependency on the illumination conditions and therefore consistent and repeatable measurements can be made of a surface [10, 11]. Additionally, 3D metrology equipment captures the actual surface topography of interest instead of an image that is a complex function of lighting conditions, surface topography, surface reflectivity, surface color, and imaging system characteristics. These features of 3D metrology address recommendations made by the National Academies and have resulted in efforts to promote the use of 3D metrology instruments for tool mark examination. It is important to note that 2D images still comprise a significant portion of forensic databases, such as the National Integrated Ballistics Information Network (NIBIN), and the algorithm presented here may be capable of analyzing the similarity of 2D images as well.

2. Methods

2.1. Surface map comparisons

To begin the development of a comparison algorithm it is important to understand the metric used for determining similarity of surface maps. Numerous

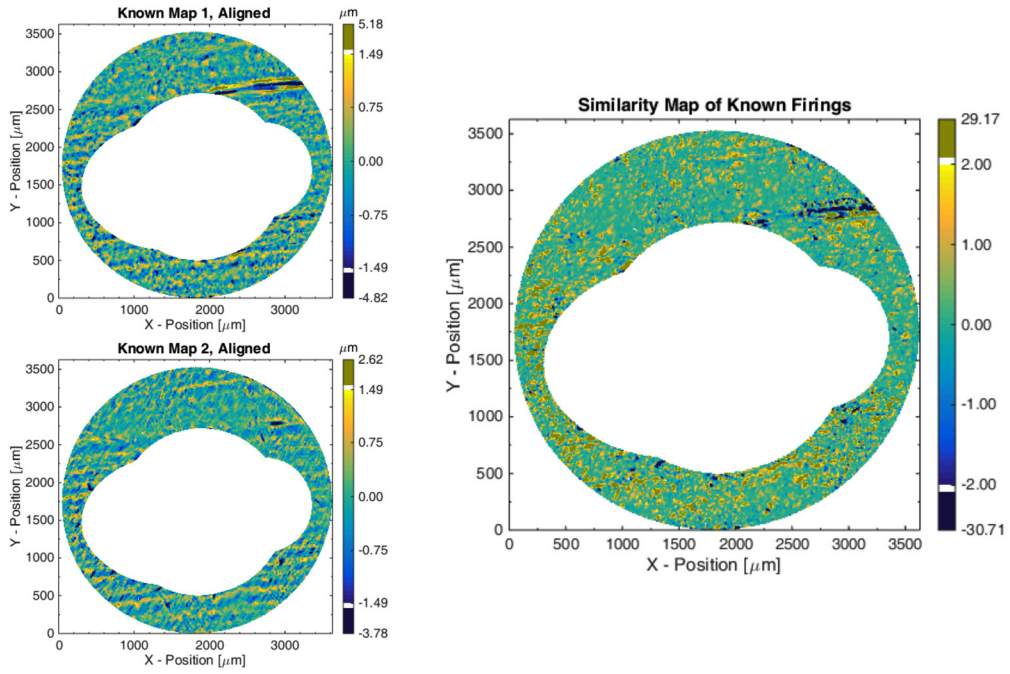

Figure 1. Two surface maps of breech face impressions (left) generated by the same firearm have been registered such that they produce maximum ACCF. The registered maps are then used with equation (2) to create the unitless similarity map (right).

metrics are available to quantify the similarity of two images [12], but the most commonly used is the normalized areal cross correlation function (ACCF) [13], that is the Pearson correlation coefficient generalized to 2D data arrays as opposed to one-dimensional (1D) arrays. The coefficient equals the covariance of the two compared data arrays divided by the product of the respective standard deviations. It has a value of 1 for two maps that are the same, except for a linear scale factor, and -1 for two maps that are scaled mirror copies. The calculation of the ACCF between maps $\boldsymbol{A}$ and $\boldsymbol{B}$ is shown in equation (1), where $\boldsymbol{A}$ and $\boldsymbol{B}$ are arrays containing the surface heights at overlapping points, μ is the respective mean surface map height, and σ is the uncorrected sample standard deviation of the surface map heights. Bold variables represent multi-dimensional arrays and non-bold variables are scalar quantities. This metric describes the average similarity of the entire overlapping area of maps $\boldsymbol{A}$ and $\boldsymbol{B}$. But, as was already pointed out, there is always some variability between firings leading to regions where poor, spurious, or misrepresentative marks are present that may have a large effect on both the average similarity metric and image registration. This is exacerbated by the products of map heights in the ACCF definition, which amplify the impact of large surface height values. Other metrics and algorithms have been proposed to reduce the effects of these invalid regions [14–20].

$$\text{ACCF} = \frac{\text{Mean}\left[(\boldsymbol{A} - \mu_A) \cdot (\boldsymbol{B} - \mu_B)\right]}{\sigma_A \sigma_B}. \tag{1}$$

The method proposed here seeks to identify regions of a surface map that consistently and significantly contribute to a similarity score. For the ACCF, the key component of the similarity score is a pointwise multiplication of the two centered and normalized surface map heights. The ACCF value is then obtained as the mean of these pointwise multiplication values. By removing the calculation of the mean value from equation (1), it is possible to generate a map showing the regions that contribute or detract from the ACCF score as shown in equation (2). This similarity map is indicative of the pointwise similarity of the two surfaces and can be used to visually highlight the most similar regions. Intuitively, the map is constructed by pointwise multiplication of two centered variables such that two aligned peaks or two aligned valleys will produce a positive contribution to the correlation, or a similarity, and a peak aligned with a valley will produce a negative contribution to the correlation, or a dissimilarity. The additional normalization serves to scale the map such that its mean equals the ACCF value.

$$\textbf{Similarity Map} = \frac{\boldsymbol{A} - \mu_A}{\sigma_A} \cdot \frac{\boldsymbol{B} - \mu_B}{\sigma_B}. \tag{2}$$

The similarity map is constructed by first registering the surface maps in translation and rotation to produce a maximum value of the ACCF, after which the similarity map is calculated. Figure 1 shows two breech face impressions from the same firearm (see section 3) which have been registered using an ACCF optimization. Note that the regions adjacent to the breech face impressions on the primer cup (firing pin impression, firing pin drag mark, primer edge roll-off, and primer flowback) have been manually trimmed and are represented by white regions in the figure.

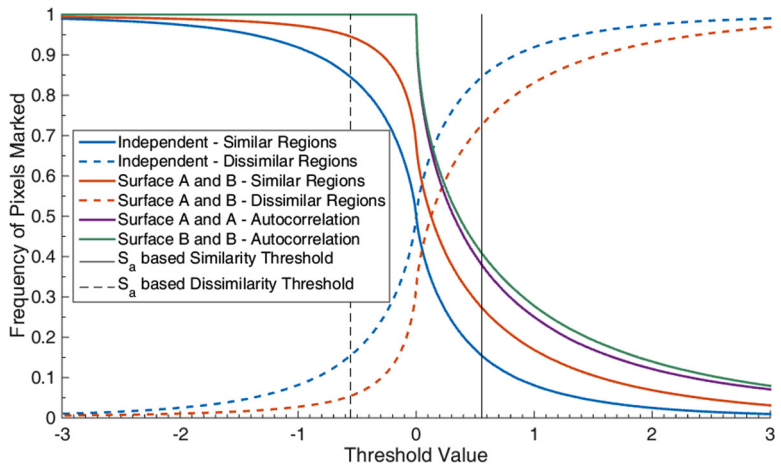

Figure 2. The fraction of the similarity map which is identified as similar or dissimilar based on the threshold value applied to the map. The curves represent four different similarity maps calculated using maps ***A*** and ***B*** of figure 1: the surface similarity map shown in figure 1 (red), two autocorrelation maps which compare each map to itself (purple and green), and a map comparing randomly sorted versions of ***A*** and ***B*** to represent the behavior of uncorrelated surfaces (blue). The fraction of each map that is marked as similar is shown by the solid curves and the fraction marked as dissimilar is shown with dashed curves.

Furthermore, a bandpass filter was applied to attenuate noise and features with large spatial wavelengths as will be discussed in section 3. The associated similarity map of the two maps is also shown in the figure. The false-color scale of the topography images is in units of micrometers and the similarity map is unitless due to the normalization. The color scales used in this figure, and throughout the paper, consist of two parts to improve visibility of the height variations. The linear central part of the color scale bar covers the majority of the surface height data (for surface maps the data within three standard deviations from the mean). This range is bounded by solid white on the color scale bar. Any values outside of this range are colored using the same muted tones at the top and bottom of the color scale bar and the values listed indicate the maximum and minimum values present in the surface map. The scale for the similarity map is not arbitrary, as the average value of the map gives the ACCF value of 0.415, but it is arbitrary in terms of its use as a visualization tool. The effects of measurement and analysis repeatability on ACCF, including instrument repeatability, sample setup, manual trimming, and image processing and registration, were evaluated by comparing repeated measurements of the same sample, resulting in ACCF values larger than 0.98. This implies that the low ACCF value can be primarily attributed to variation between firings rather than variations due to the measurement and analysis process.

The similarity map represents the contribution to the similarity metric of each measured point or pixel in the overlapping image domain. Similarity maps can be generated using other similarity criteria, even if they are not based on area similarity such as the ACCF. Lilien [21] presents a similarity map based on the areal density of matching features. It is possible to isolate the most important regions on a surface map by defining appropriate similarity and dissimilarity threshold values. In this context, 'important' means that the respective regions contribute significantly to the similarity metric. For the ACCF metric, this requires that both the heights of the respective features are significant relative to the standard deviation of the sample surface map heights and that the compared features are similar (or dissimilar), i.e. have relative heights of the same sign (or opposite sign). Areas of the similarity map that are larger than or equal to the similarity threshold value are marked as highly similar. Areas where the similarity map values are smaller than or equal to the dissimilarity threshold value are marked as highly dissimilar. The threshold itself can be defined in several ways, which will affect the type of features selected as well as the overall percentage of the map identified as important. In firearms impression evidence the important features are those that match significantly more than the similarities present in samples from different firearms. Ideally the threshold can be used to highlight these key features in a visually meaningful way.

For a given threshold value applied to the similarity map of equation (2), the fraction of the surface that will be marked as highly similar depends on both the similarity of the compared map areas and the distribution of the normalized height values. Figure 2 provides a useful illustration of the effect of the threshold value on the overall fraction of a surface that will be highlighted. To construct this figure, the two known matching maps of figure 1 were registered to each other and the similarity map was calculated. For each threshold value, the solid curves represent the fraction of the surface marked as highly similar, and the dashed curves the fraction marked as highly dissimilar. The respective red curves were calculated from the similarity map of the compared maps. The blue curves represent the marked fraction for two uncorrelated surface maps with the same height distribution, and surface roughness height parameters, as the measured maps. They

were obtained by calculating the similarity map of the two surface maps after randomizing the order of the elements in the measurement value arrays. Thus, the blue curves can be interpreted as a baseline result when comparing different-source samples that have the same statistical distribution of surface heights as maps ***A*** and ***B***. Note that the maximum difference between the red and blue curves occurs for a threshold value of zero. However, in that case, half of the similarity map would have been marked for two uncorrelated maps, which reduces the usefulness of the marked areas. The green and purple curves were obtained from the similarity map of each surface map with itself, i.e. the autocorrelation similarity map. They represent the result obtained when the compared surface maps are the same or only differ in height scale. The red solid curve will approach the green and purple curves when the surface maps are highly similar and the red curves will approach the blue curves when the surface maps are less similar.

The shape of the curves in figure 2 changes depending on the surface features present on a measured sample. Even the variability of the marks that are produced by the same firearm will produce different behavior of the curves. Therefore, it is necessary to define a consistent threshold for selecting regions of similarity or dissimilarity. There are several ways to achieve this; any of which may be suitable for particular applications of this concept. For example, based on the red curves of figure 2, the threshold value can be adjusted for each comparison to produce a fixed percentage of highlighting of similarity or dissimilarity on the surface map. Choosing the ACCF value as the threshold highlights areas with a more than average contribution to the ACCF value. For both of these approaches, there is no correlation between the size of the highlighted area and the surface similarity. Alternatively, the blue curves of figure 2 can be used to select a threshold value that yields an acceptable fraction of highlighted surface features when the two maps are uncorrelated. It is also possible to select a fixed threshold which represents the minimum necessary point-wise similarity. The dissimilarity threshold can be defined independently using any of these concepts or by defining it as the negated value of a positive similarity threshold, which is how it will be defined in this manuscript.

The approach used here to define a threshold is based on the properties of a similarity map generated by two maps that are identical except for a linear scaling factor. In this case, the effect of the threshold will follow the trends exhibited by the autocorrelation comparisons shown in figure 2. For this scenario, the surface points that are marked as contributing significantly to the similarity metric are those whose absolute map height, normalized by the σ value of the map, exceeds the square root of the threshold value. The choice of the threshold now becomes a matter of determining the normalized map height value above which a feature is deemed significant. In surface metrology, there are numerous parameters to characterize the height distribution of a surface map [22]. Common parameters for surface height variation are the surface roughness average (S_a), which is the mean of the absolute height differences from the mean of the surface map, and the root mean square surface map height (S_q), which is the square root of the mean of the squared height differences from the mean and is equivalent to σ. Therefore, if the threshold equals the product of the S_a values of the compared maps, divided by the product of the S_q values (σ), the points whose absolute surface map height exceeds S_a will be marked. Alternatively, if S_q is used in the numerator, the threshold will select points whose absolute value exceeds S_q. Keep in mind that the discussion up until this point has been in regards to identical surface maps that only differ by a linear scaling factor. In actual comparisons, there will be variations between the surfaces, even for the case of tool marks generated by the same source [5]. This will result in a different ratio of S_a/S_q for each surface map. Furthermore, the interpretation of marked points will no longer relate directly to areas of a compared surface map that exceed S_a or S_q. However, the interpretation for the case of identical surfaces (or inverted copies), provides a direct link between the height of a feature and the similarity (or dissimilarity) significance threshold.

The S_q value of a map is more sensitive to single large peaks or valleys because of the squaring operation and is equal to or larger than S_a. Breech face impressions can have regions with large peaks or valleys whose occurrence may not be repeatable due to variations in firing conditions, wear, pre-fire marks, or contaminants. Furthermore, the measurement process can sometimes result in large spurious values. For these maps, the S_a parameter is more robust and yields similarity maps that are more consistent. In terms of the threshold itself, using S_a values in the numerator will create a threshold that varies from comparison to comparison, whereas using S_q values in the numerator will produce a threshold of unity due to the equivalence of σ and S_q. Returning to the example of figure 2, the threshold based on the product of the S_a values of the compared maps is 0.556 (for dissimilarity, the threshold is -0.556). A higher percentage of the similarity map will be highlighted using this threshold compared to using S_q, which would result in a threshold of 1.

In considering the various options for defining an appropriate threshold for similarity maps, the threshold based on S_a values was chosen for its robustness and its ability to identify a practically useful fraction of significant features for visualization purposes, similar to those that an examiner would identify. Figure 3 shows the similarity map from figure 1 shaded to identify areas that are highly similar and dissimilar based on this threshold. Locations of figure 1 where the similarity value is greater than or equal to the threshold are considered highly similar, and these are colored red in regions where the surface maps are positive (peaks)

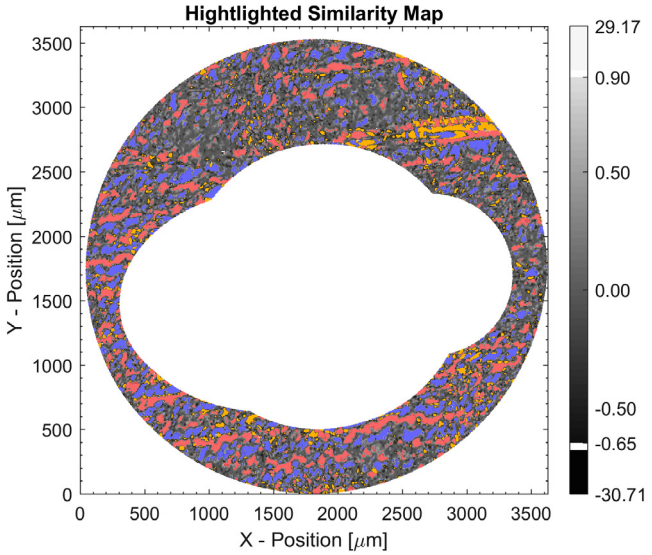

Figure 3. The unitless similarity map from figure 1 shaded with a linear intensity palette. An S_a—based threshold of 0.556 is used to identify regions of importance on the similarity map. A red hue is applied to highlight the regions in the image containing similar peaks, a blue hue to similar valleys, and a yellow hue to regions of dissimilarity. The intensity of the colored regions still corresponds to the intensity scale associated with the figure. The less influential regions contain no additional hue and appear with a simple grayscale intensity.

and blue in regions where the surface maps are negative (valleys). Locations where the similarity is less than or equal to the negated threshold are considered highly dissimilar regions and these areas are colored yellow. The regions with similarity values between these threshold values are classified as lesser contributors to the ACCF value and are shaded in grayscale.

The similarity map is a useful visual tool for identifying the regions that might traditionally be identified by an examiner during a visual comparison of the samples. The colored areas in figure 3 should be similar to the regions that an examiner would highlight manually if given the appropriate tools. In this sense, the similarity map draws a connection between the qualitative comparisons made by human examiners and the quantitative comparisons made with computer algorithms. The similarity map could be applied in several ways: (1) to quickly identify regions of interest for further detailed examination by a firearms examiner, (2) to support conclusions presented by an examiner after an independent visual comparison, (3) as a quality control check during independent peer review of a comparison conclusion, or (4) to evaluate whether possible sub-class features, identified by an examiner, contribute significantly to the similarity value obtained by an algorithm. In these applications, it is important to implement procedures that minimize the possibility of confirmation bias.

2.2. Determining persistent and characteristic features from several surface maps

The ability of the similarity map to identify regions of interest in tool mark comparisons can be leveraged to create an algorithm that attenuates the effects of tool mark variability. The goal of the algorithm is to combine images of multiple samples fired from the same firearm into one image, the characteristic surface map, that highlights features that are both reproducible and significant. Combining multiple instances of images before a comparison, even as an average, is fundamentally different from averaging the results of individual sample comparisons. The latter mean similarity score yields reduced variability of comparison results, but the averaging does not affect the mean difference between the similarity scores of same-source and different-source comparisons. For example, uncorrelated measurement noise reduces the absolute value of the ACCF similarity score. This bias cannot be reduced by simply averaging the ACCF values of multiple comparisons.

An alternative approach for improving comparison results by combining maps of multiple same-source samples was recently presented by Planka [23]. In Planka's marks step integration (MSI) approach, each region on the combined or composite surface map represents the measured data from that sample which is determined to be most representative or least damaged in that particular region. The algorithm presented here is different in that points on the composite surface map are calculated based on the surface maps from all the same-source cartridge cases. In principle, this approach reduces variability even for regions that are undamaged or 'well-marked'. Since practical considerations limit the number of samples used, the challenge is to identify sample areas with major discrepancies and improve upon the attenuation of their effect on the composite image that is achieved by simple averaging.

The goal of the presented algorithm is to reduce the variability of comparison scores while also improving

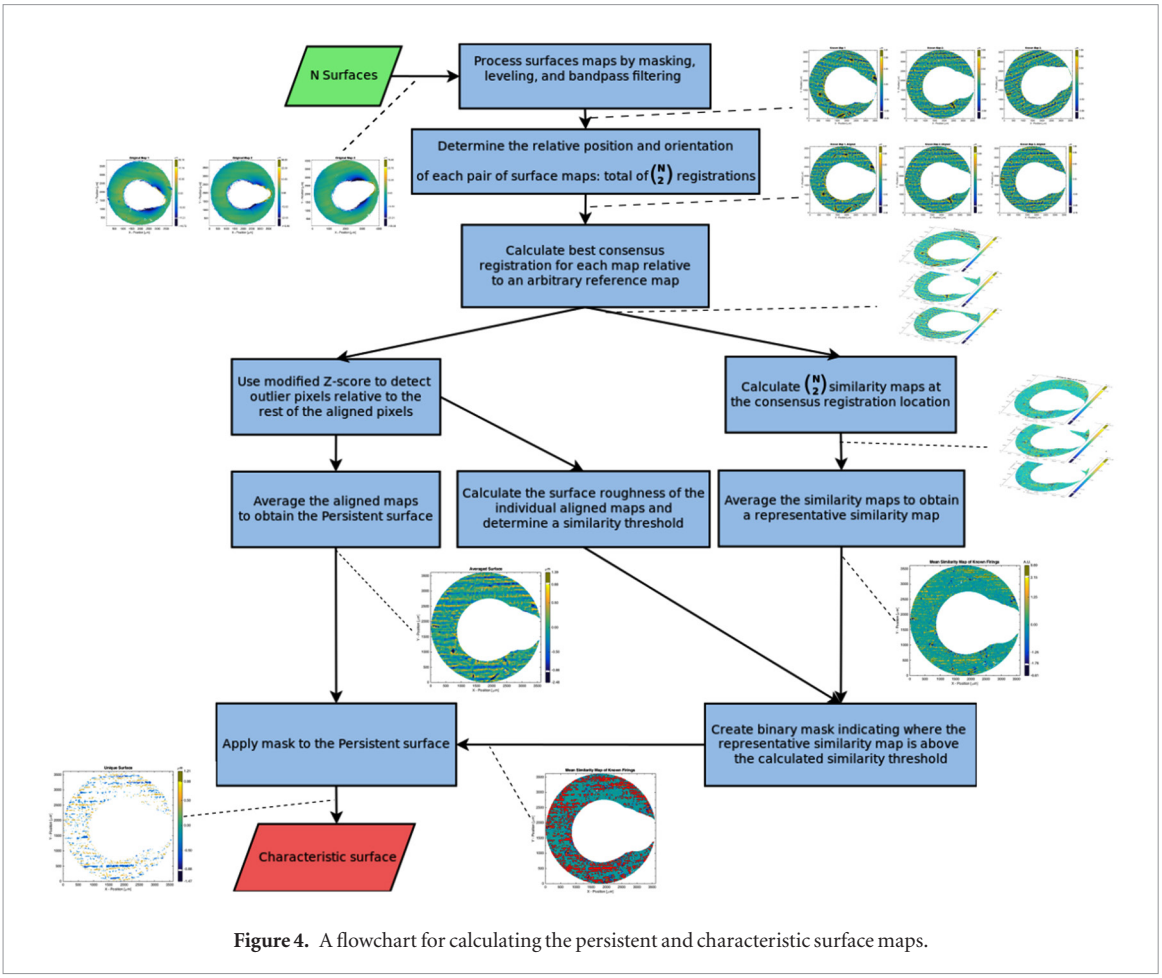

Figure 4. A flowchart for calculating the persistent and characteristic surface maps.

the discrimination between scores of known matching comparisons and known non-matching comparisons. The algorithm employs three mechanisms to reduce the effects of same-source sample to sample variability: outlier rejection, averaging, and similarity maps. The algorithm is summarized in the flow chart in figure 4, where relevant maps are inset for the case of three maps being used to create a representation of the characteristic marks generated by this individual firearm. After trimming, filtering, and leveling, as detailed in section 3, each surface map or image is registered to all other maps from the same source, typically by maximizing the respective normalized correlation coefficients. A best consensus registration is estimated based on the results of all pairwise registrations, yielding a stack of registered surface maps. Outliers, e.g. due to a scratch in one image that is not present in all other images, are detected using a modified Z-score criterion [24] applied to the stack of image values at each pixel location. The modified Z-score is shown in equation (3) where x_i is a pixel value of the ith image under test. The denominator in the equation represents the median absolute deviation (MAD) and the factor 1.4826 is included to make it a consistent estimator [24] of the respective standard deviation under the assumption that the measured surface heights of same-source samples at a single location roughly obey a normal distribution. If the absolute value of Z_i is greater than 3.5, the respective image pixel value is determined to be an outlier and it is excluded from further analysis. Although the assumption of an approximate normal distribution seems reasonable for this purpose, considering the large number of influence factors that cause variations in same-source map values, this assumption requires further study. Modifications to this correction factor will scale the cutoff for values that are considered outliers at each pixel, but will not fundamentally change the procedure. The registered maps with outliers removed are averaged to obtain a description of the typical surface impressions made at the time of firing. This averaged map is called the persistent surface map in that it represents an improved estimate for the surface topography that persists across multiple firings. An alternative approach to address outliers, which in our case yielded roughly similar results, is to obtain the pixel wise median of the stack of registered maps.

$$Z_i = \frac{x_i - \text{median}(\boldsymbol{x})}{1.4826 \cdot \text{median}(|\boldsymbol{x} - \text{median}(\boldsymbol{x})|)}. \quad (3)$$

It is important to note that the surface area of each impression is not identical between firings. Therefore, in the stack of registered images, there may be locations where certain surface maps do not have any height information. This is due to variations in the location and size of the masked (trimmed) areas: the primer roll-off area, the firing pin impression, the material flow back region, aperture shear area, and firing pin drag mark. Furthermore,

drop-outs, i.e. sample points where the instrument was unable to obtain a measurement value, can vary between samples. Drop-outs can, for example, occur in areas whose the surface slope exceeds the measurement capability of the instrument or areas where reflectivity is too high or too low. In the algorithm presented here, these non-overlapping regions were excluded from the analysis. This reduces the overall area of the persistent surface map but, in this particular example, the reduction was not substantial. For applications where the reduction in area would be problematic, it is possible to include these regions by implementing a stitching algorithm of the maps. In this case, the relative offset and tilt of each image would be estimated by minimizing the variance of overlapping pixel values, after which the overlapping pixel values would be averaged. Blending weights can be applied at the image edges to avoid sharp height variations at these edges in the stitched map. However, variations in the number of averaged measurement values would still cause variations in the attenuation of non-repeatable surface values across the domain. This stitching method was not implemented in the algorithm presented here but could be considered if applying the algorithm to images with substantially varying surface areas.

The next step is to determine the features that consistently and significantly contribute to similarity. At the consensus registration location, every possible pairwise combination of the N surface maps is used to generate a similarity map. The result is $\binom{N}{2}$ similarity maps representing the pointwise magnitude of similarity for each of the comparisons. This stack of images is then averaged to create a representative similarity map. This process is equivalent to averaging the similarity maps for the comparison of each image with the mean of all other images. The representative similarity map is then used to identify areas that are consistently, highly similar by applying an appropriate threshold. A mask is created that identifies the regions where the similarity map is above the threshold. This mask is then applied to the persistent surface map that was obtained through averaging of the registered surface topography maps. The result is a characteristic surface map, that is, a sparse surface map that contains only the most significant and consistent surface features imparted by a given tool.

The threshold used to define highly similar areas on the representative similarity map is a slightly modified version of the threshold discussed above. In this case, the surface roughness average S_a for every surface map used to construct the characteristic surface map is calculated. For each map comparison, the threshold value for that comparison is calculated, which yields an array of $\binom{N}{2}$ threshold values. The mean of this array is the threshold value for the representative similarity map.

3. Results

3.1. Applying the algorithm to measured data

To demonstrate the use of this algorithm in practice, we analyzed a set of Winchester 9 mm luger 115 grain cartridge cases which were fired from two M&P9 Smith & Wesson[2] handguns. These cartridge cases are part of a test set produced by the Federal Bureau of Investigation (FBI) where 11 firearms with consecutively manufactured slides were each fired 100 times. Only a selection of the cartridge cases have been measured for this study. This includes 75 cases from the eleventh firearm of the set and 72 from the seventh firearm of the set. The selection of the firearms was chosen at random and the difference in sample numbers is due to three missing samples in the set collected for the eleventh firearm. The breech face impression was measured using a spinning disk confocal microscope with a 10× objective and a nominal lateral pixel spacing of 3.125 μm. The root-mean-square instrument noise, tested by measuring an optical flat, was approximately 12 nm. The set of unprocessed measurements is available on the NIST Ballistics Toolmark Research Database [25]. To prepare the measurements for use in the comparison algorithm, the measured breech face impressions were first trimmed to remove roll off at the edges of the surface map, the firing pin impression, firing pin drag marks, and aperture shear. The trimmed surface maps were further processed by identification of dropouts and outliers, leveling, and application of a robust Gaussian bandpass regression filter with cutoff lengths of 25 μm and 250 μm to attenuate noise, waviness, and surface form [26].

The measured and processed breech face impressions were used to calculate numerous persistent and characteristic maps, as described in figure 4, using N random selections (without replacement) of reference images for different values of N. These surface maps were then compared to all of the remaining breech face impression measurements from the same firearm as well as the measurements from the other consecutively manufactured firearm. The similarity was calculated by finding the maximum ACCF between the questioned surface map and either the persistent surface map or the sparse characteristic surface map. Then a distribution of comparison scores was obtained for both the persistent and the characteristic surface maps. An example of the score distributions is shown in figure 5. In this example, the persistent map and the characteristic map were calculated using a set of three randomly selected firings from the seventh firearm

[2]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

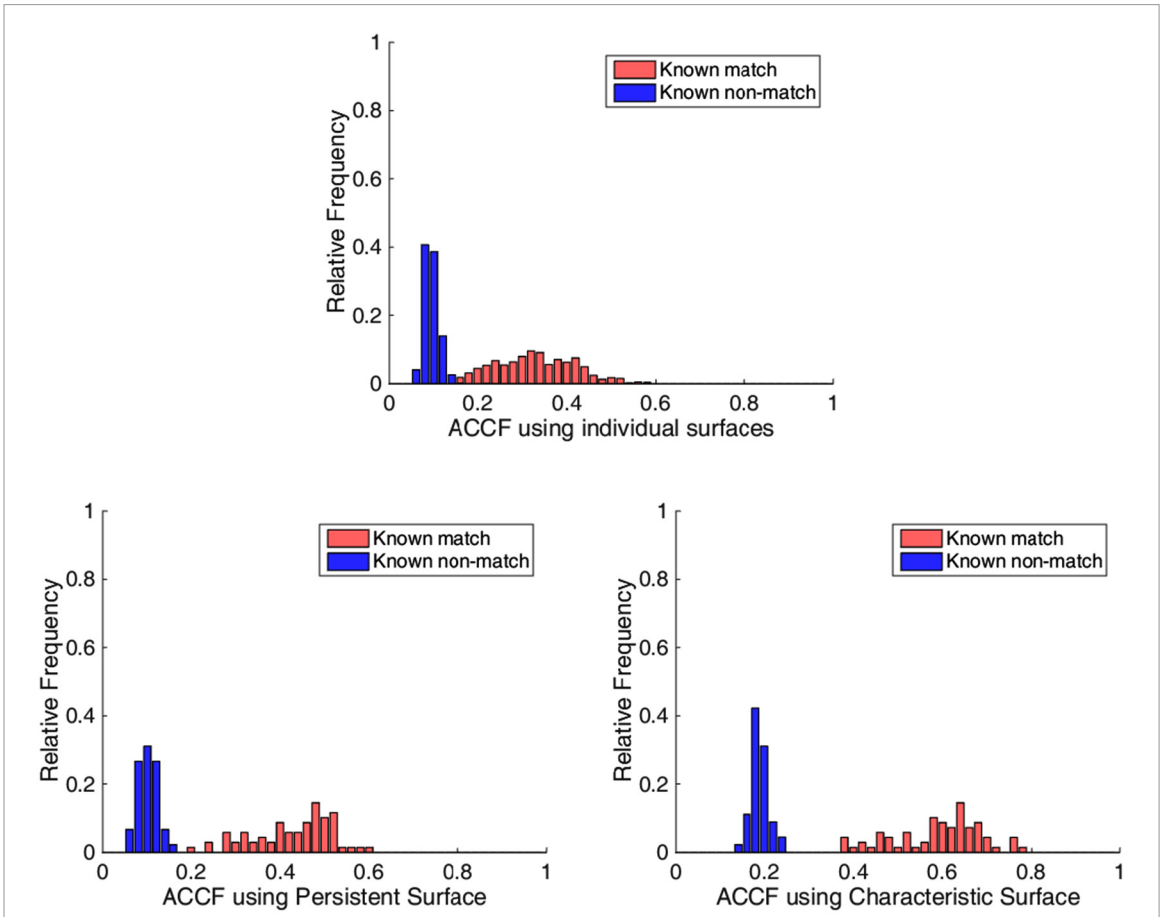

Figure 5. An example of the distribution of similarity scores when using different averaging methods. At the top are the ACCF similarity scores produced by comparing individual breech face impressions. The bottom charts show the score distributions obtained using three randomly selected breech face impressions from Firearm 7 to create persistent (left) and characteristic maps (right). The use of the characteristic surface map increases all of the ACCF scores and creates a better separation between the known matching and the known non-matching score distributions.

which were then compared to all of the remaining firings from Firearm 7 to generate the known matching distribution and all of the firings from Firearm 11 to generate the known non-matching distribution. This figure demonstrates that the use of the characteristic surface map increases the value of the similarity metric for both matching and non-matching comparisons, but it has a greater effect on the matching distributions yielding better discrimination between matching and non-matching scores.

3.2. Assessing variability in breech face impressions

The distributions calculated in the previous section were combined to demonstrate the overall effect of using composite images. For each value of N, the distributions of known matching and known non-matching scores associated with each of the independently selected sets of reference images were combined into a single distribution. The first moment (mean), the lower 5% bound, and the upper 95% bound of each combined distribution were calculated to understand the effect of N on the discriminatory power and the variability in comparison scores for this comparison algorithm. Although the distributions cannot be fully described by their first moment and percentile bounds, these values do give a preliminary understanding of how the distributions vary and allow the results to be concisely depicted

Figure 6 shows the first moment of the score distributions generated using Firearm 7 as the source of the persistent and characteristic surface maps. The bars indicate the lower 5% and upper 95% bounds of the distributions. First, $N = 1$ was used to form a baseline for the distributions. In this case, it is not possible to average or identify characteristic features, which means that this part of the graph simply describes the ACCF scores obtained when comparing two images directly, such as in [11]. As the number of combined surface maps increases, the mean value of the comparison scores also increases for known matches but stays relatively low for the non-matching comparisons. In addition, the spread of the distribution, indicated by the bars, remains relatively constant as N is increased. The lack of a significant reduction in the dispersion with N is mainly due to the unchanged variability of the individual questioned or evidence images used in the comparisons. In general, a wider separation between the mean values of the known matches and known non-matches, coupled with small dispersion bars for all distributions, indicates a more

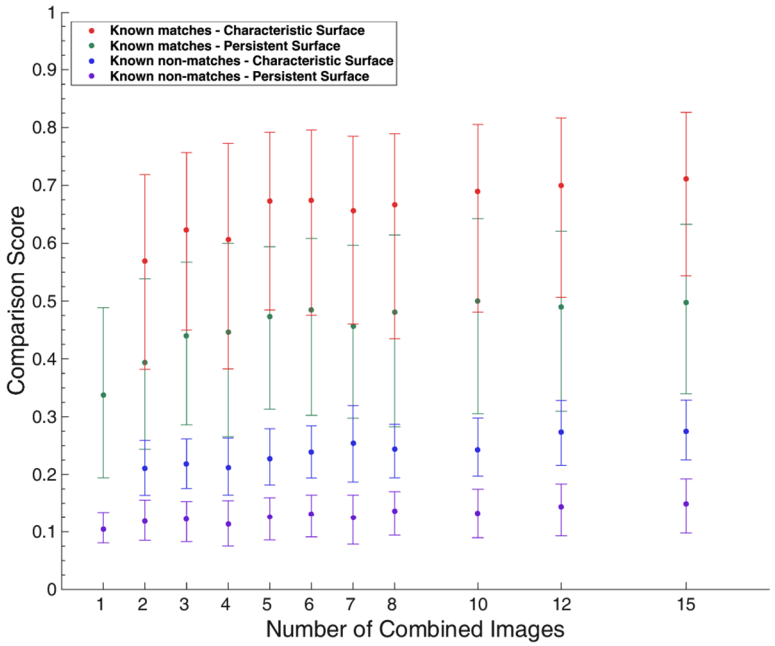

Figure 6. Characteristics of the ACCF score distributions for matching and non-matching comparisons as a function of the number of images used for the persistent reference surface map or the characteristic reference surface map for Firearm 7. The means of the distributions are shown together with error bars indicating the lower 5% and upper 95% bounds.

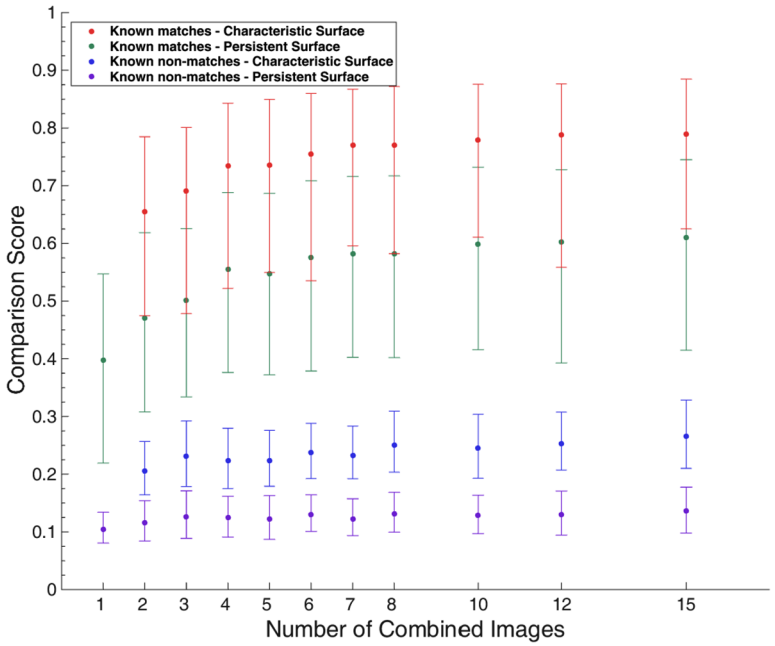

Figure 7. Characteristics of the ACCF score distributions for matching and non-matching comparisons as a function of the number of images used for the persistent reference surface map or the characteristic reference surface map for Firearm 11. The means of the distributions are shown together with error bars indicating lower 5% and upper 95% bounds.

discriminatory system. In comparing the scores of the characteristic and persistent maps, there is an increase of all scores when using the characteristic map but the increase for known matching comparisons is greater than the increase of the known non-matching scores. Use of a larger number of combined surface maps provides a better discriminatory system because non-repeatable features in the reference are attenuated. The comparisons using the characteristic surface map give an additional improvement to source discrimination by focusing on areas that both consistently and significantly contribute to the similarity metric of known matching reference samples. In the absence of subclass features, which we do not see evidence of in these cartridge cases, these improvements mainly affect same-source comparisons.

The same process was repeated using Firearm 11 as a reference, and the results of these comparisons are shown in figure 7. In this case, the overall ACCF scores were typically higher for the known matches and the

overall discrimination for this firearm was better. The same general trends that were seen in figure 6 are also present in this figure. There is an increase in discrimination when averaging a larger number of surface maps and the comparisons using the characteristic surface map show an improvement over the same comparisons with the persistent surface map. The asymptotic improvement with increased N in both cases indicates that most of the improved discrimination can be achieved with approximately five averaged samples; beyond this, there are relatively small improvements. It is important to note that these are consecutively manufactured firearms. In general, subclass characteristics may increase the ACCF score for known non-matching comparisons when images are combined. An evaluation of the samples in this study, by one of the authors who is an experienced firearm and tool mark examiner, revealed that there were no apparent visible indicators of sub-class features that would significantly influence the identification conclusions. The lack of a significant increase with N of the ACCF values for non-matching comparisons of the persistent surface map seems to confirm this. The large initial jump between $N = 1$ and $N = 2$ in the mean of the comparison scores for the characteristic reference map may be explained by the significant reduction of the reference map domain to areas with features that have a significant height. This typically provides more opportunities for the image registration process to increase the ACCF comparison value by shifting the position and orientation of the compared map.

4. Conclusions

The proposed method for calculating a persistent or characteristic surface map provides an improved estimate for the reference surface map for specific-source identification scenarios where a firearm is available. However, in most scenarios, it is not possible to account for the shot to shot variability present in the questioned crime-scene sample. Even when multiple crime-scene samples have been found, they typically must be analyzed independently as one cannot assume that they were fired from the same firearm. This variability limits the maximum value for the similarity metric that can be attained for same-source comparisons. By creating figures like figures 6 and 7, it is possible to estimate the benefits of combining multiple firings from a firearm to produce a representation of the persistent or characteristic tool marks. The trends present in figures 6 and 7 depend on the same-source reproducibility of the breech face impression data for a given firearm. For these particular Smith & Wesson firearms, significant benefits were obtained by averaging roughly five reference files. Averaging more samples did not produce significant additional improvements. The possible applications of this approach relate to understanding the consistency of firings and the ability to develop a robust approximation of the respective persistent and significant components. In addition to increasing the discriminatory power of a comparison system, the ability to quantify variability between firings is useful for expressing confidence in comparison scores. In a research setting, the approach may be applied to both reference and compared samples from different consecutively manufactured firearms to identify sub-class features and quantify their effect. Finally, the approach may be applied to persistence studies where a large number of firings are made from a single firearm. Numerous firings are collected and measured at discrete intervals within the firing sequence and then compared to see how the impression changes over time. Persistent features can be used to obtain a better understanding of a typical surface impression during each of the intervals, resulting in a better representation of the change in important features after a large number of firings.

In conclusion, a new method for analyzing impressed tool marks has been introduced which can extract and visualize consistent *and* significant features. The identified features persist over repeated firings and are used to improve similarity conclusions by 'ignoring' sample regions whose features are less significant or less reproducible. The resulting persistent surface map, along with highlighting of prominent features (figure 3), serves as a useful tool in supporting the conclusions of examiners by providing a link between subjective visual similarity and objective similarity metrics. The ability to visualize areas that are contributing significantly to the similarity score will also aid examiners in determining if sub-class characteristics are mistakenly being used by the comparison algorithm. The new comparison method was applied to a set of breech face impressions and improved discrimination was observed between non-matching and matching scores. We provided an initial approach for estimating the number of averaged tool markings that yield significant improvements to the discriminatory power of comparisons, providing insight into the required number of known firings to be collected in a forensic investigation. With further validation, these concepts will support the goal of providing more objective and discriminatory comparison methods in firearm and tool mark examination.

Acknowledgments

We would like to thank Jennifer Stephenson and Erich Smith at the Federal Bureau of Investigation (FBI) for providing the collection of test fires used in this paper. The funding for this project is provided by the Special Programs Office (SPO) of NIST.

ORCID iDs

Daniel Ott https://orcid.org/0000-0001-5221-8819
Robert Thompson https://orcid.org/0000-0003-2578-8845

References

[1] Committee for the Advancement of the Science of Firearm and Toolmark Identification 2011 Theory of identification as it relates to toolmarks *AFTE J.* **43** 287

[2] The National Research Council 2009 *Strengthening Forensic Science in the United States—a Path Forward* (Washington, DC: The National Academies Press) pp 153–5

[3] The National Research Council 2008 *Ballistic Imaging* (Washington, DC: The National Academies Press) p 3

[4] Thompson R M 2010 *Firearm Identification in the Forensic Science Laboratory* (Alexandria, VA: National District Attorneys Association) (https://doi.org/10.13140/rg.2.2.16250.59846)

[5] Senin N, Groppetti R, Garofano L, Fratini P and Pierni M 2006 Three-dimensional surface topography acquisition and analysis for firearm identification *J. Forensic Sci.* **51** 282–95

[6] McClarin D 2015 Adding an objective component to routine casework : use of confocal microscopy for the analysis of 9 mm caliber bullets *AFTE J.* **47** 161–70

[7] Vorburger T V, Rhee H G, Renegar T B, Song J F and Zheng A 2007 Comparison of optical and stylus methods for measurement of surface texture *Int. J. Adv. Manuf. Technol.* **33** 110–8

[8] Gambino C *et al* 2011 Forensic surface metrology: tool mark evidence *Scanning* **33** 272–8

[9] Vorburger T V, Song J and Petraco N 2016 Topography measurements and applications in ballistics and tool mark identifications *Surf. Topogr.: Metrol. Prop.* **4** 013002

[10] Banno A, Masuda T and Ikeuchi K 2004 Three dimensional visualization and comparison of impressions on fired bullets *Forensic Sci. Int.* **140** 233–40

[11] Vorburger T *et al* 2007 Surface topography analysis for a feasibility assessment of a national ballistics imaging database *NISTIR* 7362 NIST, Gaithersburg, MD (https://www.nist.gov/publications/surface-topography-analysis-feasibility-assessment-national-ballistics-imaging-database)

[12] Zitová B and Flusser J 2003 Image registration methods: a survey *Image Vis. Comput.* **21** 977–1000

[13] Song J and Vorburger T 2000 Proposed bullet signature comparisons using autocorrelation functions *Proc. of NCSL (Toronto, July 2000)*

[14] Song J 2015 Proposed 'congruent matching cells (CMC)' method for ballistic identification and error rate estimation *AFTE J.* **47** 177–85

[15] Weller T, Brubaker M, Duez P and Lilien R 2015 Introduction and initial evaluation of a novel three-dimensional imaging and analysis system for firearm forensics *AFTE J.* **47** 198–208

[16] Spotts R, Chumbley L, Ekstrand L, Zhang S and Kreiser J 2015 Optimization of a statistical algorithm for objective comparison of toolmarks *J. Forensic Sci.* **60** 303–14

[17] Roth J, Carriveau A, Liu X and Jain A K 2015 Learning-based ballistic breech face impression image matching *IEEE 7th Int. Conf. on Biometrics Theory, Applications and Systems (BTAS) (Arlington, VA)* pp 1–8

[18] Hamby J, Norris S and Petraco N 2016 Evaluation of GLOCK 9 mm firing pin aperture shear mark individuality based on 1632 different pistols by traditional pattern matching and IBIS pattern recognition *J. Forensic Sci.* **61** 170–6

[19] Riva F and Champod C 2014 Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases *J. Forensic Sci.* **59** 637–47

[20] Petraco N K *et al* 2012 Application of machine learning to tool marks: statistically based methods for impression pattern comparisons *NCJ Report* 239048 National Institute of Justice, Washington, DC (https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=261107)

[21] Lilien R 2015 Applied research and development of a three-dimensional topography system for imaging and analysis of striated and impressed tool marks for firearm identification using GelSight *NCJ Report* 248962 (National Institute of Justice, Washington, DC) (https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=271102)

[22] ISO 25178-2:2012 2012 *Geometrical Product Specifications (GPS)—Surface Texture: Areal—Part 2: Terms, Definitions and Surface Texture Parameters* (Geneva: International Organization for Standardization) (https://www.iso.org/standard/42785.html)

[23] Planka B 2015 Increasing the efficiency of automated ballistic imaging search systems using the marks step integration method *AFTE J.* **47** 209–14

[24] Iglewicz B and Hoaglin D 1993 *The ASQC Basic References in Quality Control: Statistical Techniques* (*How to Detect and Handle Outliers* vol 16) ed E F Mykytka (Milwaukee, WI: ASQC Quality Press)

[25] National Institute of Standards and Technology, NIST Ballistics Research Database (NBRTD) (Online) (Available: www.nist.gov/forensics/ballisticsdb (Accessed: 2017))

[26] Brinkman S and Bodschwinna H 2003 Advanced Gaussian filters *Advanced Techniques for Assessment Surface Topography* ed L Blunt and X Jiang (Amsterdam: Elsevier) ch 4