International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 33, No. 11 (2019) 1940021 (34 pages)
© The Author(s)
DOI: 10.1142/S0218001419400214

Objective Identification of Bullets Based on 3D Pattern Matching and Line Counting Scores

Danny Roberge*, Alain Beauchamp† and Serge Lévesque‡

Research Department, Ultra Electronics Forensic Technology
5757 Cavendish Blvd., Suite 200
Montréal, Québec, H4W 2W8, Canada
**danny.roberge@ultra-ft.com*
†*alain.beauchamp@ultra-ft.com*
‡*serge.levesque@ultra-ft.com*

Received 12 July 2018
Accepted 20 January 2019
Published 13 March 2019

In firearm identification, a firearm examiner looks at a pair of fired bullets or cartridge cases using a comparison microscope and determines from this visual analysis if they were both fired from the same firearm. In the particular case of fired bullets, the individual firearm signature takes the form of a striated pattern. Over the time, the firearm examiner's community developed two distinct approaches for bullet identification: pattern matching and line counting. More recently, the emergence of technology enabling the capture of surface topographies down to a submicron depth resolution has been a catalyst for the field of computerized objective ballistic identification. Objectiveness is achieved through the statistical analysis of various scores of known matches and known nonmatches exhibit pair comparison, which in turn implies the capture of large quantities of bullets and cartridge cases topographies. The main goal of this study was to develop an objective identification method for bullets fired from conventionally rifled barrels, and to test this method on public and proprietary bullet 3D image datasets captured at different lateral resolutions. Two newly developed bullet identification scores, the Line Counting Score (LCS) and the Pattern Matching Score, computed on 3D topographies yielded perfect match versus nonmatch separation for three different sets used in the standard Hamby–Brundage Test. A similar analysis performed using a larger, more-realistic set, enabled us to define a discriminative line at a false match rate of 1/10 000 on a 2D plot that shows both identification scores for matches and nonmatches. The LCS is shown to produce a better sensitivity than the standard consecutive matching striae criteria for the more-realistic dataset. A likelihood function was also computed from a linear combination of both scores, and a conservative approach based on extreme value theory is proposed to extrapolate this function in the score domain where nonmatch data are not available. This study also provides a better understanding of the limitations of studies that involve very few firearms.

Keywords: Forensic science; firearm identification; bullet; topography measurements; consecutive matching striae (CMS); false match rate; likelihood ratio.

† Corresponding author.

1. Introduction

Firearm identification is a discipline of forensic science that involves determining if fired bullets or cartridge cases had a particular firearm as a common source, by analyzing the unique marks left on their surfaces during firing. This discipline is often referred to as firearm fingerprinting because it is analogous to the identification of individual fingerprints.

The fact that the interaction of a tool with a tooled surface produces, apart from the desired effect of the tool, microscopic random marks on the tooled surface constitutes the fundamental hypothesis that supports this discipline.[2] For instance, the various gun barrel manufacturing processes leave tiny marks inside the barrel that, in turn, produce a unique microscopic stria pattern on the surfaces of bullets fired from that firearm. The uniqueness of the pattern is the consequence of the random nature of the manufacturing marks left on the surfaces inside the firearm.

Unfortunately, some random flaws or incidents during manufacturing, such as an accidental dent on a tool edge, can impart unintended marks on several consecutively manufactured parts of the firearm. Some manufacturing processes, such as broaching or metal injection molding, are also prone to producing marks that transfer over to several tooled surfaces (Ref. 27, Chap. 3). These marks are not the result of random processes occurring during the interaction of the tool with the tooled material, and their traces left on bullets or cartridge cases should be identified as such during the firearm identification process. However, it is impossible for a layperson to distinguish random marks from subclass characteristics.[a] The ability to segregate specific marks from subclass characteristics constitutes one of the many successfully resolved challenges of forensic firearm identification.

In order to confirm the fundamental hypothesis that supports the discipline and to validate the firearm identification process, firearm examiners have developed several types of proficiency tests. One of the most popular types consists of collecting a group of consecutively manufactured barrels directly from the production line. Several bullets are fired from each barrel and the associated bullets and cartridge cases are carefully identified. Usually, the test set would consist of 20 pairs of known matches (KMs) from 10 different barrels and 10 to 15 unknowns that need to be compared and associated with the correct barrel.

Many different versions of this type of test, involving fired bullets or cartridge cases, have been performed by thousands of firearm examiners around the world with extremely high levels of success.[18,27] This high success rate constitutes a validation of the random and unique nature of the marks left on fired bullets and cartridge cases. Since the barrels were manufactured consecutively, it also validates the process by which the firearm examiners identify subclass characteristics among the truly random marks and discard them from the identification process. In the community of firearm examiners, it is widely held that consecutively manufactured barrel

[a] Subclass characteristics is the name given to the nonrandom marks which are characteristics to a subset of a particular class of firearms.

tests involve the most difficult firearm identification possible, and that their successful completion guarantees the validity of all other, inherently easier, firearm identifications.

The main instrument used for forensic firearm identification is the comparison microscope, invented by Goddard in the mid-1920s. It consists of two microscopes connected by an optical bridge, which results in a split view window enabling two separate objects to be viewed simultaneously. The observer does not have to rely on memory when comparing two objects as he or she would have to do when using a conventional microscope.

During training, a firearm examiner spends thousands of hours using the comparison microscope to observe patterns of numerous KM and known nonmatch (KNM) pairs of fired bullets and cartridge cases. The training also demands the acquisition of in-depth knowledge of various manufacturing processes and tooling used by different firearm manufacturers, along with the standard firearm identification procedures developed by generations of firearm examiners. Since manufacturing processes are always evolving, and because new challenges continue to emerge, even a confirmed firearm examiner must undergo periodic training to improve and update his or her expertise.[27]

This model, whereby a trained expert states the origin of ballistic evidence based on his or her experience, and observations using only a comparison microscope, has served the judiciary and criminalistics communities quite well over the years. However, this model has two drawbacks. First, although the model works well for criminal cases in which a suspected firearm is readily identified and can be analyzed by a firearm examiner, it is not appropriate in helping to solve cold cases, where the police investigator does not have much information about the criminal shooting. Second, even if the firearm identification is based on scientific grounds, the act of identifying a fired bullet or cartridge case as being from a particular firearm by visually observing its striae or impressed patterns is subjective in nature.[2] In particular, it is not possible for the firearm examiner, based on his or her knowledge and experience, to objectively evaluate or compute the probability that two different firearms would produce a *specific* random pattern that he or she believes to have come from the same firearm.

Since the mid-1990s, with the emergence of electronic imaging, several automated ballistic identification systems (ABISs) were developed to automate firearm identification. Their purpose was to help solve cold cases by comparing images of bullets or cartridge cases, either from test fires from seized weapons or exhibits recovered from crime scenes, with those stored in the databases of cold case information. The systems output lists of the most-relevant matching candidates for a particular exhibit and still rely on the expertise of firearm examiner to confirm a potential match.

The early versions of the various ABISs relied on reflectance (2D) imaging. Due to the inherent specular nature of ballistic evidence, performance was limited by the large image variability associated with relatively small differences on the object's surface. This limitation was removed with the advent of 3D topographic imaging which is associated with a major leap in ABIS performance. Today, previously unknown matches, later confirmed by a firearm examiner, are routinely found by law enforcement agencies in datasets of tens of thousands of ballistic exhibits.[16]

Unlike a reflectance 2D image, which is highly dependent on the lighting geometry, the 3D topographic image of a ballistic exhibit provides a true representation of its surface. Over the last decade, efforts have been made to develop ballistic identification methods using 3D topographic images with the objective of providing reliable error rates or likelihood ratio (LR) with the comparison results.[5,19,30,32,35]

Several studies have been conducted using the aforementioned consecutively manufactured barrel standard tests. The belief that these tests involve the most difficult firearm identification problems possibly compensates for the small size of their dataset (i.e. about 35 exhibits from 10 guns). Furthermore, for such small dataset, there is often no overlap at all between the score distributions of the K and KNM, making it difficult to objectively evaluate error rates, which is the goal of these studies. Most of the time, a standard distribution (a normal, binomial, or beta-binomial distribution, depending on the nature of the comparison algorithm), would be fit on the available score distributions. The distribution would then be extended to score values that are arbitrarily far from its bulk.

The primary objective of this paper is to present an objective identification method suitable for bullets fired from conventionally rifled barrels[b] based on a dataset that is larger than those used in other studies; we use standard consecutively manufactured barrel sets to train the feature extraction used in our model. Objective scores computed from the features are trained and validated with a proprietary dataset consisting of 406 bullets fired from 136 pistols of different makes and models (all pistols share the same class characteristics of 9 mm caliber with six lands and grooves with right twist). The dataset is representative of the variability encountered in the context of casework in a ballistic laboratory. We will also shed light on the implicit hypothesis that the shape of the bulk of the empirical nonmatch score distribution function dictates the behavior of the distribution tail that is used for most of the published error rates or LR evaluations. This paper is the first to apply extreme statistics in this context. Furthermore, our proposed method constitutes a combination of both large classes of objective identification methods for bullets: pattern matching[38] and line counting.[12]

The second objective is to propose a visual representation of the results of the objective identification methods that can be easily interpreted by firearm examiners. However, the visual analysis of multi-dimensional feature vectors, like those used by machine learning (ML) algorithms, is a well-known challenge in more than three (and sometimes two) dimensions. We adopted the strict approach of defining features in a 2D space. With pattern matching and the consecutive matching striae (CMS) method being two different, but successful, viewpoints, it seemed reasonable to adopt a pattern matching score (PMS) and a line counting score (LCS). By

[b] Polygonal and unrifled bullets are excluded from this study.

working with only two types of scores, finding a match versus nonmatch discriminative line or curve, associated with a pre-defined false match rate (FMR), would be feasible using visual inspection, without any support from the more abstract ML machinery. Also, since the nature of the work performed by firearm examiners is visually oriented, we expect a better acceptability of such a method if valid conclusions can effectively be drawn from 2D score plots. While imposing a 2D feature vector is a severe constraint, it can be relaxed by allowing each score to be a function of several complementary measures.

The third objective of the paper is to show that the objective identification of fired bullets is possible with ABISs such as IBIS® TRAX-HD3D™ | BULLETTRAX™,[c] which usually capture images at a slightly lower lateral resolution than those captured with devices used in other studies. As it is necessary to validate any objective identification methods on real, very large datasets, the usage of fully automated and fast 3D topographic devices such as BULLETTRAX is mandatory. The topographic measurements of a pristine bullet and of a bullet fragment (not used in this study) are shown in Fig. 1.

The fourth objective is to develop a new line counting method. While line extraction methods have been developed, published similarity measures based on line counting[12,19] have been tested on small datasets. Our new line counting measure is tested on a realistic dataset, captured at a relevant resolution.

The fifth objective of this paper is to validate, or invalidate, the hypothesis that consecutively manufactured barrel tests likely provide the most similar KNM pairs. With access to a large set of bullets from firearms of various brands, we will be in the position to know how common, or uncommon, it is to find a random pair of

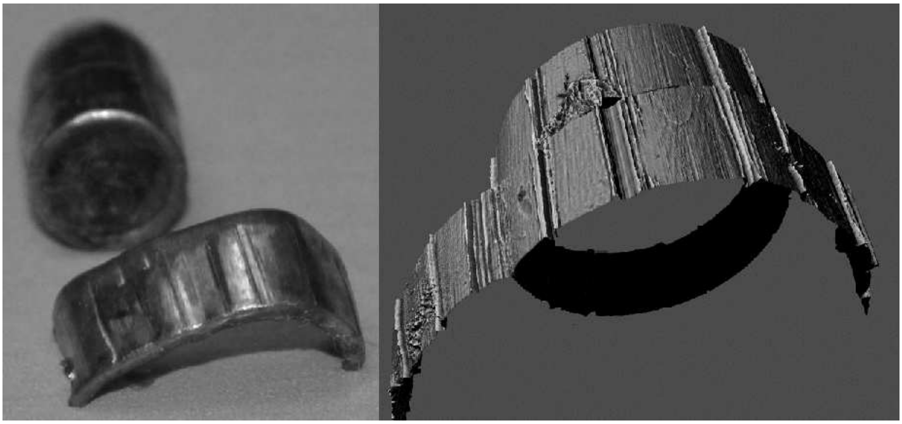


Fig. 1. Pristine and deformed bullets (left) with corresponding topographic measurements (right).

[c] BULLETTRAX is a bullet acquisition station developed by Ultra Electronics Forensic Technology Inc., specialized for the entry of bullet information onto an Integrated Ballistic Identification System (IBIS) network.

fired bullets that show more similarities, according to our objective identification score, than the best KNM pair from a consecutively manufactured bullet test.

The rest of this paper is structured as follows. Section 2 summarizes the previous work regarding 3D approaches for bullets and objective identification in the context of ballistic analysis. We describe our main strategy, and our profile and line extraction algorithms in Sec. 3. In Sec. 4, we list the four datasets used in this study, define the PMS and the LCS, and explain the analysis of the FMR and the LR based on scores in the PMS–LCS space.

2. Related Work

2.1. *High resolution 3D imaging of bullets*

Several research projects based on imaging and comparison algorithms for bullets have been published over the last two decades. Some of them involved 2D imaging[9,20–25,31]; others presented prototype systems based on the acquisition of 3D surface topography but did not provide quantitative results from automated comparison algorithms.[6,15,33,37] In this section, we will concentrate on the relatively small set of studies which involve bullet comparison algorithms from high resolution 3D data. Furthermore, we summarize two specific studies on objective identification of cartridge cases.

The first 3D bullet studies were realized using systems that captured a series of high resolution profiles along the circumference of bullets. De Kinder and Bonfanti[14] used a laser profilometer that performed more than 150 scans along the circumference of the bullets in their study, thus generating 3D surfaces. Their test set consisted of two groups of three bullets, fired by a Beretta 92S and a Mauser P08 pistols, respectively. The authors had already performed the basic operations that would appear in several subsequent studies in order to extract relevant microscopic information: high pass filtering that removes the form of the bullet, the selection of regions of interest from the land impressions, and the computation of a feature vector for each land impression by averaging several pre-aligned profiles. Their comparison method was a standard correlation algorithm. The authors found that the results obtained were consistent with the quantity and quality of marks present. They also established that a vertical resolution better than 1 micron is required for automated 3D bullet analysis.

Bachrach[4] presented the results generated by SCICLOPS$^{\mathrm{TM}}$, an automated system which used confocal microscopy. The system captured five profiles along the circumference of each bullet in the study, and the feature vector was defined as the mean of these, properly aligned, five profiles. The similarity score between two bullets was quantified as the cross-correlation function (CCF) of their respective feature vector. Because the results were based on only three firearms, the performance assessment of this 3D comparison algorithm was of course preliminary. However, the gap between the KM and KNM score distributions already demonstrated that high resolution 3D was a promising alternative to 2D imaging.

Chu *et al.*[11] presented an automated bullet signature identification based on topography images using confocal microscopy. The authors selected two barrels for each six firearm brands and fired four bullets from each firearm. A feature vector, called a signature, was computed by finding effective correlation areas and averaging pre-aligned profiles in these areas. Each bullet was correlated against its three matches and the four bullets fired from the other barrel in the pair from the same manufacturer. The expected outcome was to find the three matches at the top of the sorted list of seven scores for 48 lists. The correlation results showed a 9.3% higher accuracy rate compared to conventional 2D imaging.

In Ref. 38, the input data were the high-resolution 3D topographic images from one of the Hamby–Brundage Test sets,[18] consisting of 35 bullets fired from 10 consecutively manufactured rifled Ruger P85 pistol barrels, grouped as 10 pairs of known training bullets and 15 unknowns for comparison. The cylindrical form component was removed with filtering, and an edge detector was applied to estimate the orientation of the marks and to identify areas with strong striae. The profiles in those areas were properly aligned and averaged. Perfect separation of the match and nonmatch empirical score distribution was achieved with the CCF.

The studies listed above were all based on profile similarity measures. Chu *et al.*[12] innovated by implementing a comparison algorithm based on CMS. The conventional CMS method is a line counting method that is used by a firearm examiner, comparing either two physical bullets installed on a comparison microscope or comparing two photographs of bullets. The firearm examiner registers and manually counts groups of CMS when comparing two striated toolmarks. This method is based on empirical observations dating back to Biasotti.[7] In 1997, Biasotti and Murdock[8] stated their quantitative CMS method criteria for the objective confirmation of the common source of toolmarks. For 3D toolmarks (which include striae on fired bullets), there is an identification when "at least two different groups of at least three CMS appear in the same relative position, or one group of six CMS is in agreement in an evidence toolmark compared to a test toolmark". In terms of CMS groups, these criteria translate to a minimum of two groups of at least three lines, or a single group of at least six lines.

Chu *et al.*[12] also used a Hamby–Brundage Test set.[18] Image processing techniques were applied to create a mask of relevant points for stria detection and generation of a representative profile for each land impression. Peaks were then extracted from the profile. When comparing a pair of profiles, two peaks were said to match if they satisfy some tolerance criteria based on the stria definition parameters, and groups of CMS were computed. The authors found that none of the known nonmatching land impression pairs satisfied the CMS method criteria, while 48% of the known matching land impression pairs did. Furthermore, the 10 known bullet match pairs were objectively identified.[d]

[d] For this particular set, each bullet contains six land impression toolmarks, yielding 36 land impression pair comparisons when comparing two bullets. A single land impression pair comparison that satisfies the CMS criteria is sufficient to state that the two bullets were fired from the same firearm.

Hare *et al.*[19] used random forests, a specific ML algorithm, to successfully classify a Hamby–Brundage Test set of bullets from the NIST Ballistics Toolmark Research Database, a public high-resolution 3D ballistic database.[28] The training data were a set of 7D feature vectors, some elements of which were profile similarity measures, including the CCF, while others were based on line counting, like the CMS method. In this original approach, the feature vectors for ML were not generated from the set of individual profiles, but from the comparison of every pair of profiles. An advantage of random forests compared to some other nonlinear ML algorithms is that they provide an assessment of the importance of each type of input feature. The authors found that the two most relevant features for classification of the matches and nonmatches are the CCF and the total number of matching lines in a land impression pair comparison. Interestingly, the consecutiveness of the lines, which is at the core of the CMS method, has a much lower importance. Instead, it is the total number of matching lines — regardless of their relative position or grouping — that matters the most.

A characteristic that is shared by the aforementioned bullet studies is the small number of firearms considered in the analysis, generally 10 or less. Research on bullet toolmarks has been hampered by the challenge of capturing the topography of the bullets due to their approximate cylindrical and sometimes arbitrary shape; such capturing requires advanced surface tracking methods coupled with image stitching algorithms and controlled rotation and translation motors. Only two datasets, from NIST, are currently available for public research, both composed of 35 bullets fired from the same 10 guns. In this context, it is safe to say that, to date, there is no standard bullet database that is representative of the variability observed in forensic laboratories; the conclusions drawn from such datasets must be confirmed using representative datasets.

2.2. *Objective identification*

Objective identification methods provide meaningful scores associated with a probability measure, yet to be determined, that can reinforce the legal admissibility of expert conclusions. The most relevant probability measure for objective identification is the LR.[1] In the context of forensic firearm identification, the LR is the probability of being a match divided by the probability of being a nonmatch, based on a set of meaningful scores obtained from the comparison of a pair of toolmarks. When computed from a single score s, the LR is the ratio of the match probability density to the nonmatch probability density, given the score:

$$\mathrm{LR}(s) = \frac{f_{\mathrm{Match}}(s)}{f_{\mathrm{NonMatch}}(s)}, \tag{1}$$

where f is a probability density function (PDF).

When comparing two particular bullets (herein called the reference bullet and the test bullet), the strict application of the LR definition imposes severe constraints on the population used to build both statistical distributions. In particular, the match score distribution must be the result of comparisons between the reference bullet and a sufficiently large set of test bullets fired from the same gun as the reference bullet.[e] For a study that involves only small groups of KMs (pairs, triplets), the match score distribution of each firearm cannot be used to draw statistically valid conclusions. The remaining option is to study the typical behavior of the LR by combining all the available reference bullet match scores from various firearms into a single average match score distribution (and doing so for the nonmatch scores). In practice, it is still a challenge to build representative average match score distributions. For example, the set of pairwise comparisons over a sample of 100 pairs of KMs generates nearly 20 000 nonmatch scores, but only 100 match scores.

A modest and conservative alternative is to adopt an FMR and define a boundary that discriminates the matching and nonmatching regions in the score space in a way that is consistent with this FMR, a methodology based solely on the nonmatch score distribution. In the context of firearm identification, the FMR is the probability that two bullets (or cartridge cases) will be erroneously identified as having been fired from the same firearm. For an automated system that makes a binary decision (match or nonmatch) based on a single score, the score space is 1D and the discriminative boundary is a predefined threshold t. Assuming that typical match scores are higher than nonmatch scores, the FMR is the probability that a nonmatch comparison yields a score that is higher than t, that is, the area of the nonmatch distribution over score values that are higher than t:

$$\mathrm{FMR}(t) = \int_t^\infty dx\ f_{\mathrm{NonMatch}}(x) = 1 - C_{\mathrm{NonMatch}}(t), \tag{2}$$

where C is the nonmatch cumulative distribution function. A score that is higher than t implies that the observed score is exceptional for a nonmatch pair but draws no conclusion regarding its probability of being a match. The latter is the purpose of the LR.

No LR based on 3D score distributions of bullets has been published so far. However, this is a very active area of research for cartridge cases.[30,32,35] Riva and Champod[32] developed an automated comparison system of cartridge cases based on 3D measurements of the breech face and firing pin marks. Three scores were defined for each mark. Their analysis was based on a sample composed of three distinct groups of cartridge cases fired from Sig Sauer 9 mm firearms: a dataset of 79 cartridge cases fired by different firearms and two groups of 60 cartridge cases fired by two particular firearms. These last two groups were used to build distinct match score distributions that characterized individual firearm: nearly 1800 scores in each distribution, which is orders of magnitude larger than in any other studies. The 79 cartridge cases were also correlated against themselves, thus providing an empirical

[e] Note that when the firearm is not available for both the reference bullet and the test bullets (for example, when two bullets are recovered from distinct crime scenes, but there is no firearm), there is no way to build this distribution.

nonmatch distribution. The author found that the two KM distributions were significantly different thus demonstrating that every firearm is characterized by its own match score distribution. By approximating the score distributions with multivariate normal, they computed a mean LR of the order of 10^{20} and 10^{23} for the two KM populations.

Song[35] applied NIST's recently developed Congruent Matching Cell (CMC) method for image comparisons of breech face marks. The CMC method divides the compared topography images into squared-shaped cells and the similarity score is the number of cell pairs that meets similarity and congruency requirements. False positive and false negative error rates were computed by fitting a beta-binomial function on the discrete empirical match and nonmatch distributions. They tested the method on two datasets: 40 cartridge cases from 10 consecutively manufactured Ruger 9 mm pistols and 100 cartridge cases from 11 barrels. The empirical match and nonmatch distributions did not overlap and extremely low error rates were found. By converting the authors identification probability R_1 into an LR, we find a corresponding LR value as high as 10^{35} for the lower score extreme of the match distribution. This implies even larger values in the bulk of this distribution. As the authors state, for realistic databases, the overlap of the KM and KNM distributions can become significant and the error rates will likely increase significantly.

Current objective identification models predict LRs larger than 10^{20}, which is orders of magnitude larger than the number of firearms in the world. Such high numbers clearly support 3D imaging technologies and newly developed comparison algorithms for automated firearm identification. However, they also raise questions about the methodologies involved when modeling empirical score distributions, especially for small datasets. LRs are computed by fitting standard distributions (univariate or multivariate normals, binomial, beta-binomials, etc.) and extrapolating them beyond their valid regime.

Research on objective identification of bullets would benefit from a sufficiently larger and realistic dataset, representative of the variability encountered in the context of laboratory casework. Such dataset would also allow the use of extreme statistics[13] in order to compute matching probabilities associated with score values far from the bulk of the empirical distributions.

3. Methods

3.1. *Strategy*

Our proposed method combines two comparison scores: one associated with pattern matching and one related to line counting. We take advantage of the fact that stria patterns can be reduced to a single dimension signal. Section 3.2 describes the steps that convert a full 3D acquisition to a single depth profile for each region of interest of a bullet. This section also describes how these profiles are compared.

In Sec. 3.3, we describe the method that is used to transform the depth profiles into a binary signal representing the striae, along with the method that is used to compare two such binary signals. Three different line counting measures are defined for this comparison process. It is worth noting that all the parameters used to define a line (or stria) were tuned on the three Hamby–Brundage Test sets described in more detail in Sec. 4.1.

Because a bullet surface can be divided into several regions of interest, the comparison between two bullets involves the comparison of several pairs of regions of interest. Section 4.2 explains how a single score is computed from these multiple comparisons.

Section 4.3 presents some aspects of the behavior of the CCF and CMS comparison methods using two different topographic image resolutions. These observations govern the construction of the PMS and LCS, which are described in Secs. 4.4 and 4.5, respectively. These scores are constructed from the pattern matching and line counting measures discussed in Sec. 3.

Sections 4.6 and 4.7 provide the FMR and LR calculations. The visual representation of bullet comparisons is discussed in Sec. 4.8, with use cases from a standard dataset. We then show, in Sec. 4.9, how the method can be applied, with minor modifications, to deal with bullet fragments. In Sec. 4.10, we compare the behavior of the score distribution of a standard dataset, made of bullets fired from consecutively manufactured barrels, to that of the proprietary large dataset.

3.2. *Profile extraction and comparison*

The process that we used to extract a profile from our bullet 3D images is similar in essence to some of those used in the previously published studies,[12,19] the difference being that the process starts with the topography of a 360° band for bullets captured using BULLETTRAX. The bullet's local curvature is removed from the band, yielding a roughness image. A straightening algorithm determines the dominant orientation of the toolmarks and applies a geometric transformation to align them perpendicularly to the direction of the band. Another algorithm automatically detects the shoulders and splits the roughness image into individual land and groove images (Fig. 2).[f]

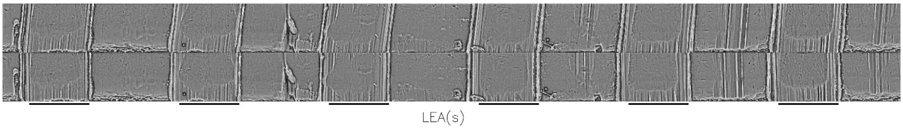


Fig. 2. Image processing over a 360° band. Original roughness image (top); straightening operation (bottom). Location of individual LEAs is shown in black segments.

[f] These algorithms are proprietary to Ultra Electronics Forensic Technology.

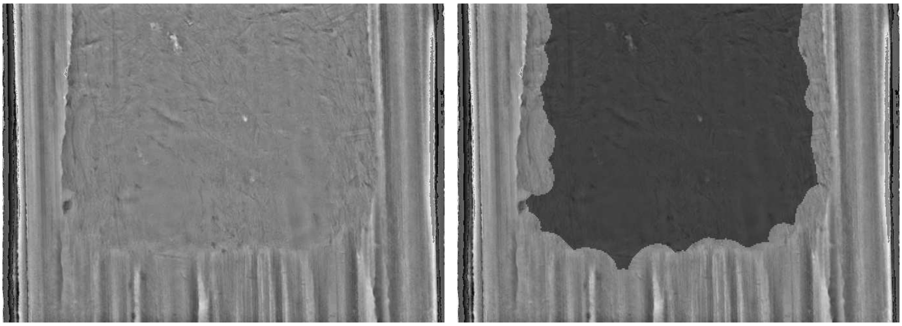


Fig. 3. Relevant areas for the profile computation in a land impression. Original image (left); original image mixed with the binary mask of valid areas (right).

As shown in Fig. 3, the toolmarks do not always extend over the whole image. In most cases, the majority of toolmarks are concentrated very near the base of the bullet. The relevant areas for objective identification are automatically emphasized using a binary mask based on a measure of local coherence and profile representation of the topography within the mask is computed. At this stage, the comparison process is reduced to computing a similarity measure between such profiles (an example of profiles is shown in Fig. 5).

Two similarity measures were evaluated. One is the global maximum of the CCF, which has proven its usefulness for objective identification of bullet profiles[11] and cartridge case areas[29,36,38]:

$$\mathrm{CCF} = \mathrm{MAX}_{\Delta}\left(\frac{1}{N}\sum_{i}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\frac{Y_{i-\Delta}-\mu_Y}{\sigma_Y}\right)\right), \tag{3}$$

where N is the number of elements in the X and Y profiles, μ_X is the average, and σ_X is the standard deviation of the X profile (and similarly for the Y profile). The final CCF score is the highest score among the set of CCF values computed over several horizontal translations (Δ) of one vector with respect to the other.

The CCF is invariant under a global change of the (vertical) scale of any of the two vectors. Other similarity measures are not scale invariant, and might then complement the CCF. We selected the Absolute Normalized Difference (AND):

$$\mathrm{AND} = 1 - \frac{\sum_i |(X_i-\mu_X)-(Y_{i-\Delta}-\mu_Y)|}{\sum_i |(X_i-\mu_X)+(Y_{i-\Delta}-\mu_Y)|}, \tag{4}$$

where Δ is determined from the CCF measure. The AND yields +1 only when the mean-corrected profiles are rigorously identical. A vanishing denominator is theoretically possible in this equation, but it was not encountered in practice.

3.3. *Line extraction and comparison*

In this study, we were interested in finding a single score based on line counting that would then be computed independently for peaks and valleys, and combined.

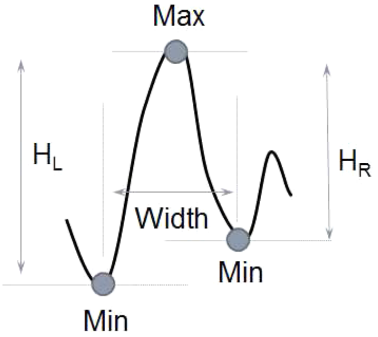


Fig. 4. Parameters of a peak: horizontal position (Max), width, left height (H_L), and right height (H_R).

The definition of the peak/valley of a profile must be part of the training process of the matching algorithm. Similar to Chu *et al.*,[12] we defined four parameters for peaks: horizontal position, width, left height, and right height (Fig. 4). The peak detection criteria based on these values were fine-tuned from the data in order to optimize objective identification performance. The valleys are characterized by the same set of parameters, computed from the reversed profile (i.e. from the profile that is the result of a reflection of the original profile across the horizontal axis).

To our knowledge, it has been tacitly assumed that the manual line counting methodology would provide comparable performance for objective identification regardless of whether the location of peaks or valleys is used when matching profiles. This might not be the case since two peaks can indeed coincide (within the required tolerance for matching them) while neighboring valleys are misaligned, and vice versa. Peaks and valleys thus provide complementary information, and the LCS to be defined must consider this fact. Hare *et al.*[19] combined peaks and valleys into a single line counting value. We used a slightly different approach, by computing two distinct LCSs, one for peaks and one for valleys.

Figure 5 summarizes the process of line counting for a reference bullet profile and a test bullet profile. The first step in the line counting process (for peaks) is to align the reference profile and the test profile according to the relative position found by the pattern matching algorithm (CCF). Each profile is then converted into an idealized binary profile. Binary operations (AND and OR) are then applied to the pair of binary profiles, and a binary vector P (for position) is created, where values of 1 indicate matching lines and values of 0 indicate nonmatching lines which are present in either one of the input profiles but not in both. The peak or valley shape of the toolmarks is not used in the comparison process, only their location is used.

Line counting analysis is easily performed by reading the P vector from left to right. A CMS group corresponds to a set of consecutive 1s that is interrupted by a 0 on its right side or by the end of the vector. The whole comparison is then summarized by the number of groups of different lengths, such as, in the

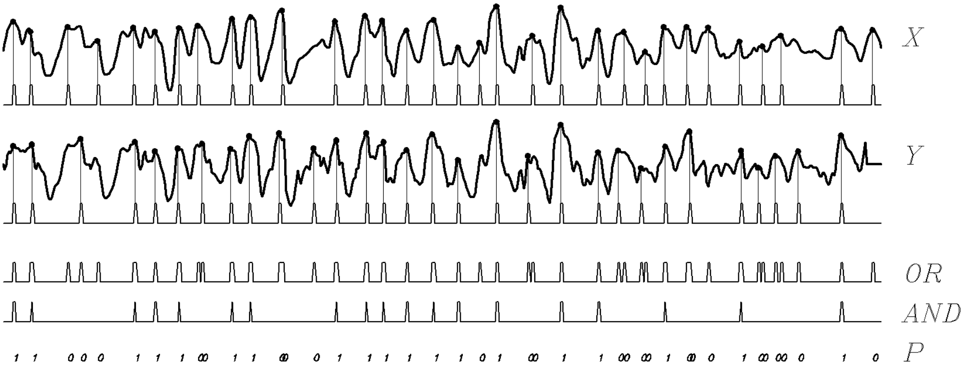


Fig. 5. Peak detection and comparison, and generation of the P vector (bottom) from a reference bullet profile X and a test bullet profile Y.

above example: four single lines, three groups of 2 CMS, one group of three and one group of six.

One possibility was to define a CMS score equal to the size of the largest CMS group (for example, a CMS score of 6 is obtained from the P vector in Fig. 5). However, no criterion based on this single score was found to be equivalent to the Biasotti and Murdock[8] criteria which are a combination of two conditions (two groups of at least 3 CMS or one group of at least 6 CMS). A better choice was to count the number of striae that belong to groups of at least three striae (i.e. neglecting isolated matching lines and pairs of matching lines):

$$S_{\mathrm{CMS}} = \sum_{i=3}^{N_{\mathrm{MAX}}} i g_i, \tag{5}$$

where N_{MAX} is the size of the largest CMS group and g_i is the number of CMS groups of length i. With this definition, a single criterion (S_{CMS} of 6 or higher) is equivalent to the two Biasotti and Murdock criteria for objective identification. Some other parameters, computed from the P vector, will be used in the next section. These are the total number of matching striae, T ($= 19$ from the example in Fig. 5); the number of consecutively matching pairs, T_2 ($= 10$); the number of CMS groups, N_G ($= 9$). Results from the S_{CMS}, T and T_2 similarity measures are discussed in Sec. 4.5.

There is an important distinction to observe at this point. In regular CMS analysis, lines are identified and counted on 2D images of the toolmark surface by a firearm examiner trained in this particular technique. Line counting methods performed by a computer, as described in Chu *et al.*[12] and presented here, rely on a linear profile averaged over a 3D recording of the toolmark topography. The parameters that distinguish a valid line from noise are usually adjusted to optimize the overall objective identification performance. Therefore, it is normal to observe some differences in the CMS counts determined by a human observer and a computer, and even those determined by different computer programs.

4. Results and Discussion

4.1. *Data*

The optimal range in the lateral and depth resolution of the 3D images used for automated objective identification is yet to be determined but, intuitively, a higher resolution carries more information and should be preferred. This is critical for a fully automated system, which requires advanced surface tracking methods coupled with controlled rotation and translation motors in order to build the full 360° band of a pristine bullet or the arbitrary shape of a fragment.

In practice, higher lateral resolution is obtained at the cost of a reduced field of view, which translates into a need to capture a larger number of images, thus increasing acquisition time. The resulting band is also narrower, which limits the area available for firearm identification. A larger number of images also raises the possibility of errors caused by bullet manipulation and/or image stitching when the acquisition process is manual. Furthermore, a smaller field of view negatively affects the automated image acquisition and bullet-tracking processes, since the base of some pristine bullets is far from being flat, and deformed bullets and fragments can be much more difficult to manage.

The data used in this study consisted of four different sets of pristine bullets as follows:

- Two different Hamby–Brundage Test sets from the NIST public database[28]: Each set corresponds to 10 pairs of bullets fired from consecutively manufactured barrels and 15 unknowns. Both image sets consist of individual land impressions captured using a 3D confocal microscope equipped with a 20× objective at 1.5625 μm/pixel resolution. One of these two sets has been used in the previous studies.[12,19]
- A third Hamby–Brundage Test set, captured using BULLETTRAX and the same type of 3D confocal microscope, but equipped with a 10× objective at 3.125 μm/pixel resolution. Bullet images are 360° bands built by stitching together several images that are captured by an automated surface tracking acquisition algorithm. The physical bullets are distinct from those in the NIST database, but belong to the same test (i.e. fired from the same set of barrels).
- The fourth set consisted of a dataset of pristine, copper jacketed, bullets from different ammunition manufacturers captured using BULLETTRAX. The 9 mm bullets of the dataset (six grooves, right twist) were fired from 136 firearms (from pairs to quintuples, for a total of 406 bullets) from different firearm manufacturers that are representative of the population of hand guns in an urban US environment (Table I).

Before any computations were performed, every KM pair of the fourth set was inspected using a 3D virtual comparison microscope, and was labeled as visually matching or not. In this study, we define the subjective visual matching criterion as a minimum of two distinct land impression pair comparisons in a common phase[g]

[g] The concept of phase will be defined in Sec. 3.1.

where sufficient agreement is observed. The fraction of confirmed visually matching pairs, called the Visual Matching Ratio (VMR) in this paper, is 55% for this dataset. This dataset is therefore more representative of real casework than the Hamby–Brundage Test sets, where every KM pair can be matched visually according to the visual matching criterion mentioned above. The purpose of this labeling was to focus on reasonably good matches when training the algorithms. In our opinion, there is no point in trying to match bullets that cannot be matched visually.

Bachrach[5] also concluded that a high percentage of barrels does not effectively impart the individual characteristics left on the inside of the barrel during the manufacturing processes onto fired bullets. In their ambitious study, they used the SCICLOPS$^{\text{TM}}$ system to analyze bullets fired from nearly 80 different barrels from 8 different manufacturers. They determined that the barrel manufacturer is the most dominant factor in both the individuality and classification performance of the bullets fired from it. For example, bullets fired from Ruger, Beretta, and Smith & Wesson barrels could be identified with very low probability of error, while at the other extreme, firearm identification from bullets fired from Hi-Point or SIG Sauer barrels was very limited. The second dominant factor was the ammunition brand. Furthermore, significant variations were observed on the quality of the individual marks even within a given barrel model and using the same ammunition.

To our knowledge, the VMR is not considered in most studies. This might stem from the fact that several firearm identification tests, such as the Hamby–Brundage Test used here, have a VMR = 100%. Therefore, all KM pairs are conclusive matches. An example of good and bad matching pairs from our proprietary dataset is shown in Fig. 6.

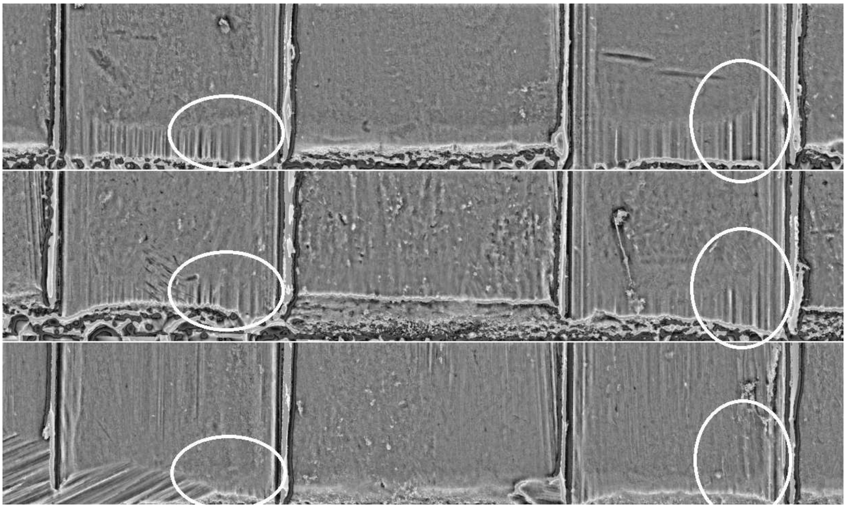


Fig. 6. Portion of a 360° band from three known matching bullets aligned at their common optimal phase. The top two bands match visually within the areas indicated by the ellipses. The bottom band is a KM but cannot be visually confirmed to match either of the other two.

4.2. *Bullet comparison score strategy*

Each land of a firearm barrel is like a unique tool, independent of the other lands, leaving distinctive toolmarks, called land engraved areas (LEAs), on the fired bullet's surface. The comparison of two pristine bullets showing N LEAs (i.e. not fragments) leads to the possibility of N^2 LEA-to-LEA comparisons. However, since the sequence of the LEAs is fixed inside the barrel, these N^2 possibilities can be arranged into N groups, called phases, of N LEA-to-LEA comparisons with consistent ordering.

Finding the correct phase, by considering either the phase with the best overall LEA-to-LEA comparison or the phase which shows the best average agreement, is usually the first step in bullet identification. Firearm examiners rely on pattern matching techniques to conduct this step; the CMS method is being used only as a quantitative measure of the pattern agreement after the two compared LEAs have been properly aligned. In this study, the best average CCF value was used to find the correct phase; the profile alignment (i.e. the Δ value in Eq. (3)) that yielded the CCF score for each LEA-to-LEA comparison was used for all line counting-type scores.

What remained was the question of how many LEAs, or distinct toolmarks, must be considered in the construction of an objective score for the bullet pair as a whole. Our choice had to consider two elements: (i) firearm examiners, in practice, rarely draw final conclusions based only on one LEA-to-LEA comparison (they look for similarities over the whole circumference of the bullet), and (ii) bullet fragments from crime scenes can have missing LEAs. Consequently, developing a similarity measure based on the comparison of all possible LEA-to-LEA comparisons at the best phase from datasets of pristine bullets can be of no practical value for the objective identification of crime evidence.

As will be discussed as follows, the average of the best two LEA-to-LEA comparisons offers a good compromise. Therefore, for the new types of scores being presented in this paper (i.e. the newly developed PMS and LCS), the score value represents the average of the best two LEA-to-LEA comparisons within the selected phase. The selection of the LEAs to be part of the score relies on the PMS values. The LCS is computed for those LEA-to-LEA comparisons with optimal alignment obtained from the CCF. This best two-score approach is also consistent with our visual matching criterion described earlier.

4.3. *Behavior of the CCF and CMS on the Hamby–Brundage test sets*

We began by evaluating the performance of the most common quantitative comparison measures, namely the CCF and the CMS, applied on the three Hamby–Brundage sets.

For every pair of bullets, the representative CCF score was first computed as the highest CCF value from the N LEA-to-LEA comparisons at the best phase. Figure 7(top) shows the distribution of the match and nonmatch CCF scores for

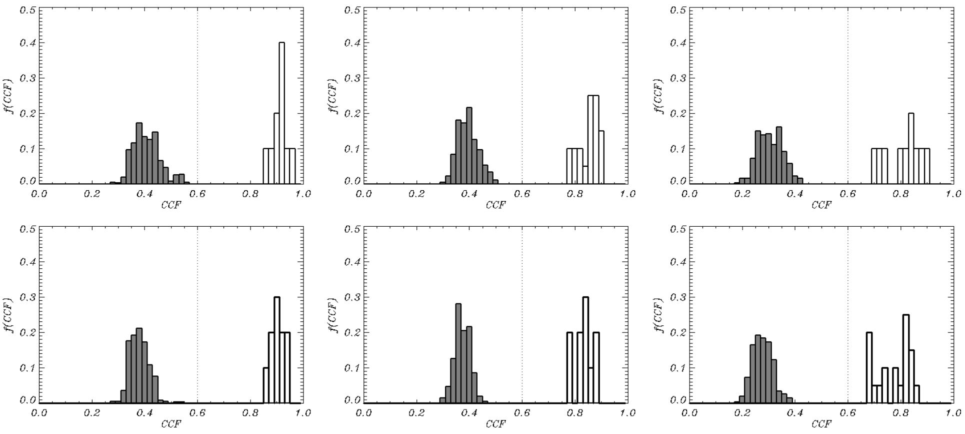


Fig. 7. Distribution of the CCF for the three Hamby–Brundage sets: sets 1 and 2 at 20×, and set 3 at 10×, from left to right. Best LEA-to-LEA comparison from each bullet pair (top); average of the best two LEA-to-LEA comparisons at the same phase from each bullet pair (bottom).

the three Hamby–Brundage Test sets in this scenario. The CCF unequivocally discriminates the two classes: match and nonmatch. The gap between the match and nonmatch score distributions, as measured by the Fisher ratio,[h] is 6.9 and 5.9 for the two sets captured with a 20× objective. This gap is reduced slightly to 4.3 for the set captured with a 10× objective, which already shows that pixel resolution has some impact on the discriminative power of the CCF. Next, by redefining the representative CCF score as the average of the best two out of N LEA-to-LEA comparisons at the best phase, the gap between both distributions improves significantly (Fig. 7(bottom)). The Fisher ratio is now 9.1 and 6.8 for the 20× sets and 4.7 for the 10× set.

As discussed previously, we have defined the CMS score (S_{CMS}) for a LEA-to-LEA comparison as the number of striae that belongs to groups of at least three CMS. Following the standard methodology for CMS, the representative score value of the bullet pair comparison is the highest among the N CMS scores computed at the best phase.

Since the computed CMS score is a function of several parameters (see Figs. 4 and 5 for peak definition and alignment tolerance when creating the P vector), a systematic search in the parameter space was first performed in order to find the combination that optimizes performance in the three Hamby–Brundage Test sets. With the optimal parameters for the definition of a stria, the CMS method criteria (equivalent to $S_{\text{CMS}} \geq 6$) applied to peaks yield a 0% FMR for the three sets. The sensitivity of the CMS method (i.e. the proportion of the KMs that satisfies the criteria) varies significantly over the three Hamby–Brundage Test sets: 100% for the two sets of the NIST public database and 70% for the set captured using BULLETTRAX.

[h] The definition of the Fisher ratio used is the absolute difference of the averages divided by the sum of the standard deviations.

This sensitivity is obtained for datasets that yield a 100% VMR and that show clear class separation from pattern matching analysis.

Taken at face value, objective identification with CCF and CMS score distributions favors pixel resolution consistent with the 20× magnification rather than 10×. However, as discussed previously, improving the pixel resolution cannot be done without significant costs in terms of image capture automation and duration, stitching and manipulation accuracy, and usable width of the captured topography.

An alternative to selecting 20× magnification is to focus research on quantitative measures of similarity that better segregate matches and nonmatches than CCF for data captured at 10× magnification, and that show better sensitivity than the CMS method for bullet pairs that can be visually matched. The next sections describe the PMS and the LCS that have been developed to achieve this goal.

4.4. *Pattern matching score*

Our PMS is defined as the combination of the CCF and the AND that optimizes the separation between the match and nonmatch distributions for our large dataset derived from a population of 136 firearms of various makes and models. The representative value of the CCF and the AND scores is computed as the respective average value over the LEAs that yield the two highest CCF scores at the best phase.

The score distribution of the visually confirmed matches lies along a straight line that passes close to the origin (correlation coefficient = 0.96) in the CCF–AND plane while the nonmatch is nearly gaussian with significant variance along the direction perpendicular to its main axis (Fig. 8). It is then natural to define the PMS as a linear

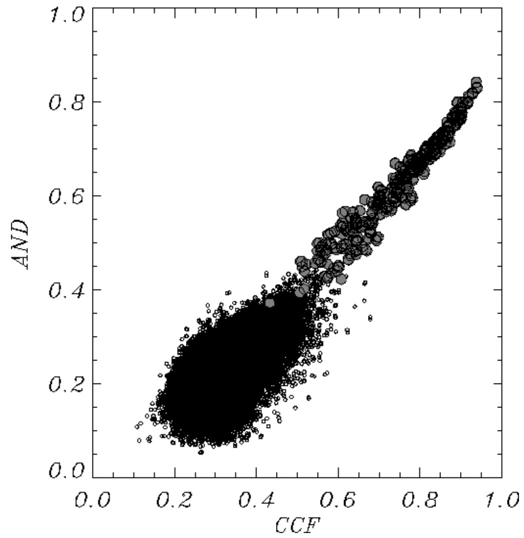


Fig. 8. 2D statistical distribution of the CCF and AND. Visually confirmed KM scores (large gray circles, top right) and nonmatch scores (small black circles, bottom left).

combination of the CCF and the AND, with their respective weight computed from the eigenvector of the main principal component of the match distribution:

$$\text{PMS} = w\ \text{CCF} + (1 - w)\ \text{AND}, \tag{6}$$

where the weights add up to 1 to ensure that the upper bound of the PMS is +1. In practice, both weights are very close to 0.5.

4.5. *Line Counting Score*

Figure 9 shows the distributions of the S_{CMS}, for peaks, and for visually confirmed matching and nonmatching bullet pairs of our proprietary dataset. Consistent with the published studies[7,8,26] and algorithms,[12] the FMR is very small with acceptance of the CMS threshold at 6: 0.37% for CMS_{PEAK} for 81 793 comparisons. However, the sensitivity of the method is rather low, with only 58% of the visually confirmed matches satisfying the CMS method criteria based on the S_{CMS}.

However, results from Ref. 19 suggest that the total number of matching striae, regardless of their consecutiveness, is the most useful line counting-type score for objective bullet matching using a random forest algorithm. This score would then simply be expressed as T, the sum of the binary elements in the P vector.

The 2D (peak and valley) visually confirmed match and nonmatch T distributions strongly overlap, which makes T unsuited for objective identification in this dataset (Fig. 10(left)). However, dividing T by N_P, the number of elements in the P vector, significantly reduces the overlap. This binary similarity measure is the Jaccard

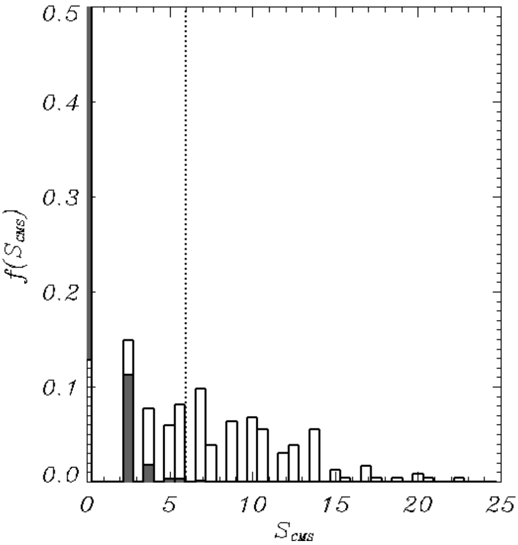


Fig. 9. Statistical distribution of S_{CMS} based on peaks for visually confirmed matches (white) and nonmatches (gray). The CMS threshold is shown as a vertical dotted line.

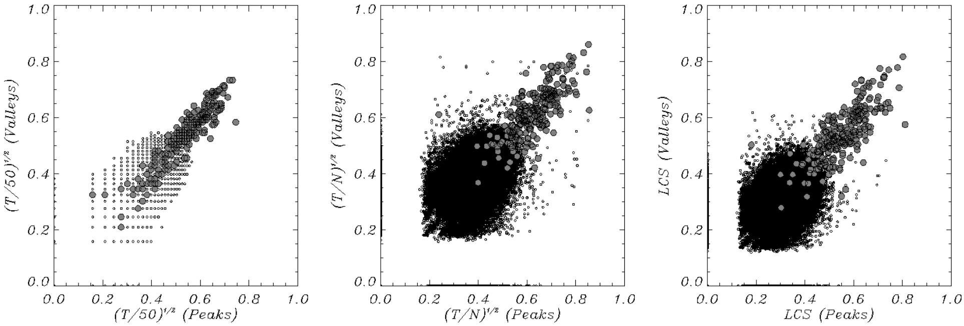


Fig. 10. 2D statistical distributions of three types of LCSs for visually confirmed matches (large gray circles) and nonmatches (small black circles). Square root of $T/50$ (left), T/N_P (center), and $(T + T_2)/2N_P$ (right).

coefficient,[10] which always lies between 0 and 1. Figure 10 shows the distributions for $\sqrt{T/50}$ and $\sqrt{T/N_P}$, again, the average of the best two matching LEA-to-LEA comparisons at the best phase. A constant rescaling factor ($= 50$) is applied to T in this figure to ensure that both measures can be shown on the same scale, thus helping visual comparison. The square root is added for convenience in order to obtain a threshold that segregates matches and nonmatches near 0.5.

Since the total number of striae does not consider the potentially positive contribution from consecutiveness, we tested different linear combinations of T and a new term, T_2, defined as the number of consecutively matching pairs in the P vector (i.e. the number of pairs of 1s). Similar to the individual striae used to compute T, the pairs can be distributed anywhere in the P vector, that is, they may be CMS pairs or parts of a larger CMS group. It was found that the quantity $(T + T_2)/N_P$ slightly improves the compactness of the nonmatch distribution by shifting some outliers toward the center of mass of the nonmatch distribution (Fig. 10(right)). Other numerical experiments with similar contributions from larger series of consecutive 1s in the P vector $(T_3, T_4, \ldots)$ did not improve the separation between matches and nonmatches.

As a final refinement, we analyzed the nonmatching LEA-to-LEA comparisons which yield score values (Peak or Valley) greater than 0.5; those are outsiders in Fig. 10(right). Most of them were characterized by a small number of striae on both LEAs, i.e. small N_P, which makes them irrelevant for firearm identification. We thus applied a penalty multiplicative factor which reduces the score when N_P is smaller than some integer N_{MIN}. The optimal value of N_{MIN} was found to be equal to 6 by experiment.

We adopted the following LCS for peaks and valleys:

$$\text{LCS}_{\substack{\text{PEAK}\\ \text{VALLEY}}} = \sqrt{\frac{1}{2}\left(\frac{T + T_2}{N'_P}\right)_{\substack{\text{PEAK}\\ \text{VALLEY}}}} = \sqrt{\frac{T}{N'_P}\left(1 - \frac{N_G}{2T}\right)_{\substack{\text{PEAK}\\ \text{VALLEY}}}}\,, \tag{7}$$

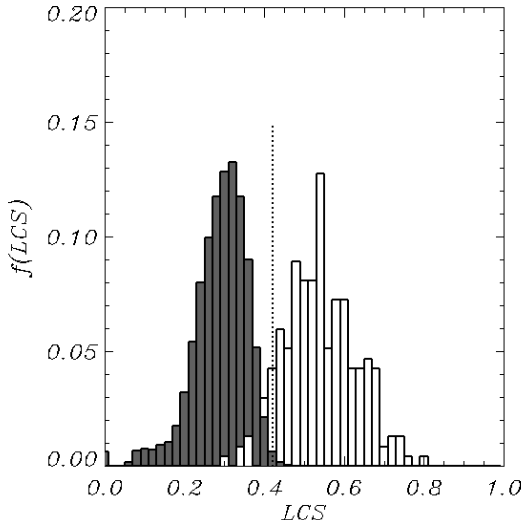


Fig. 11. Statistical distribution of the LCS for visually confirmed matches (white) and nonmatches (gray).

where $N'_P \equiv \text{Max}(N_P, N_{\text{MIN}})$; the $\frac{1}{2}$ factor yields an upper bound near unity and the square root ensures that the threshold segregating matches and nonmatches is about 0.5; it can be shown that both mathematical expressions in (7) are equivalent from the identity $T_2 = T - N_G$. The rightmost one can be interpreted as the (square root) product of two contributions: the normalized number of matching striae and a correction term which penalizes a high number of CMS groups, N_G, for a given T.

Following the same reasoning as for the PMS, the final LCS is defined as a weighted sum of contributions from peaks and valleys:

$$\text{LCS} = \frac{1}{2}\text{LCS}_{\text{PEAK}} + \frac{1}{2}\text{LCS}_{\text{VALLEY}}, \tag{8}$$

with equal weights in this case, as computed from the behavior of the visually confirmed match distribution (Fig. 10(right)).

When using an LCS threshold value that yields a similar FMR as with the regular CMS method criteria (i.e. below 1%), this new LCS shows a much better sensitivity of 93% (Fig. 11), compared to 58% for the CMS score (Fig. 9).

4.6. *PMS and LCS: FMR*

The PMS and the LCS were computed for every pair of bullets in the large proprietary dataset. Figure 12 shows the scores for all known matching and nonmatching bullet pairs, that is, 422 match scores and 81 793 nonmatch scores, and for the subset that includes only the 235 visually confirmed match pairs. The PMS and the LCS are highly correlated for the visually confirmed known matching pairs, and follow a straight line that passes close to the origin. The line that best fits the match scores distribution is found with PCA analysis.

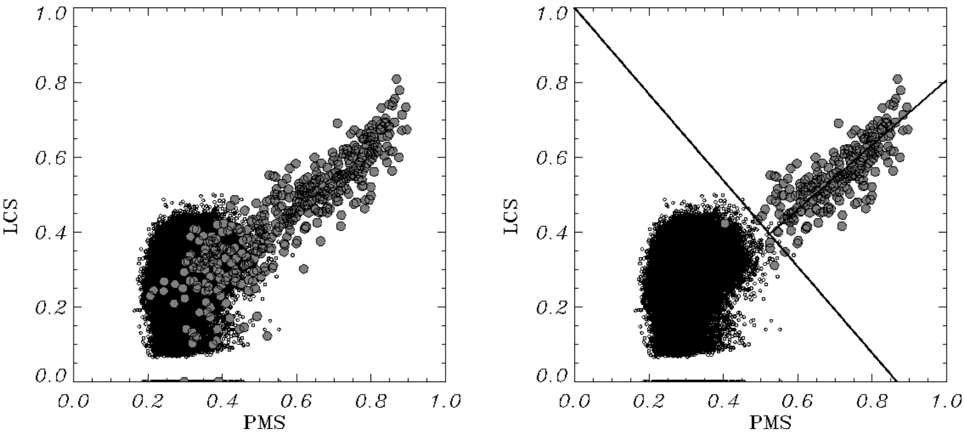


Fig. 12. 2D statistical distributions of the PMS and LCS. All nonmatches and all matches (left); all nonmatches and visually confirmed matches, with the best-fitting line (right). The linear decision boundary corresponds to an FMR of 1/10 000.

An FMR function is obtained by performing an orthogonal projection of the nonmatch 2D scores on the best-fitting line, in order to obtain a distribution of the projected distance D between the projected points and a fixed point on the line. In this process, we are in fact defining a single score D that linearly combines the PMS and the LCS. This score can be used to define an FMR as discussed as follows, but it cannot be interpreted easily by firearm examiners since it is the weighted sum of contributions from pattern matching and line counting methods. We thus proposed to keep the 2D representation for ease of interpretation, but used their linear combination D for the computation of error rates. The D score match and nonmatch probability densities are shown in Fig. 13.

The FMR associated with a given projected distance D is the area of the nonmatch distribution for distances larger than D:

$$\mathrm{FMR}(D) = \int_D^\infty dD' f(D')_{\mathrm{NonMatch}} = 1 - C(D)_{\mathrm{NonMatch}}, \tag{9}$$

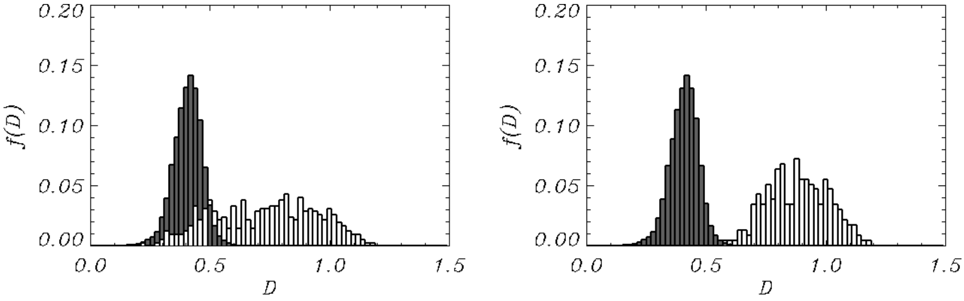


Fig. 13. Probability density of the nonmatch (dark gray) and match distributions (pale gray) for the projected distance D. All match scores (left) and scores from visually confirmed match pairs only (right).

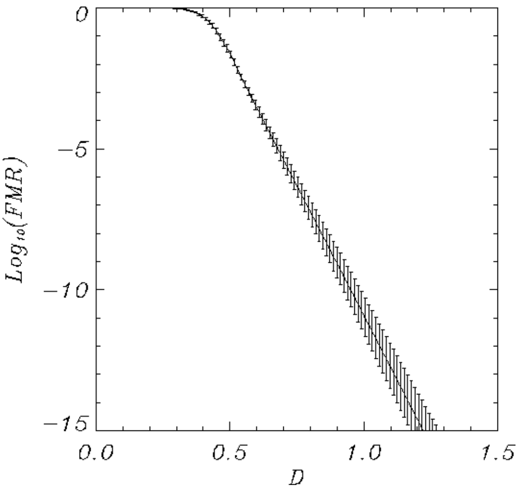


Fig. 14. FMR function versus projected distance D with bootstrapped 95% error interval. The extrapolation is applied for $\log_{10}(\mathrm{FMR}) < -4$.

where $f(D)$ is the nonmatch distribution density of D, and $C(D)$ is its cumulative. A linear decision boundary corresponding to an FMR of 1/10 000 is shown in Fig. 12 (right). The corresponding sensitivity (i.e. the proportion of the visually confirmed KMs that satisfy the criterion) is 98%. The sensitivity of the CCF alone at the same FMR value is 89%.

As a reminder, these results were obtained by averaging the LCSs and PMSs over the best two LEA-to-LEA comparisons. This analysis was repeated by considering only the single best LEA-to-LEA comparison. Very similar results were obtained, with a slightly lower sensitivity of 93%, thus supporting the choice of the two-score averaging process.

The computation of 95% error intervals of the FMR function was realized using bootstrapping (Fig. 14). For each bootstrap iteration, the nonmatch and visually confirmed match distributions were resampled with repetition; the weights of the linear combination of the PMS and LCS was recalculated from the first principal component of the bootstrap match distribution, and an FMR function was recomputed based on these weights and the bootstrap nonmatch distribution. For FMRs smaller than 1/10 000, there is very little data from the nonmatch distribution; the FMR is extrapolated based on the observed linear behavior of $\log_{10}(\mathrm{FMR})$ in the $(-4, -2)$ range, that is, from the highest 1% nonmatch scores (nearly 800 scores). Bootstrapping was also used to determine a 95% interval for the extrapolated values. For D values larger than about 1, the error bar spans a range that is over one order of magnitude. The extrapolation is discussed in the next section.

4.7. *LR function and extrapolation of the nonmatch distribution*

An LR function characterizing our larger dataset can be computed as the ratio of the match and nonmatch probability densities of the projected distance score D.

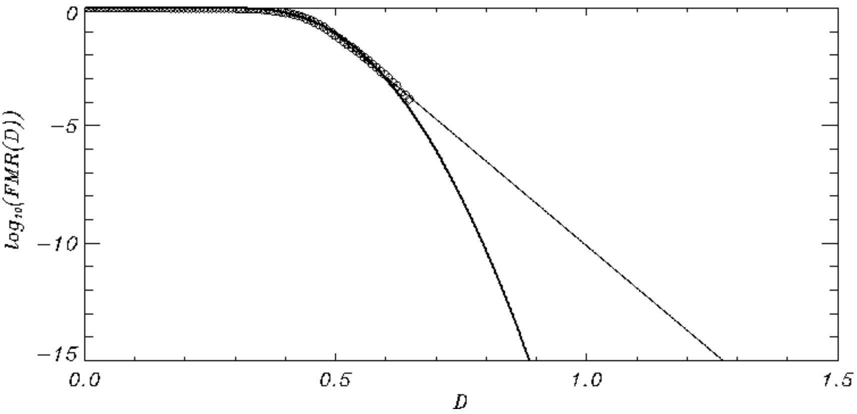


Fig. 15. Extrapolation of the FMR function based on linear fitting (thin) over extreme scores and global Gaussian fitting (thick). The experimental data (dark circles) and the Gaussian fit start to diverge for $\log_{10}(1-C)$ near -2.

For D values of interest, the nonmatch distribution must be extrapolated beyond the highest available score values. Visual inspection of Fig. 13 highly suggests that the nonmatch distribution has a Gaussian shape; a reasonable approximation then seems to fit a normal distribution over the whole set of nonmatch scores and use the resulting analytical function for extrapolating over high scores.

However, a Gaussian shape is not consistent with the observed behavior of the nonmatch cumulative distribution for high scores. As shown in Fig. 15, the right wing of the nonmatch distribution and the Gaussian that best fits the whole distribution diverges for 1% highest scores; the logarithm of the complementary cumulative distribution $1-C$ (or, equivalently, of $\mathrm{FMR}(D)$, the FMR function of D) becomes linear while the Gaussian function has an approximate quadratic asymptotic form.

An alternative to global fitting is to perform statistical modeling of the highest scores of the experimental distribution, which is the object of extreme value theory. For univariate distributions, four families of extreme value distributions are commonly used to model block maxima or right tails[13]: the Gumbel, the Fréchet, the (Reversed) Weibull and the generalized Pareto distributions. The observed linear behavior of $\log(1-C)$ for large scores is consistent with modeling the wing of the experimental distribution by an exponential distribution:

$$\log_e(1-C(x)) = -\frac{x-\mu}{\sigma} + c \rightarrow P(X<x|X>\mu) = 1 - e^{-\frac{x-\mu}{\sigma}}, \tag{10}$$

where μ is some predefined threshold, and σ and c are free parameters to be determined from a fitting procedure. The exponential distribution is a limiting case of the generalized Pareto distribution:

$$\mathrm{GPD}(x;\mu,\sigma,\xi) = 1 - \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}, \tag{11}$$

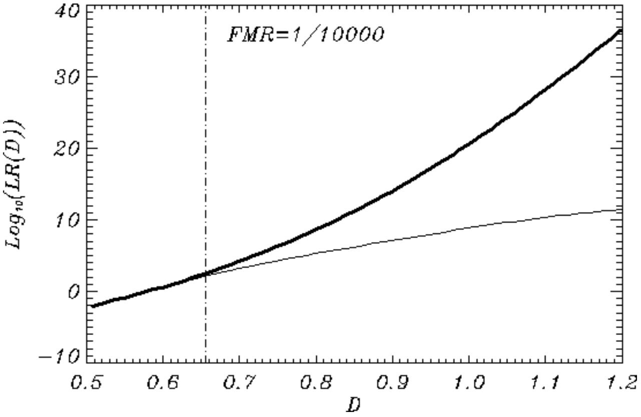


Fig. 16. LR function: linear extrapolation (thin), Gaussian fitting of the whole distribution (thick). The D value corresponding to FMR = 1/10 000 (0.66) is along the vertical dotted line.

in the limit where ξ converges to 0. Such a linear fitting yields the extrapolation of the data presented in Figs. 14 and 15.

The LR function is computed by approximating the nonmatch distribution by two analytical functions: the Gaussian fit for D values smaller than 0.55 and the PDF associated with the linear extrapolation previously discussed, for larger D values. The Parzen-window technique is used to smooth the match distribution and takes the whole set of matches (including those that cannot be visually confirmed) into account. The LR value corresponding to FMR = 1/10 000 is $10^{1.95}$ (≈ 90). Using only the Gaussian fit would overestimate the LR of the visual matches by several orders of magnitude (Fig. 16).

4.8. *Use cases*

The linear boundary with FMR = 1/10 000 in the PMS-LCS plane and the LR(D) function could be a valuable quantitative criterion in the context of proficiency tests.

Figure 17 summarizes the output of the method on the PMS–LCS plane for two of the 15 bullets considered as unknown in the Hamby–Brundage Test set captured at 10× (similar results are obtained from every bullet of this set). Each unknown bullet is compared with 10 known matching pairs. The 20 score vectors are shown as black dots, among which two are above the discriminative line; this pair and the tested bullet are connected to the same barrel.[17] All nonmatches are concentrated below the line. The figure also displays the empirical distribution of nonmatches and visually confirmed matches from our dataset, to be used as a visual reference. Isolines of ($-\log_{10}$) FMR and ($\log_{10}$) LRs computed from our dataset provide a quantitative assessment of the similarity between the matches. The 30 LR values computed from the projected distance D between each unknown and its best matching pair are 10^3 or greater (Fig. 18).

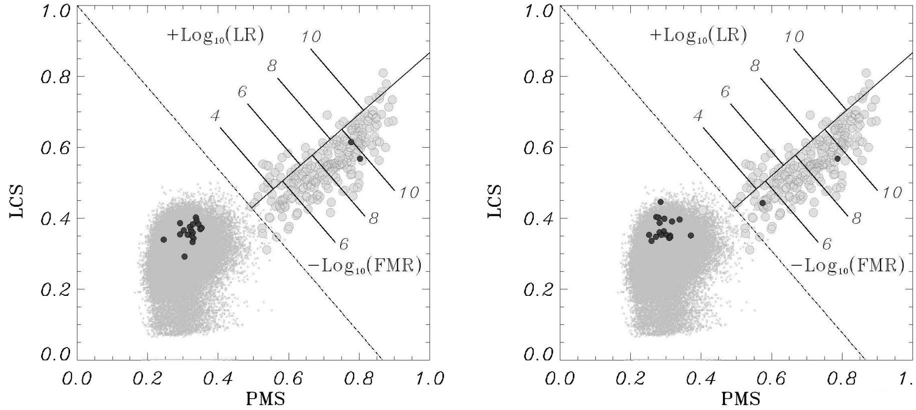


Fig. 17. For two unknowns of a Hamby–Brundage set, scores against the set of 10 pairs of KMs (black) superimposed on the score distributions of the large dataset (gray) and isolines of ($-\log_{10}$) FMR and ($\log_{10}$) LR. The matching pairs stand out.

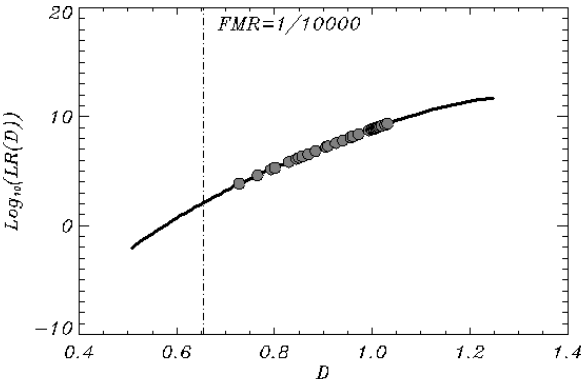


Fig. 18. For the 15 unknowns of the Hamby–Brundage Test set, LR values (logarithmic scale) against their respective matching pairs of KMs superimposed on the LR function. The D value corresponding to FMR $= 1/10\,000$ is along the vertical dotted line.

We also show the proposed methodology for a group of five visually confirmed matches from our large dataset (Table 1). In Fig. 19, the five wrap-around images can be compared. While the agreement between every pairs of bullets is excellent, it is seen that bullet #5 (at the bottom) has a smaller number of well-matching striae.

Table 1. Description of the fourth set of bullets.

KM Group Size	Number of Firearms	Number of Bullets
2	17	34
3	107	321
4	9	36
5	3	15
Total:	136	406

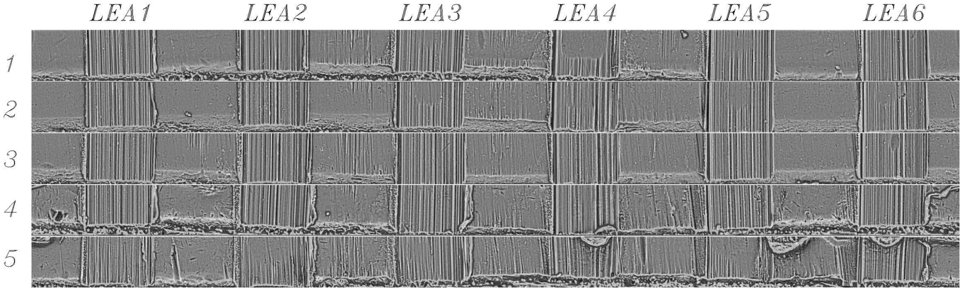


Fig. 19. High-resolution 3D roughness images for a matching quintuple.

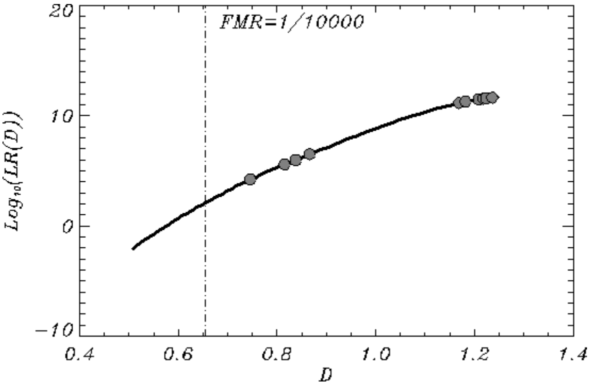


Fig. 20. For the 10 possible pairs of a visually confirmed matching quintuple, LR values superimposed on the LR function.

The LR values of the four pairs where bullet #5 is involved are in the range of 10^4–10^6; the remaining six pairs have LR values over 10^{10} (Fig. 20).

4.9. *Application of the presented method on bullet fragments*

As discussed in Sec. 4.2, the presented method is based on a two-score approach for which all computed scores (CCF, AND, $\mathrm{LCS}_{\mathrm{PEAK}}$, $\mathrm{LCS}_{\mathrm{VALLEY}}$) represent the average of the best two LEA-to-LEA comparisons within the selected phase. This method is obviously not directly applicable to bullet fragments with only one available land impression because the average cannot be calculated.

Using a score computed from a single available land impression on a fragment (as if it were the average of the best two values) would generate a biased result since the statistical models in this study were built from experimental score distributions based on the two-score approach. The solution, for this particular case, would be to build a second PMS–LCS statistical model using the best score approach (which, however, would yield a slightly lower sensitivity, as discussed in Sec. 4.6). The probabilities associated with bullet fragments with only one land impression could then be computed using this second model.

4.10. *The challenge of consecutively manufactured barrels*

Consecutively manufactured barrels are sometimes described as the most challenging, or worst case, part of ballistic toolmark analysis.[34] Such barrels sometimes share common microscopic marks, called subclass characteristics, which are transferred by the common tool that machined these barrels. In the 6th version of the AFTE glossary, subclass characteristics are defined as "features that may be produced during manufacture that are consistent among items fabricated by the same tool in the same approximate state of wear".[3] It is possible that subclass characteristics be erroneously interpreted as individual marks (i.e. characteristic of a given barrel) by a firearm examiner not trained on subclass characteristics, or by an algorithm.

In the Hamby–Brundage Test, subclass characteristics were not an issue for 502 individuals who examined the bullets using a comparison microscope.[18] This is consistent with our score results. Figure 21 compares the match and nonmatch score distributions computed from the 10 known pairs in our third set of images of a Hamby–Brundage Test and the large proprietary dataset (both sets of images have the same lateral resolution). The match scores of the Hamby–Brundage Test are all above the discriminating line. More importantly, all nonmatch scores of this set are below the line. Furthermore, several nonmatch scores of the proprietary dataset are much nearer to the discriminating line than any nonmatch of the Hamby–Brundage Test. According to the comparison algorithm, the Hamby–Brundage Test, while being made of bullets fired from consecutively manufactured barrel, is less challenging than a realistic set with no consecutively made barrels. However, this test still confirms the random nature of observed marks.

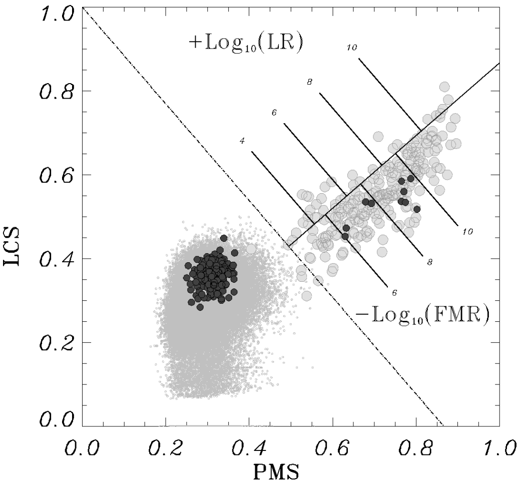


Fig. 21. For the 10 known pairs of a Hamby–Brundage set, match and nonmatch scores (black) superimposed on the score distributions of the large dataset (gray) and isolines of $(-\log_{10})$ FMR and $(\log_{10})$ LR.

5. Conclusion

Using line counting and pattern matching methods, two new objective identification scores have been developed based on high-resolution 3D topography for bullets fired from conventionally rifled barrels. The PMS is a linear combination of the CCF and the AND similarity measure; the LCS is the arithmetic average of a quantity calculated separately for peaks and valleys, and is based on the normalized number of matching striae and a correction term of order unity sensitive to consecutiveness. The two scores, treated as a 2D feature vector, yield perfect match versus nonmatch separation of the two standard Hamby–Brundage Test sets, and a third set captured at twice this resolution.

A similar analysis done on a larger, more realistic, set captured with a lateral resolution of 3.125 μm per pixel, allowed us to define a discriminative line at the FMR $= 1/10\,000$ level in a 2D plot that shows both objective identification scores for matches and nonmatches. The proportion of the visually confirmed KMs that satisfy the criterion (sensitivity) is 98%, which is larger than the corresponding proportion with the CCF alone, 89%. The LCS alone is also found to have a better sensitivity than CMS, 93% versus 58%.

The larger level of lateral resolution used in our study (3.125 μm per pixel) provides sufficient information for the objective identification of bullets fired from conventionally rifled barrels. This lateral resolution also inherently provides a larger field-of-view, which facilitates acquisition automation and visual analysis using a virtual comparison microscope.

A linear combination of the PMS and the LCS yields a more abstract univariate score, called the projected distance D. Its nonmatch probability density looks like a perfect Gaussian distribution. However, its right wing diverges significantly from the Gaussian behavior, and the highest 1% of scores better fit an exponential distribution, which is a special case of the generalized Pareto distributions used in extreme value theory. This fitting was used to compute the LR for high D scores not sampled by the experimental nonmatch distribution.

Any FMR or LR computed in the high score regime is evidently highly speculative due to missing data from the nonmatch distribution. For extrapolations beyond the nonmatch distribution, the huge difference between the prediction of the Gaussian and exponential fits emphasizes the importance of using large datasets. The exponential fit was based on the highest 1% of more than 80 000 nonmatch scores, that is, 800 values, with a behavior that significantly differs from the asymptotic right wing of the normal distribution. Conversely, this suggests that LR estimations, based on analytic extrapolation PDFs fitted on an entire empirical distribution of a few hundred KNM scores, are flawed.

The result of a bullet pair comparison can be displayed using a visualization tool that shows the strength of the pair's PMS and LCS relative to the isolines of FMR and LRs, with a discriminative line at FMR $= 1/10\,000$, and with representative distributions of match and nonmatch score values.

An analysis of Fig. 21 reveals that pairs of bullets fired from different brands of firearms can show better comparison scores than the best KNM scores obtained from pairs of bullets from consecutively manufactured bullets of the Hamby–Brundage Test set. Consequently, deriving LR estimations and objective identification methods based on such limited datasets is misleading.

The consecutively manufactured barrels used for the construction of the Hamby–Brundage Test sets do not show a significant quantity of subclass characteristics, such as striae that extend over the whole barrel length and carry over to the next manufactured barrel. It would be interesting to perform the same analysis on a set of Ruger LC9 consecutively manufactured barrels, which are known to produce a high quantity of subclass characteristics, to evaluate the robustness of the new scores under such conditions and possibly develop strategies to overcome the influence of these subclass characteristics.

Acknowledgments

We are grateful to the Allegheny County Forensic Laboratory Division for providing the bullets for the large dataset in this study; to Jim Hamby, who provided the physical bullets from a Hamby–Brundage set; to Michel Paradis and Marlene Reed for reviewing the manuscript and providing valuable suggestions; and to Debbra Desrosiers, who performed all bullet 3D digital image captures.

This study was supported by Ultra Electronic Forensic Technology.

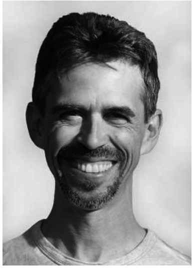
References

1. C. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, 2nd edn. (John Wiley & Sons Inc., West Sussex, 2004).
2. Association of Firearm and Tool Mark Examiners (AFTE) Criteria for Identification Committee, Theory of identification, range of striae comparison reports and modified glossary definitions, *AFTE J.* **24** (1992) 336–340.
3. Association of Firearm and Tool Mark Examiners (AFTE) Standardization and Training Committee, AFTE Glossary, sixth edn. Version 6.030317 (AFTE, Albuquerque, USA, 2013).
4. B. Bachrach, Development of a 3D-based automated firearms evidence comparison system, *J. Forensic Sci.* **47**(6) (2002) 1253–1264.
5. B. Bachrach, A statistical validation of the individuality of guns using 3D images of bullets, National Institute of Justice, Washington DC, NCJ No. 213674 (2006).
6. A. Banno, T. Masuda and I. Katsushi, Three-dimensional visualization and comparison of impressions on fired bullets, *Forensic Sci. Int.* **140** (2004) 233–240.
7. A. A. Biasotti, A statistical study of the individual characteristics of fired bullets, *J. Forensic Sci.* **4** (1959) 34–50.
8. A. A. Biasotti and J. E. Murdock, Firearms and toolmark identification: Legal issues and scientific status, in *Modern Scientific Evidence: The Law and Science of Expert Testimony*, eds. D. L. Faigman, D. H. Kaye, M. J. Saks, J. Sanders, 1st edn. (West Publishing Co., St. Paul, 1997).

9. S. Bigdeli, H. Danandeh and M. Ebrahimi, A correlation-based bullet identification method using empirical mode decomposition, *Forensic Sci. Int.* **278** (2017) 351–360.
10. S. S. Choi, S. H. Cha, and C. Tappert, A survey of binary similarity and distance measures, *J. Syst. Cybern. Inf.* **8** (2010) 43–48.
11. W. Chu, J. Song, T. Vorburger, J. Yen, S. Ballou, and B. Bachrach, Pilot study of automated bullet signature identification based on topography measurements and correlations, *J. Forensic Sci.* **55** (2010) 341–347.
12. W. Chu, R. Thompson, J. Song, and T. Vorburger, Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria, *Forensic Sci. Int.* **231** (2013) 137–141.
13. S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer-Verlag, London, 2001).
14. J. De Kinder and M. Bonfanti, Automated comparisons of bullet striations based on 3D topography, *Forensic Sci. Int.* **101** (1999) 85–93.
15. R. Fisher and C. Vielhauer, Forensic ballistic analysis using a 3D sensor device, in *MM&Sec 12 Proc. Multimedia and Security* (ACM, New York, 2012), pp. 67–76.
16. S. R. Garten and T. L. Burrows, IBIS Brass TRAX correlation review techniques, *AFTE J.* **49** (2017) 104–110.
17. J. Hamby, private communication (2005).
18. J. Hamby, D. Brundage and J. Thorpe, The identification of bullets fired from 10 consecutively rifled 9 mm Ruger pistol barrels: A research project involving 507 participants from 20 countries, *AFTE J.* **41** (2009) 99–110.
19. E. Hare, H. Hofmann and A. Carriquiry, Automatic matching of bullets land impressions, *Ann. Appl. Stat.* **11** (2017) 2332–2356.
20. M. Heizmann and F. Puente Leon, Imaging and analysis of forensic striation marks, *Opt. Eng.* **42** (2003) 3423–3432.
21. Z. Huang and J. Leng, An online ballistics imaging system for firearm identification, in *2010 2nd Int. Conf. Signal Processing Systems* (IEEE, Dalian, China, 2010), pp. 68–72.
22. J. Kong, D. Li and C. Zhao, An automatic analysis system for firearm identification based on ballistics projectile, in *Rough Sets and Current Trends in Computing*, eds. S. Tsumoto, R. Sowiski, J. Komorowski, J. W. Grzymaa-Busse, Lecture Notes in Computer Science, Vol. 3066 (Springer, Berlin, Heidelberg, 2004).
23. D. Li, Ballistics projectile image analysis for firearm identification, *IEEE Trans. Image Process.* **15** (2006) 2857–2865.
24. D. Li, A novel ballistics imaging system for firearm identification, in *Technological Developments in Networking, Education and Automation* eds. K. Elleithy, T. Sobh, M. Iskander, V. Kapila, M. Karim and A. Mahmood (Springer, Dordrecht, 2010).
25. J. Monkres, C. Luckie, N. D. K. Petraco and A. Milam, Comparison and statistical analysis of land impressions from consecutively rifled barrels, *AFTE J.* **45** (2013) 3–20.
26. M. Neel and M. Well, A comprehensive statistical analysis of striated tool mark examinations, Part 1: Comparing known matches and known non-matches, *AFTE J.* **39** (2007) 176–198.
27. R. Nichols, *Firearm and Toolmark Identification. The Scientific Reliability of the Forensic Science Discipline*, 1st edn. (Academic Press, London, 2018).
28. NIST ballistics toolmark research database, Available at https://tsapps.nist.gov/NRBTD, accessed on 31 May 2017.
29. D. Ott, R. Thompson and J. Song, Applying 3D measurements and computer matching algorithms to two firearm examination proficiency tests, *Forensic Sci. Int.* **271** (2017) 98–106.

30. N. D. K. Petraco, P. Shenkin, J. Speir, P. Diaczuk, P. Pizzola and C. Gambino, Addressing the National Academy of Sciences' challenge: A method for statistical pattern comparison of striated tool marks, *J. Forensic Sci.* **57** (2012) 900–911.
31. F. Puente Leon, Automatic comparison of firearm bullets, *Forensic Sci. Int.* **156** (2006) 40–50.
32. F. Riva and C. Champod, Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases, *J. Forensic Sci.* **59** (2014) 637–647.
33. N. Senin, R. Groppetti, L. Garofano, P. Fratini and M. Pierni, Three-dimensional surface topography acquisition and analysis for firearm identification, *J. Forensic Sci.* **51**(2) (2006) 282–295.
34. J. Song, W. Chu, T. V. Vorburger, R. Thompson, T. B. Renegar, A. Zheng, J. Yen, R. Silver and M. Ols, Development of ballistics identification from image comparison to topography measurement in surface metrology, *Meas. Sci. Technol.* **23** (2012) 1–6.
35. J. Song, T. V. Vorburger, W. Chu, J. Yen, J. A. Soons, D. B. Ott and N. F. Zhang, Estimating error rates for firearm evidence identifications in forensic science, *Forensic Sci Int.* **284** (2018) 15–32.
36. T. Weller, A. Zheng, R. Thompson and F. Tulleners, Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides, *J. Forensic Sci.* **57** (2012) 912–917.
37. F. Xie, S. Xiao, S. L. Blunt, W. Zheng and X. Jiang, Automated bullet-identification system based on surface topography techniques, *Wear* **266** (2009) 518–522.
38. X. Zheng, J. Soons, T. Vorburger, J. Song, T. Renegar and R. Thompson, Applications of surface metrology in firearm identification, *Surf. Topogr.* **2**(1) (2014) 014012, doi: 10.1088/2051-672X/2/1/014012.

Danny Roberge received BSc and MSc Degrees in Physics Engineering from École Polytechnique de Montréal (Montréal, Canada) in 1989 and 1991, respectively. In 1995, he completed his PhD in Physics (optical pattern recognition) at Université Laval (Québec City, Canada). He was subsequently awarded a fellowship for a post-doctoral project in Toronto, Canada, that was related to optical fingerprint identification and biometric encryption. In 2000, he joined the Research and Prototyping Group at Ultra Electronics Forensic Technology as a Senior Scientist, and has developed and implemented algorithms related to the acquisition and correlation of bullet and cartridge case images for IBIS. He is the author and coauthor of several patents.



Alain Beauchamp received his BSc in Physics in 1988 from McGill University (Montréal, Canada). In 1991 and 1995, respectively, at Université de Montréal (Montréal, Canada), he completed his MSc and his PhD in Astrophysics (on software simulation of model atmospheres of white dwarf stars). In 1997, he received the Plaskett Medal Award from the Canadian Astronomical Society. He is currently a Chief Scientist of the Research and Prototyping Group at Ultra Electronics Forensic Technology. Since 2000, he and his team have developed and implemented IBIS algorithms dedicated to automated acquisition and correlation of bullet and cartridge case images. He is the author and coauthor of several patents.

Serge Lévesque received BSc and MSc degrees in Physics from Université de Montréal (Montréal, Canada) in 1987 and 1991, respectively. In 1998, he completed his PhD in Physics Engineering, in the field of atomic laser spectroscopy, at École Polytechnique de Montréal (Montréal, Canada). In 2004, he joined the Research and Prototyping Group at Ultra Electronics Forensic Technology as a Senior Scientist, and has developed original methods for the high-definition topography capture of fired bullets and cartridge cases in IBIS. His work also includes the analysis of 2D and 3D images in the context of forensic firearm and toolmark identification. He is the author and coauthor of several patents.