Adapting the Chumbley Score to match striae on Land Engraved Areas of bullets

Abstract

The same-source problem remains a major challenge in forensic toolmark and firearm examination. Here, we investigate the applicability of the Chumbley method(1)(10), developed for screwdriver markings, for same-source identification of striations on bullet LEAs.

The Hamby datasets 44 and 252 measured by NIST and CSAFE (high-resolution scans) are used here. We provide methods to identify parameters that minimize error rates for matching of LEAs, and a remedial algorithm to alleviate the problem of failed tests, while increasing the power of the test and reducing error rates.

For 85,491 land-to-land comparisons (84,235 known non-matches and 1256 known matches), the adapted test does not provide a result in 176 situations (originally more than 500). The Type I and Type II error rates are 7.2% (6105 out of 84235) and 21.4% (271 out of 1256) respectively.

This puts the proposed method on similar footing as other single feature matching approaches in the literature.

Keywords: forensic science, toolmark, cross-correlation, Mann-Whitney U statistic, land engraved areas (LEAs), algorithm, signatures, same-source problem

Same-source analyses are a major part of a Forensic Toolmark Examiner's job. In current practice, examiners make these comparisons by visual inspection under a comparison microscope and come to one of the following four conclusions: identification, inconclusive, elimination or unsuitable for examination (2). These conclusions are made on the basis of "unique surface contours" of the two toolmarks being in "sufficient agreement" (2). AFTE describes the term "sufficient agreement" as the possibility of another tool producing the markings under comparison, as practically impossible (2). Potential subject bias in the assessment as well as the lack of specified error rates are the main points of criticisms first raised by the National Research Council in 2009 (3) and later emphasized further by the President's Council of Advisors on Science and Technology (4).

Technological advances, such as profilometers and confocal microscopy, have made it possible to capture 3D surfaces in a high-resolution digitized form. This technology has become more accessible over the last decade, and has made its way into topological images of ballistics evidence, such as bullet lands and breech faces (5; 6; 7; 8). Digitized images of 3D surfaces form the basis of statistical analysis of toolmarks. Statistical approaches based on this data remove both subjectivity from the assessment and allow a quantification of error rates for false positive and false negative identifications.

Methods for matching striated marks for a variety of tools have been studied in the literature (see Table 1 for an overview): in (9) and (10) digitized screwdriver marks have been analyzed using a profilometer; in (11) 3D marks from screwdriver, tongue and groove pliers captured using a confocal microscope have been investigated; digitized marks from slip-joint pliers generated by a surface profilometer have been investigated in (12).

This data forms the basis of a statistical analysis that allows us to quantify similarity of markings and serves as a basis for error rate calculations. In (11) a relative distance metric is defined and used as similarity measure between two toolmarks. This approach is expanded in Faden et al. (9): a set of small segments in the markings of two toolmarks are extracted and similarity is compared using a maximum Pearson correlation coefficient. The Chumbley scoring method, first introduced in (10), uses a similar but more extensive framework based on a Mann-Whitney U test of the resulting correlation coefficients. This approach is non-deterministic, because segments are chosen randomly. In (1) the score is fixed to be deterministic for each pair of toolmarks by choosing

segments for comparison systematically. This approach also ensures independence between segments of striae.

In this paper, we are investigating the applicability of the Chumbley scoring method by Hadler and Morris (1) to assess striation marks on bullet lands for same-source identification. Striation marks on bullets are made by surface imperfections in the barrel. As the bullet travels through the barrel, these imperfections leave "scratches" on the bullet surface (see top of Figure 1). Typically, only striation marks in the land engraved areas (LEAs) are considered (13). Bullet lands are depressed areas between the grooves made by the rifling action of the barrel. Compared to toolmarks made by screwdrivers, striation marks on bullets are typically much smaller, both in length and in width. Bullets also have a curved cross-sectional topography.

In same-source comparisons this curvature is usually removed using some form of Gaussian filter (14) or non-parametric smoothing (15). An overview of some of the error rates reported in the literature on bullet matching is given in Table 2. Chu et al. (16) use an automatic method for counting consecutive matching striae (CMS). The authors report an error rate of 52% for known same-source land comparisons to be (incorrectly) identified as different-source (false negative) and zero false positives for known different-source lands. Ma et al. (14) and Vorburger et al. (17) discuss CCF (cross-correlation function) and its discriminating power and applicability for same-source analyses of bullets, but do not provide any error rates in their discussion. Hare et al. (15) use multiple features, such as CCF, CMS, D (distance measure), etc. in a random forest based method and compare every land against every other land of digitized versions of Hamby 252 and Hamby 44 (18) published on the NIST Ballistics Database (19). The authors report an out-of-bag overall error rate of 0.46%, comprised of an error rate of 30.05% of same-source pairs that were not identified and an error rate of 0.026% of different-source pairs that were incorrectly identified as same-source.

The Chumbley score provides us with another approach in the same-source assessment of bullet striation marks. Chumbley et al. (10) compare two toolmarks for same-source. The data for this study was obtained from 50 sequentially manufactured screwdriver tips. Chumbley et al. (10) report error rates for markings made by the tips at different angles. For markings made at 30 degree the authors report an average false negative error rate of 8.9% and an average false positive error rate of 2.3%. For marks made under angles of 60 and 85 degrees, respectively, the false negatives

error rate is 9% while the rate of false positives decreases to 1%. The paper by Hadler and Morris (1) is based on the same data but the authors focus on markings made under the same angle. The error rates associated with the deterministic version of the score are reported as 6% for false negatives and 0% for false positives.

In this paper we evaluate the adaptability of the Chumbley score as a measure to quantify similarity in land engraved areas (LEAs) on bullets. For that we briefly introduce the deterministic method suggested by Hadler and Morris (1) in the methods section of this paper. In the process we provide methods to identify parameters that minimize the error rates. We then investigate persistent scenarios in which the method proposed by Hadler and Morris (1) fails to come to a result. We go on to provide a solution to the failed tests problem, consequently increasing the power of the test and reducing error rates in the process. We set up a testing framework to compare the performance of the two algorithms in the testing setup section and finally discuss results.

Methods

Scans for land engraved areas

Comparisons of striae from bullets are usually based on comparisons of striae in land engraved areas, which are extracted in form of cross sections, called *profiles* (15; 14). Bullet striae are most pronounced at the base of the bullet (because the base typically has the most contact with the inside of the barrel). However, these areas are also affected by the break off due to friction effects between barrel and the bullet. An optimal cross section is chosen orthogonally to the striae, close to the base while avoiding break off as shown in Figure 1, see also (15) for mathematical details of the extraction.

Bullet *signatures* (16; 15) are extracted from *profiles* as residuals of a LOESS fit or Gaussian filter. This effectively removes topographic structure from the data in the attempt to increase the signal to noise ratio. Figure 1 shows how the signature from a bullet land (bottom) lines up with the image of the land (top) from which it was extracted. We can see in the figure how the depth and relative position of the striation markings seen in the image are interpreted as peaks and valleys in the signature.

There are two sources of scans for sets from the Hamby study available to us: scans of Hamby 44 and Hamby 252 are available from the NIST database (19). The physical Hamby 44 set has also been made available to us and has been scanned locally for CSAFE at the Roy J. Carver High Resolution Microscopy Facility using a Sensofar confocal light micro- scope. Scans in the NIST database are made with a NanoFocus at 20x magnification. The resolutions of the two instruments are different: the NIST scans are taken at a resolution of 1.5625 μm per pixel, while the CSAFE scans are available at a resolution of 0.645 μm per pixel. The length of an average bullet land from Hamby (9 mm Ruger P85) is about 2 millimeter, resulting in signatures of about 1200 pixels for NIST scans, and about 3000 pixels for CSAFE scans.

In comparison, scans from the profilometer used by Chumbley et al. (10); Hadler and Morris (1) were taken at a resolution of about 0.73 μm per pixel. The screw driver toolmarks are about 7 mm in length (9), for a total of over 9000 pixels for the width of these scans. This severe limitation in the amount of available data might pose the main challenge in adapting the Chumbley score to matching bullet lands, because of the potential loss in discriminating power. This is the main question that we want to investigate with our case study.

The Chumbley Score Test

A digitized toolmark forms a spatial process $z(t)$ with location indexed by t. 't' here, denotes equally spaced pixel locations for the striation marks under consideration. For a toolmark consisting of t pixels, $t = 1, ..., T$. Let further $z^s(t)$ denote a vector of markings of length s starting in location t.

The Chumbley score algorithm takes input in form of two digitized toolmarks:

Let $x(t_1)$, $t_1 = 1, 2, ...T_1$ and $y(t_2)$, $t_2 = 1, 2...T_2$ be two digitized toolmarks (where T_1 and T_2, the lengths of the two marks, are not necessarily equal). The toolmarks under consideration are potentially from two different-sources or the same-source.

In a pre-processing step the two markings are smoothed using a LOWESS (20) with coarseness parameter c. Originally, this smoothing is intended to remove drift and (sub)class characteristics from individual markings, however, in the setting of matching bullet striae, we can also make use of this mechanism to separate bullet curvature in profiles from signatures before matching

signatures. Figure 2 shows an example of a bullet land profile (left) and the corresponding signature (right).

The implementation of the Chumbley score in Hadler and Morris (1) uses a normalization step before going into the optimization and validation step described below. Normalization is done by using a LOWESS smooth to reduce extraneous structure in the markings, such as a drift, or spatial trends introduced during the barreling. To a degree, this normalization can also be used to address problems stemming from sub-class characteristics, i.e. markings in a pattern that are not unique to a single barrel but shared across a group of barrels introduced by specifics in the manufacturing process.

After normalizing profiles, the Chumbley score is calculated in two steps: an optimization step and a validation step. In the optimization step, the two markings are aligned horizontally such that within a pre-defined window of length w_o the correlation between $x(t_1)$ and $y(t_2)$ is maximized:

$$(t_1^o, t_2^o) = \underset{\{1 \le t_1 \le T_1 - w_o, 1 \le t_2 \le T_2 - w_o\}}{\arg\max} cor((x^{\{w_o\}}(t_1), y^{\{w_o\}}(t_2)))$$

This results in an optimal vertical (in-phase) shift of $t_1^o - t_2^o$ for aligning the two markings. We will denote the relative optimal locations as t_1^* and t_2^* where $t_k^* = t_1^o/(T_k - w_o))$ for k = 1, 2, such that $t_1^*, t_2^* \in [0,1]$. After profiles are normalized, the relative optimal locations should be distributed according to a uniform distribution in [0, 1].

In the validation step, two sets of windows of size w_v are chosen from both markings, see Figure 4. In the first set, pairs of windows are extracted from the two markings using the optimal vertical shift as determined in the first step, whereas for the second set the windows are extracted using a different (out-of-phase) shift.

More precisely, let us define starting points $s_i^{(k)}$ for each signature k = 1, 2 as

$$s_i^{(k)} = \begin{cases} t_k^* + i\, w_v & for\ i < 0 \\ t_k^* + w_o + i\, w_v & for\ i \ge 0 \end{cases} \qquad (1)$$

for integer values of i with $0 < s_i^{(k)} \le T_k - w_v$.

Same-shift pairs of length w_v are defined in Hadler and Morris (1) as all pairs $\left(s_i^{(1)}, s_i^{(2)}\right)$ with integer values *i* for which both $s_i^{(1)}$ and $s_i^{(2)}$ are defined. Similarly, different-shift pairs are defined as $\left(s_i^{(1)}, s_{-i-1}^{(2)}\right)$ for all i where both $s_i^{(1)}$ and $s_{-i-1}^{(2)}$ are defined (see Figure 3).

For both same- and different-shift pairs, correlations between the markings are calculated. The intuition here is that for two markings from the same-source the correlation for the in-phase sample should be high, while the correlations of the out-of-phase sample provide a measure for the base-level correlation for non-matching marks of a given length w_v. More specifically, the null hypothesis of the Chumbley score test is stated as H_o**: the markings come from two different sources** with the alternative given as H_a**: the markings come from the same source**. A *p*-value for the Chumbley score test is then computed as a Mann Whitney U statistic to compare between in-phase sample and out-of-phase sample. A low *p*-value is interpreted as a rejection of the null hypothesis in favor of the alternative hypothesis of *same-source*.

In the original method proposed in Chumbley et al. (10) both in-phase and out-of-phase sample are extracted randomly, whereas Hadler and Morris (1) proposed the above specified deterministic rules for both samples to make the resulting score deterministic while simultaneously avoiding overlaps within selected marks to ensure independence.

A problem with failed tests

Looking closer at Equation 1, we see that by definition, some number of tests will fail to produce a result. Note that this problem is different from erroneous test results. The problem of failed tests is first mentioned in Grieve et al. (12). Unfortunately, the authors do not provide any percentage of how many tests failed for their data. The algorithm fails to produce for two reasons: either the number of eligible same-shift pairs is zero, or the number of different-shift pairs is zero. Section 1 in the Appendix discusses scenarios of failed tests in more detail.

The number of same-shift pairs will be zero, if the optimal locations t_1^o and t_2^o are so far apart, that no segments of size w_v are left on the same sides of the optimal locations, i.e. $t_1^o < w_v$ and $t_2^o > T_2 - w_o - w_v$ or $t_1^o < T_1 - w_o - w_v$ and $t_2^o < w_v$ i.e. we have a failure rate of

$$P(t_1^o < w_v \cap t_2^o > T_2 - w_o - w_v) + P(t_1^o < T_1 - w_o - w_v \cap t_2^o < w_v)$$

While we can assume for normalized profiles, that optimal locations t_1^o and t_2^o are uniformly distributed across the length of the profile, we cannot assume that t_1^o and t_2^o are independent of each other. In particular, for same-source profiles, we would expect a strong dependency between these locations, in which case a large difference between locations in unlikely. However, for different-source matches, we can assume that locations are independent. In that case, we expect a test to fail with a probability of $2w_v^2/(T_1 - w_o)(T_2 - w_o)$. For an average length of T_i of 1200 pixels, w_o = 120 pixels and w_v = 30 pixels this probability is about 0.0015.

The number of possible different-shift pairs also depends on the location of the optimal locations t_1^o and t_2^o. Whenever the optimal locations are close to the boundaries, the number of possible pairings decreases and reaches zero, if $t_i^{(o)} < w_v$ and $t_i^{(o)} > T_i - w_o - w_v$. Assuming a correlation between optimal locations t_1^o and t_2^o of close to one for same-source profiles, this results in an expected rate of failure of $2w_v/(T_i - w_o)$, or about 5.6% for an average length of T_i of 1200 pixels, w_o = 120 pixels and w_v = 30 pixels. Assuming independence in the optimal locations for different-source profiles the expected probability for a failed test is, again, $2w_v^2/(T_1 - w_o)(T_2 - w_o)$.

A modified approach

While failures due to missing correlations from same-shift pairs are unavoidable by definition of the Chumbley score, failures due to missing correlations from different-shift pairs can be prevented by using a different strategy in assigning pairs.

Using the same notation as in Equation 1, we define same-shift pairs identical to Hadler and Morris (1) as pairs as pairs $\left(s_i^{(1)}, s_i^{(2)}\right)$ for all *i* where the boundary conditions of both sequences are met simultaneously. Let us assume that this results in I pairs. Define $s_{(1)}^{(k)}$ to be the j^{th} starting location in sequence *k* = 1, 2, i.e. $s_{(1)}^{(k)} < s_{(2)}^{(k)} < \cdots < s_{(I)}^{(k)}$.

We then define the pairs for different-shifts by matching up windows from opposite ends of the markings, i.e. the first pair consists of a matchup of the first window on the first marking and the last window on the second marking, the second pair consists of the second window on the first marking and the second to last pair on the second marking, and so on. In case of an odd number of pairs we need to be careful to exclude the middle pair from this assignment:

the middle pair is already part of the same-shift pair; therefore, we cannot re-use the same pair as part of the different-shift pairs.

Mathematically, this assignment of pairs is written as:

$$\left(s_{(j)}^{(1)}, s_{(I-j+1)}^{(2)}\right) for\ j = \begin{cases} 1, \ldots, I & for\ even\ \ I \\ 1, \ldots, (I-1)/2, (I-1)/2\ +\ 2, \ldots, I & for\ odd\ \ I \end{cases} \quad (2)$$

Note that for an odd number of same-shift correlations, we skip the middle pair for the different-shift correlations (see also Figure 5). This pairing ensures that the number of different-shift pairings is the same or at most one less than the number of same-shift pairings in all tests. In the remainder of the paper, we will refer to the algorithm defined by Hadler and Morris (1) as **(CS1)** and the suggested modified algorithm as **(CS2)** and compare their performance on the available scans of the Hamby study.

Performance of tests is measured with respect to the errors a test makes in situations where ground truth is known. We distinguish between two error rates: (1) false positive and (2) false negative rate. *False positives* (or *false identifications*) are situations where the test indicates a match (i.e. the test falsely rejects) but the markings come from different sources. This is also known as a *Type I* error. *False negatives (*or *missed identifications*) are situations where the test fails to reject, i.e. the test indicates that the markings come from different sources, but in fact the markings come from the same source. This is a Type *II* error. In both cases error rates are calculated as the ratio of the number of errors observed and the number of tests executed.

Note that in all of the following land-to-land comparisons only lands are compared that are suitable for a comparison, i.e. a signature can be extracted from the scan. In particular, lands which exhibited “tank rash” (random tool marks on the fired bullet surface caused by the impact with the interior surfaces of the bullet capture tank) were removed from comparison (15).

Testing setup

The Data

Lands for all Hamby-44 and Hamby-252 scans are made available through the NIST ballistics database (19) and are considered, here. Both of these sets of scans are part of the larger

Hamby study (18). Each set consists of twenty known bullets (two each from ten consecutively rifled Ruger P85 barrels) and fifteen questioned bullets (each matching one of the ten barrels). Ground truth for both of these Hamby sets is known and was used to assess correctness of the tests results.

Profiles for each bullet land were extracted from scans close to the base of the bullet while avoiding break-off as described in Hare et al. (15).

Setup

Both algorithms (CS1) and (CS2) are implemented in R (21). (CS1) is available from package toolmaRk (22), (CS2) is available from a modified version of the "toolmaRk" package available from GitHub (https://github.com/heike/toolmaRk). We applied both methods to all pairwise land-to-land comparisons of the Hamby scans provided by NIST for a total of 85,491 land-to-land comparisons.

Results

Failed Tests

As described above, the Chumbley-score is based on three parameters: coarseness c and the sizes of the optimization window w_o and validation window w_v. In a first run of results, we applied default settings for the parameters, as suggested in Hadler and Morris (1): w_o = 120 pixels or about 190 μm (ten percent of the average length of profiles) and coarseness c = 0.25, and varied the size of the validation window w_v in steps of 10 from 10 pixels to 60 pixels. Based on a significance level α of 5% for the test, this results in a correct identification of same-source and different-source toolmarks of 93.5% to 94.1%, corresponding to a rate of false negatives between 28% and 36% and a rate of false positives between 5% and 6%. However, the most prominent result we encountered, are the high number of failed tests, i.e. the number of instances, in which CS1 did not return any result. Figure 6 shows the percentage of failed tests among the 85,491 land-to-land comparisons of the NIST data for different values of the validation window size w_v. For same-source lands up to 12.5% of the tests fail using CS1. The highest percentage of failed tests under CS2 is 1.3% for different-source tests using a validation window size w_v of 60 pixels. Rates of expected failures are based on

simulation runs using covariances between locations of same-source profiles of 85.4%, and 12% for locations from different-source profiles, matching observed covariances for the Hamby scans. Observed failure rates are higher than expected. This might be due to remaining structure at a coarseness of 0.25 resulting in a distribution of optimal locations different from the assumed uniform.

Coarseness

The purpose of the coarseness parameter is to remove extraneous structure from profiles before comparisons for matching. Hadler and Morris (1) suggest a coarseness parameter of 0.25 in the setting of toolmark comparisons. For bullet lands, coarseness might need to be adjusted because of the strong effect bullet curvature has on profiles.

Figure 7 gives an overview of the effect of different coarseness parameters: from left to right, coarseness levels c are varied in steps of 0.05 from 0.1 to 0.3. The top row shows resulting signatures after smoothing the profile shown in Figure 2 with different levels of coarseness. The histograms in the bottom row show the relative optimal location t^*.

Optimal locations should be distributed uniformly once profiles are normalized.

However, for coarseness values of $c > 0.20$ we see quite distinct boundary effects: optimal locations t^* are found at the very extreme ends of a profile more often than one would expect based on a uniform distribution.

The key effect of the optimal locations and thereby the coarseness is seen in the number of failed tests. Irrespective of whether CS1 or CS2 is being used, if the relative optimal locations are at the boundaries we will see an increase in the number of failed tests. A balance is therefore needed in the selection of the coarseness parameter which reduces the boundary effect but does not remove important individual characteristics. Based on Figure 7 a coarseness value of $c = 0.15$ seems to be best suited to strike this balance for this example. For the remainder of the analysis, we will use this value for c.

Error rate assessment

Figure 8 gives an overview of ROC (Receiver operating characteristic) curves for methods CS1 and CS2 over a range of different optimization window sizes w_o and two sizes for the validation window w_v (shape). The different color hues represent the two methods CS1 (red) and CS2 (blue). The ROC curves show the superior performance of CS2 over CS1. Generally, an optimization window w_o of 150 pixels or more leads to the best performance with respect to ROC curves. Results based on a validation window of size $w_v = 30$ are generally better than results for $w_v = 50$.

Figure 9 shows a comparison of the performance of the two methods CS1 and CS2 with respect to EER (equal error rate) and AUC (area under the curve) corresponding to the ROC curves shown in Figure 8. Equal error rates are reduced using method CS2, while area under the curve significantly increases (at a significance level α of 5%) compared to method CS1.

The results from Figures 8 and 9 are summarized in numbers in Table 3. Equal error rates (EER), rates for false positives (FPR) and false negatives (FNR) are shown side by side with the area under the curve (AUR) for both methods for a set of different optimization windows w_o and a validation window w_v of 30 pixels. The rate of false positive same-source identifications is equal to the statistical type I error, which is set to $\alpha = 5\%$ for this example. The rate of false negatives are missed same-source markings. This rate is also known as the type II error rate. A detailed plot on the type II error rates for CS1 and CS2 can be found in the section 2 of the Appendix. Area under the curve (AUC) is shown with confidence intervals as given by DeLong et al. (23). CS2 significantly outperforms CS1 with respect to its predictive power in most situations.

Observed versus Nominal Type I error rates

Figure 10 shows the percentages of observed type I errors (%FP) across a range of optimization windows w_o. Generally, observed type I errors are higher than expected. Method CS1 shows in this instance slightly better performance than method CS2, but for both an increase in the size of the optimization window leads to a decrease in the observed type I errors.

High resolution Hamby 44 scans

The high-resolution scans of Hamby set 44 are capturing images at a resolution of $0.645\mu m$ per pixel. On average, land engraved areas are 3000 pixels in length. A coarseness of c = 0.125 seemed to be sufficient in removing any bullet curvature. Both methods have a failed test rate of less than 0.6%, indicating, again, that the larger number of pixels alleviates the problem of test failures. Figure 11 shows the resulting EER and AUC for methods CS1 and CS2 based on two sizes of validation windows ($w_v \in 75, 125$) and optimization window sizes around 300 pixels (10 percent of the average length), with corresponding errors shown numerically in Table 4. Both methods show an increase in performance around w_o = 300 pixels. CS2 out-performs CS1 in all scenarios, but the difference is not significant (using DeLong's confidence intervals). Interestingly, the overall performance of both CS1 and CS2 is a lot lower for the high-resolution version of Hamby-44 than for the lowerresolution scans. The area under the curve overall is significantly lower for the high-resolution scans than for the previous set of scans. Partly, this might be due to the particular choice of the parameters, partly the higher-resolution scans might be picking up on real differences between the lands that the lower-resolution scans fail to detect. Reassuringly, the observed error rate of false positives for CS2 is closer to the nominal rate of 5% than for the lower- resolution scans.

Conclusions

In assessing the suitability of the (deterministic) Chumbley Score for matching striae on bullet lands we have gained valuable insights into the process: method CS1 as proposed by Hadler and Morris (1) has a strong dependency on the specific choice of parameters; the defaults suggested by Hadler and Morris (1) for screw drivers are not directly applicable for the smaller bullet lands. The coarseness parameter in particular has a strong impact on the performance of the test. However, we were able to suggest some heuristics based on the assumption that for normalized profiles, optimal locations are distributed uniformly across the profile. For bullet lands we found a coarseness value of c = 0.15 to be suitable for the low-resolution scans from NIST and a value of c = 0.125 suitable for the higher-resolution scans from CSAFE. Sizes for optimization windows w_o were based on cross-validation to minimize overall type 2 error rates. Ideally, the exact values for parameters will be determined in a large study

incorporating different types of firearms and brands of ammunition. Results of this paper therefore do not transfer immediately to case work, where a forensic examiner would only deal with a few identifications.

Method CS1 proposed by Hadler and Morris (1) has a minimal type 2 error of 27.2% for an optimized window size of 140 pixels – which is considerably higher than the error rates achieved on matching toolmarks, but, is similar to other single-feature methods proposed for bullet matching. Unfortunately, method CS1 also has a high rate of failed tests – situations, in which the algorithm does not provide a result, due to the way different-shift pairs are constructed. Algorithm CS2 is introduced here as a remedy for failed tests by introducing an alternate version of choosing different-shift pairs. Algorithm CS2 is constructed in a way that achieves on average a ten-fold reduction in the number of failures. While reducing the failure rate, the algorithm also shows an increase in the power of the test. Type II error of CS2 reach a minimum of 21.7% for an optimized window size of 130 pixels. This increase in power of CS2 over CS1 should also apply to previous studies on toolmarks. It would be interesting to see these results using the adjusted algorithm. Unfortunately, none of the studies have made the data publicly accessible.

While significantly reduced over CS1, CS2 still has type 2 error on bullet lands that are higher than the error achieved on the –much larger– toolmarks. Applying these methods to the high-resolution scans provided by CSAFE shows that better scanning methodology does not guarantee a better matching performance.

Different avenues for improving the performance of the Chumbley-score are still open: (i) When making the original Chumbley Score deterministic, Hadler and Morris (1) rely on an optimal shift based on the maximum cross-correlation between two markings. An erroneous decision in identifying the optimial shift between markings leads almost always to a missed identification. CS2 is still susceptible to this source of errors. (ii) Bullets usually have multiple lands – in the case of Ruger P85s as used in the Hamby study, there are six lands for each bullet. As shown by Chu et al. (24) in the example of the cross-correlation between lands, we will be able to get more power out of the test by adapting the algorithm to make use of the relative position of a land on the bullet to combine multiple land-to-land comparisons into a single bullet-to-bullet comparison.

References

1. Hadler JR, Morris MD. An Improved Version of a Tool Mark Comparison Algorithm. J Forensic Sci 2018;63:849–855.

2. AFTE Glossary. Theory of Identification as it Relates to Toolmarks. AFTE Journal 1998;30:86–88.

3. National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press; 2009. https://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward (accessed 7-August-2018).

4. President's Council of Advisors on Science and Technology. Report on Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods;2016.https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed 2018-08-07).

5. De Kinder J, Prevot P, Pirlot M, Nys B. Surface topology of bullet striations: an innovating technique. AFTE Journal 1998;30(2):294–299.

6. De Kinder J, Bonifanti M. Automated comparison of bullet striations based on 3D topography. Forensic Sci Int 1999;101:85–93.

7. Bachrach B. Development of a 3D-based Automated Firearms Evidence Comparison System. J Forensic Sci 2002;47(6):1253–1264.

8. Vorburger TV, Song J, Petraco N. Topography measurements and applications in ballistics and tool mark identifications. Surface topography: metrology and properties 2016;4(1):013002.

9. Faden D, Kidd J, Craft J, Chumbley LS, Morris MD, Genalo LJ, et al. Statistical Confirmation of Empirical Observations Concerning Toolmark Striae. AFTE Journal 2007;39(2):205–214.

10. Chumbley LS, Morris MD, Kreiser MJ, Fisher C, Craft J, Genalo LJ, et al. Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm. J Forensic Sci 2010;55(4):953–961.

11. Bachrach B, Jain A, Jung S, Koons R. A Statistical Validation of the Individuality and Repeatability of Striated Tool Marks: Screwdrivers and Tongue and Groove Pliers. J Forensic Sci 2010;55(2):348–357.

12. Grieve T, Chumbley LS, Kreiser J, Ekstrand L, Morris M, Zhang S. Objective Comparison of Toolmarks from the Cutting Surfaces of Slip-Joint Pliers. AFTE Journal 2014;46(2):176–185.

13. AFTE Criteria for Identification Committee. Theory of identification, range striae comparison reports and modified glossary definitions. AFTE Journal 1992;24:336–340.

14. Ma L, Song J, Whitenton E, Zheng A, Vorburger TV, Zhou J. NIST bullet signature measurement system for RM (Reference Material) 8240 standard bullets. J Forensic Sci 2004;49:649–659.

15. Hare E, Hofmann H, Carriquiry A. Automatic matching of bullet land impressions. Ann Appl Stat 2017 12;11(4):2332–2356.

16. Chu W, Thompson RM, Song J, Vorburger TV. Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria. Forensic Sci Int 2013;231:137–141.

17. Vorburger TV, Song J, Chu W, Ma L, Bui SH, Zheng A, et al. Applications of cross-correlation functions. Wear 2011;271(3-4).

18. Hamby JE, Brundage DJ, Thorpe JW. The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries. AFTE Journal 2009;41(2):99–110.

19. Zheng XA. NIST Ballistics Toolmark Research Database (NBTRB); 2016. https://tsapps.nist.gov/NRBTD/Studies/Search (accessed 2018-08-07).

20. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. JASA 1979;74(368):829–836.

21. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018. https://www.r-project.org/ .

22. Hadler J. toolmaRk: Tests for Same-Source of Toolmarks; 2017. R package version 0.0.1. https://github.com/heike/toolmaRk (accessed 7-August-2018).

23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 1988;44(3):837–845.

24. Chu W, Song J, Vorburger TV, Yen J, Ballou S, Bachrach B. Pilot Study of Automated Bullet Signature Identification Based on Topography Measurements and Correlations. J Forensic Sci 2010;55:341–347.

List of Tables

Research paper	Method	Data Source	False Positives	False Negatives
Faden et al. (2007) (9)	Maximum Pearson Correlation	Screwdrivers	-	-
Chumbley et al. (2010) (10) (Same-Surface Same-Angle)	Randomized Chumbley Score	Screwdrivers	2.3%	8.9%
Grieve et al. (2014) (12)	Randomized Chumbley Score	Slip-joint	-	-
Hadler & Morris (2017) (1) (Same-Surface Same-Angle)	Deterministic Chumbley Score	Screwdrivers	0%	6%
Bachrach et al. (2010) (11) (Different Surfaces-same angle)	Similarity Measure Relative Distance Metric	Screwdrivers	5.9%	9.4%
(Same Surfaces-same angle)			0.22%	0%

Table 1: Error Rates in same-source analyses for striated toolmarks reported in the literature. The top four reported papers use some variation of the Chumbley score method.

Method	Data Source	False Positives	False Negatives
Hare et al. (15)	Hamby 252 (NIST scans)		
Consecutive Matching Bullets Striae (CMS)		6.25%	33.85%
Consecutive Non-matching Striae (CNMS)		6.25%	35.42%
Average Distance (D)		6.25%	45.83%
Cross-correlation Function (CCF)		6.25%	17.71%
Sum of Peaks (S)		6.25%	18.23%
Chu et al. (16)	Hamby 252 (NIST scans)		
Consecutive Matching Bullets Striae (CMS)		0%	52%
Chu et al. (24)	6 types of firearms, 2 types of ammunition		
Ma et al. (14)	NIST standard bullet SRM 8240		
Cross-Correlation Function (CCF)		-	-

Table 2: Error Rates in same-source single feature land-to-land analysis reported in the literature. Note that only error rates from the same data source can be compared directly between methods. Error rates from different sources are influenced by type of firearm and ammunition used.

	CS1			CS2		
w_o	%FP	%FN	AUC (95% C.I.)	%FP	%FN	AUC (95% C.I.)
90	6.8	33.0	0.850 (0.835, 0.865)	7.5	24.2	0.877 (0.863, 0.890)
120	6.5	30.9	0.864 (0.850, 0.878)	7.2	23.6	0.890 (0.877, 0.903)
130	6.5	31.7	0.863 (0.850, 0.877)	7.2	22.4	0.898 (0.886, 0.910)
140	6.5	30.4	0.873 (0.859, 0.886)	7.1	22.9	0.902 (0.891, 0.914)
150	6.7	32.6	0.863 (0.850, 0.877)	7.2	22.5	0.904 (0.892, 0.916)
180	6.1	32.6	0.877 (0.864, 0.890)	7.0	22.9	0.907 (0.896, 0.919)
210	6.4	34.2	0.865 (0.851, 0.878)	6.6	23.1	0.906 (0.895, 0.918)

Table 3: Overview of results as shown in Figures 8 and 9. %FP is the observed percentage of false positives (for a fixed level $\alpha = 0.05$), %FN is the percentage of false negatives. Area under the curve (AUC) is shown with confidence intervals.

	CS1			CS2		
w_o	**%FP**	**%FN**	**AUC (95% C.I.)**	**%FP**	**%FN**	**AUC (95% C.I.)**
210	8.6	41.3	0.780 (0.740, 0.821)	5.5	38.7	0.803 (0.763, 0.843)
240	8.5	42.7	0.784 (0.744, 0.824)	5.5	36.5	0.801 (0.760, 0.841)
270	8.2	41.5	0.782 (0.741, 0.823)	4.8	37.4	0.808 (0.769, 0.848)
300	8.2	40.5	0.791 (0.751, 0.832)	5.1	38.3	0.814 (0.776, 0.852)
330	8.2	42.8	0.791 (0.753, 0.830)	5.0	38.3	0.804 (0.765, 0.844)
360	8.3	42.7	0.788 (0.747, 0.828)	4.9	40.0	0.803 (0.765, 0.842)
390	8.5	40.0	0.789 (0.748, 0.830)	5.0	38.3	0.810 (0.771, 0.849)

Table 4: Overview of high-resolution scan results as shown in Figures 11. %FP is the percentage of observed false positives (for a fixed level $\alpha = 0.05$), %FN is the percentage of false negatives. Area under the curve (AUC) is shown with 95% confidence intervals.