Final Technical Report

Quantification of Toolmarks

L. S. Chumbley
Ames Laboratory / Iowa State University
August 25, 2009

Report Title: Quantification of Toolmarks

Award Number: This work was conducted at Ames Laboratory, Iowa State University under the Department of Justice Interagency Agreement number 2004-R-IJ-088, M-003, dated September 2004, Quantification of Toolmarks.

Author(s): L.Scott Chumbley

This report describes research conducted at Ames Laboratory / Iowa State University involving characterization of toolmarks using a quantitative, statistical analysis method. Persons involved in the project include ISU faculty members L. Scott Chumbley, Max Morris, Shana Smith, Song Zhang and Larry Genalo. Graduate students involved in the research include David Faden, Julie Kidd, and Jeremy Craft. Undergraduate assistants employed during the course of the work include Craig Bossard, Stephen Davis, Charles Fisher, and Anne Saxton. The project was also assisted by Mr. David Eisenmann, staff scientist at the Center for Nondestructive Evaluation at Iowa State University and Mr. Jim Kreiser, retired forensic examiner for the State of Illinois.

Abstract

The goal of this research was to develop a methodology to characterize toolmarks using quantitative measures of the 3-dimensional nature of the mark, rather than a two-dimensional image of the mark. Once such a methodology was developed, objective comparisons between two toolmarks could be made to determine whether marks made from similar tools could be distinguished quantitatively from marks made using other tools.

The toolmarks studied were produced using 50 sequentially manufactured screwdriver tips that had not seen service. Marks were made at angles of 30, 60, and 85 degrees by a qualified toolmark examiner using a special jig. Four replicas were made of each mark, the marks being characterized using a stylus profilometer. Ten traces were taken from each mark, yielding a database of 12,000 potential data files. Initial efforts to use stereomicroscopy to obtain similar 3-dimensional information failed due to the inability to produce suitable registry between the SEM images. This problem became evident as attempts were made to "stitch" the numerous images together to obtain a continuous trace of the surface at a magnification high enough to allow quantitative measurement of the fine detail of the surface. An optical profilometer was eventually purchased that allowed suitable data to be obtained not involving touching the sample surface.

The algorithm developed to allow comparison of two scans in an objective, quantitative manner mimics the procedure used by forensic examiners in that it compares the 3-d information contained in a user-specified "window" from one file to any other selected data file. The region of best fit between the two files is found, then a verification sequence is run that compares corresponding regions in the two files, selected at random, which are translated rigid, fixed distances from the region of best fit. The quality of these comparisons is evaluated using a t-statistic. If these comparisons also have a good correlation between each other a high t-statistic value is returned, indicating a high probability of a match. If the value from the rigid translation comparisons is low, the likelihood of a match is also low. A baseline for the values is established by taking totally random comparisons between the two datafiles.

Initial results showed that datafiles of known matches (i.e. marks made using the same screwdriver) could be identified on average 95% of the time with a false positive error rate of 1% and a false negative error rate of 9%. These numbers change as the angle of the mark changed, since higher angles typically produced higher quality marks.

In an effort to improve these results a study involving actual examiners was conducted at the 2008 AFTE meeting. In this study 20 samples, including matches and nonmatches both correctly and incorrectly identified by the algorithm, were selected and shown to examiners in a blind test. The examiners yielded a much higher degree of success than the algorithm, with no false positives and only a limited number of false negatives. (Note: Examiners are trained to only make a positive ID when absolutely certain. Thus, a false negative in this case means an examiner was not completely convinced that a match existed.) This study revealed that contextual information plays a large role in the examiner's decision-making process. Such information was purposefully omitted from the initial trials of the algorithm in order to make it's operation as general as possible. A final study including contextual information was conducted, and the results were vastly superior to those obtained when this information was omitted.

From these studies it can be concluded that an objective method of comparison is feasible as a screening process for toolmark comparisons. Inclusion of contextual information should be included in this screening process. However, an experienced examiner is still essential in verifying the actual results of any computer-based algorithm.

Table of Contents

Executive Summary

This project has sought to answer the question: Can a series of toolmarks be obtained and compared in an automated manner to yield objective, statistically valid matches when toolmarks related to a particular tool (and only that tool) are compared to each other? While optical characterization between tools and toolmarks through a method of comparative matching has been utilized for nearly a century the assumption inherent in the method is that each mark represents unique characteristics of the tool that created it. The 1993 Daubert v. Merrell Dow Pharmaceuticals created a higher standard for federal courts to accept expert witness testimony; the new standard calling for scientific knowledge with a basis in the scientific method to be the foundation for testimony of expert witnesses (in this field, toolmark examiners). Thus, development of a method of analysis that reduces the subjective nature of comparative evaluation and provides statistical confirmation of a match, with known error rates and confidence intervals, is desirable. The scientific method involving observation, hypothesis formation, hypothesis testing through experimentation, and result analysis was followed in this study.

This study involves an examination of sequentially manufactured screwdriver tips that had yet to see service. Three distinct hypotheses were tested, the hypotheses being based on previous observations by forensic examiners. The hypotheses tested are summarized in Table I.

Hypothesis 1:	*The 50 sequentially produced screwdrivers examined in this study all produce uniquely identifiable tool marks*
Hypothesis 2:	*In order to be identifiable, tool marks from an individual screwdriver must be compared at similar angles.*
Hypothesis 3:	*Different sides of a flat-bladed screwdriver produce different uniquely identifiable marks.*

Table I: Hypotheses tested.

The initial experimental design consisted of the acquisition of a series of ostensibly identical screwdriver tips; making marks with the tips; characterization of the marks thus made; and comparison of the produced marks using a statistical algorithm in accordance with the stated hypotheses. Subsequent studies were conducted involving actual forensic examiners at the 2008 AFTE convention, and using a non-contact optical profilometer instead of the stylus profilometer used for the majority of the study.

Fifty sequentially produced tips were obtained from Omega Company and used by Mr. Jim Kreiser, former head toolmark examiner for the State of Illinois, to mark lead samples at angles of 30º, 60º, and 85º using a jig to maintain the selected angle for the toolmark. Both sides of the screwdriver were used and four replicates were made of each toolmark. A Homelwerk Surface Stylus Profilometer was used to measure surface roughness on all toolmarks, 10 traces being taken on each mark, yielding a total data set of 12,000 possible scans.

A computer algorithm was developed to match data along the one-dimensional profilometer data traces. The data consisted of surface height (z) as a function of distance (x) along a linear trace taken perpendicular to the striations present in a typical tool mark. Important assumptions in the analysis are that the values of z are reported at equal increments of distance along the trace and

that the traces are taken as nearly perpendicular to the striations as possible. The algorithm then allows comparison of two such linear traces.

The algorithm works by identifying a region of best agreement for a selected window size (Optimization Step) and then conducting a series of comparisons (Validation Step) to determine whether the specific window identified as the region of best fit has any statistical meaning or simply exists due to random chance. Optimization occurs simply by finding the maximum correlation statistic, or "R-value", associated with the windows for the scans under comparison. Validation involves comparing a series of corresponding windows of equal size selected at randomly chosen, but common distances from, the previously identified regions of best fit. The correlation statistic for these pairs are determined and evaluated. The assumption behind the Validation step is that if a match truly does exist, correlations between these shifted window pairs will also be reasonably large because they will correspond to common sections of the tool surface. In other words, if a match exists at one point along the scan length (high R-value), there should be fairly large correlations between corresponding pairs of windows along their entire length. However, if a high R-value is found between the comparison windows of two nonmatch samples simply by accident, there is no reason to believe that the accidental match will hold up at other points along the scan length. In this case rigid-shift pairs of windows will likely not result in especially large correlation values.

The correlation values computed from these segment-pairs can be judged to be "large" or "small" only if a baseline can be established for each of the sample comparisons. This is achieved by identifying a second set of paired windows (i.e. data segments), again randomly selected along the length of each trace, but in this case, without the constraint that they represent equal rigid-shifts from their respective regions of best fit. In other words, for this second set of comparisons the shifts are selected at random and independently from each other – any segment of the selected length from one specimen has an equal probability of being compared to any segment from the other.

The Validation step concludes with a comparison of the two sets of correlation values just described, one set from windows of common random rigid-shifts from their respective regions of best agreement, and one set from the independently selected windows. If the assumption of similarity between corresponding points for a match is true the correlation values of the first set of windows should tend to be larger than those in the second. In other words, the rigid-shift window pairs should result in higher correlation values than the independently selected, totally random pairs. In the case of a nonmatch, since the identification of a region of best agreement is simply a random event and there truly is no similarity between corresponding points along the trace, the correlations in the two comparison sets should be very similar.

A nonparametric Mann-Whitney U-statistic is generated for the comparison. Where the correlation values of the two comparison sets are similar, T1 takes values near zero, supporting a null hypothesis of "no match". If the correlations from the first rigid-shift sample are systematically larger than the independently selected shifts, the resulting values of T1 are larger, supporting an alternative hypothesis of "match".

Analysis of the data, then involves simply looking at the resulting T1 values to see if any particular hypotheses tested is supported by the data. Values at or near 0 will support the null hypothesis, i.e., there is no relationship between the comparison pairs. A non-zero value says there is a relationship, the greater the separation from 0 the higher probability of a "match" actually existing.

When the data obtained in this study are plotted and analyzed, support for all three Hypotheses is evident. When comparing tool marks made at similar angles with different tools, the resulting T1 values cluster near zero, which indicates no relationship between the marks, supporting Hypothesis 1. When the same tool is used to make marks at similar angles, the T1 distributions are substantially larger values, again giving support for Hypothesis 1. Support for Hypothesis 2 is demonstrated since even among same-tool marks, only those made at the same angle produce large T1 values. The last hypothesis considered, is that when comparing tool marks made from screwdriver tips, the marks must be made from the same side of the screwdriver was supported when it was found the T1 values cluster around 0 regardless of the angles used in making the marks, indicating no relationship between the opposite sides of the screwdriver.

Examination of the data indicates that the algorithm operates best using data obtained at higher angles than lower angles, i.e. the separation between match and nonmatch T1 values is more defined for the 85 degree data than, for example, the 30 degree data. This is believed related to the quality of the mark being higher at the higher angles. Algorithm performance also appears more efficient at reducing false positives than it does in eliminating false negatives.

Threshold values for establishing "Positive ID" and "Positive Elimination," T1 values were chosen based on a K-fold cross validation using 95% one-sided Bayes credible intervals. Specifically, the lower threshold is a lower 95% bound on the 5^{th} percentile of T1 values associated with nonmatching specimen pairs, and the upper threshold is an upper 95% bound on the 95^{th} percentile of T1 values associated with matching specimen pairs. The region between these two threshold values is considered "Inconclusive". A Markov Chain-Monte Carlo simulation was used to determine potential error rates.

Using this method comparisons made at 30 degrees have an estimated probability of a false positive (i.e. a high T1 value for a known nonmatch comparison) of 0.023. In other words there is a possibility of slightly over two false positives for approximately every 100 comparisons. The estimated probability of a false negative is 0.089, or almost 9 true matches having a low T1 value per every 100 comparisons. The cross-validation method used ensures that all the data have similar error rates, and the rates found for the 60 and 85 degree data are approximately 0.01 and 0.09 for false positives and false negatives, respectively. It would be possible to shift these error rates, i.e. produce fewer false negatives at the expense of more false positives, by altering the percentiles used in the estimation procedure.

In order to compare the effectiveness of the algorithm to human examiners, and potentially identify areas where the algorithm might be enhanced or improved, a double-blind study was conducted during the 2008 Association of Forearms and Tool mark Examiners Training Seminar. A series of 20 comparison pairs covering a range of T1 values from low to high were selected that covered the possible range of algorithm results, i.e. five correctly identified matched sets; five correctly eliminated nonmatchs; five incorrectly eliminated matches; and five incorrectly identified nonmatches. Examiners were asked to assess each pair of samples twice, initially with paper blinders in place to restrict the view to the same area characterized by the profilometer, then secondly based on a view of the entire sample. In each case, examiners were asked to render an opinion as to whether they were viewing a positive identification, a positive elimination, or inconclusive.

It should be recognized that the conditions under which the examiners rendered an opinion would ordinarily be regarded as restrictive or even professionally unacceptable. Without having

the tool in hand, or without being permitted to make the actual mark for comparison, tool mark examiners were forced to make assumptions they would not make in an actual investigation. This caused examiners to be more conservative in their willingness to declare a positive identification or elimination. Several examiners commented that typical lab protocol would require them to have physical access to the subject tool before rendering a "positive identification" judgment. Examiners also do not typically employ the terms used to denote the three regions identified for error analysis. Thus, while useful for the purpose of this research, the conducted study should not be taken as an absolute assessment of examiner performance.

In a small number of cases (12 out of 252 comparisons), when examining the entire tool mark after first viewing only the restricted area, examiners changed their opinion from inconclusive to either positive ID or positive elimination. This indicates that algorithm performance might be improved simply by increasing the amount of data processed. This may be achieved by ensuring that adequate data is taken to characterize the mark.

In five cases, comparisons between specimens made by the same screwdriver that were not conclusively identified as such by the algorithm also presented problems for the examiners. Thus, while examiners in general were vastly superior to the algorithm in picking out the matches, both the algorithm and the examiners had more trouble with some true matches than with others.

Consideration of these results pointed out weaknesses in the AFTE study and in the laboratory tests of the algorithm, namely the examiners (and the algorithm) having a lack of a point of reference or registry of the mark for the comparison. Such reference is usually available to examiners when they are responsible for making the mark themselves, where they can visually verify the feature of the toolmark that represents the edge of the screwdriver. This information was lacking to them during the AFTE study, and the algorithm has no way of distinguishing this point either since the data used is a simple file containing z versus x measurements. This type of information, called contextual information, is of critical importance to an examiner when making a determination.

Incorporating contextual information into the algorithm is difficult, but it can be done in a limited sense. As written the algorithm treats all possible pairs of windows the same way and functions under the assumption that the marks can be compared without regard to how the mark was made. This clearly is not the case. Differentiation between the left side of a screwdriver mark as opposed to the right side – information that is used by examiners in making a determination – appears critical in improving the performance. (Note: Left and right terms refer to the z data obtained from a single side of the screwdriver, just at different locations on the toolmark, which can be thought of as progressing in a linear fashion from left to right. It does NOT refer to comparing data obtained from marks made using opposite sides of a screwdriver.) The best way to enhance algorithm performance at this time is to ensure that most-similar windows found at the trace edges are used as a basis for match identification only if they are found at the same end of their respective traces. In a sense this restriction adds in the necessary "contextual information" available to examiners.

As a test of this theory the AFTE samples were re-evaluated, with two major changes being made. Firstly, data was obtained form the samples using an optical profilometer that had been purchased for this purpose. Such an instrument allows non-contact data to be obtained from the sample and enables z data to be obtained from surfaces at high angles with respect to each other. Secondly, the algorithm was altered slightly so that the data files were only compared in one

direction, i.e. the left side data was compared to the left side, the right side data compared to the right side etc. With these adjustments the performance of the algorithm for the AFTE samples was seen to improve dramatically for the incorrectly identified samples used in the AFTE study. Slight improvement was also seen in the correctly identified samples, attesting to the quality of the optical data.

In conclusion, the question posed as the goal of the study, "Can a series of toolmarks be obtained and compared in an automated manner to yield objective, statistically valid matches when toolmarks related to a particular tool (and only that tool) are compared to each other?" has been answered in the affirmative given the right conditions. Factors affecting a correct identification include the quality of the marking, suitable quantification of the mark, and suitable design and use of the automated routine. The major drawback of most simple automated routines is that the data is analyzed in the absence of any context. When this information is incorporated, even in a basic manner, the results of any algorithm developed are expected to improve dramatically.

Given that the tools examined in this study should have been as identical as possible to one another implies that unique marks do exist for every tool manufactured, at least by this manufacturer using the techniques and tooling currently employed. The question of which of the factors listed above is most critical, or can all variables be addressed in a satisfactory manner, is another matter. If a poor quality marking exists an unambiguous determination may be impossible.

Additional testing of the algorithm on other types of tool marks would be appropriate to determine the applicability of the program to other marks. A study concerning the variance in data would also be of value. This would involve having multiple persons acquire data from the same samples then employ the algorithm to see how the results compare between operators. This study would yield valuable statistical information concerning algorithm performance and robustness.

A study involving development of a "virtual tool" that would allow any mark to be simulated, whether full or partial, is also of considerable interest. The simulation could be used to provide data concerning exactly how the tool had to have been used to create the mark in question, yielding information concerning the angle of attack, the applied pressure, twist of the tool, etc. Such a study would address the question of uniqueness since only a virtual tool that adequately simulates the actual could be manipulated in the manner needed to produce a satisfactory comparison.

I. Introduction

A. Statement of the Problem

This project has sought to answer the following question: Can a series of toolmarks be obtained and compared in an automated manner to yield objective, statistically valid matches when toolmarks related to a particular tool (and only that tool) are compared to each other? In other words, can it be said that toolmarks obtained from a particular tool yield characteristics that are unique enough to that tool (and only that tool) to allow an objective identification algorithm to be employed for identification purposes? Providing answers to these questions based upon quantitative techniques rather than subjective analysis removes uncertainties raised by the Daubert decision, and reinforces statistical efforts involving measurement of consecutive matching striations in placing toolmark identification on a sound scientific footing.

B. Literature Citations and Review

Optical characterization between tools and toolmarks through a method of comparative matching is a technique that has been utilized for nearly a century. Experience has shown that tools, generally accepted to possess unique surface characteristics, can be accurately paired with toolmarks, i.e., marks made on softer surfaces by the tool. Marks are often left on metal when a tensile, shear, or compressive force is applied. Comparative identifications of tools and corresponding toolmarks have been used to prove that a particular tool was responsible for a mark in criminal investigations [1], with the assumption that each mark represents unique characteristics of the tool that created it. A similar assumption of uniqueness has been held true of fingerprints. In 1993, the case of Daubert v. Merrell Dow Pharmaceuticals created a higher standard for federal courts to accept expert witness testimony; the new standard calling for scientific knowledge with a basis in the scientific method to be the foundation for testimony of expert witnesses (in this field, toolmark examiners). The field of toolmark examination has therefore been forced to examine the validity of the basic assumption that toolmarks are unique. Development of a method of analysis that reduces the subjective nature of comparative evaluation and provides statistical confirmation of a match, with known error rates and confidence intervals, is desirable.

B.1 Tools and Toolmarks:

The Association of Firearm and Toolmark Examiners (AFTE) defines a tool as “an object used to gain mechanical advantage; also thought of as the harder of two objects which when brought into contact with each other, results in the softer one being marked.” [2] This definition allows for a broad range of objects to be classified as tools. The area of the tool that comes into contact with the softer material to leave behind a mark is known as the working surface of the tool. Toolmarks, which can be made by virtually any object, are created when the tool’s working surface comes into contact with a softer material, and leaves a representation of its surface.

Comparisons of tools and toolmarks fall into two key categories according to Biasotti and Murdock: pattern fit and pattern transfer. Pattern fit, also described as a physical match or a fracture match, is a term describing the unique features of surfaces fitting together uniquely; the more contours a surface possesses the higher the probability of a unique match. For example, if a piece of glass was fractured into two pieces and the pieces were fit perfectly back together, a pattern fit would have been made. Pattern transfer is not as simple as pattern fit; it involves the impressions and striations of two and three dimensional marks. [3] Toolmarks are considered pattern transfer, Figure 1. Impressions are created when force and motion applied to the tool are perpendicular to the surface being marked. For example, a hammer impact is an impression.

Contours created when force and motion are applied parallel to the surface being marked are known as striations. Scraping a surface with a pry bar creates a striated toolmark. [2]

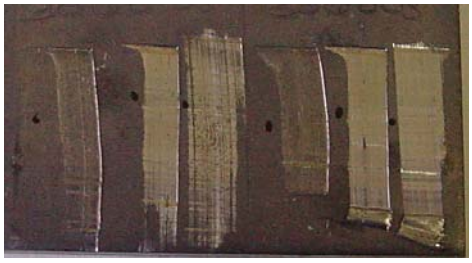


Figure 1. Striated toolmark

B.2 Toolmark Characteristics:

Individual characteristics are a completely unique series of features on a surface "produced by the random imperfections or irregularities of tool surfaces. These random imperfections or irregularities are produced incidental to manufacture or are caused by use, corrosion or damage," according to F. Taroni, author of "Statistics: A Future in the Toolmarks Comparison?" [3] Individual characteristics are unique and distinguish it from all other tools of similar type.

Class characteristics are indicative of the source of the tool; they are marks characteristic of the class of tools, often resulting from the tool design. Class characteristics are typically more macroscopic in nature. For example, in the area of firearms class characteristics are related to the matching of caliber of the firearm and cartridge or bullet, and the rifling pattern contained in the barrel of the firearm as it is transferred to a bullet. [4]

Subclass characteristics are more distinctly defined—they are related to manufacture, have, a narrow source of origin, and are ever changing. An example of a subclass characteristic would be a tool produced from a common master that shares characteristics present only in other tools produced by the same master. In 1949, Churchman observed subclass characteristics in a series of bullets fired from consecutively made rifle barrels. [5] Twenty-six years later, in 1975, Skolrood's observations were similar to Churchman's; he detected subclass characteristics when he examined three similar rifle barrels. [6] Nichols explains what qualifies a characteristic as a subclass characteristic:

> If one were to examine a cast of the bore of a firearm, such characteristics would have to exist for the entire length of the cut surface. If a certain characteristic appeared after the cut surface had already started, then it would be an imperfection caused by the current process. If it disappeared before the end of the cut surface, then it is gone and by definition of its absence cannot be passed onto the next cut surface. Therefore, the only characteristics capable of being defined a subclass would be those that persist for the entire length of the cut surface. [6]

Examiners also have found class and subclass characteristics in toolmarks. In 1968, Burd and Kirk's study of screwdrivers that had not experienced finishing work had the potential to show subclass characteristics. [7] Miller documents research into subclass characteristics present in tongue and groove pliers, nails, metal punches, metal snips, and screwdrivers. In each instance, subclass characteristics were present and yet experienced toolmark examiners were able to distinguish between different tools used to create the marks. [2]

Subclass characteristics are partly defined by their ability to evolve over time. The evolution of subclass characteristics in firearms is attributed to use—this may include cleaning, handling, or dismantling. The barrel interior is affected primarily by erosion, corrosion, and deposition of particles. Bonfanti's review of literature explores the lifetime of a subclass characteristic; she emphasizes differences in subclass characteristics from weapon to weapon and the need for the subjective interpretation of photographic evidence by a toolmark examiner. [8] Even in consecutively made toolmarks from the same tool, differences in individual surfaces may be present. However, the slow change of tool surfaces does not prohibit identification criteria to be established and positive identifications to be made. [6] Experienced examiners, those who understand the differences between class characteristics and individual characteristics, are crucial to distinguishing true matches. [5]

B.3 Toolmarks as they Relate to Firearms:
Toolmarks created by firearms have been extensively studied with the purpose of determining whether a specific weapon fired a specific bullet or cartridge. The bore of a firearm consistently creates unique toolmarks on bullets as a result of compressive and tensile forces. The bores of firearms are rifled to allow the bullet to spin in a more controlled manner, increasing stability and accuracy, and it is the markings created by this rifling that are transferred to the bullet when it is fired. Similarly, markings on the firing pin, breech, and ejector mechanism can be transferred consistently to the case of each cartridge as it is fired. Thus, a large number of markings exist in a firearm investigation.

If a weapon is suspected of being a match with a piece of evidence and the firearm is in the examiner's possession, marks are relatively easy to produce. The examiner simply has to load the firearm with the appropriate type of cartridge and fire the weapon in a controlled manner to enable recovery of both the case and the expended round. Toolmarks generated from the firing can then be observed to see if marks characteristic of the weapon either have or have not been transferred to the case or load of the test cartridge used.

B.4 Toolmarks as they Relate to Tools:
As described earlier, toolmarks are primarily classified as impressions and striations. To prepare to make a comparison from a specific tool, test marks must be created. Toolmark identification, in comparison to firearm identifications, faces different challenges—no standard shape or size can be expected from a toolmark. Toolmarks will vary as a function of how they were made—pressure applied, angle of the tool and twisting all introduce variations to the mark. Tools are more often subjected to abuse (using the tool for something other than its intended purpose) than firearms, creating the possibility of significantly altering the original surface. It is important to try to replicate the conditions that made the evidence mark to make the comparison as similar as possible.

When a tool and a toolmark are suspected of having a correlation, the tool is used to make a series of marks in an attempt to produce a mark similar to the evidence mark. The test marks are generally made in lead or a similar soft material. Because tools lack a standard method of use, many marks may need to be made with varying angles and pressures to replicate the found toolmark. According to Biasotti and Murdock, preparing a comprehensive set of test marks yields the best chance of positive identification, if one does indeed exist, and reduces the probability of a false inclusion. The series of preliminary marks would be compared to one another to demonstrate reproducibility of the tool. The best representative sample would then be compared to the evidence toolmark to form an opinion as to whether the tool in question did

make the mark. Biasotti and Murdock expect the following four items to be true if the tool did create the toolmark in question: the tool was used to make the evidence mark; the working surface had not been altered since the creation of the evidence toolmark; the evidence toolmark is characterized by unique features; and the surface is not simply a subclass surface or a surface that simply possesses characteristics of other similarly created tools. [3]

B.5 How Toolmark Comparisons are Made:

Toolmark comparisons, for both firearms and toolmark identifications, are typically made using an optical system such as a comparison microscope. The angle of illumination used produces regions of high and low reflectance due to the uneven surface, yielding an effectively two-dimensional method of comparison [9]. A comparative microscope employs dual stages that allow two samples to be compared simultaneously. The comparative microscope image, Figure 2, compares a section of an 'evidence' sample on left to a 'standard' sample on right—the region of match is highlighted by the box.

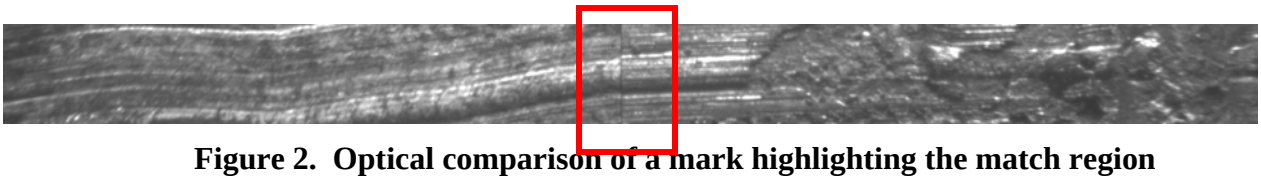


Figure 2. Optical comparison of a mark highlighting the match region

It should be noted that the two-dimensional optical striations have intrinsic three dimensional roughness. While evaluation by light microscopy is easily accomplished, it is time intensive and is dependent upon the light source and view point. The image can change substantially as the lighting is changed, while the three dimensional roughness remains constant.

The AFTE Theory of Identification as it Relates to Toolmarks [10] remains qualitative, describing a match as a "sufficient agreement." "Sufficient agreement" refers to the duplication of the surface contours of the tool in the toolmark. Two or more sets of contours are compared utilizing features of heights of peaks, widths of peaks and valleys, and curvatures. Agreement is considered "significant" when it exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools. The agreement must also be self-consistent, i.e. toolmarks known to have been produced by the same tool and be identifiable and consistent with each other. The statement 'sufficient agreement' exists between two toolmarks means that the optical agreement between the two patterns displayed by the marks is so exact that the likelihood another tool could have made the mark is considered a remote possibility. [10]

A mark examined with a comparison microscope is classified as: identification, inconclusive, eliminated or unsuitable. Identification is defined as characteristics whose agreement exceeds those of known non-matches and is consistent with agreements between known matches. Inconclusive comparisons may demonstrate: i) agreement of all characteristics but agreement of individual characteristics is insufficient to declare a match; ii) agreement of all class characteristics without agreement in individual characteristics; or iii) agreement of class characteristics with or without agreement of individual characteristics. Elimination occurs when there is significant disagreement between class and/or individual characteristics. A classification of unsuitable indicates that the sample is not appropriate for comparison on the comparative microscope. [2]

Biasotti and Murdock recommend the following be considered until a method of matching is developed: studies of consecutively manufactured tools; importance of method of manufacture

on working surfaces, mechanical and mathematical models of consecutive marks, quantity and quality of match agreements among known non-matches. [3]

B.6 Consecutive Matching Striae:

Historically matches have been made using a method described as pattern match. A conservative criteria for identification includes the use of consecutive matching striae (CMS), a method that introduces a more statistical approach to identifying matches. While pattern matching techniques seem qualitative, patterns assessed often possess features that can be quantified. The following are examples of features sought to make a match that can be quantified: positions of striae relative to a reference point; height and width of striations; and consecutive series of known height and width striae. CMS may be best described as a method to determine the best non-match observed; the understanding that matching striations occur in known non-matches creates a standard comparison within the examiner's experience for the minimum number of matching striae for an examiner to confidently declare a match.

Pattern match and CMS represent the same science but use different methods to describe it. Additionally Nichols believes an examiner who utilizes CMS (described as a line counter) may appear more impartial because the method used to describe the work is more likely to be understood by a lay person and they are able to utilize the best non-match from someone else's observation to supplement their own training and experience. [11]

A variety of known matches are reviewed and evaluated by Ronald Nichols using the conservative criteria for identification. Different firearms and toolmarks are examined and each author puts forth the minimum number of CMS they determine through their examinations. Examination of these works allows examiners to call upon other examiners' experiences to help them evaluate specific cases of evidence before them. As a point of reference, Biasotti and Murdock suggest that for three dimensional toolmarks (only ridges are counted), at least two separate groups of three CMS appear relative to one another or six CMS in a single group when compared to a test mark; for two dimensional toolmarks (striae that match exactly with respect to width and relative position), two groups of five CMS appear relative to one another, or a single group of eight CMS when compared to the test mark. Biasotti and Murdock's conservative recommendation for CMS identification has been supported by approximately 4,800 known non-match comparisons; these tests reported include no false inclusions based on their criteria. [11]

B.7 Uniqueness with Respect to Firearms:

The forensic study of markings of bullets and shell casings is an area of toolmark examination that has been widely studied, with standards in place for evaluation. These methods form the basis of a standard method of examination for other types of toolmarks. Bullets carry unique marks inherent from their manufacture and from the path they experience during firing. Optical methods are used to compare striations on test fired bullets to corresponding marks on evidence bullets. [12]

B.8 Uniqueness with Respect to Toolmarks:

The theory that tools have and make unique marks has been the premise for many investigations. Over the last century many studies have been conducted in an attempt to prove this theory; the majority find that toolmarks are indeed unique. The preponderance of this research, as well as more thorough investigations into toolmark evaluations, has been performed with firearms, making these findings a valuable resource when investigating other types of toolmarks.

After a critical toolmark match was deemed inadmissible by the State of Florida in Ramierz v. State of Florida due to the fact that the uniqueness of toolmarks had not been specifically evaluated, evidence was gathered to support the fact that toolmarks are indeed unique. As early as 1926, Calvin Goddard reported that every pistol barrel (a ballistic tool), even those newly manufactured, has unique characteristics on the barrel's surface. [13]Cassidy in his examination into tongue and groove pliers concluded that the wear and damaged areas on workings surfaces of tools over the course of years of use make them unique. Due to this, a tool will leave marks that only that tool can produce. [14] Butcher and Pugh examined successively manufactured bolt cutters and Watson consecutively manufactured knife blades and crimping dies—from which it was hypothesized that toolmarks are indeed unique, their uniqueness rising from defects in materials and manufacturing. [15] Biasotti and Murdock agree with the theory of uniqueness stating that "it is possible to individualize toolmarks because there are practical probability limits to: (1) the number of randomly distributed consecutive matching striae, and (2) the number of randomly distributed matching individual characteristics in impression toolmarks in known non-match positions." [3]Thus, they suggest a quantitative match criterion based on two groups of three consecutive striae in three dimensional toolmarks in positions relative to each other, or a single group of six consecutive striae. In two-dimensional toolmarks, they suggest two groups of five striae or a single group of eight striae when compared to the test toolmark. [3]

Not all reports support the uniqueness of all tools. Extensive testing reviewed by Bonafanti and DeKinder call attention to the few cases tested where positive identification could not be made solely from bullets fired from a specific gun. In the majority of the cases positive identification is possible—however, that does not diminish the significance of the few for which identification was not possible. It should be noted, however, that these studies really did not address whether marks were unique. Rather, they addressed the question concerning the ability to distinguish between marks, which is a more practical question for a forensic examiner. Assuming uniqueness, Bonafanti and DeKinder reviewed the reproducibility of marks made by firearms after extensive firing. Some tests were able to correctly identify the 5000^{th} bullet to the first bullet while others were only able to match the 50^{th} bullet to the first. A trend also existed chronologically; tests performed more recently had higher success rates than their forbearers. This may be attributable to the advances in imaging technology. [16]

Testing the uniqueness of firearms is simple compared to the effort needed to replicate toolmarks made with other sorts of tools. To discharge a firearm, a predictable, controlled series of events occurs. This is not the case with many simple tools. A screwdriver, for instance, may not simply be used to drive screws, it may also be used to pry or scrape, making the wear pattern difficult to predict. For example, a large number of people could use a firearm and the markings produced on the bullet and cartridge could be exactly the same. However, if two different people were asked to use a screwdriver to pry open a locked door, the number and variety of marks could vary widely due to applied pressure, twist of the tool, angle at which the person held the tool, etc. While it is true that marks made at the same angle, twist, and pressure should be similar, the sheer number of potential marks greatly increases the complexity of the match. The conclusion one can draw from examination of these studies is that uniqueness, while supported by the literature, has not been definitively proven using a controlled testing of scientific hypotheses.

C. Statement of Hypotheses

This study involves an examination of a series of 50 sequentially manufactured screwdriver tips that had yet to see service. In this study three distinct hypotheses were tested. These are shown in Table I.

Hypothesis 1:	*The 50 sequentially produced screwdrivers examined in this study all produce uniquely identifiable tool marks*
Hypothesis 2:	*In order to be identifiable, tool marks from an individual screwdriver must be compared at similar angles.*
Hypothesis 3:	*Different sides of a flat-bladed screwdriver produce different uniquely identifiable marks.*

Table I: Hypotheses tested.

These hypotheses were tested using the procedures outlined below.

II. Methods

The experimental design consisted of the acquisition of a series of ostensibly identical screwdriver tips; making marks with the tips; characterization of the marks thus made; and comparison of the produced marks using a statistical algorithm in accordance with the stated hypotheses. Details of these steps are provided in the following subsections.

A. Materials:

Fifty identical screwdrivers produced by a single manufacturer, Omega Company, were obtained. The screwdrivers were certified by Omega to have been produced sequentially although the exact order was not known. Specifics of the manufacturing were not provided by the manufacturer. Mr. Jim Kreiser, former head toolmark examiner for the State of Illinois, believes this set of tools was formed from a hexagonal rod; turned on a lathe to the appropriate shape and size; and cut off, likely with a screw machine. Examination with a stereoscope reveals circular cutting marks on the end of the tip, along the shank of the tip, and in the notches cut into the hexagonal shaft.

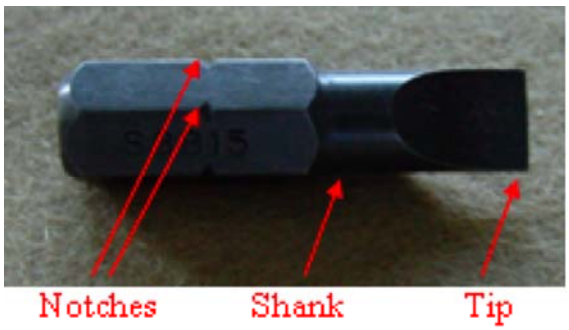


Figure 3: Typical screwdriver tip used in this study.

The sides of the tip were likely formed by grinding. The edges of the tip that create striations and are the features of interest for the SEM study are at the intersection of the cut circular surface and the ground surface [17]. The sequential production of the screwdrivers should ensure that they are as practically identical as is possible. Though the population is limited, repetition in the comparison of these tools will provide an adequate sample base, enough to form a small database that could be used to generate blind test matching.

B. Sample Production:

Mr. Kreiser used the screwdrivers to mark lead samples at angles of 30º, 60º, and 85º using a jig to maintain the selected angle for the toolmark, as illustrated in Figures 4 and 5. Both sides of the screwdriver were used. Four replicates were made of each toolmark. After the marks were made, the toolmarks themselves were characterized using a surface profilometer.

Figure 4. Jig to make toolmarks.

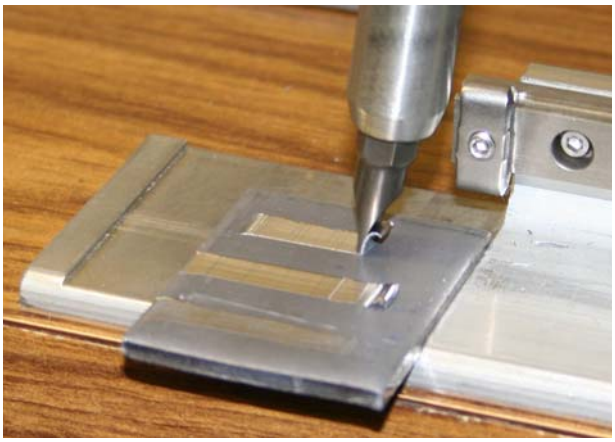


Figure 5. Detail of Jig Making Marks.

C. Profilometery

A Homelwerk Surface Stylus Profilometer was used to measure surface roughness on all toolmarks and selected tools. The advantage of a stylus profilometer is that it mechanically measures the surface, eliminating possible distortions generated from reflected light. The disadvantage is that the surface is affected by the passage of the stylus, albeit only slightly. Scans on each toolmark were performed in a region where the mark was found to be most complete by visual examination. Each scan consisted of ten separate traces run perpendicular to the toolmark striations, illustrated in Figure 5 by the red lines on the right trace in the image, with each trace sampling 9600 points along a line approximately 7 mm in length. (Note: ten traces were run simply to provide a number of traces for comparison if desired.) The vertical resolution of this device is 0.005 microns. [18].

Figure 6. Marked lead sample with red lines representing profilometer traces.

During the latter stages of the study an Alicona Infinite Focus optical profilometer was obtained for characterization, Figure 1. This instrument eliminates the danger of introducing artifacts due to having a tip in physical contact with the lead surface. It is capable of scanning with a resolution of up to 800 nm in the z axis at 5x magnification, and up to a resolution of 10 nm in the z axis at 100x magnification, over an extended x-y range of 100 mm by 76 mm respectively at 5x magnification, and 5 mm by 4 mm respectively at 100x magnification. Rough surfaces can be easily quantified with accurate measurement of Ra, Rq and Rz where Ra is the arithmetical mean

roughness of a measured surface, Rq is the root mean square roughness, and Rz is a result of ISO 9000 standards and specifically is measured over 5 peaks and valleys at 10 points on the part. Measurement of roughness, waviness and contour all conform to recognized international ISO standards. This instrument also allows for the accurate measurement of surfaces at steep angles of up to 80 degrees from the x- y plane.

Figure 7: Alicona Infinite Focus Instrument.

D. Statistics:

The algorithm developed and used matches data along the one-dimensional profilometer data traces. The data examined in this analysis are of the type collected by a surface profilometer that records surface height (z) as a function of distance (x) along a linear trace taken perpendicular to the striations present in a typical tool mark. Important assumptions in the analysis are that the values of z are reported at equal increments of distance along the trace and that the traces are taken as nearly perpendicular to the striations as possible. The algorithm then allows comparison of two such linear traces.

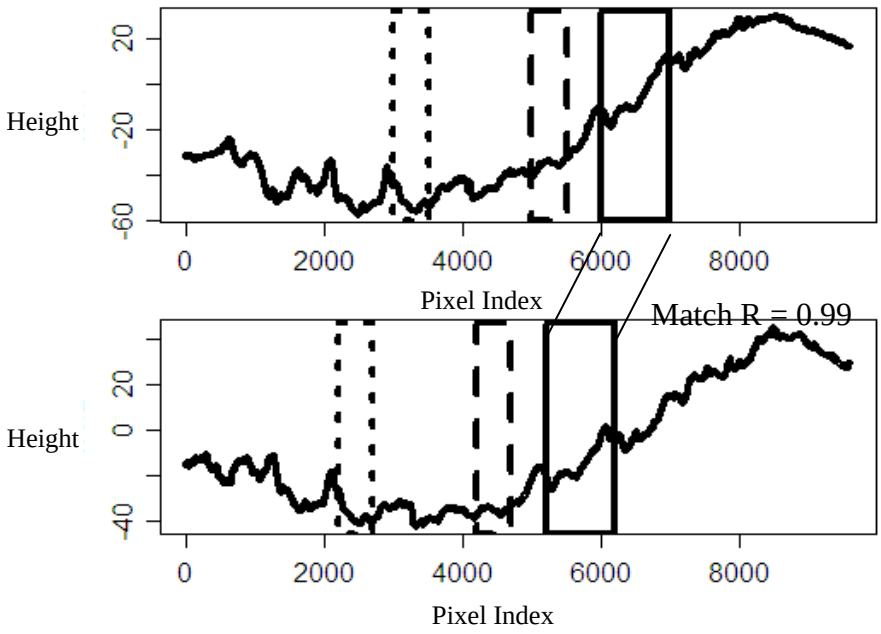
The first step taken by the algorithm, referred to as Optimization, is to identify a region of best agreement in each of the two data sets for the specified size of the comparison window. (Note: the window size is user-defined. The comparison window size used in this study was selected after a series of experiments to determine which size gave the highest percentage of correctly identified matches in cases where it was known a match did exist.) This is determined by the maximum correlation statistic, hereafter referenced as an “R-value”, and described in [19]. By way of illustration, two different possibilities are shown in Figure 8. The schematic of Figure 8a shows the comparison of a true match, i.e. profilometer recordings from two specimens made with the same tool, while Figure 8b shows data from a true nonmatch pair of specimens (i.e. two marks from two different tools). In each case, the matched regions marked with solid rectangles

are the comparison windows denoting the trace segments over which the ordinary linear correlation coefficient is largest. Note that in both cases the R-value returned is very close to 1, the largest numerical value a correlation coefficient can take. In the first instance, this is so because a match does in fact exist, and the algorithm has succeeded in finding trace segments that were made by a common section of the tool surface. In the second case, the large R-value is primarily a result of the very large number of correlations calculated in finding the best match. Even for true nonmatches, there will be short trace segments that will be very similar, and it is almost inevitable that the algorithm will find at least one pair of such segments when computing the R-value. It is primarily for this reason that the R-values cannot be interpreted in the same way that simple correlations are generally evaluated in most statistical settings.

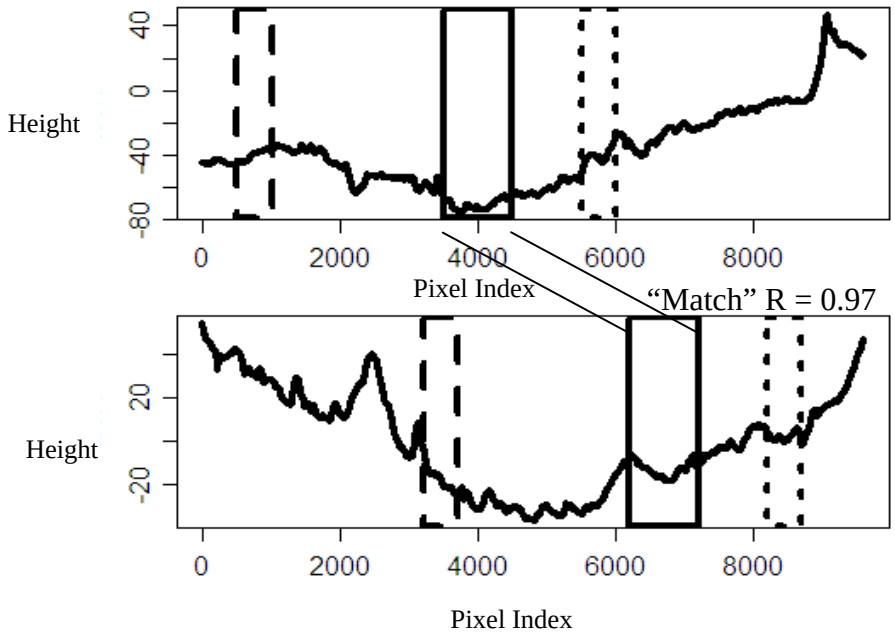
For the reasons described above, the algorithm now conducts a second step in the comparison process called Validation. In this step a series of corresponding windows of equal size are selected at randomly chosen, but common distances from the previously identified regions of best fit. For example, a randomly determined shift of 326 pixels to the left, corresponding to the dashed rectangles in Figure 8a, might be selected. The correlation for this pair of corresponding regions is now determined. Note that this correlation must be lower than the R-value, since the latter has already been determined as being the largest of all possible correlations determined in the Optimization step. The assumption behind the Validation step is that if a match truly does exist, correlations between these shifted window pairs will also be reasonably large because they will correspond to common sections of the tool surface. In other words, if a match exists at one point along the scan length (high R-value), there should be fairly large correlations between corresponding pairs of windows along their entire length. However, if a high R-value is found between the comparison windows of two nonmatch samples simply by accident, there is no reason to believe that the accidental match will hold up at other points along the scan length. In this case rigid-shift pairs of windows will likely not result in especially large correlation values.

During the Validation step a fixed number of such segment pairs is identified, corresponding to a number of different randomly drawn shifts, and the correlation coefficient for each pair is computed. (Note: For this study the number of segment pairs selected was rather arbitrarily chosen as 100. Increasing this number significantly did not materially affect the results.) Dotted and dashed rectangles displayed in Figure 8 illustrate schematically the selection of two such pairs of shifted data segments; in actual operation the algorithm chooses many such pairs. In the case of the true match the regions within the corresponding dashed windows of Figure 8a do appear somewhat similar, and can be expected to return fairly large correlation values. However, when similar corresponding pairs of windows are taken from the nonmatch comparison of Figure 8b, the shape of the scans within the windows is seen to differ drastically. Lower correlation values will be obtained in this case.

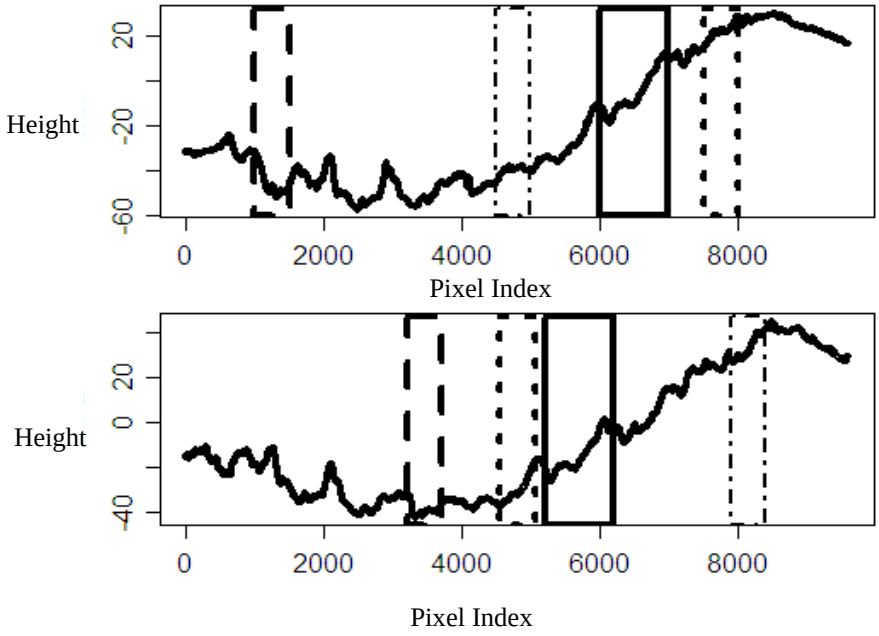
The correlation values computed from these segment-pairs can be judged to be “large” or “small” only if a baseline can be established for each of the sample comparisons. This is achieved by identifying a second set of paired windows (i.e. data segments), again randomly selected along the length of each trace, but in this case, without the constraint that they represent equal rigid-shifts from their respective regions of best fit. In other words, for this second set of comparisons the shifts are selected at random and independently from each other – any segment of the selected length from one specimen has an equal probability of being compared to any segment from the other. This is illustrated in Figure 8c for three pairs of windows, denoted by the dashed rectangles, the dotted rectangles, and the dot-and-dash rectangles.



a.



b.



c.

Figure 8: a) Comparison pair showing a true match. Best region of fit shown in solid rectangle with corresponding R value. Note the similarity of the regions within the two possible sets of validation windows (dashed and dotted rectangles). b) Comparison pair showing a true nonmatch. While a high R value is still found between "Match" segments, the validation windows are distinctly different from one another. c) Validation windows (dashed, dotted, and dot-and-dash rectangles) selected at random for the comparison pair shown in a) to establish a baseline value.

The Validation step concludes with a comparison of the two sets of correlation values just described, one set from windows of common random rigid-shifts from their respective regions of best agreement (Note: the regions of best agreement are excluded from this step), and one set from the independently selected windows. If the assumption of similarity between corresponding points for a match is true the correlation values of the first set of windows should tend to be larger than those in the second. In other words, the rigid-shift window pairs should result in higher correlation values than the independently selected, totally random pairs. In the case of a nonmatch, since the identification of a region of best agreement is simply a random event and there truly is no similarity between corresponding points along the trace, the correlations in the two comparison sets should be very similar.

A nonparametric Mann-Whitney U-statistic (referred to in this paper as T1), computed from the joint ranks of all correlations computed from both samples, is generated for the comparison. Where the correlation values of the two comparison sets are similar, T1 takes values near zero, supporting a null hypothesis of "no match". If the correlations from the first rigid-shift sample are systematically larger than the independently selected shifts, the resulting values of T1 are larger, supporting an alternative hypothesis of "match".

E. AFTE Study

In order to compare the effectiveness of the algorithm to human examiners, and potentially identify areas where the algorithm might be enhanced or improved, a blind study was conducted during the 2008 Association of Forearms and Tool mark Examiners Training Seminar. During

the course of this meeting 50 different volunteers rendered over 250 opinions on some of the sample pairs used for this study and evaluated by the algorithm.

A series of 20 comparison pairs covering a range of T1 values from low to high were selected from the tool marks produced at the 85 degree comparison angle. Of the twenty comparison pairs, five were from samples where the algorithm correctly identified a matched set (high T1); five were correctly eliminated nonmatch comparisons (low T1); five were incorrectly eliminated matched sets (T1 values in the low or inconclusive range); and five were incorrectly identified nonmatches (intermediate or high T1). Examiners were asked to assess each pair of samples twice. For the initial observation, paper blinders were placed on the samples so that examiners were restricted in their view to the same general area where the profilometer data were collected, Figure 9. After making an initial assessment, the blinders were removed and the examiner was given the opportunity to make a second assessment based on a view the entire sample. In each case, examiners were asked to render an opinion as to whether they were viewing a positive identification, a positive elimination, or inconclusive, for reasons that will become apparent.

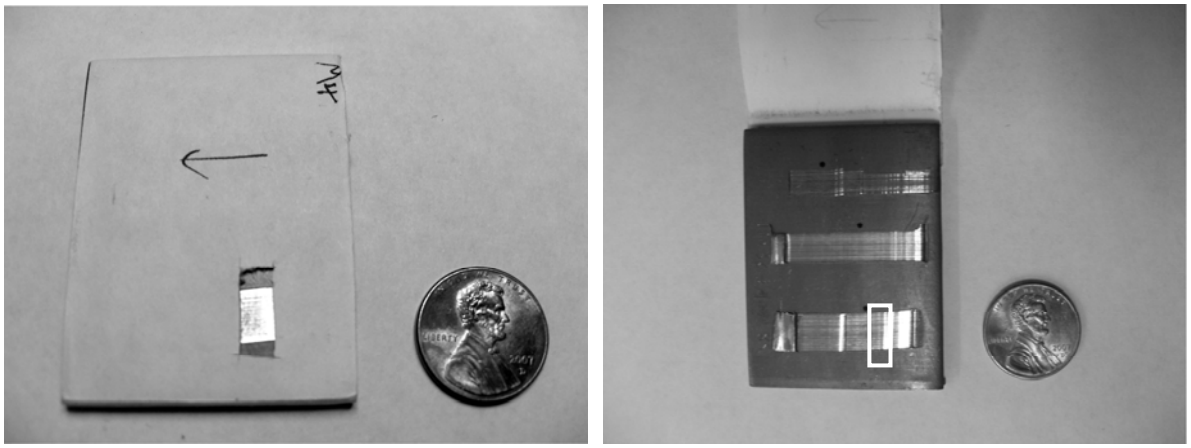


Figure 9: Image of a tool marked plate with a) blinder in place and b) removed, showing the entire mark. Area shown in a) indicated by white rectangle.

Names of examiners were not recorded, although demographic data was collected concerning the experience and training of the volunteers. Of the 50 volunteers all except five were court qualified firearm and tool mark examiners. Of the remaining five, two were firearms (but not tool mark) qualified, two were in training, and one was a foreign national where a court qualification rating does not exist. Volunteers were required to do a minimum of two comparison pairs, and could do as many as they wished. Several chose to do the maximum number of comparisons possible. Numbers were assigned to identify each volunteer during data collection; afterwards the ID numbers were randomly mixed to preserve anonymity.

Examiners were asked to use whatever methodology they employed in their respective labs. This caused some confusion initially and placed constraints on the volunteers since some labs never use the term “positive elimination”, while others are reluctant to use the term “positive identification” unless the examiner personally either makes the marks or knows more information about them than what could be supplied in this study. After understanding this the examiners were told the direction of the tool when making the mark and that the tool marks were all made at the same angle from similar, sequentially made, flat blade screwdriver tips. Also,

examiners were told that for the purposes of the study they could consider the terms of "positive elimination" or "inconclusive" to be essentially interchangeable.

F. Optical vs. Stylus Profilometer

Once the optical profilometer was acquired, a study was initiated to compare the results obtained from using this instrument when contextual information was included as part of the search criteria. The samples chosen to be re-evaluated were those examined at the 2008 AFTE convention. The resultant data were analyzed using the same computer algorithm discussed above and the T-statistic indexes determined. The samples examined fall into four distinct classes: True match samples were the algorithm returned a high T1 value; true Match samples where the algorithm returned a low T1 (indicative of a nonmatch); true nonmatch samples where the algorithm returned a low T1; true nonmatch samples where the algorithm returned a high T1 (indicative of a match).

III. Results

A. Hypotheses Testing:

The data obtained from the profilometer was used to test the hypotheses of Table I, which are commonly held as being true by tool mark examiners. The first and most fundamental assumption, that all tool marks are unique, was tested by a comparison of marks made by different screwdriver tips at the angles of 30, 60 and 85 degrees with respect to horizontal. The T1 statistic values are shown in Figure 10 as a function of angular comparison. The data is plotted as box plots, the boxes indicating where 50% of the data falls with the spread of the outlying 25% at each end of the distribution shown as dashed lines. As stated previously, when using a T1 statistic a value relatively close to 0 indicates that there is essentially no evidence in the data to support a relationship between markings. For pairs of samples made with different screwdrivers (Figure 10) the majority of the index T1 values produced by the algorithm fall near the 0 value; only 3 outlier comparisons had a T1 value greater than ± 4.

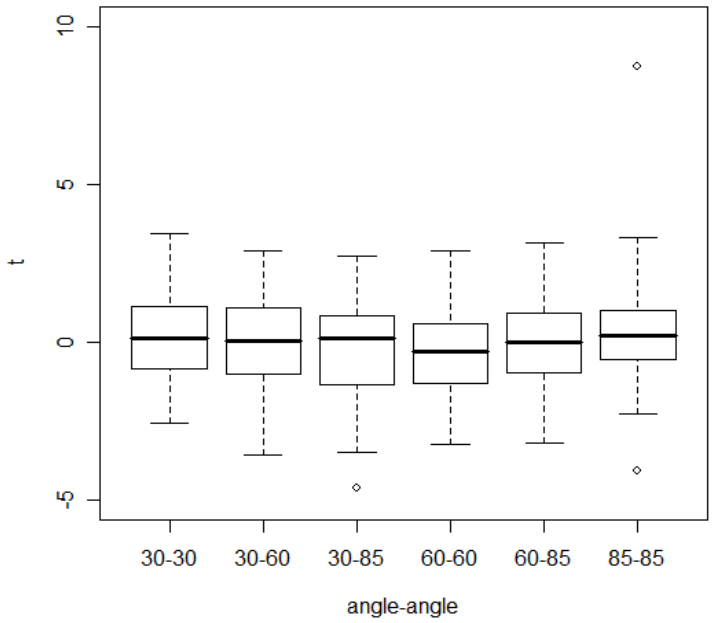


Figure 10: Box plots showing T1 results when comparing marks from different screwdrivers.

In comparison, Figure 11 displays indices computed using the algorithm from profilometer scans of marks made by the same side of the same tool and compared as a function of angle. While marks made at different angles still produce index values near 0, the T1 statistic jumps dramatically when marks made at similar angles are considered. Clear separation is seen between the 50% boxes, although overlap still exists when the outliers are considered.

Taken together, Figures 10 and 11 support Hypotheses 1 and 2. When comparing tool marks made at similar angles with different tools, the resulting T1 values cluster near zero (Figure 10), but when the same tool is used to make marks at similar angles, the T1 distributions are on substantially larger values, giving support for Hypothesis 1. Support for Hypothesis 2 is demonstrated by Figure 11 alone, since even among same-tool marks, only those made at the same angle produce large T1 values.

The last hypothesis considered was that when comparing tool marks made from screwdriver tips, the marks must be made from the same side of the screwdriver; marks made using different sides of the screwdriver appear as if they have come from two different screwdrivers. These results are shown in Figure 12. The hypothesis is again supported because, as in Figure 10, the T1 values cluster around 0 regardless of the angles used in making the marks, indicating no relationship between the samples.

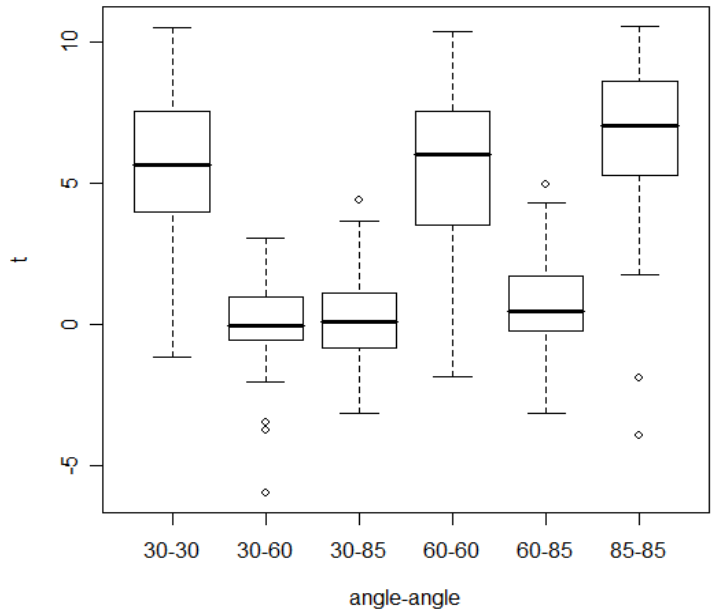


Figure 11: Box plots showing T1 results when comparing marks obtained from the same side of the same screwdrivers.

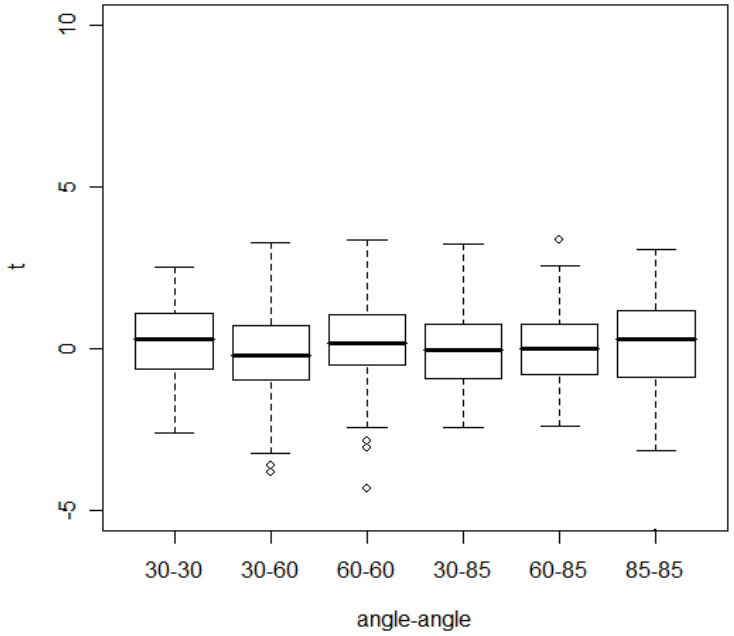


Figure 12: Box plots showing T1 results when comparing marks made from different sides of the same screwdrivers.

The T1 values summarized in Figures 10 and 11 are replotted in Figure 13, with the density of data in the y-axis plotted as a function of T1 value to make it easier to view the data separation. Examination of these plots indicates that the algorithm operates best using data obtained at higher angles than lower angles, i.e. the separation between match and nonmatch is more defined for the 85 degree data than, for example, the 30 degree data. This is believed related to the quality of the mark. As the angle of attack of the screwdriver with the plate increased the quality

of the mark increased. It was common to obtain marks that represented the entire screwdriver tip at high angles, while marks at lower angles were often incomplete [5]. Algorithm performance also appears more effective at reducing false positives than it does in eliminating false negatives. At all angles known matches were found with very low T1 values, while nonmatches with high T1 values were very limited. The fact that false negatives are more common than false positives is most likely related to the fact that a poor quality tool mark will often result in a false negative. On the other hand, a poorly created tool mark would almost never result on a false positive.

While T1 is a much more stable index of match quality than R-value, problems still remain in establishing an effective, objective standard for separating true matches from nonmatches. Ideally, when employing standard U-statistic theory the critical T1 values separating the regions of known matches (black lines) and known nonmatches (gray lines) should remain constant for all data sets. Examination of Figure 13 shows that this is not the case. For example, reasonable separation for the 30 and 60 degree data appears to be somewhere around a T1 value less than 5, but rises to approximately 7 for the 85 degree data. This variation is most likely due to lack of complete independence among the correlations computed in each sample, arising from the finite length of each profilometer trace.

For the reasons discussed above, assigned threshold values indicating “Positive ID” and “Positive Elimination,” and denoted by black lines on the graphs of Figure 13, were chosen based on a K-fold cross validation using 95% one-sided Bayes credible intervals. Specifically, the lower threshold is a lower 95% bound on the 5^{th} percentile of T1 values associated with nonmatching specimen pairs, and the upper threshold is an upper 95% bound on the 95^{th} percentile of T1 values associated with matching specimen pairs. The region between these two threshold values is labeled “Inconclusive”. A Markov Chain-Monte Carlo simulation was used to determine potential error rates.

Using this method the estimated error rates are as follow. For comparisons made at 30 degrees the estimated probability of a false positive (i.e. a high T1 value for a known nonmatch comparison) is 0.023. In other words there is a possibility of slightly over two false positives for approximately every 100 comparisons. The estimated probability of a false negative is 0.089, or almost 9 true matches having a low T1 value per every 100 comparisons. The cross-validation method used ensures that all the data have similar error rates, and the rates found for the 60 and 85 degree data are approximately 0.01 and 0.09 for false positives and false negatives, respectively. What is most noticeable is that the T1 lower threshold value for the 85 degree data is much larger than for the 30 and 60 degree data, being 4.10, 1.34 and 1.51, respectively. This suggests that a more distinct difference is required to classify nonmatches for the 30 and 60 degree cases than is true for the 85 degree case. This, in turn, results in a corresponding increase for the estimated inconclusive error rates, which are 0.103, 0.298, and 0.295 for the 85, 60 and 30 degree data, respectively. It would, of course, be possible to shift these error rates, i.e. produce fewer false negatives at the expense of more false positives, by altering the percentiles used in our estimation procedure.

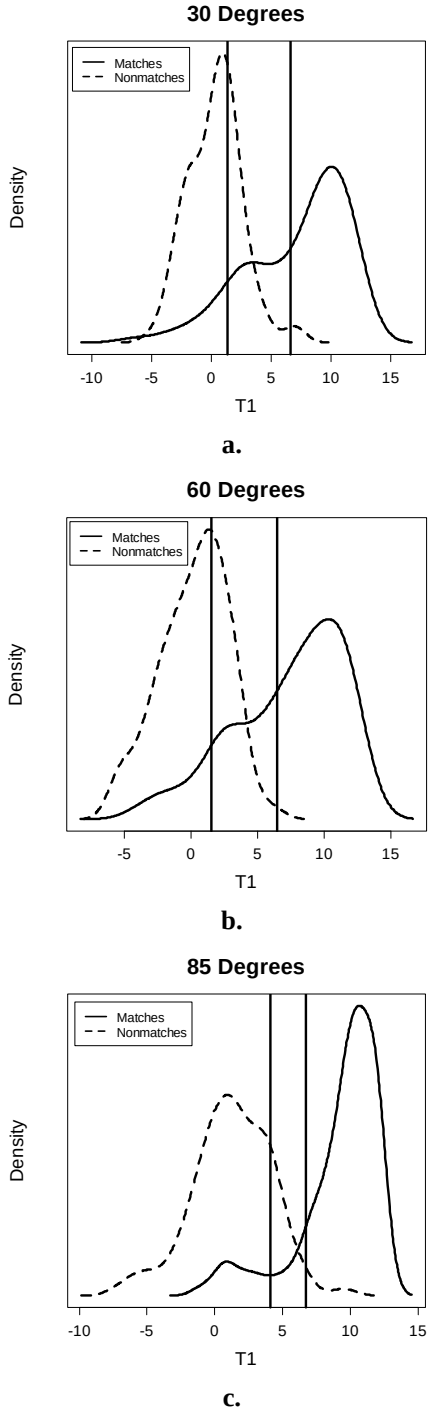


Figure 13: Summation of the T1 values from comparisons made at a) 30 degrees; b) 60 degrees; and c) 85 degrees.

B. Results of AFTE Study:

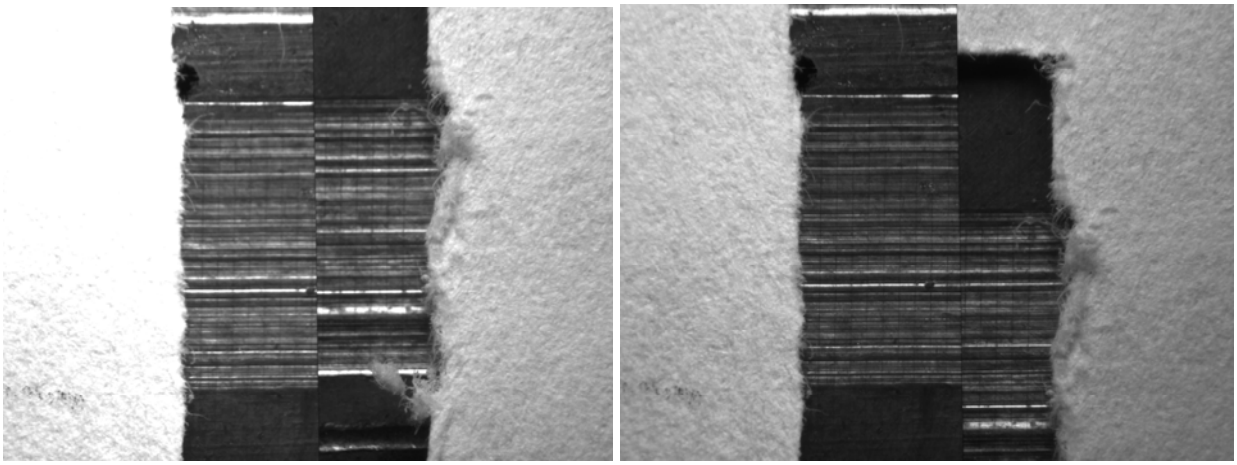
Results of the computerized analysis of specimen pairs was compared to expert evaluations of the same samples made by volunteer examiners at the 2008 Association of Firearm and Tool mark Examiners seminar. However, before the algorithm performance can be discussed in comparison to the data obtained at the Association of Firearm and Tool mark Examiners seminar

using human volunteers, a brief consideration of the constraints experienced by the examiners is in order. Firstly, it should be recognized that the conditions under which the examiners rendered an opinion would ordinarily be regarded as restrictive or even professionally unacceptable. Without having the tool in hand, or without being permitted to make the actual mark for comparison, tool mark examiners were forced to make assumptions they would not make in an actual investigation. For example, without having the screwdriver tip in hand the examiners did not know whether the mark they observed represented the entire width or only a portion of the screwdriver blade. Secondly, given this uncertainty about how the specimen was made, examiners tended to be more conservative in their willingness to declare a positive identification or elimination. During the course of the Association of Firearm and Tool mark Examiners study several examiners commented that typical lab protocol would require them to have physical access to the subject tool before rendering a "positive identification" judgment. Finally, examiners do not typically employ the terms used to denote the three regions identified for error analysis. Thus, while privately saying they felt a comparison was a "positive elimination" (given their knowledge of the test being conducted), lab protocol required an opinion of "inconclusive" to be rendered. Such policies are put in place since the signature of a tool may so change during use that a mark made at one point in time may not resemble a mark made with the same tool at a different point in time, e.g., after the tip has been broken and/or re-ground. In such cases positive elimination is only allowed if the class characteristics of the marks are different.

When viewed in light of these constraints, some interesting observations concerning the algorithm performance are apparent. In a small number of cases (12 out of 252 comparisons), when examining the entire tool mark after first viewing only the restricted area where the profilometer scans were obtained, examiners changed their opinion from inconclusive to either positive ID or positive elimination. This indicates that algorithm performance might be improved simply by increasing the amount of data processed. This may be achieved, for example, by ensuring that the profilometer scans span the entire width of the mark or possibly by considering a number of scans taken at locations dispersed along the entire length of the available mark.

In a slightly smaller number of cases, comparisons between specimens made by the same screwdriver that were not conclusively identified as such by the algorithm also presented problems for the examiners. Five true matches that received low T1 values and were classified as a positive elimination by the algorithm were examined during the Association of Firearm and Tool mark Examiners study. Three of the five were given ratings of "inconclusive" or "positive elimination" on one occasion, and one particular comparison sample (designated MW4) was rated this way seven times. Thus, while examiners in general were vastly superior to the algorithm in picking out the matches, both the algorithm and the examiners had more trouble with some true matches than with others.

Close examination of the sample that was most often problematic for examiners (i.e. MW4) was conducted and the images obtained are shown in Figure 14. Figure 14a shows the side-by-side comparison of the marks, where no match is seen. Note that the mark width matches extremely well, and the entire mark seems to be present. Figure 14b shows the samples positioned where the true match is evident. It can be seen that each mark only represents a portion of the screwdriver blade width, predominately from the two sides of the tip. A match is only possible if the marks are offset, allowing the opposing "edge" sections (which actually were produced by the middle of the screwdriver blade) to be viewed side-by-side.



a. **b.**

Figure 14: Sample MW4. a) Tool marks placed so that assumed edges align. b) Correct placement required for positive identification.

This sample points out weaknesses in the study conducted at Association of Firearm and Tool mark Examiners as well as in the laboratory tests of the algorithm. In a screwdriver mark comparison it is common for examiners to use the edges of the marks as initial registration points for the start of an examination. Since examiners make the comparison marks themselves they are well aware of the edge markings, if not for the evidence marks, at least for the marks they produced. In the Association of Firearm and Tool mark Examiners study, such information was not provided and may have led to some false assumptions. For example, in the majority of cases the volunteers were under some pressure to quickly conduct a comparison before, e.g. the next meeting session started, or so that another examiner could use the equipment, etc. Due to these time constraints, samples often were placed on the stages of the comparison microscope for the volunteer, giving the examiner little or no time to observe the macroscopic appearance of the mark. Without the benefit of seeing the size of the entire mark, and given the identical widths of the two partial marks for sample MW4 when initially viewed using the comparison microscope, the assumption that the entire width of the screwdriver blade was represented would be a natural one. However, such an assumption could easily lead to an inconclusive conclusion, especially if the examiner was being conservative due to lack of information concerning the sample. Such a situation would never happen in actual practice since the examiners themselves employ the tool in question to make test marks for examination.

The problem described above essentially relates to the examiners having a lack of a point of reference or registry of the mark for the comparison. The same could be said of the algorithm and the manner in which it performs, since no point of registry exists to indicate when the data being acquired is actually coming from a tool-marked region or from the unmarked plate. All of the profilometer scans analyzed by the algorithm were run using the same set of sampling parameters. However, the initial positioning of the stylus was inexact. For incomplete marks, large regions of the unaffected lead plate were also scanned in order to keep the file sizes consistent and this lack of registry could have affected algorithm performance. This is not immediately evident if one examines the raw profilometer traces, Figure 15. In this figure the top and bottom traces show the entire scans while the two middle traces show the matched details found within the two corresponding solid rectangles superimposed on the top and bottom

traces. At first sight the two scans do appear quite different, as the offset in the scans, revealed during examination at Association of Firearm and Tool mark Examiners, is not immediately evident in the data files. Given observation of Figure 14, one can mark the approximate location of the region that is common between the two traces; this is shown in Figure 15 by the dashed rectangles. In this case, paired validation windows, displaced equal amounts in either direction may return a low T1 value since the majority of either scan is not held in common with the other. In other words, there is a high probability that the validation windows fall in regions where no correspondence between plates exists (see Figure 14b). Thus, what should be a match is rated as a nonmatch.

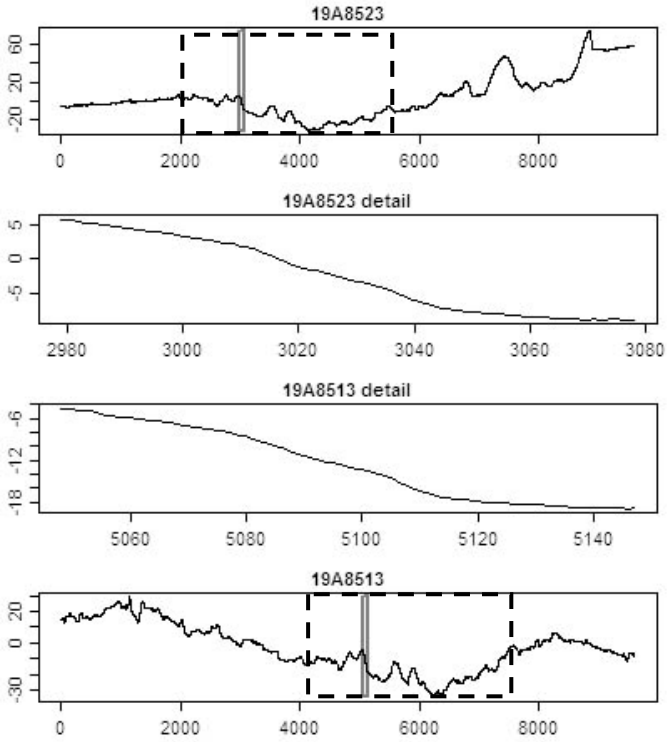


Figure 15: Profilometer data showing results from comparison MW4. The match region is shown by the solid rectangle. Dashed rectangles show the approximate location of the common region revealed in Figure 14.

A somewhat different problem is revealed when traces from true nonmatch samples are examined, Figure 16. In these instances, the optimization step may identify windows at extreme edges of the two traces as being most similar. Given the nearness of the match to the ends of the traces, the random selection of paired, rigid-shift windows during the validation step is severely constrained. For the example shown in Figure 16a the match region (denoted by solid rectangles) falls at the extreme right ends of the data files. This means that the rigid translations taken for each pair of verification windows must always fall to the left of the match region. While this may be less than desirable, the entire mark is still available for validation and a large number of rigid-shift windows spaced across the entire length of the file should be sufficient to produce good separation between this accidental match and the T1 values of a true match. However, this is not true for the true non-match shown in Figure 16b. In this case the windows identified in the optimization step as being most similar are at opposite ends of the compared

data traces. The distances of possible rigid translations are constrained to a short distance to the left of the top profile and a short distance to the right of the bottom profile. Thus, the majority of the mark cannot be used in the validation step for this accidental match. If the regions in the immediate vicinity of the accidental match are also similar, high T1 values may be returned due to the constrained sampling parameters, giving results that cannot be separated from a true match.

Once these difficulties were realized the optimization windows used in the validation step were changed to specifically address the problems encountered in these select samples. While the algorithm could be made to correctly identify these difficult comparisons, it came at the expense of creating problems with other comparison pairs. This shows that while algorithm performance could be tailored to address and correctly identify any specific pair, setting up a general set of operating parameters to encompass all possible comparison sets will most likely always result in a finite amount of error.

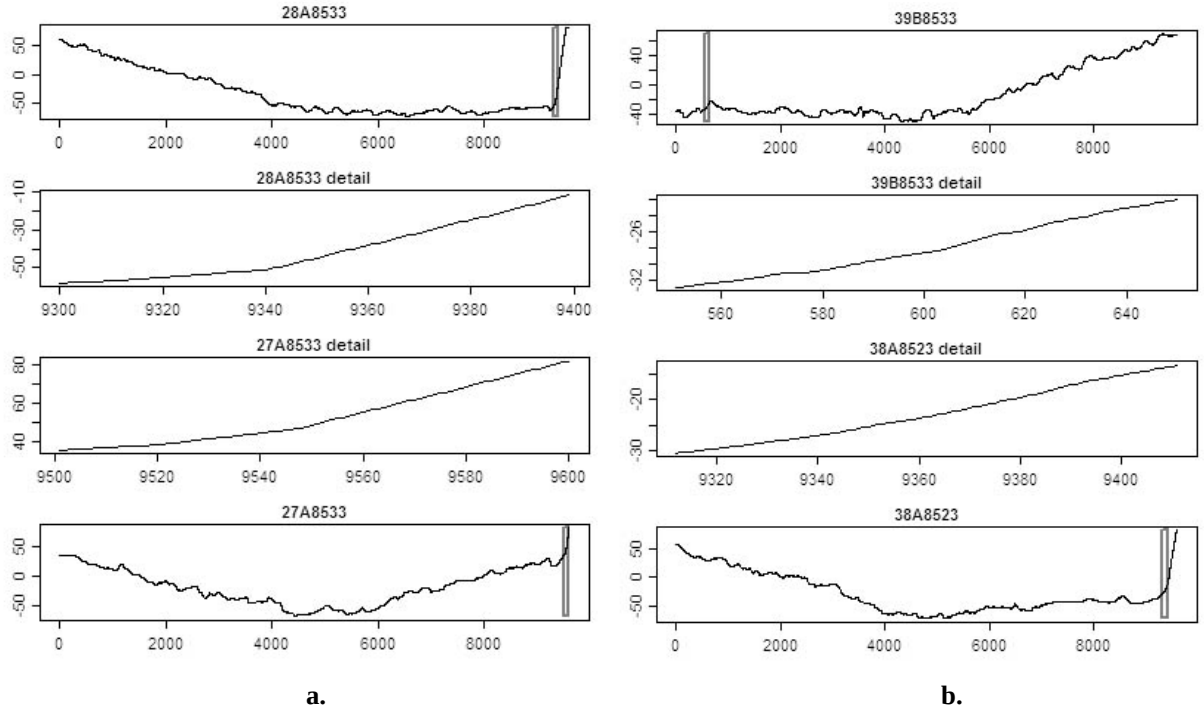


Figure 16: Comparisons of traces obtained from four different screwdrivers that were rated as possible matches by the algorithm. Match areas denoted by thin rectangles. Good agreement found at a) similar and b) opposite ends of the traces resulted in high T1 numbers for known non-matching pairs.

The above discussion suggests that further development of the algorithm to incorporate additional data concerning the region of the profilometer trace that is actually tool-marked and/or the location of the tool edge might improve its performance. While tool mark examiners do not directly use features such as these as a basis for identification, they do use it indirectly in establishing a context for the comparison. Such information, routinely and quickly noted by a human examiner, is unavailable to the current algorithm. The algorithm treats all possible pairs of trace windows the same way and functions under the assumption that all marks analyzed are essentially the same, i.e., it assumes the screwdriver tip has completely marked the lead plate and

that no unmarked regions exist. This clearly is not the case. At this time it appears the best way to enhance algorithm performance is to ensure that all comparison windows (i.e. Match and Validation) are taken from regions representing the true marked surface of the lead, and that most-similar windows found at the trace edges are used as a basis for match identification only if they are found at the same end of their respective traces.

As a final comment, it should be noted that all types of volunteers (practicing examiners, trainees, retired examiners) were involved in the study, with records kept as to the experience of the participant. The majority of volunteers were currently practicing examiners, followed by those who were either recently retired or no longer actively conducting examinations, with a limited number of trainees (less than 5) taking part. Examination of the demographic data in relation to the results showed no significant difference between experienced examiners and rather newly qualified examiners or those in training; all performed equally well.

C. Results of Optical Profilometer Study

The results of samples evaluated using the optical profilometer and analyzed using the algorithm are shown in Figure 17. In this figure five samples were analyzed in each of the four categories, the x-axis showing the sample number with the y-axis showing the corresponding T1 values measured for that sample using the two techniques. Note that all of the optical data was taken immediately adjacent to the trace of the prfilometer stylus and that a high T1 value is indicative of a match, while values near to 0 indicate little similarity between the comparison scans. (It should be noted that the consistent increase in T1 values for the stylus scans is merely a function of which samples were selected for the study.)

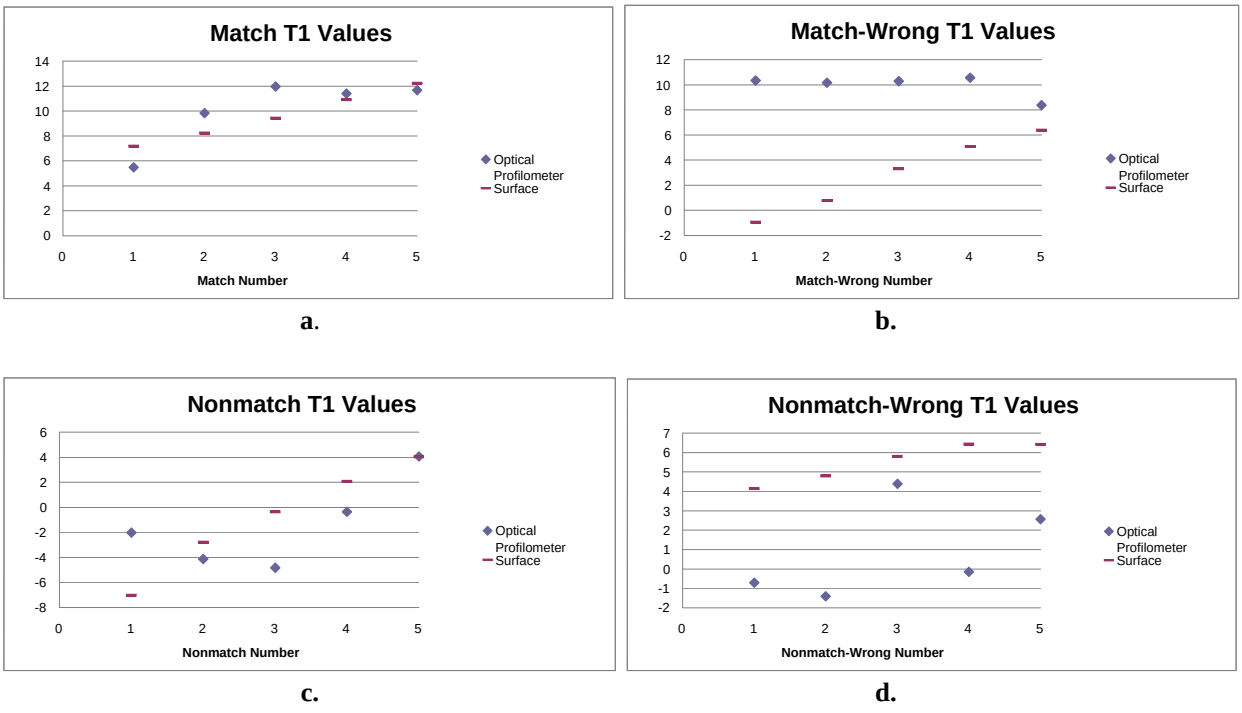


Figure 17: Comparison of T1 values for data obtained using a stylus profilometer vs. the optical IFM. a) true matches properly classify as such; b) true matches improperly classified as nonmatches; c) true nonmatches properly classified as such; d) true nonmatches improperly classified as matches.

In almost all cases the T1 values obtained using the optical data are superior to those from the profilometer. For the correctly identified matches of Figure 17a the optical T1 values are similar to the stylus data, being only slightly higher on average. However, a dramatic change is seen when the data from the previously incorrectly analyzed match samples of Figure 17b is considered. All of these samples now clearly would be classified as matches. Similarly, the nonmatch samples correctly identified as nonmatches have similar values for the optical and stylus data (Fig. 17c); the nonmatches incorrected identified as matches using the stylus data correctly result in low T1 values in four out of five cases when the optical data is used (Fig. 17d).

While it is encouraging that the optical data out performs the stylus data, care should be taken in interpreting this results as being solely due to the application of the IFM. For the initial study using surface profilometer data, contextual information was not taken into account. In that study the algorithm was allowed to compare the linear marks without regard to which side of the mark corresponded to the left and/or right side of the screwdriver, in other words, it was assumed the profiles could be "flipped" so each pair of profiles was compared in two possible relative orientations. As was pointed out by the AFTE study, ensuring that this type of contextual information was used was one manner in which performance might be enhanced. This information was included for the comparisons of the optical data. However, it is clear that the optical data obtained using the IFM is of excellent quality.

IV. Conclusions

Discussion of Findings: The goal of this study was to answer the following question: Can a series of toolmarks be obtained and compared in an automated manner to yield objective, statistically valid matches when toolmarks related to a particular tool (and only that tool) are compared to each other? This question has definitely been answered in the affirmative for the population of tools examined. The ability to discriminate between markings is based on a number of factors, including:

1. The quality of the marking itself.
2. The ability to quantify the marking that exists.
3. The manner in which the objective, automated routine is designed to operate.

The importance of having high quality markings is supported both by the objective results of the study and the study undertaken using practicing examiners. As angle of the markings increased, mark quality increased, allowing higher T1 numbers for correct matches.

If high quality markings do exist, using suitable methods to quantify the results and incorporating contextual information into the analysis of the data greatly increase the ability of an automated routine to separate marks made from ostensibly identical tools. The optical profilometer performed very well in producing high quality data files, at least as well as the stylus profilometer. This implies that the acquired evidence need not be subjected to a characterization method that involves contact of the sample in order for a suitable comparison to be made.

The drawback of automated routines is that the data is analyzed in the absence of any context. For example, the computer routine currently cannot tell whether a mark represents the edge of

the screwdriver, whether the entire screwdriver width is present, which side of the marking relates to the corresponding screwdriver edge, etc. The AFTE study showed that this contextural evidence is routinely a part of an examiners comparison methodology. When incorporated in a basic manner into the algorithm by restricting the random approach of the matching step to one that more nearly mimics examiners, the results of the algorithm improved dramatically. This shows the value of having an experienced examiner involved in algorithm development and operation.

Implications for Policy and Practice: Given that the tools examined in this study should have been as identical as possible to one another implies that unique markers do exist for every tool manufactured by this company using the tools currently employed for their manufacture. The question then becomes, of the factors listed above, what elements must be addressed to yield a fully automated, objective result? If a poor quality marking exists an unambiguous determination may be impossible. The level to which the toolmark(s) must be examined then becomes a matter of question and this level was not determined in this study. Certainly the level used in this study would appear sufficient if contextual information is included. Finally, the exact manner in which the algorithm operates, and the manner in which data is acquired, becomes a critical question.

Implications for Further Research: Testing of the algorithm on other types of tool marks would be appropriate to determine the applicability of the program to other marks. Currently, the algorithm is set up to evaluate striated marks; whether this can be generalized to other marks is unknown.

A second area of research is the question concerning the cut-off values used to qualify a comparison as either a match or nonmatch. The T1 values identified for the angles tested were seen to vary. This is most likely related to the quality of the marking. Some quantitative measure of mark quality could possibly be developed that would give an indication of when the results of a comparison are likely to be valid. Related to this, a study concerning the variance in data would be of value. This would involve having multiple persons acquire data from the same samples then employ the algorithm to see how the results compare between operators. This study would yield valuable statistical information concerning algorithm performance and robustness.

Finally, a third type of study suggested involves characterization of the toolmark with the goal being to “reverse engineer” the mark to create the tool from which the mark must have come. In this research scenario a “virtual tool” would be created that would allow any mark to be simulated, whether full or partial. The simulation would then provide data concerning exactly how the tool had to have been used to create the mark in question, yielding information concerning the angle of attack, the applied pressure, twist of the tool, etc. Such a study would also address the question of uniqueness since only a virtual tool that adequately simulates the actual could be manipulated in the manner needed to produce a satisfactory comparison. The authors of this report have begun initial experiments on this idea and have produced encouraging results, indicated that such a study is possible.

V. References

1. Meyers, Charles. “Firearms and Toolmark Identification An Introduction” AFTE Journal. Vol. 25, No. 4, 1993, pp. 381-385.

2. Miller, Jerry. "An Introduction to the Forensic Examination of Toolmarks." AFTE Journal. Vol.33, No.3, 2001, pp. 233-248.

3. Biasotti, Alfred and John Murdock. "The Scientific Basis of Firearms and Toolmark Identification." Chapter 23, Section 23-2.0, Modern Scientific Evidence: The Law and Science of Expert Testimony. (D.L. Faigman, D.H. Kaye, M.J. Saks, and J. Sanders eds., West Publishing Co., 1997) Vol. 2, pp 131-151.

4. "Fundamentals of Firearms ID." An Introduction to Forensic Firearm Identification. 2005. Accessed 8 March 2007 at: <http://www.firearmsid.com/A_FirearmsID.htm>

5. Churchman, J.A. "The Reproduction of Characteristics of Signatures of Cooey Rifles." AFTE Journal, Vol. 13, No. 1, 1981, pp. 46-52.

6. Nichols, Ronald. "The Scientific Foundations of Firearms and Tool Mark Identification—A Response to Recent Challenges." An Introduction to Forensic Firearm Identification. 2005. Accessed 7 March 2007 at: <http://www.firearmsid.com/Feature%20Articles/nichols060915/AS%20Response%20110805.pdf>

7. Nichols, R. "Firearm and Tool Mark Identification: The Scientific Reliability and Validity of the AFTE Theory of Identification Discussed Within the Framework of a Study of Ten Consecutively Manufactured Extractors." AFTE Journal Vol. 36, No. 1, 2004, pp.67-88.

8. Bonfanti, M.S. and J. DeKinder. "The Influence of the Use of Firearms on their Characteristic Marks." AFTE Journal. Vol. 31, No. 3, 1999, pp. 318-323.

9. De Kinder, Jan, Pascal Prevot, Marc Pirlot, Bart Nys. "Surface topology of bullet striations: An innovating technique" AFTE Journal. Vol. 30, No. 2, 1998, pp. 294-299.

10. "Theory of Identification, Range of Striae Comparison Reports, and Modified Glossary Definitions—An AFTE Criteria for Identification Committee Report." AFTE Journal. Vol. 24, No. 2, 1992, pp. 336-340.

11. Nichols, Ronald. "Consecutive Matching Striations (CMS): Its Definition, Study and Application in the Discipline of Firearms and Tool Mark Identification." AFTE Journal Vol. 35, No. 3, 2003, pp. 298-306.

12. Atsuhiko Banno, Tomoito Masuda, Katsushi Ikeuchi. "Three dimensional visualization and comparison of impressions on fired bullets." Forensic Science International. Vol. 140, No. 2-3, 2004, pp. 233-240.

13. Miller, Jerry and Michael McLean "Criteria for Identification of Toolmarks" AFTE Journal. Vol. 30, No. 1, 1998 pp. 15-61.

14. Cassidy, F. H. "Examination of Toolmarks from Sequentially Manufactured Tongue and Groove Pliers" Journal of Forensic Sciences. Vol. 25, No, 4, 1980, pp 796-809.

15. 401: Validation for Toolmark Identification: Manufacturing Methods. Ramirez v. State of Florida SC 92975 2000

16. Bonafanti, M.S. and J. DeKinder. "The Influences of Manufacturing Processes on the Identification of Bullets and Cartridge Cases—A Review of Literature." Science and Justice. Vol. 39, 1999, pp. 3-10.

17. Discussion with Mr. Jim Kreiser

18. Hommelwerk Company Literature

19. Faden D, Kidd J, Craft J, Chumbley LS, Morris M, Genalo L,. Kreiser J, Davis S. Statistical Confirmation of Empirical Observations Concerning Tool mark Striae. Association of Firearm and Tool mark Examiners Journal 2007 Summer;39;3:205-214.

VI. Dissemination of Research Findings

A. Refereed Publications

D. Faden, J. Kidd, J. Craft, L.S. Chumbley, M. Morris, L. Genalo, J. Kreiser, S. Davis, "Statistical Confirmation of Empirical Observations Concerning Toolmark Striae," AFTE Journal, 39, 3, 205-214, Summer 2007.

L.S. Chumbley, J. Kreiser, C. Fisher, J. Craft, M. Morris, L. Genalo, S. Davis, D. Faden, J. Kidd, "Validation of Toolmark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm," accepted, Journal of Forensic Science.

B. Theses Written

Julie Ann Kidd, "Comparison of screwdriver tips to the resultant toolmarks," Master of Science Thesis, Iowa State University.

C. Non-Refereed Publications

L.S. Chumbley, D.J. Eisenmann, M. Morris, S. Zhang, J. Craft, C. Fisher and A. Saxton, "Use of a Scanning Optical Profilometer for Toolmark Characterization", to be published, Proceedings, Scanning '09, Monterrey, CA.

D. Presentations

L.S. Chumbley, L. Genalo, C. Bossard, "Characterization of Tool Marks," Midwest Forensics Resource Center, Ames, Iowa, June 5, 2003.

L.S. Chumbley, C. Bossard, L. Genalo, J. Kreiser, "Quantification of Toolmarks, " annual meeting AAFS, Dallas, TX, 2004.

L.S. Chumbley, L. Genalo, M. Morris, C. Bossard, M.J. Kreiser, D. Faden, J. Kidd, S. Davis, C. Pfau, I. Overton, "Quantitative Characterization of Toolmarks for Comparative Identification," Invited poster presentation, NIJ Showcase, AAFS, New Orleans, LA, Feb. 2005.

L.S. Chumbley, L. Genalo, M. Morris, J. Kidd, D. Faden. S Davis, “Quantification of Toolmarks”, Invited presentation, NIJ Workshop, Washington, D.C., May 2006.

M. Morris, L.S. Chumbley, L. Genalo, , J. Kidd, D. Faden. S Davis, “Quantification of Toolmarks”, NIJ Grantees Meeting, Boston, MA, July 2006.

L.S. Chumbley, J. Kidd, J. Craft, M. Morris, L. Genalo, J. Kreiser, “Quantitative Analysis of Toolmarks Using Stereoimaging ” invited presentation, Scanning 2007, April 10-12, Monterey, 2007.

L.S. Chumbley, J. Kidd, J. Craft, M. Morris, L. Genalo, and J. Kreiser, “Statistical Analysis of toolmark Striations,” 2007 General Forensics R&D Grantees Meeting, IAI 92nd International Education Conference, San Diego, July, 2007.

L.S. Chumbley, J. Kidd, J. Kraft, M. Morris, L. Genalo, D. Faden, J. Kreiser, C. Fisher, “Statistical Analysis of Toolmark Striations”, AFTE, Honolulu, May, 2008.