Document Title:	**Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons**
Author:	**Nicholas D. K. Petraco, Ph.D.; Helen Chan, B.A.; Peter R. De Forest, D.Crim.; Peter Diaczuk, M.S.; Carol Gambino, M.S., James Hamby, Ph.D.; Frani L. Kammerman, M.S.; Brooke W. Kammrath, M.A., M.S; Thomas A. Kubic, M.S., J.D., Ph.D.; Loretta Kuo, M.S.; Patrick McLaughlin; Gerard Petillo, B.A.; Nicholas Petraco, M.S.; Elizabeth W. Phelps, M.S.; Peter A. Pizzola, Ph.D.; Dale K. Purcell, M.S.; Peter Shenkin, Ph.D.**
Document No.:	**239048**
Date Received:	**July 2012**
Award Number:	**2009-DN-BX-K041**

Report Title:

Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons

Award Number:

2009-DN-BX-K041

Authors:

Nicholas D. K. Petraco[a,b], Ph.D.; Helen Chan[a], B.A.; Peter R. De Forest[a], D.Crim.; Peter Diaczuk[a,b], M.S.; Carol Gambino[a,c] M.S., James Hamby[d], Ph.D.; Frani L. Kammerman[a], M.S.; Brooke W. Kammrath[a,b], M.A., M.S; Thomas A. Kubic[a,b], M.S., J.D., Ph.D.; Loretta Kuo[a], M.S.; Patrick Mc Laughlin[a,e]; Gerard Petillo[f], B.A.; Nicholas Petraco[a,e], M.S.; Elizabeth W. Phelps[a], M.S.; Peter A. Pizzola[g], Ph.D.; Dale K. Purcell[a,b], M.S.; Peter Shenkin[a], Ph.D.

[a]John Jay College of Criminal Justice, City University of New York
[b]The Graduate Center, City University of New York
[c]Borough of Manhattan Community College, City University of New York
[d]International Forensic Science Laboratory & Training Centre
[e]New York City Police Department
[f]Independent Firearms Examiner
[g]New York City Office of the Chief Medical Examiner

Abstract

Over the last decade, forensic firearms and toolmark examiners have encountered harsh criticism that there is no accepted methodology to generate numerical "proof" that independently corroborates their morphological conclusions. This project strives to answer that criticism and focuses on:

a. The collection of 3D quantitative surface topographies of toolmarks by confocal microscopy;
b. Identification of relevant modern multivariate machine learning methods for tool-toolmark associations and estimations of identification error rates; and
c. Dissemination of toolmark surface data and software generated for the project to aid further research.

A database was assembled which consists of 3D striation and impression patterns on Glock fired cartridge cases, screwdriver and chisel striation patterns. The database is now available to registered users. Statistical studies were carried out on a large portion of the primer shears (cartridge cases) and screwdriver striation patterns collected thus far. Principal component analysis, canonical variate analysis and support vector machine methodology was used to objectively associate these toolmarks with the tools that created them. Estimated toolmark identification error rates were on the order of 1% using these algorithmic methods. Conformal prediction theory was used to assign confidence levels to each toolmark identification and is suggested as a useful measure in gauging the quality of a toolmark "match" for a multivariate

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

classification system. The findings of this objective and quantitative scientific research reinforce the general conclusions codified in the AFTE theory of identification.

Table of Contents

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Executive Summary

1. Introduction

Forensic science has come under increased scrutiny in recent years. In February 2009, the National Academy of Sciences (NAS) released their report on the forensic sciences in the United States. The report, entitled "Strengthening Forensic Science in the United States: A Path Forward," states that "much forensic evidence— including, for example, bite marks and firearm and toolmark identifications—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline" (p. 3-18). The NAS report further contends that "sufficient studies have not been done to understand the reliability and repeatability of the methods (p. 5-21)" and, as a result, "additional studies should be performed to make the process of individualization more precise and repeatable" (p. 5-21). This experiment sought to develop a statistical foundation for assessing the likelihood that one tool is the source of a given toolmark to the exclusion of all other tools.

Impression evidence has received the brunt of attack, and while some of the criticism is justified, much of it is naive and based on misunderstandings. Impression evidence is a broad category of important, commonly encountered, and valuable physical evidence. It includes fingerprints, toolmarks, footwear impressions, tire tracks, and those impressions associated with firearms identification (i.e. microstriae in land impressions on bullets, breech face impressions, firing pin impressions, and other marks on cartridge cases). Although impression evidence of various types has been used successfully for decades, its examination has lacked a well-articulated scientific basis. This research seeks to place the analysis of impression evidence, specifically those made by tools and firearms, on a sound scientific foundation by laying down, testing, and fully publishing methodological statistical foundations for toolmark impression pattern recognition and comparison

2. Scope of the project

This study focuses on striation patterns left by tools and on cartridge casings imparted by firearms. All impressions made by tools and firearms can be viewed as mathematical patterns composed of features. In order to recognize variations in these patterns, we used the mathematics of multivariate statistical analysis. In a computational pattern recognition context, this is called

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

machine learning. The mathematical details of machine learning can give what Moran calls "…the quantitative difference between an identification and non-identification" (Moran 2002). They also enable the estimation of extrapolated identification error rates and even in some cases, the calculation of rigorous, universal random match probabilities (Duda 2001; Fukunaga, 1990; Theodoridis 2006; Kennedy 2003; Kennedy 2005).

The overarching aim of this research is to lay down, test and publish multivariate statistical foundations for tool mark impression pattern recognition and comparison. In order to realize this overarching goal, the project is divided into three main initiatives:

1. Toolmark pattern collection and archiving
2. Database and web interface construction for the distribution of tool mark data, and software developed for this project.
3. Identification/exploitation of multivariate machine learning methods relevant to the analysis of collected toolmarks, striation patterns in particular.

3. Conclusions

This research outlines a set of objective and testable methods to associate toolmark impression evidence with the tools and firearms that generated them. Striation patterns are the focus. The results complement previous univariate based toolmark discrimination studies and are consistent with and buttress the qualitative conclusions of the forensic firearms and toolmark examination community.

Three dimensional confocal microscopy, surface metrology and multivariate statistical methods lie at the heart of the approach presented in this project. Through the studies described, practitioners can see how a surface metrological-statistical scheme can provide an investigative aid and estimate algorithmically based identification error rates for firearm and toolmark comparisons.

Striated toolmarks were collected from screwdrivers and chisels. Striated and impressed toolmarks were collected from cartridge cases. Quantitative confocal images of the surface topographies of all toolmarks examined have been included in a database. The forensic research and practitioner community can access information in the database at the website URL: http://toolmarkstatistics.no-ip.org/ . Data from this project is being made available for further research by the academic and practitioner communities and for interested practitioners to

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

construct images for court exhibits. Several pieces of software, including software for visualization of/measurement on the toolmark surfaces in the database, were generated in the course of the project. All software and R statistical analysis scripts used are available on the website.

The reasonably complete striation patterns from screwdrivers and the primer shear from 9mm Glock fired cartridge cases could be summarized as multivariate feature vectors in the form of mean profiles. These mean profiles were used with standard multivariate machine learning methods in order to estimate identification error rates from such an algorithmic regime. A combination of principal component analysis (PCA), canonical variate analysis (CVA) and support vector machines (SVM) proved most effective for accomplishing this task with low identification error rate estimates, generally ~1% with 95% confidence intervals ~[0%,3%]. Bootstrap resampling was used to estimate these identification error rates and confidence intervals. Conformal prediction theory (CPT) was used to assign rigorous levels of confidence to all PCA-CVA-SVM toolmark identifications. Such levels of confidence can help a judge or jury assess the quality of an algorithmic association of a tool to a toolmark. The CPT classifiers proved to be reasonably efficient, producing only small multi-label confidence regions and only at relatively low rates. Uninformative confidence regions were not observed. Note that bootstrapping methods, PCA, CVA, SVM and CPT have very few underlying assumptions built in, and this was a major reason why they were chosen. This is a major advantage to their use in a courtroom setting where their results will be far more likely to stand up to adversarial scrutiny and be less open to attack.

Unfortunately the three-dimensional impressed toolmarks and the “patchy” chisel striation patterns proved too complicated for our current suite of developed software to analyze at this time. (This is another reason why we are making the data collected for the project available to the wider research community.) Development of open source software for the machine learning analysis of complete three-dimensional impression patterns and incomplete toolmarks will be the subject of future research.

That said, practitioners could apply the machine learning regime presented here, to any set of reasonably complete striation patterns (i.e. of reasonable quality), and generate tool-toolmark association error rate estimates and identifications at a chosen level of confidence. Given the findings of the studies presented above as well as those of previous univariate based

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

projects, it is surmised that the results will be consistent with the theory that no two striation patterns derived from different tools are identical.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Final Technical Report

I. Introduction

1. Statement of the problem

Forensic science has come under increased scrutiny in recent years. In February 2009, the National Academy of Sciences (NAS) released their report on the forensic sciences in the United States. The report, entitled "Strengthening Forensic Science in the United States: A Path Forward," states that "much forensic evidence— including, for example, bite marks and firearm and toolmark identifications—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline" (p. 3-18). The NAS report further contends that "sufficient studies have not been done to understand the reliability and repeatability of the methods (p. 5-21)" and, as a result, "additional studies should be performed to make the process of individualization more precise and repeatable" (p. 5-21). This experiment sought to develop a statistical foundation for assessing the likelihood that one tool is the source of a given toolmark to the exclusion of all other tools.

Impression evidence has received the brunt of attack, and while some of the criticism is justified, much of it is naive and based on misunderstandings. Impression evidence is a broad category of important, commonly encountered, and valuable physical evidence. It includes fingerprints, toolmarks, footwear impressions, tire tracks, and those impressions associated with firearms identification (i.e. microstriae in land impressions on bullets, breech face impressions, firing pin impressions, and other marks on cartridge cases). Although impression evidence of various types has been used successfully for decades, its examination has lacked a well-articulated scientific basis. This research seeks to place the analysis of impression evidence, specifically those made by tools and firearms, on a sound scientific foundation by laying down, testing, and fully publishing methodological statistical foundations for toolmark impression pattern recognition and comparison.

2. Review of relevant literature

The NAS report (2009) contends that "many forensic tests—such as those used to infer the source of toolmarks or bite marks—have never been exposed to stringent scientific scrutiny" (p. 1-6). Critics of toolmark examination argue that the field has no scientific basis, that error rates

are unknown and incalculable, and that comparisons are subjective. However, there have been numerous studies on the topic of toolmarks and toolmark examination, especially in the area of firearms, which examine the reproducibility of toolmarks, individualization of toolmarks, reliability of methods, as well as method validation.

2.1 Introduction to toolmarks and toolmark examination

Toolmarks are generated when a hard object (tool) comes into contact with a relatively softer object (De Forest, Gaensslen, and Lee, 1983). "When two objects come into contact, the harder object may mark the surface of the softer object. The tool is the harder object. The relative hardness of the two objects, the pressures and movements, and the nature of the microscopic irregularities on the tool are all factors that influence the character of the toolmarks produced" (Biasotti, Murdock, & Moran, 2008). There are three categories of toolmarks: (1) imprints, which are two-dimensional contact markings; (2) indentations, which are three-dimensional contact markings; and (3) striations, which are sliding contact patterns, in which one or both surfaces move. Imprints involve a transfer of material to (or removal from) some surface, while indentations are produced when a harder material leaves an impression of its surface features and contours in a softer one.

Forensic science often involves matching a questioned piece of evidence to a known item, by analyzing class characteristics, subclass characteristics, and individual characteristics. Class characteristics are properties that all members of a certain class of objects have in common. They are produced from design factors and are determined prior to manufacture. Subclass characteristics are distinct surface features of an object that are more restrictive than class characteristics, but not as unique as individual characteristics. Subclass characteristics are produced incidental to manufacture and can arise from a source that changes over time. Subclass characteristics are significant because they relate to a subset of the class to which they belong. Individual characteristics are also produced incidental to the manufacturing process and are typically at the microscopic level. These characteristics are produced by the random imperfections or irregularities on the surfaces of the tools used to manufacture the object. Class characteristics guide forensic scientists in identifying a piece of evidence, while individual characteristics aid in individualizing a piece of evidence. Miller (1997) states that "even though hundreds of barrels may have been produced consecutively, the manufacturing process and

subsequent use often permits the identification of the individual barrel that the bullet was fired from" (p. 282 – 283).

Individualization of crime scene evidence to its unique source is a common goal in forensic science, although this may be difficult to achieve. Forensic science is the application of natural sciences to matters of law; it is different from traditional sciences in the sense that we need to know where an item came from, not only what the item is. For forensic science, the "where it came from" may be a critical part in solving a case. Knowing that the object is a screwdriver is insufficient; we need to know if this is the screwdriver used in the crime, and that this screwdriver belonged to the suspect. Literature to date suggests that it is possible to individualize a particular mark or a bullet to a specific tool or firearm.

According to the Association of Firearm and Tool Mark Examiners (AFTE) Theory of Identification (1998), there are four categories of examination outcomes typically used by toolmark examiners: (1) Identification; (2) Inconclusive; (3) Elimination; and (4) Unsuitable for comparison. AFTE defines "identification" as an "agreement of a combination of individual characteristics and all discernable class characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool" (p. 86). An inconclusive outcome is declared when there is: (1) some agreement of individual characteristics and all discernable class characteristics, but insufficient for identification, (2) agreement of all discernable class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility, or (3) agreement of all discernable class characteristics and disagreement of individual characteristics, but insufficient for an elimination (p. 87). Elimination occurs when there is a significant disagreement of discernable class characteristics and/or individual characteristics (p. 87). The final possible outcome, "Unsuitable for comparison," occurs when there are no microscopic marks of value for comparison. Toolmark examiners are impartial observers attempting to determine whether a toolmark and a particular tool match. They offer their opinion based on their examination of the evidence. Toolmark examiners obtain information about a piece of evidence so that it may be combined with other facts and assumptions to form a theory of what happened.

2.2 Individualization of toolmarks

Individualization is the process of determining whether two objects have a common origin. "The individualization of firearms and toolmarks involves the physical comparison of one solid object with another solid object to determine through pattern recognition whether or not they were: (1) once part of the same object; (2) in contact with each other; or (3) share similar class or individual characteristics" (Biasotti, Murdock, & Moran, 2008). As Burd and Gilmore (1968) state, "identifying a toolmark produced by a specific tool requires finding sufficient correspondence in both class and individual characteristics in the mark and on the tool surfaces" (p. 390). They go on to explain that the mass production of tools often results in repetition of structural details, especially when tools were formed in a mold, die stamped, or die forged. When Burd and Gilmore (1968) analyzed several mass-produced screwdrivers of the same model, they concluded that even though the tools had similar surface features, the abrasion markings made by each screwdriver were distinct. As a result, identification of toolmarks produced by a specific tool was possible because the screwdriver tips were individual and unique. Nevertheless, they acknowledge that certain types of structure can resemble accidental characteristics that could be mistaken for individual characteristics.

Miller (1998a; 1998b) conducted two experiments to observe if there were any changes to the tool working surfaces and their effect on subclass characteristics. He analyzed the production of cut nails at various stages of manufacturing and explained that the "manufacturing process imparts toolmarks to various areas of the nails. The toolmarks are reproducible on many nails, and a microscopic examination of the nails shows identifiable toolmarks on the head, flat, and edge" (Miller, 1998a, p. 493). In the first experiment, he collected six samples of consecutively manufactured 4d cut masonry nails every 30 minutes for 9 hours from a single machine, totaling 32,400 nails. Miller (1998a) concluded that all of the nails exhibited toolmarks, which could be identified to the tool producing them. At 3,600 nails, the toolmarks present on the edge were not as well defined as those present on the first six nails. However, Miller (1998a) explains that this did not preclude an identification. The last six nails were also compared to the first six nails and it was determined that the toolmarks observed on the nail flat, nail edge, and nail head could still be identified to the tool which produced the toolmark. In the second experiment (Miller, 1998b), six sample nails were collected every 1000 nails from an entire production run of nails. Miller (1998b) concluded that, as the tool wears, striated toolmarks would change more quickly than

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

impressed toolmarks. Furthermore, these groups of nails acquired subclass characteristics in the manufacturing process; they had identifiable and reproducible toolmarks, but could not be identified to nails produced before or after this group in the run.

Brundage (1998) obtained ten consecutively rifled Ruger P-85 pistol barrels, both standards and unknowns, for examination by thirty firearm examiners from nationally accredited laboratories. He sought to determine if the forensic firearm examiners could accurately (1) distinguish between two or more multiple gun barrels that were consecutively rifled or (2) differentiate individual characteristics of bullets fired from gun barrels that were consecutively rifled. Each test set consisted of thirty-five bullets for analysis (fifteen unknown bullets and twenty test standards). Of the results collected, there were no incorrect answers (inconclusive answers were not considered incorrect). Each examiner properly associated each gun barrel and all unknown bullets. However, one laboratory did not have an answer for one of the barrels, but also had one bullet that was not identified to any of the barrels. From the results, Brundage (1998) concluded that properly trained firearm examiners could distinguish between two or more bullets fired from consecutively rifled gun barrels, as well as accurately differentiate the individual characteristics of test shots fired from consecutively rifled gun barrels. Furthermore, he determined that, not only are consecutively rifled gun barrels different from each other, they are unique and can be differentiated from each other.

Hamby, Brundage, and Thorpe (2009a) extended Brundage's 1998 study to address the following issues:

1. To determine if a firearm and toolmark examiner has the ability to correctly associate test fired bullets to the correct consecutively rifled gun barrels;
2. To expand the test data base from the original 67 participants to participants in laboratories worldwide;
3. To provide test sets of known bullet pairs and unknown test bullets from the 10 consecutively rifled barrels for laboratories to use in their organizational training programs;
4. To evaluate the issue of subclass characteristics on bullets fired from consecutively rifled barrels;
5. To provide information to counter various legal challenges concerning the ability of firearm and toolmark examiners to identify bullets to firearms;

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

6. To provide examiners with examples of best known nonmatch (KNM) bullets. (p. 104)

The authors sought to determine if trained firearm and toolmark examiners could identify unknown fired bullets to the rifled barrels. Ten consecutively rifled Ruger P-85 pistol barrels were obtained from the manufacturer and test fired to produce "known" bullets and "unknown" bullets. These known and unknown bullets were provided to firearms examiners around the world for comparison. Of the 7,605 unknown fired bullets examined, only three of the bullets were considered unsatisfactory for microscopic examination due to damage. Two firearm and toolmark examiner trainees were unable to match five of the unknown fired bullets to the known samples. The remaining 7,597 unknown fired bullets were correctly identified by participants to the provided known bullets. Hamby et al. (2009a) explained that the test procedure used to ascribe bullets fired from consecutively rifled barrels is reproducible on a worldwide basis because there were no actual errors.

> "Based on the results of this research, having fired bullets in good condition and properly trained firearm and toolmark examiners, the identification process has an extremely low estimated error rate. In circumstances where bullets are deformed or fragmented, the comparison process may be more difficult and the error rate may increase. This study also shows that various statements made about the inability of examiners to associate fired bullets to consecutively rifled barrels were incorrect." (p. 107)

In summary, Hamby et al. (2009a) concluded there were identifiable surface features on fired bullets that allow the individualization of a fired bullet to the gun that fired it. From the literature, it is clear that it is possible to individualize a particular toolmark to the tool that produced it. Now that we know individualization is possible, we move on to the different materials and methods of toolmark examination.

2.3 Materials for experimentation

Contrary to criticisms of toolmark examination, there has been much quality scientific research into the methods and techniques for toolmark examination. In terms of producing or replicating toolmarks, Cowles and Dodge (1948) found that polished aluminum was a good material for making test toolmarks. They also found that the angle at which a tool is held may alter the toolmark significantly. Grodsky (1999) determined that Elmer's Glue was an

inexpensive and non-destructive technique to replicate a toolmark. Du Pasquiera, Hebrardb, Margota, and Ineichen (1996) evaluated and compared various elastomers and plasters as casting materials based on (1) practical features (ease of use), (2) hardening time, (3) viscosity, (4) dimensional stability (molding should accurately reproduce the impression dimensions), (5) elastic memory (cast does not return to its original shape), (6) temperature dependence, (7) conservation (preservation of a cast so it should not deteriorate), and (8) cost. The test materials (Sta Seal®, Xantopren®, Coltoflax®, Express®, Imprint®, Mikrosil®) were assessed regarding their dimensional behavior. The researchers concluded that while all of the tested products had disadvantages as well, Mikrosil was the best choice for crime scene work, Xantopren was the best choice for lab work, and Sta Seal could be used for either crime scene or lab work. Greene and Burd (1950) discussed using plastic casting of die impressions to reproduce toolmarks and using magnesium smoke treatment to reveal toolmarks.

Petraco, Petraco, and Pizzola (2005) explain that test toolmarks were generally made on soft metal or metal alloys, such as lead, because they are soft enough to make test marks without damaging the tool's working surface. Because the soft metals are malleable, it is easier and may create more accurate reproductions of a tool's working surface. However, the reproduction of several identical test toolmarks can be difficult to achieve with soft metal test materials. Because of the health hazards certain soft metals pose to the examiner, Petraco et al. (2005) proposed jewelry modeling or carving waxes as alternative materials for the preparation of test toolmarks for comparison microscopy. In their experiment, a test tool was applied to a piece of wax. The authors explained that the replicas obtained were exact, highly detailed, 1:1, negative impressions of the exemplar tool's working surface and were suitable for use in toolmark examination and comparison cases. Jeweler's carving wax was an ideal material for producing test toolmarks, since their initial purpose was to produce highly detailed, intricate carvings to be cast into jewelry. Moreover, the jewelry wax did not shrink and was applicable to any category of tools (i.e. – hand tools, power tools, etc.). In addition to being inexpensive and readily available, the wax is available in many sizes, shapes, flexibility, etc., and could be stored at room temperature without drastic changes. An important aspect for toolmark examiners is that if the wax was packaged properly, it could be transported easily without breakage and had a long stable shelf-life.

Petraco, Petraco, Faber, and Pizzola (2009) provide a summary of various jewelry

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

modeling waxes commercially available for preparing toolmark standards. They explain the process of creating toolmark standards with the exemplar tool and jewelry modeling wax:

1. An appropriate piece of modeling wax is selected and the toolmark standards are then prepared;
2. Excess wax is removed as necessary both prior to and after making the toolmarks;
3. Any veil of wax obscuring the toolmark standards is removed by treatment with a solvent as necessary; and
4. Each toolmark standard is marked for identification. (p. 356)

As in the Petraco et al. (2005) study, the authors explained that the replicas obtained were exact, highly detailed, 1:1, negative impressions of the exemplar tools working surface and were suitable for use in toolmark examination and comparison cases.

2.4 Two schools of thought

The basic elements of toolmark examination and comparison include the reproduction of toolmarks resembling the questioned toolmark, and comparison of the toolmarks. Methods and materials for the reproduction of toolmarks were described previously. However, the analysis and comparison of toolmarks seems to be divided into two schools – those comprised of "pattern matchers" and those comprised of "line counters." Traditional toolmark examination uses a comparison microscope, which gave the examiner the ability to observe and compare two objects at the same time under magnification. These examiners would compare the test toolmark with the questioned toolmark simultaneously and determine if the pattern on both objects matched (see Figures 1 and 2).

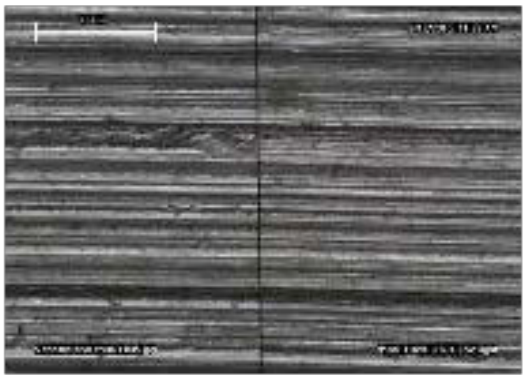


FIGURE 1. Images from a comparison Microscope of a known-match. Photograph courtesy of Gerard Petillo.

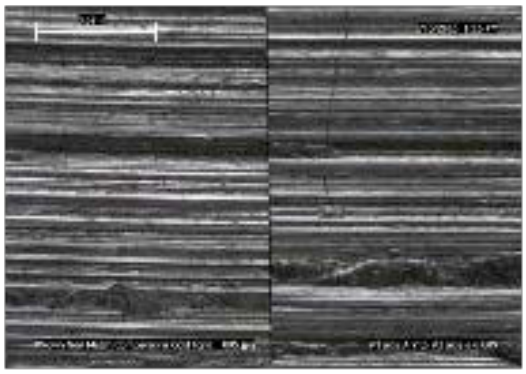


FIGURE 2. Images from a comparison Microscope of a known non-match. Photograph courtesy of Gerard Petillo.

As can be seen in Figure 1, it is clear that the striations line up in the known-match. On the other hand, the striations do not line up in the known non-match in Figure 2. However, this method of toolmark examination has been criticized as subjective because it relies on the toolmark examiner's knowledge and experience.

The other school of thought focuses on consecutively matching striae (CMS), which are striae within an array of striated markings that agree in their spatial relationship, their width, and their morphology. Using this method, examiners focus not only on the number of striations, but also on the position and relative height and width of the striations. According to Biasotti's 1959 study, a line is defined as a "striation appearing on the bullet as the result of being engraved by the individual irregularities or characteristics of the barrel, plus any foreign material present in the barrel capable of engraving the bullet. Biasotti (1959) goes on to explain that two lines are considered matching when: (1) the bullets are in phase; (2) their angle lies between the long axis of the bullet and the angle of the twist; and (3) the lines appear to be similar in contour and of common origin (p. 37). Biasotti (1959) methodically quantified the patterns he observed. In essence, CMS is a numerical description of a toolmark. CMS is often misunderstood to be a method of pattern matching, when in fact it is a quantitative method of describing an observed pattern.

Biasotti, Murdock, and Moran (2008) set out guidelines regarding consecutive matching striae (CMS). They define two-dimensional (2D) striated toolmarks as "any impressed or striated toolmarks that lacks discernable depth or: (1) occupies only the very surface of a recording

medium in which the toolmark appears; (2) has been made in a recording medium that is very thin or; (3) results from the application of the tool to the medium in such a way that only superficial markings are produced" (p. 616). In 2D striated toolmarks, "when at least two groups of at least five consecutive matching striae appear in the same relative position, or one group of eight consecutive matching striae are in agreement in an evidence toolmark compared to a test toolmark" (p. 621). Three- dimensional (3D) striated toolmarks are defined as "any impressed or striated toolmark that displays discernable contour because the medium of the toolmark is in has been displaced" (p. 616). In 3D striated toolmarks, "when at least two different groups of at least three consecutive matching striae appear in the same relative position, or one group of six consecutive matching striae are in agreement in an evidence toolmark compared to a test toolmark" (p. 621). (See Figure 3).

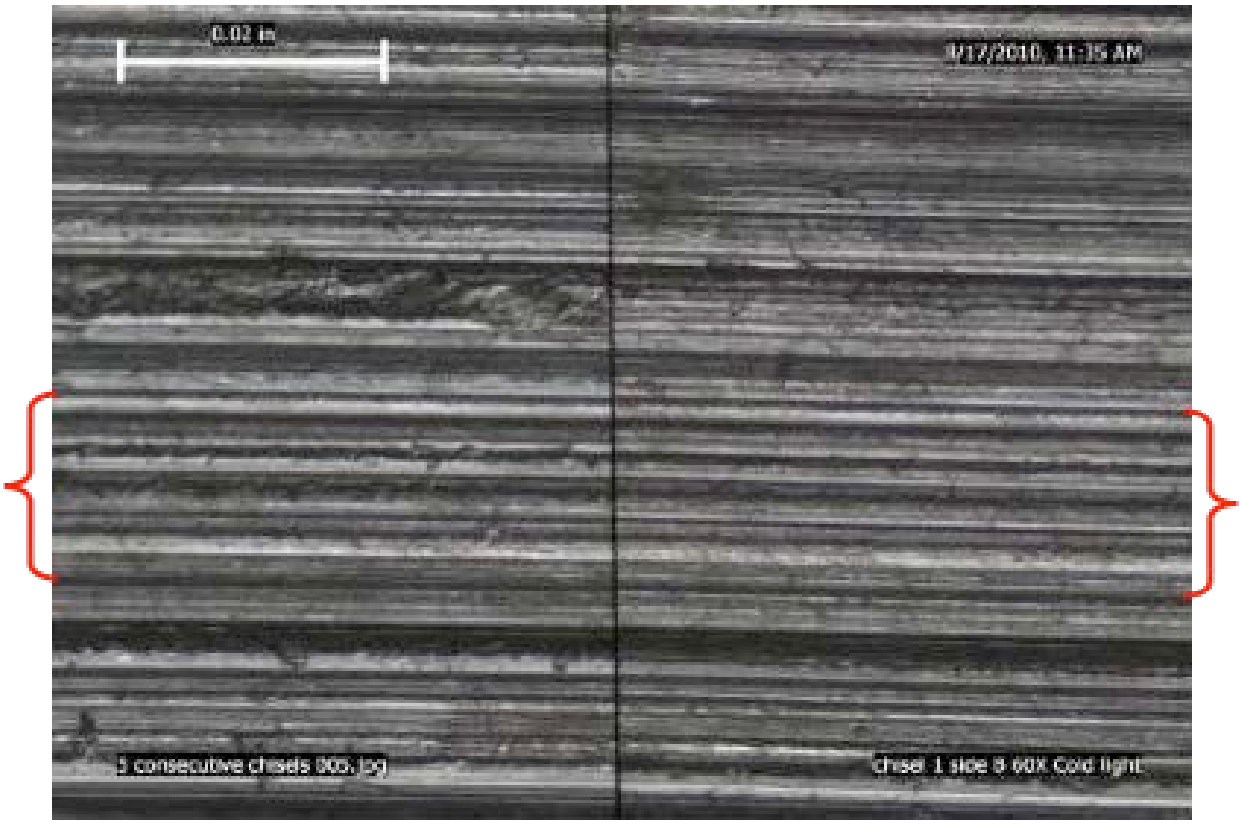


FIGURE 3. Images from a comparison microscope of a known-match (KM) with one group of consecutive matching striae marked. Photograph courtesy of Gerard Petillo.

Bunch (2000) explained that the Biasotti-style CMS counting method was testable and "inherently more scientific than the subjective regime currently used by the vast majority of examiners and thus perhaps more likely to successfully pass as a scientific theory or technique at a Daubert hearing" (p. 958). Bunch (2000) went on to delineate some criticisms of counting CMS, including the criticism that counting striations is subjective. However, he described that:

> "With consistent, national training, individual judgments on the quality and quantity of striations should converge; but they will never be unanimous. This simply means that examiners would sometimes report different LRs (likelihood ratios) for the same evidence bullet. This is not so bad as it might first appear. It is merely the analogue of examiners, using traditional methods, drawing two different conclusions about the same bullet: identification or inconclusive. Differing LRs simply reflect the fact that even objective regimes can contain subjective elements." (p. 959)

Bunch (2000) referred to the CMS model as a probability model, not an identification model, and the problem with this probability model is its inability to deal rigorously with barrel changes. While research bullets are oftentimes fired in new, clean barrels, questioned bullets retrieved from crime scenes are not. Gun barrels change over time, affecting the striation patterns on fired bullets.

Nichols (2003) countered Bunch's criticism that the CMS model is a probability model and explained CMS is merely a method to determine the minimum number of matching lines to conclude a match. The data from a CMS model is better suited for statistical analysis because numbers are actually generated, as opposed to the traditional pattern matching method. Nichols (2003) also argued that "an examiner who utilizes the CMS regime can rely on numerous studies that have been performed to show that the criterion for identification is supposed by the work of others and is not based solely in his or her own training and experience" (p. 304). In essence, there have been other studies conducted on CMS that shows that the CMS is empirically valid.

2.5 Methods and techniques of toolmark examination

Geradts, Keijer, and Keereweer (1994) created a database for toolmarks (TRAX) with video-images and data about toolmarks (width of toolmark, type of tool, microscope magnification, etc.). A video camera on a comparison microscope is connected to a computer,

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

which is used to scan the striation patterns and digitize the image. They developed an algorithm for the automatic comparison of digitized striation patterns. A comparison screen in TRAX makes it possible to compare images of toolmarks. The system was tested with ten screwdrivers of the same brand and all striation marks were identified with the correct screwdriver.

Tontarski and Thompson (1998) provided a technical overview of the Integrated Ballistic Identification System (IBIS) image acquisition hardware, image storage, case data input, ‘‘surface signature’’ analysis, and correlation scoring to an image database.

> “The IBIS standardizes a number of the steps that normally consume a firearms examiner’s time. Specimens are automatically kept in focus by the laser diode system, lighting is fixed and optimized to view bullet striations, and the computer/image capture system consistently (and tirelessly) compares the bullets’ images. In a similar manner, IBIS aids the user in cartridge casing image acquisition by automatically determining the margins of firing pin and breech face impressions on the cartridge casing primer, by gauging the lighting for more consistent images, and has precise magnification settings for an additional measure of consistency of images in the database. The system can be run by a technician, freeing the examiner for more complex and skilled tasks.” (p. 642)

Images were digitally captured on the Data Acquisition Station (DAS) and the Systems Analysis Station (SAS) derives a mathematical “signature” based on characteristics of the captured image. These signatures are entered into a database where they are correlated and compared, resulting in a “candidate list.” After reviewing the candidate list, the operator selects the indicated potential matches for visual comparison. The initial concern of the authors was whether different examiners could enter the images consistently for the database to locate a match. However, the system (image capturing and algorithm matching) eliminated operator variability. Furthermore, the modified microscope’s features reduce the potential for operator error. In summary, Tontarski and Thompson (1998) determined that IBIS was easy to operate and capable of capturing consistent, high-quality images, which could be shared and compared with other laboratories. While the system is excellent screening tool, it is not a substitute for experienced firearms examiners.

A system known as BulletTRAX-3D™ aids forensic firearms examiners in the comparison process. This system uses three-dimensional sensory technology, allowing operators to capture

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

2D digital images and to create 3D topographic models of the bullet's surface area. Roberge and Beauchamp (2006) decided to apply the Evan Thompson's test to BulletTRAX-3D and determine if the system was able to correctly match each numbered pair to a unique lettered pair. The Evan Thompson's test, named for a firearms examiner from the Washington State Police Crime Laboratory, involves the comparison of twenty-one pairs of 9mm Luger Hi-Point bullets fired from ten consecutively manufactured Hi-Point barrels. In the Roberge and Beauchamp paper, all pairs of bullets in the test were imaged with BulletTRAX-3D, which computed a score that quantifies the similarity of standard and test bullets. BulletTRAX-3D was able to accurately match each of the numbered and lettered pairs, showing that the system could reproduce what firearms examiners would do manually.

Brinck (2008) attempted to determine whether newer 3D imaging technology was better than 2D technology by evaluating the abilities of IBIS and BulletTRAX-3D. In his experiment, bullets from ten consecutively manufactured barrels were fired into a water recovery tank. One pair of copper-jacketed bullets and one pair of lead bullets were selected from those generated and uploaded into IBIS and BulletTRAX-3D by the same operator. Brinck (2008) concluded that, although IBIS is an effective tool for the identification of copper-jacketed bullets, BulletTRAX-3D was better at identifying all bullet types tested (copper-jacketed, lead, and inter-composition bullets).

De Kinder and Bonfanti (1999) developed a system capable of performing automated comparisons between striation marks on bullets, using laser profilometry, a non-contact laser scanning technique that records the topography of a bullet. The system was able to obtain a one-dimensional array of characteristics out of the recorded data (a feature vector) and compare it to similar quantities from other bullets using a correlation technique. Bachrach (2002) discussed the development of SciClops, an automated microscope comparison system using a 3D characterization of a bullet's surface. Preliminary tests were conducted to evaluate the ability of the system to identify and distinguish bullets. It was determined that it was possible to acquire reliable characterizations of a bullet's surface, to accurately identify similarities between bullets fired by the same gun, and to accurately discriminate between bullets fired by different guns. In Banno, Masuda, and Ikeuchi's study (2004), they presented an algorithm for a shape comparison of impressions on bullets using 3D shape data. A confocal microscope was used to obtain 3D data of striated surfaces and to visualize virtual impressions. Then they aligned the 3D data to

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

compare the shapes of the striations by computing a distance between two surfaces for alignment.

Senin, Groppetti, Garofano, Fratini, and Pierni (2006) introduced a 3D virtual comparison microscope to compare two specimens through their virtual 3D reconstructions. The authors determined that systems based on 3D surface topography can aid in the visual comparison process, as well as in making quantitative measurement over shape data. Furthermore, algorithms were also used to generate artificially enhanced images. They concluded that visual enhancement tools and quantitative measurement of shape properties could help a firearm examiner in comparing toolmarks. Neel and Wells (2007) compared 4000 striated toolmarks and concluded that there was a statistically significant difference between known matches (KM) and known non-matches (KNM). In essence, with 3D toolmarks, KM and KNM could be statistically distinguished from one another.

Chu et al. (2010) looked at the land impressions of 48 bullets. The barrel of a firearm may be smooth or rifled. Almost all modern handguns and rifles have rifled barrels. A rifled barrel contains grooves in its inner surface. The raised area between the grooves is called the land (see Figure 4). The land and grooves together constitute the rifling. The lands dig into the bullet and cause it to rotate on its longitudinal axis as it passes through the barrel. This rotation gives the bullet stability in flight and prevents it from tumbling, similar to how a quarterback would spiral a football down a football field.

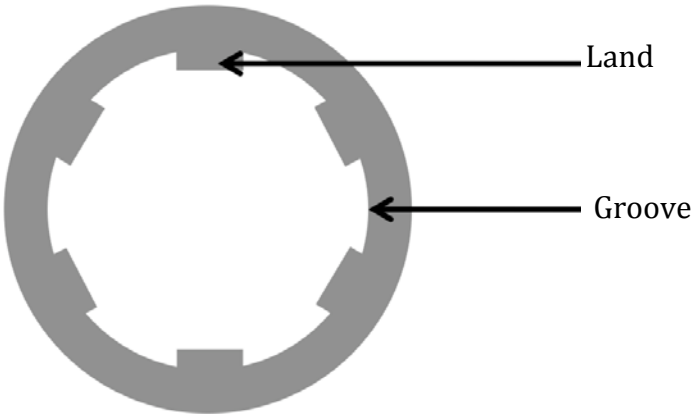


FIGURE 4. Land and Groove Rifling

Chu et al. (2010) estimated the width of lands for 48 bullets using confocal microscopy. In

their study, each barrel had six lands; as a result, 288 land engraved area (LEA) widths were calculated from each topography image. The 48 bullets were classified into different groups based on the width class characteristic for each LEA. Once the average profile is determined for each LEA image, cross-correlation values were computed between the LEAs of two bullets and a list of the best candidates is generated. For all 48 lists, the average number of correct matching bullets was about 9.3% higher than that obtained using current optical reflection systems. Furthermore, the error rate was about 24% smaller with confocal microscopy.

2.6 Reliability of toolmark examination

Prior to *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993), the *Frye* test (1923) determined the admissibility of scientific evidence. According to the *Frye* test (1923), a test or procedure is admissible in court if it is generally accepted in the particular field. However, the U.S. Supreme Court held in *Daubert* (1993) that the Federal Rule of Evidence 702 (2009) superseded Frye (1923). *Daubert* (1993), and its progeny, *Kumho Tire Company, Ltd. v. Carmichael* (1998) and *General Electric Company v. Joiner* (1997), serve as the criteria for expert witness testimony in courts. In *Daubert* (1993), the trial court serves as a "gatekeeper" of the evidence and must decide whether the proposed expert testimony meets the requirements of relevance and reliability. Rule 702 states that:

> "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case."

Under the *Daubert* test (1993), the court considers (1) whether the theory can be or has been tested, (2) whether the theory has been subjected to peer review or publication, (3) the theory's known or potential rate of error and whether there are standards that control its operation, and (4) the degree to which the relevant scientific community has accepted the theory. *Kumho Tire Company* (1998) applied the *Daubert* standard (1993) to expert testimony from non-scientists, while *General Electric Company* (1997) held that an abuse-of-discretion standard of review was

the proper standard for determining whether expert testimony should be admitted.

Collaborative Testing Service (CTS) developed a proficiency testing program, which has generated error rates for the field of firearms and toolmark examination. Peterson and Markham (1995) published a summary of the CTS proficiency tests, discussing the error rate in proficiency testing of firearm and toolmark examination. Peterson and Markham (1995) summarized twelve toolmark tests between 1980 and 1991. The tests included five toolmarks made by screwdrivers, two toolmarks each with bolt/wire cutters, a stapler, fingernail clipper, crimping tool, and die stamp. In seven tests, the examiners were provided with the test and evidence marks and asked if the test toolmarks were made by the same tool that made the evidence marks. In five tests, tools were provided with the toolmarks. Three of the tests in which a tool was provided, examiners were asked if it made one or more of the toolmarks provided. Of the 1,961 comparisons reported, 74% of comparisons correctly identified the tool, while 4% were incorrect, and 17% were inconclusive.

Grzybowski and Murdock (1998) assert that identification of striations is a science and admissible under *Daubert* (1993). First, based on knowledge of manufacturing processes, we are able to determine whether individual characteristics are present on tool working surfaces. Second, unique tool working surfaces leave reproducible and unique toolmarks. Third, the techniques employed in forensic identification can be used to associate toolmarks to the object that produced them. Grzybowski and Murdock (1998) advise that studies must be done in an attempt to falsify numerical criteria. According to scientific philosopher Karl Popper, a theory is considered scientific if and only if it is falsifiable (rather than verifiable). Furthermore, Grzybowski and Murdock (1998) explain that the purpose of the proficiency tests is to "directly test the proficiency of an individual analyst and to indirectly test the validity of a particular method and protocol" (p. 9). Examiners should be prepared to describe these proficiency tests, their strengths, and limitations in courts. Grzybowski and Murdock (1998) also discuss how there have been numerous articles published in the field of firearm and toolmark examination. "Submission to the scrutiny of the scientific community is a component of 'good science,' in part because it increases the likelihood that substantive flaws in methodology will be detected" (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993). Lastly, the authors explain that there are several cases dating back to 1929 in which firearm and toolmark examination has been accepted in the scientific community. In summary, Grzybowski and Murdock (1998) state

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

> "The firearms/toolmark [examination] field has all the indicia of a science: 1) It is well grounded in scientific method; 2) it is well accepted in the relevant scientific community; 3) it has been subjected to many forms of peer review and publication; 4) it has participated in proficiency testing and published error rates; and 5) it provides objective criteria that guide the identification process." (p. 11)

As a result, Grzybowski and Murdock (1998) contend that firearms and toolmark examination satisfies the admissibility requirements by Rule 702 (2009) and the *Daubert* test (1993).

Grzybowski, Miller, Moran, Nichols, and Thompson (2003) delve again into the reliability of toolmark examination. They contend that the AFTE Theory of Identification is empirically testable using the scientific method, has been scientifically tested, continues to be tested, has not been proven false, and is therefore scientifically valid; basically, the AFTE Theory of Identification satisfies the *Daubert* (1993) criteria. The authors assert that the error rate for firearm and toolmark examination is much smaller than the error rates reported for the CTS tests. Grzybowski et al. (2003) explain that if false eliminations were excluded in the Peterson and Markham (1995) calculations, the error rates for false identifications were 0.6% for firearms and 1.5% for toolmarks. Grzybowski et al. (2003) note that there are some limitations in using Peterson and Markham's (1995) data and advise examiners to be prepared to discuss the CTS tests and their limitations. The authors go on to describe how peer review allows other experts in the field to:

1. evaluate the validity of the hypothesis;
2. evaluate how it was formulated and tested;
3. evaluate whether the scientific method was followed;
4. evaluate whether proper conclusions were reached; and
5. encourages others to repeat the processes (replication) to further the science.

Furthermore, the authors explain that courts have accepted firearms and toolmark evidence for more than one hundred years and it has been the subject of numerous publications.

2.7 Court Decisions

Supreme Court decisions, including *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993), *Kumho Tire Company, Ltd. v. Carmichael* (1998), and *General Electric Company v. Joiner*

(1997) are making it increasingly necessary to further formalize scientific evidence presented in court. "Quantitative evidence regarding the validity of the basic premise of toolmark comparison would provide additional support for the admissibility of toolmark evidence" (Bachrach, Jain, Jung, and Koons, 2010, p. 348). The purpose of conducting these experiments is to apply the theories and techniques to criminal (or civil) cases. If the methods and techniques in toolmark examination are unreliable, they will not be accepted in a court of law. The following court cases illustrate different admissibility standards or requirements for toolmark examination.

The court in *United States of America v. Darryl Green* (2005) allowed the ballistics expert to testify as to his observations, but would not allow him to conclude that the match he found was the source of the cartridge cases to "the exclusion of all other guns." In essence, firearms evidence was admissible as an aid to the jury but the expert could not render an ultimate opinion because the field of firearms analysis was not sufficiently reliable.

In *United States of America v. Amando Montiero, Valdir Fernandes, Angelo Brandao, Brina Wurie, Luis Rodrigues, Manuel Lopes* (2006), the court stated that

> "The government must ensure that its proffered firearms identification testimony comports with the established standards in the field for peer review and documentation. If the expert opinion meets these standards, the expert may testify that the cartridge cases were fired from a particular firearm to a reasonable degree of ballistic certainty. However, the expert may not testify that there is a match to an exact statistical certainty."

In *United States of America v. Edgar Diaz, Rickey Rollins, Don Johnson, Robert Calloway, Dornell Ellis, Emile Fort, Christopher Byes, Paris Ragland, Ronnie Calloway, Allen Calloway, Terrell Jackson, and Redacted Defendant No. 1 (*2007), the court allowed testimony from the firearms experts, but ordered that "the experts may not testify to their conclusions 'to the exclusion of all other firearms in the world.' They may only testify that a particular bullet or cartridge case was fired from a particular firearm to a reasonable degree of certainty in the ballistics field." In *United States of America v. Chaz Glynn* (2008), the court held that the ballistics opinions offered at the *Glynn* retrial could be stated in terms of "more likely than not," but nothing more. In *United States of America v. Donald Scott Taylor* (2009), the court held that testimony from the firearms expert was admissible under Rule 702 (2009) and *Daubert* (1993), adopting the reasoning of the courts in *Green* (2005), *Monteiro* (2006), *Diaz* (2007), and *Glynn*

(2008).

> "Because of the limitations on the reliability of firearms identification evidence discussed above, [the firearms expert] will not be permitted to testify that his methodology allows him to reach this conclusion as a matter of scientific certainty. [The firearms expert] also will not be allowed to testify that he can conclude that there is a match to the exclusion, either practical or absolute, of all other guns. He may only testify that, in his opinion, the bullet came from the suspect rifle to within a reasonable degree of certainty in the firearms examination field." (U.S. v. Taylor, 2009)

From these cases, it is clear that firearms and toolmark examination is admissible under *Daubert (*1993). However, more statistical research has to be done so that we may testify to the certainty of the method and evidence at hand.

2.8 Statistics and Toolmarks

Since *Daubert* (1993), the explanation "I know a match when I see it" is no longer sufficient for identification. The goal in individualization is to state that a particular tool made the particular toolmark, to the exclusion of all other tools. Toolmark examination is often compared to DNA analysis, in which error rates and probabilities are known. However, establishing error rates and probabilities in the area of toolmarks is fundamentally different than in DNA analysis. With DNA analysis, all the variables and parameters of a DNA strand are known and error rates can be calculated with a high degree of accuracy. However, in toolmark examination, there are too many variables that examiners cannot control, such as force, angle, the motion of the tool, the incident surface material, the material used to produce the tool, the relative hardness of each, past use of the tool, etc. Much of this information is either not known or it cannot be determined. As a result, it may not be possible to calculate realistic error rates. However, experiments in the field provide a guide or estimate of the error rates in the field.

Some forensic scientists approach the application of probability and statistics to toolmarks from a Bayesian decision theoretic perspective (Taroni 1996). The odds form of Bayes' Theorem is shown below.

Posterior odds in favor of association given test indicates inclusion

$$\frac{\Pr(S^+ \mid t^+)}{\Pr(S^- \mid t^+)} = \frac{\Pr(t^+ \mid S^+)}{\Pr(t^+ \mid S^-)} \frac{\Pr(S^+)}{\Pr(S^-)}$$

Likelihood Ratio

Odds form of Bayes' Rule

From the "forensic Bayesian" point of view (cf. Taroni 2010) it is argued that forensic scientists should be concerned with the likelihood ratio (LR, or more generally Bayes factors in a genuinely Bayesian paradigm), whereas jurors should consider the posterior odds that the tool made the mark given evidence for an association (t^+). Forensic scientists consider two hypotheses: (1) that the tool made the mark (S^+), and (2) that the tool did not make the mark (S^-).

Champod, Baldwin, Taroni, and Buckleton (2003) discuss their Bayesian approach to firearms and toolmarks. The authors also examine the CMS regime from a statistical perspective. They explain that "the CMS approach will offer added-value under certain conditions: (1) that the model is an appropriate (although incomplete) description of the variability between impressions and (2) that the concept of consecutive striations is coherent among examiners and can be reproduced" (p. 314). Champod et al. (2003) defend the forensic Bayesian approach and argue that problems associated with toolmark examination are not flaws in the approach.

Taroni, Champod, and Margot (1996) explain that statistics and probabilities are an obligatory part of any science. Any measure has uncertainties due to the quality of the instruments used, the ability of an operator, the variance of the measured attribute, etc. Statistics is used to evaluate the ability to obtain reproducible results within a given error range. Furthermore, Taroni et al. (1996) explain that the role of statistics for the forensic scientist is limited to the assessment of the value of the likelihood ratio. The examiner should only state that the evidence supports x times the hypothesis that the screwdriver produced the mark. If the expert estimates that the probability of another match is almost zero, then it is logical to declare an identification. Furthermore, Taroni et al. (1996) argue that an experienced toolmark examiner will always achieve a more discriminative comparison than a statistical approach. The authors

conclude by stating that numerical data could help the scientist to demonstrate the scientific validity of the toolmark individualization process, and to assist the examiner in the elaboration of a conclusion.

Faden et al. (2007) developed a computer program to compare toolmarks made from forty-four consecutively manufactured screwdrivers on soft lead plates. A surface profilometer was used to make height, depth, and width measurements as a function of location on the two-dimensional sample surfaces. Four marks were produced using both sides of each tool at three different angles (30°, 60°, and 85°). Pearson correlation was used to compare toolmarks involving true matches, true nonmatches, and marks made from different sides of the same tool all produced high correlation values. The results suggest that the Pearson correlation alone is not effective at determining when there is an actual match. However, there was a significant separation in correlation values between true match and true nonmatch toolmarks produced at the same angle. Although this suggests that it may be possible to identify true matches using a computer algorithm, true match and true nonmatch toolmarks were differentiated effectively only when the toolmark was produced at the same angle. Furthermore, toolmarks made from different sides of the same screwdriver tip produced separation in data and are similar to data from true nonmatches. This supports the hypothesis that different sides of a screwdriver act as different tools when producing toolmarks.

Chumbley et al. (2010) extended the Faden et al. (2007) study by comparing the effectiveness of an algorithm to human examiners. The algorithm they used first optimized the dataset, in which it identifies a region of best agreement between the two toolmark datasets being compared. Next, the algorithm validated the dataset, in which the certain corresponding areas in the region of best fit (on both toolmarks) are compared and a correlation value is calculated. If a match exists at one point along the scan length (Optimization), there should be large correlations between corresponding areas along their entire length (Validation). The authors then conducted a double-blind study in which fifty experienced toolmark examiners gave their opinions on the sample set. In the end, the authors determined that examiner performance was much better than the algorithm, but the deficiencies could now be addressed and improved upon.

In 2008, Howitt, Tulleners, Cebra, and Chen recommended formulae to answer the need for a theoretical foundation for the identification of bullets from the striae that appear on them. Attempts were made to calculate the probability “for the correspondence of the impression marks

on a subject bullet to a random distribution of a similar number of impression marks on a suspect bullet of the same type" (p. 868). Based on the measurements, it was concluded that likelihood ratios of finding a "match" by chance are possible to be estimated.

Bachrach, Jain, Jung, and Koons (2010) compared striated toolmarks from screwdrivers and tongue and groove pliers using confocal microscopy. They considered the effect of changing the substrate onto which the toolmarks were created, as well as the angle of incidence for creating the toolmark. Bachrach et al. (2010) sought to validate the basic premise of toolmark examination, namely that toolmarks exhibit a high degree of individuality. Algorithms were developed to generate toolmark signatures, while metrics were used to assess the degree of similarity between known matching and nonmatching toolmark pairs. From these similarity values, the authors determined that it was possible to evaluate "the degree to which toolmarks created by the same tool are repeatable and distinguishable from toolmarks created by other tools" (p. 349). They concluded that: (1) the striated toolmarks produced on the same medium and under the same conditions were both repeatable and specific enough to allow for reliable identification of the producing tool; (2) striated toolmarks created on different media but under the same conditions could still be identified with high reliability; (3) screwdriver striated marks depend more on the angle at which the toolmark is created than the media; (4) the probability of a pair of different tools having similar features is extremely low; and (5) the probability of error from a faulty image, not because of the tool itself, would not create repeatable and individual toolmarks. As a result, given the low probabilities of error associated with these cases, a bad toolmark image can have a significant effect.

3. Rationale for the research

Over the last several decades, forensic tool mark and ballistic examiners have struggled with the fact that, while there is accepted methodology for the qualitative comparison of questioned tool marks, firearm and other forensic impressions, there is no accepted methodology to generate numerical proof that independently corroborates morphological conclusions. In light of critics' recent charges that firearms and tool mark examination is "un-scientific" as currently practiced, this numerical corroboration issue for source association has come to center stage and must be addressed.

This study addresses the need for establishing a sound objective scientific basis for impression evidence comparisons. Recent studies have used state-of-the-art technologies to objectify the pattern information in impressions (Chu 2010, Cork, 2008; Neel & Wells 2007; Banno, 2004; Leon, 2006; Bachrach, 2002; Geradts 2001; Senin 2006; Faden 2007; DeKinder & Bonfanti, 1999; Song 2006). Still, however much work needs to be done, as has been recently recognized by the National Academy (2009). In addition, the preeminent system for automated toolmark comparisons, IBIS, only works for firearms and is proprietary. Exactly how the IBIS functions is a closely guarded business secret (Cork 2008; Forensic Technology 2001). This second point is a critical issue as concerns the *Daubert* test. If an automated toolmark comparison system is to output estimates of matching probabilities for use in court, all of its internal algorithms should be subject to peer review.

This study focuses on striation patterns left by tools and on cartridge casings imparted by firearms. All impressions made by tools and firearms can be viewed as mathematical patterns composed of features. In order to recognize variations in these impression patterns, we used the mathematics of multivariate statistical analysis. In a computational pattern recognition context, this process is called machine learning. The mathematical details of machine learning can give what Moran calls "…the quantitative difference between an identification and non-identification" (Moran 2002). They also enable the estimation of extrapolated identification error rates and even in some cases, the calculation of rigorous, universal random match probabilities (Duda 2001; Fukunaga, 1990; Theodoridis 2006; Kennedy 2003; Kennedy 2005).

The overarching aim of this research is to lay down, test and publish multivariate statistical foundations for tool mark impression pattern recognition and comparison. In order to realize this overarching goal, the project is divided into three main initiatives:

1. Tool mark pattern collection and archiving
2. Database and web interface construction for the distribution of tool mark data, and software developed for this project.
3. Identification/exploitation of multivariate machine learning methods relevant to the analysis of collected toolmarks, striation patterns in particular.

II. Materials and Methods

1. Materials

9-mm Glock Cartridge Cases

Nine-millimeter cartridges were fired from different models of Glock pistols and the cases collected for analysis. There were thirty-seven different Glocks used to record data for the database. The following number of cartridge cases collected from each firearm:

- 23 cartridge cases from Glock 1
- 11 cartridge cases from Glock 2
- 12 cartridge cases from Glock 3 through 9
- 3 cases from Glocks 10 through 23
- 2 cases from Glock 24 through 37

for a total of 186 cartridge cases from 37 firearms. Since the shear marks were not always created exactly normal to the surface of the cartridge case, the cases were mounted on a goniometer during the microscopical analysis to reduce as much tilt as possible, keeping the scanned volume (required confocal stack) to a minimum. A quick pre-scan with the 10x objective allowed evaluation and accommodation for this natural tilt.

Screwdrivers

Eight Craftsman® brand screwdrivers and ten Iron Bridge® brand screwdrivers (Figure 5) were obtained as exemplars.

FIGURE 5. Ten Iron Bridge slotted head screwdrivers

The medium used in these studies was lead because, as explained previously, lead is soft enough to make test marks without damaging the tool's working surface and provides less noisy surface data. Each lead exemplar was engraved with the appropriate label with a Dremel engraving tool (Figure 6).

FIGURE 6. Lead exemplar from Screwdriver #37, Side B-1

Five exemplar striated toolmarks were made with each side of each screwdriver (i.e. a total of 180 striation patterns).

Chisels

Five consecutively manufactured chisels (Mayhew® brand) were obtained from the reference collection of Gerard Petillo (independent firearm/tool mark examiner). Five exemplar striated toolmarks were made with each side of each chisel.

2. Methods for toolmark impression data collection and database construction

2.1 Generating reproducible toolmark impressions

Figure 7 shows the holding jig that was constructed for generating reproducible striation patterns, specifically the screwdriver toolmarks.

FIGURE 7. Tool holding jig for generating striation patterns on any media.

The jig gives the examiner good control over lateral and rotational angles with which the tool makes contact with the impression media. Also, the jig is built from components available at any hardware store. Thus it is a low cost piece of equipment available to any tool mark examiner.

The jig was set to a consistent angle of 15˚ for comparison purposes, and in each case the screwdriver was pulled toward the jig operator in order to make the impression for consistency purposes. Note that the same angle of attack was used in the screwdriver study of Bachrach et al. (Bachrach, Jain, Jung, & Koons, 2010).

2.2 Confocal Microscope

A Zeiss Axio CSM 700 confocal microscope was used to analyze the toolmarks produced for this study (see Figure 8).

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

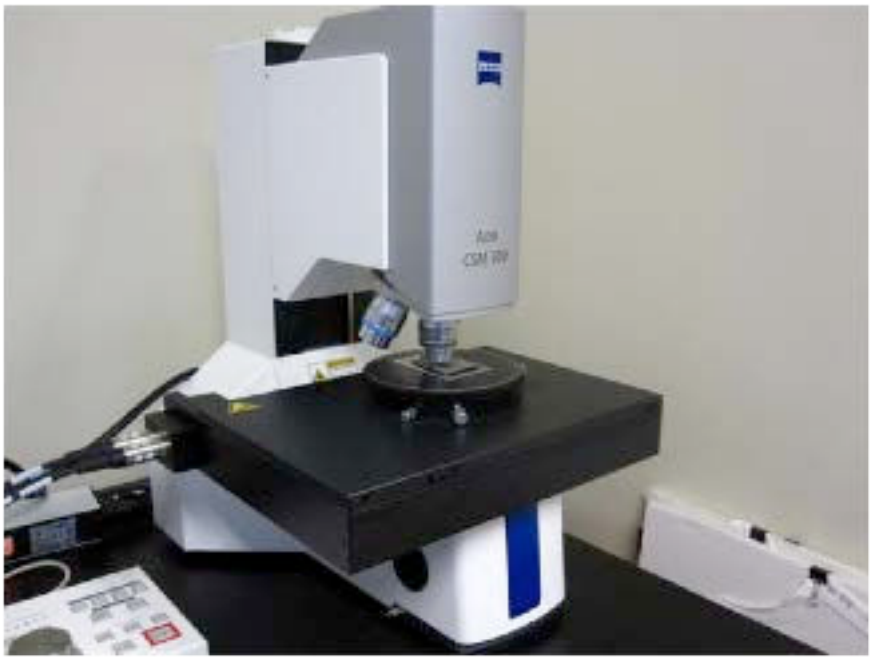


FIGURE 8. Zeiss Axio CSM 700 Confocal Microscope. Photograph courtesy of Peter Diaczuk.

Confocal Microscopy is an imaging technique that allows quantitative observation of surface microstructure details and the reconstruction three-dimensional surface topographies. The important aspect of confocal microscopy is the use of spatial filtering to eliminate out-of-focus light in samples that are thicker than the depth of focus (see Figure 9).

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

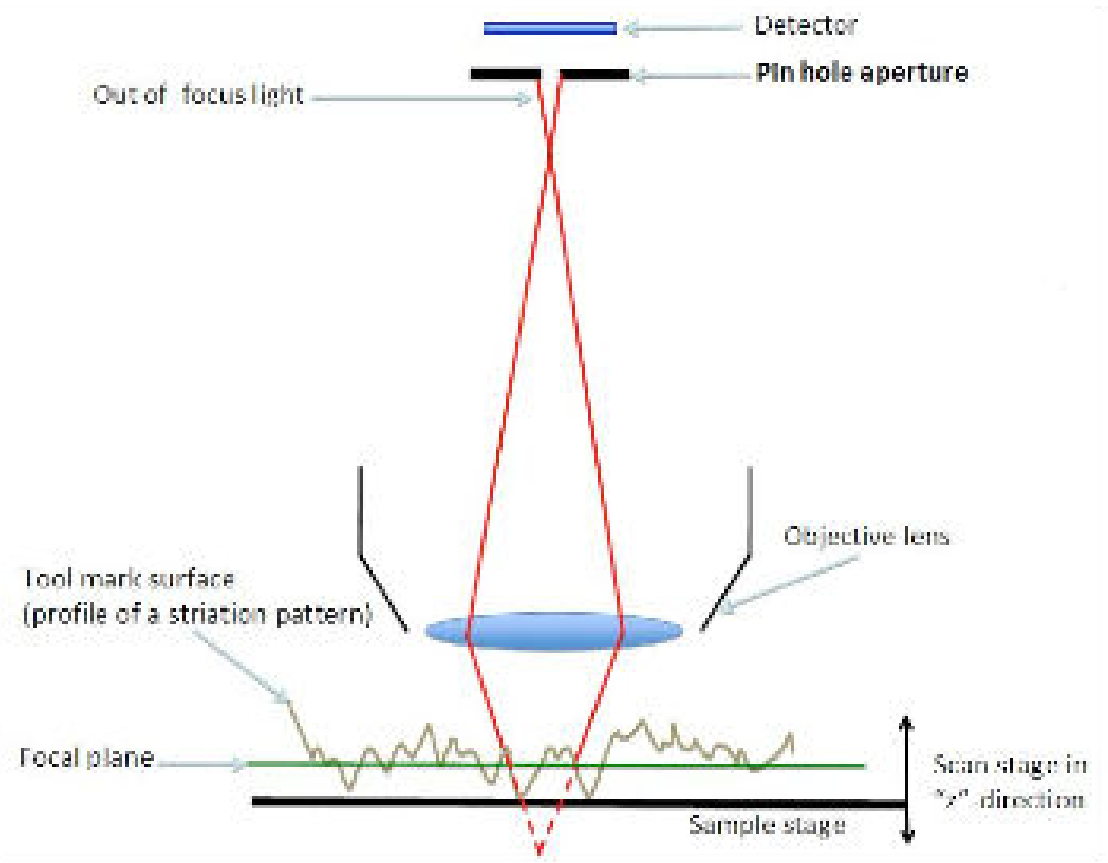


FIGURE 9. Confocal microscope eliminates out-of-focus light.

Semwogerere and Weeks (2005) explains that a confocal microscope involves point-by-point illumination of a specimen, and exclusion of out-of- focus light from the sample. The microscope for this project used white light to illuminate the sample as a series of lines through the objective lens via an epic-illumination format. While the design is somewhat unorthodox and proprietary to Zeiss/Lasertec, the net effect is the same as standard surface scanning confocal microscopy using a pinhole aperture coupled with a Nipkow disk. The light reflected from the surface of the sample passes back through the objective lens and is collected by a tube lens. One point (or a strip of points, depending on design) of the sample is observed at each moment, which increases contrast and improves the resolution of the image. (See Figure 10 for a basic schematic of how confocal microscopy works).

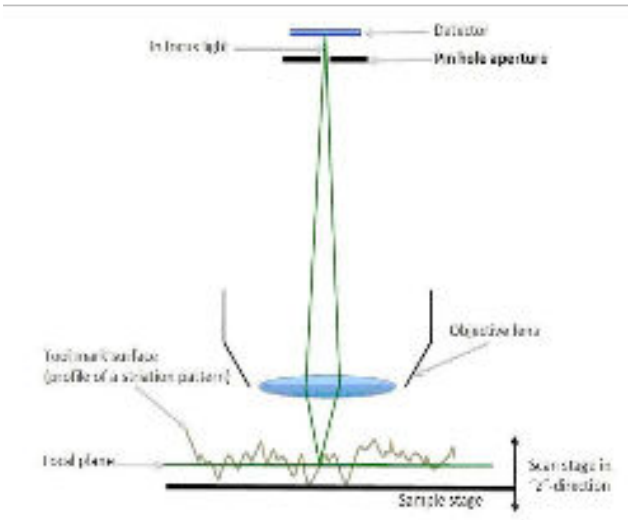


FIGURE 10. Schematic of how confocal microscopy works.

The CCD detector collects the in focus photons traveling back from the illuminated surface and software puts together the in focus cross-sections to create an all-in-focus two-dimensional image (all-in-focus meaning literally that the image is focused in all areas) (see Figure 11) or three-dimensional digital representation of the physical surface (see Figure 12).

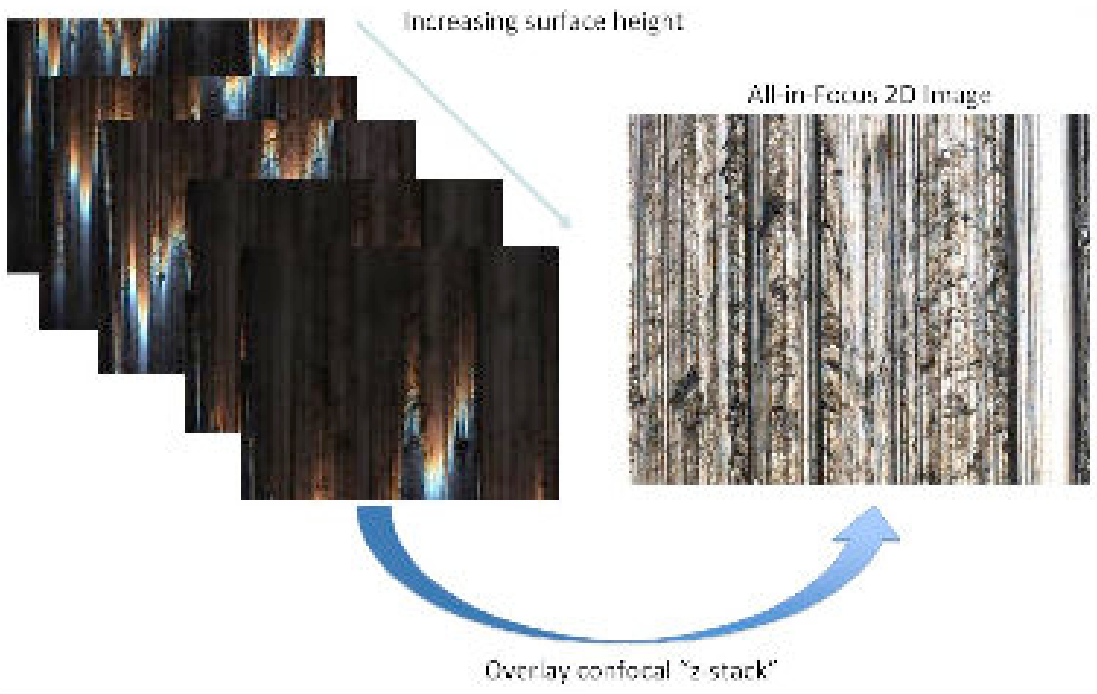


FIGURE 11. Confocal microscope stacks the images to produce an all-in-focus 2D image

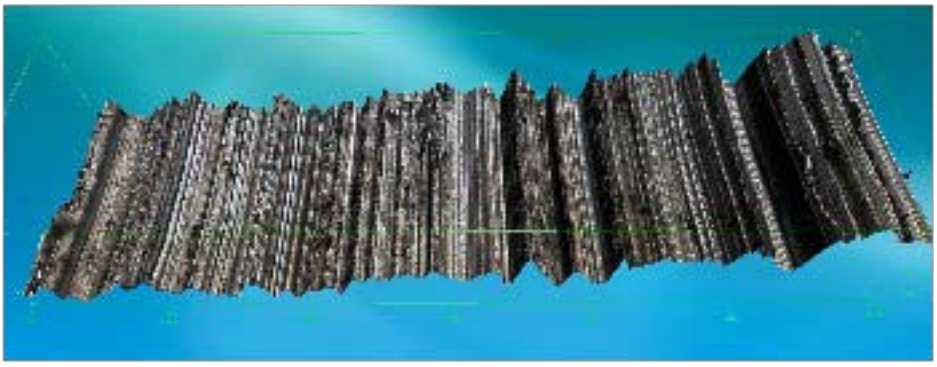


FIGURE 12. Three-dimensional image of a striated toolmark

The Zeiss Axio CSM 700 confocal microscope produces two types of images: an F image, which is the all-in-focus image (see Figure 13), and a Z image, which shows the height information of the image in varying gray levels (see Figure 14). The F images are captured from focus scan memory. The Z image is a sequence of optical sections collected at different levels perpendicular to the optical axis (the z-axis) within a sample. The Z image is captured with distance calibration data and height calibration data and expresses information on height. This image is used for measurement items that include height data, such as 3D and surface roughness measurements.

FIGURE 13. F image of a striated toolmark.

FIGURE 14. Z image of a striated toolmark.

There may be certain areas on a sample in which the white light that bounces off the sample does not make it back to the detector. The microscope's software interprets these areas as outliers (steep spikes) and dropouts (steep dips). In order to deal with these inevitable artifacts, the

microscope's software was used to threshold and locally interpolate through these areas (i.e. denoise the surface). The noise-cut method used for all the toolmarks in this experiment was Z-interpolation. In this procedure, noise spikes are removed by interpolating, or estimating, pixels based on the whole image. It is preferable to denoise the image because the toolmarks are examined on a micrometer level; the noise spikes alter the image and skew the statistics performed. (See Figures 15–18).

FIGURE 15. Original 3D image of a striated toolmark.

FIGURE 16. Denoised 3D image of a striated toolmark.

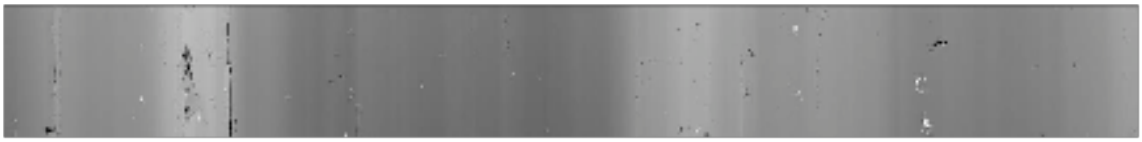


FIGURE 17. Original Z image of a striated toolmark.

FIGURE 18. Denoised Z image of a striated toolmark.

While this software is necessary (for *all forms* of 3D light microscopy, not just confocal), reliable and affects only a few points on the surface, unfortunately it is proprietary. Ultimately if 3D microscopy is going to be widely used in casework, the denoising software needs to be standardized so that all parties involved denoise their data sets in the same way.

3. Methods for direct feature vector comparison

3. 1 General striated toolmark surface preprocessing and feature vector construction

Due to gross surface warping during the toolmark formation process, *all* recorded striation patterns required form removal. Third order polynomial surface fits were used for form removals from all recorded striated surfaces. This degree polynomial was chosen because it was observed to have a minimal set of degrees of freedom to remove a majority of gross surface warp across all striated surfaces examined. An example of form removal is shown in Figure 19 with a recorded striation pattern before and after form removal.

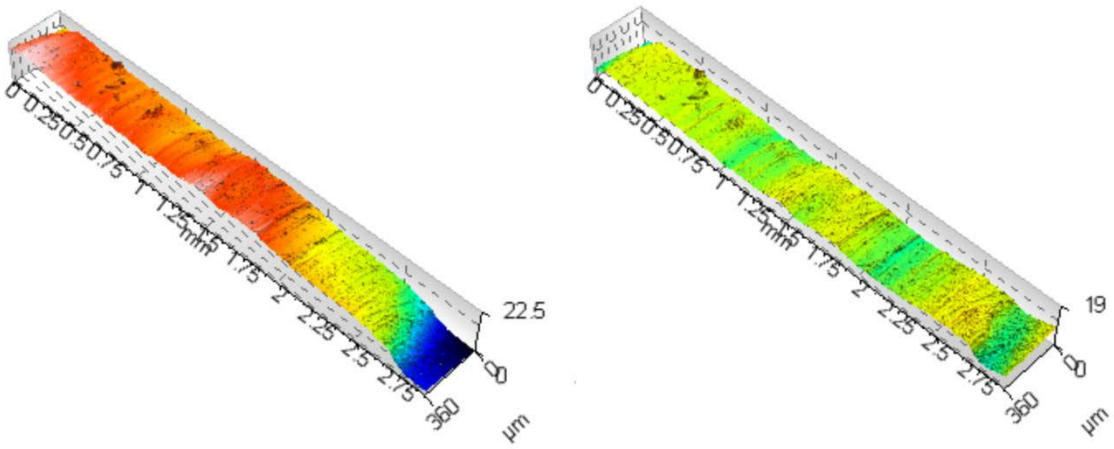


FIGURE 19. Unprocessed and processed (form removed) primer shear striation pattern from Glock #3, cartridge case 2.

The resulting form removed striation patterns were filtered into roughness and waviness components. See studies below for specific standard filters and cutoff values used.

The mean profiles of each waviness component were then computed due to the high redundancy of information found in the surface. Also, following the literature, the authors feel it is current "standard practice" to use a profile (usually the mean profile) as input into the statistical discrimination algorithms instead of the entire surface (Bachrach 2002, Chu 2010, Bachrach et al. 2010, Chumbley et al. 2010, Faden et al. 2008). It is the mean profile of the waviness surface which formed the feature vector of all the surfaces examined in this project. Note however, users of our software are not restricted to mean profiles of waviness surfaces. Median or random profiles of any surface (unfiltered, waviness or roughness) can be used as feature vectors.

Because each profile did not begin and end at the same points (see Figure 20), the profiles required alignment (*i.e.* registration) in order to be processed as multivariate feature vectors.

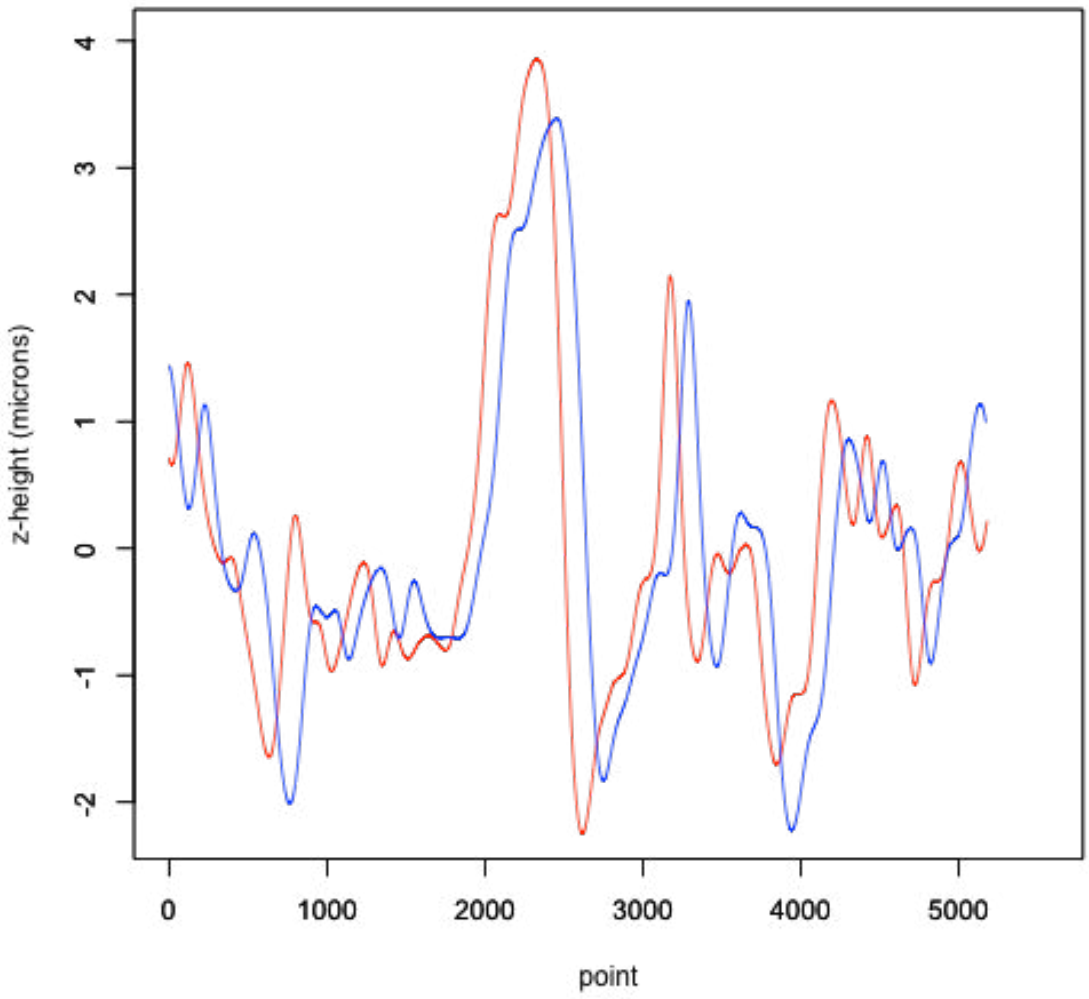


FIGURE 20. Mean waviness profiles (across section of average striation pattern) from two different cartridges fired from Glock #2.

In order to register profiles from the same experimental unit (e.g. a Glock or a screwdriver), the cross-correlation function (CCF) between two profiles from each group was computed to find the shift that yielded maximum correlation (a linear, univariate measure of similarity) (Muralikrishnan & Raja, 2009; Chu et al., 2010). That is, the lag where the maximum of the

cross-correlation function occurs tells how much to shift one profile over another so that they are maximally aligned in a "correlation" sense (see Figure 21).

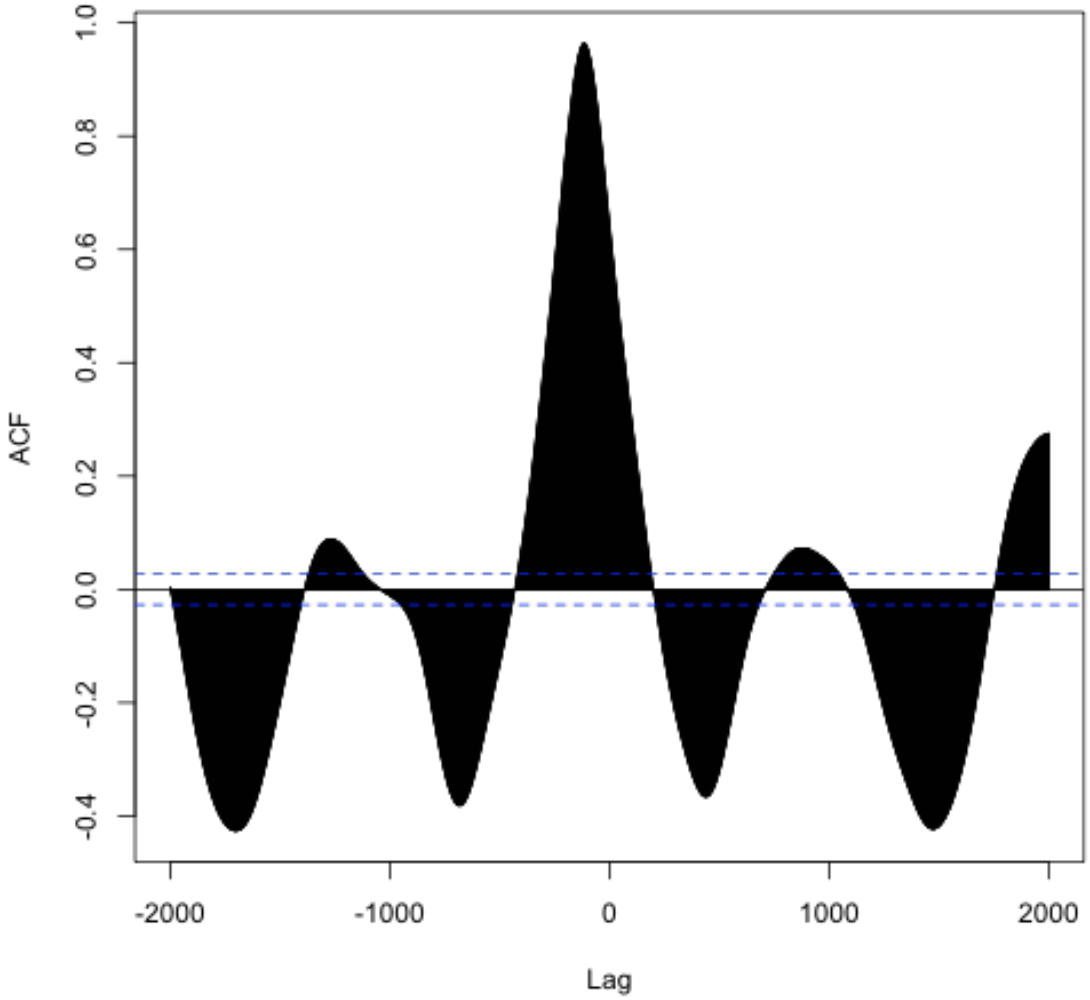


FIGURE 21. Cross-correlation function between the profiles shown in Figure 20. The graph indicates that shifting one profile backward with respect to the other by 57 units (max at lag=-57) will best align them.

Within a group of experimental units, the longest profile is chosen as a reference or "anchor profile". The remaining profiles are then maximally aligned with respect to the anchor profile. An example is shown below for Figure 22 profiles from Glock #2.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

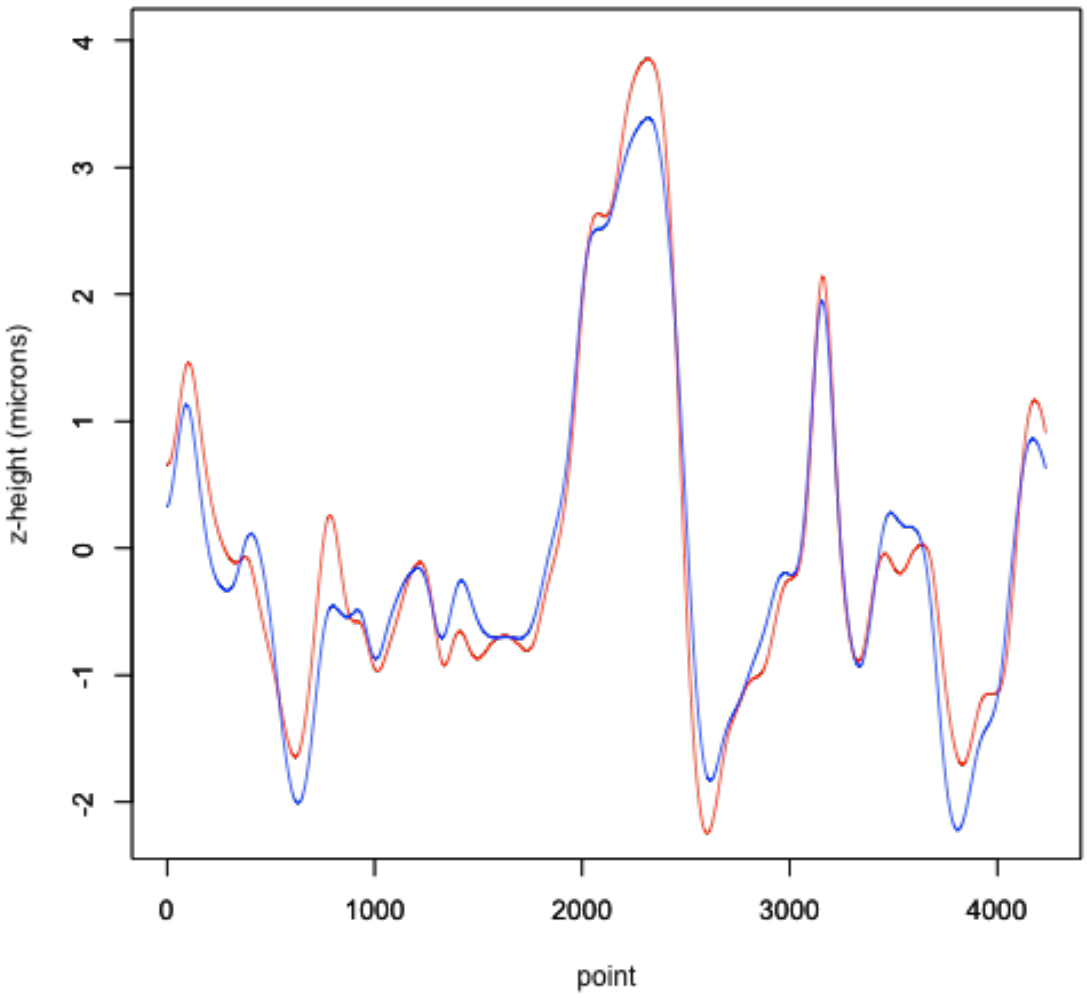


FIGURE 22. Aligned mean waviness profiles from cartridges from Glock #2.

After profiles within each experimental unit were registered, profiles between experimental units were aligned. This was done by first computing a group-mean-profile (GMP) for each of the within-group aligned profile sets. The GMP for each experimental unit served as a representation for that unit. The GMPs were then registered with respect to each other within a user defined "uncertainty window". The reason an uncertainty window was required for between group registration was that in general, no very well defined reference land mark is available for any given type of profile. Instead, knowing that all surface exemplars for a given study (e.g. Glocks, screwdrivers) were roughly recorded to within, 300μm +/-100μm from a left edge area on each striation pattern, the GMPs were aligned within this +/-100μm uncertainty window (users of the software can adjust the window size). The shift parameters produced by the registration of

group-means were used to shift all the mean profiles of the groups in blocks. That is, each group of mean profiles was shifted by the amount required to register the GMPs.

Finally, all profiles used in an analysis were rescaled such that the lowest profile point was designated 0 and the highest 1. This was done in order to minimize discrimination between experimental units due only to valley depth and peak height variation. Valley depth and peak height variation can be due solely to pressure variations in toolmark formation. Generally, this should not be information that is used in toolmark discrimination. These features can depend on the slight variations in the toolmark generating process, not the tool.

Note that, users of the software developed for this project need not normalize their data in this way. In fact, profile data can be scaled in any way the user chooses. Note also that scaling must be the same throughout a study otherwise results will not be comparable between experimental units (i.e. tools).

3.2 The Data Matrix and Principal Component Analysis

Profiles for a given study were arranged into an $n \times p$ data matrix ($\mathbf{X}$):

$$\mathbf{X} = \begin{bmatrix} X_{11} & .. & X_{1j} & .. & X_{1p} \\ : & & : & & : \\ X_{i1} & .. & X_{ij} & .. & X_{ip} \\ : & & : & & : \\ X_{n1} & .. & X_{nj} & .. & X_{np} \end{bmatrix}$$

where n is the number of profiles and p is the number of points in each profile. Each X_{ij} represents scaled z-height j in striation pattern profile i. At this point in the analysis, neighboring points in the profiles contain a great deal of redundant information. That is, proximal points in a profile are correlated. One way to capture much of the essential information within profiles is through principal component analysis (PCA) (Jolliffe 2004). PCA measures information in a data set via variance. It is generally used to reduce redundant information in a data set ($\mathbf{X}$) by taking linear combinations of the original variables to form a new set of "derived variables" ($\mathbf{Z}_{PC}$)

$$Z_{ij} = \sum_{l=1}^{p} a_{il} X_{il} .$$

In matrix form this is

$$\mathbf{Z}_{\mathrm{PC}} = \mathbf{X}\mathbf{A}_{\mathrm{PC}}^{\mathrm{T}}$$

where the superscript T indicates the transpose of $\mathbf{A}_{\mathrm{PC}}$. This transformation simply rotates the coordinate axes in feature space and the above equation is a transformation of the data (**X**) into the basis of principal components. The entire set of derived variables is equivalent to the original data (**X**). The new data set (**Z**), however orders the variables (columns) according to the amount of variance of the data set they contain, from highest to lowest. If the first few variables in **Z** contain a majority of the variance, then the remaining variables can be deleted with a minimum loss of information contained in the data. The dimensionality of the data set is then effectively reduced to include only those variables that adequately represent the data.

The matrix $\mathbf{A}_{\mathrm{PC}}$ contains the $p' < p$ principal components (depending on the study) as rows and is computed by diagonalizing the $p \times p$ maximum likelihood covariance matrix (**S**) of **X**

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}}) \otimes (\mathbf{X}_i - \bar{\mathbf{X}}) .$$

where $\otimes$ is the Kronecker product of vectors. The ratio of eigenvalues

$$\lambda_i \Big/ \sum_{j=1}^{p} \lambda_j$$

gives the proportion of variance explained by the i^{th} principal component and is useful in selecting the number of principal components required to adequately represent the data. Hold-one-out cross validation (HOO-CV, see below) was used to select a small set of PCs adequate to obtain good tool correct classification rates while minimally risking over-fitting the discrimination model.

3.3 Canonical variate analysis

Canonical variate analysis (CVA, also called Fisher discriminant analysis, linear Fisher discriminant analysis and linear discriminant analysis) seeks to characterize the ratio of between group variance (**B**) to within group variance (**W**) (Rencher 2002). Unlike PCA, canonical variate analysis requires that different samples of toolmark patterns are labeled with their identity and fed into the algorithm as groups. These a priori labeled toolmark pattern samples serve as a

training set in order to compute the canonical variates (CVs). Geometrically, the CVs define axes onto which the data is projected that best separates the samples into discrete clusters (46, 69). In *p*-dimensional space, min(*p*, *k*-1) canonical variates can be computed. However, CVA can also be used to reduce the dimensionality of the data by retaining only the first few CVs. Additionally, like PCA, CVA can be formulated as an eigenvector-eigenvalue problem with the magnitude of the eigenvalues providing a guide as to the number of CVs to be retained. The CVs, $\mathbf{A}_{cv}$ and their eigenvalues $\mathbf{\Lambda}_{cv}$ are computed by diagonalizing the matrix $\mathbf{W}^{-1}\mathbf{B}$ with

$$\mathbf{B} = \sum_{i=1}^{k} n_i (\bar{\mathbf{X}}_i - \bar{\mathbf{X}}) \otimes (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})^{\mathrm{T}}$$

and

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{X}_{i,j} - \bar{\mathbf{X}}_i) \otimes (\mathbf{X}_{i,j} - \bar{\mathbf{X}}_i)^{\mathrm{T}} .$$

A standard inversion method is used to invert non-singular $\mathbf{W}$. $\mathbf{X}_{ij}$ represents the j^{th} toolmark pattern in the i^{th} sample and $\bar{\mathbf{X}}_i$ is the average of all the toolmark patterns in the i^{th} sample. Note that there are n_i toolmark patterns in sample i. Because the eigenproblem for CVA

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{A}_{cv}^{\mathrm{T}} = \mathbf{A}_{cv}^{\mathrm{T}}\mathbf{\Lambda}_{cv}$$

is not symmetric, its eigenvectors are not guaranteed to be orthogonal. Thus unlike in PCA the CV are not necessarily at right angles to each other. The data is then transformed to the basis of (retained) CVs as

$$\mathbf{Z} = \mathbf{X}\mathbf{A}^{\mathrm{T}} .$$

3.4 Support Vector Machines

PCA itself does not classify objects into groups. In order to carry out a discrimination task PCA must be combined with a method to "learn" classification rules. Statistical learning theory and its practical application, the support vector machine (SVM) is just such a method and was developed in response to the need for reliable statistical discriminations within small to medium sample size studies (Vapnik). SVMs seek to determine efficient classification rules for

objects assuming nothing about the form of the underlying probably distribution generating the data. *This is a great advantage for application in forensic science. The fewer the decision algorithm's underlying assumptions, the less vulnerable its conclusions are to attack in court*.

Using a regularized Lagrange multiplier scheme, the SVM algorithm determines linear decision rules in a "warped" Hilbert space (kernel space) with maximum possible margins for error (Scholkopf 2002):

$$\max_{\lambda} \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to

$$\sum_{i=1}^{n} \lambda_i y_i = 0 \quad \sum_{i=1}^{n} \lambda_i = 1$$

and $0 \le \lambda_i \le C$ for all i. Given sample patters of toolmarks made by two different tools, this maximization procedure above locates the toolmark patterns $\mathbf{x}_i$ which define the "band" of data space which maximally separates the different set of patterns. Figure 23 visually depicts this process.

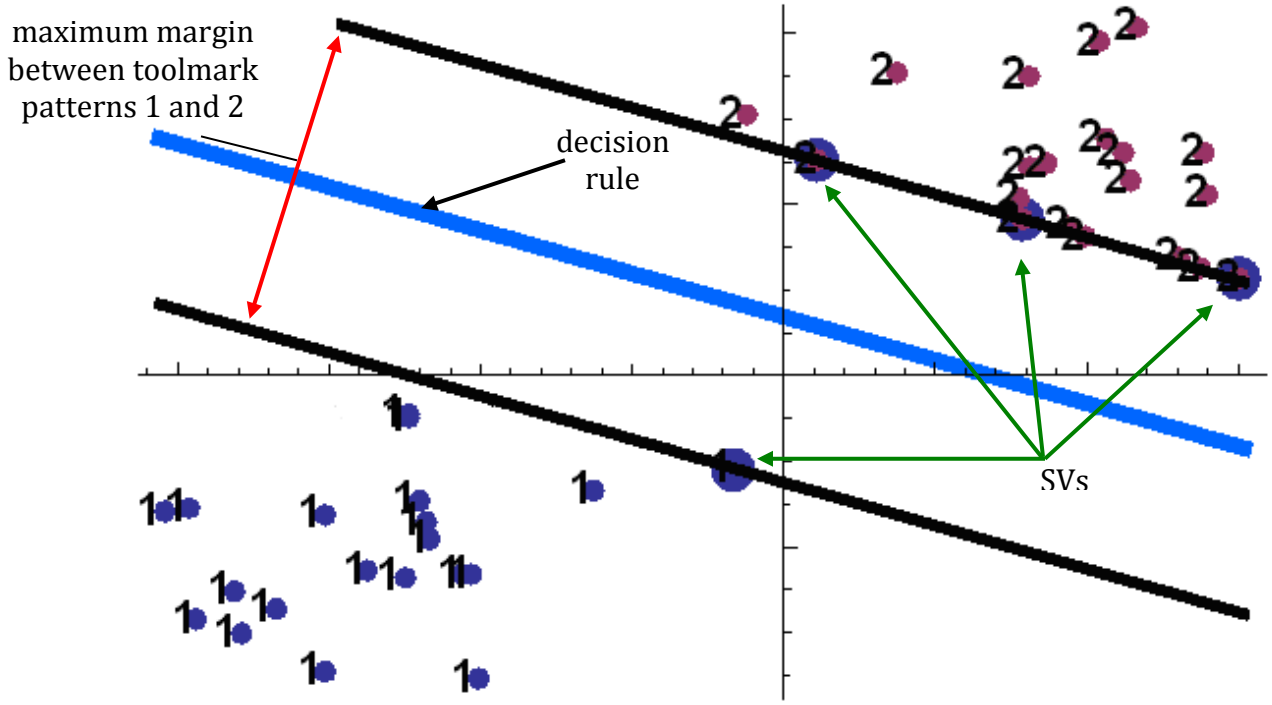


FIGURE 23. The job of the SVM algorithm is to determine the blue line, which separates the measurement data for tool 1 from tool 2. The bigger purple data points are the actual support vectors.

The SVM methodology was originally designed to separate two groups but it does have several incarnations which can handle multi-group (multicategory) problems. The most popular, because it works well in practice, is to consider all possible pairs of groups of striation patterns and determine a two-group (binary) SVM for each pair. Thus if there are k samples of striation patterns generated from k screwdrivers, $k(k\text{-}1)/2$ binary SVMs are computed and group identity of a pattern is determined by voting of the decision rules (69). This multicategory approach to SVMs is called the one-vs.-one method and is what we use in the studies presented below. Combined, the discrimination method is referred to as PCA-SVM

4. Methods to estimate error rates

4.1 Resubsitiution methods

An error is defined as a misclassification of a toolmark by the comparison algorithm. This occurs when the algorithm does not identify the unknown toolmark as having been made by the suspect tool when indeed it had, or the algorithm identifies the unknown toolmark as having been made by the suspect tool when indeed it had not. Following Wayman and Saunders et al. the former can be referred to as a false “no-match” error and the latter as a false “match” error (Wayman, Saunders 2011). Note that the term “match” here is specifically defined as an association a machine learning algorithm declares between a toolmark and a tool. It should not be taken to mean an exact match.

In order to estimate the algorithm’s error rate over a population of toolmarks, which it was not trained with (estimated error rate) several robust methods exits. For most applications of statistical pattern recognition to real world problems, the simplest method to empirically estimate the error rate is resubstitution (Smith 1947). This is simply the application of the computed classification rules to the set of data used to derive them. The percentage of misclassifications via the resubstitution method is called the apparent error rate. This is a biased estimate and tends to be overly optimistic and should be corrected. The first and simplest correction is called hold-one-out cross validation (Efron 1993, Rencher 2002). This method computes the decision rules using

all but one of the toolmark patterns in the data set. The hold-one-out procedure is repeated for each toolmark pattern in the data set and the results are averaged to compute an estimated error rate (Efron 1993)

$$\mathrm{Err}^{\text{HOO-CV}} = \frac{1}{n}\sum_{i=1}^{n} 1 - \delta_{y_i, g^{\text{hold-out } \mathbf{x}_i}(\mathbf{x}_i)} .$$

If *c* toolmark patterns are held out the resulting error rate is called *c*-fold cross-validation. Another error rate estimate, which was used in this study is the bootstrap (Efron 1993). First a set of *B* bootstrap data sets, **X*** are generated by randomly selecting (with replacement) *n* toolmark pattern feature vectors from the original data set **X**. Note that each bootstrap data set contains the same number of elements (toolmark pattern feature vectors) as the original data set, thus some patterns may be repeated. The decision rules are recomputed for each bootstrap sample (*g**) and an average error rate is computed using them on the original data (Efron 1993)

$$\mathrm{Err}^{\text{all-data}} = \frac{1}{n}\sum_{i=1}^{n} 1 - \delta_{y_i, g^*(\mathbf{x}_i)}$$

as well as the bootstrapped data used to compute them

$$\mathrm{Err}^{\text{boot-data}} = \frac{1}{n}\sum_{i=1}^{n} \#\left\{\mathbf{x}_i \in \mathbf{X}^*\right\}\left(1 - \delta_{y_i, g^*(\mathbf{x}_i)}\right).$$

The difference between these two averages is called the bootstrap estimated optimism (Efron 1993). Averaging together these optimisms gives the expected bootstrap estimated optimism,

$$\omega^{\text{boot}} = \frac{1}{B}\sum_{i=1}^{B} \mathrm{Err}_i^{\text{all-data}} - \mathrm{Err}_i^{\text{boot-data}}$$

which is then added to the apparent error rate to obtain the final error estimate.

4.2 Conformal prediction theory

Solomonoff's and Kolmogorov's algorithmic theory of randomness is a mathematically sophisticated way to gauge the amount of true information in a string of symbols (Li 2008). Recently, a method which gives confidence levels to identification of unknown patterns and control over error rates has arisen from the study of algorithmic randomness (Vovk 2005). This method, called conformal prediction can be applied to any statistical pattern comparison algorithm and holds a great deal of potential when applied to tool mark analysis. Prediction regions produced by conformal prediction can give a judge or jury an easy to understand

measure of reliability for tool mark pattern identification because the method yields confidence on a scale of 0%-100%.

The way the method works is actually very simple (Vovk 2005). Given a training set of striation patterns with known identities (called a bag) and at least one striation pattern of unknown identity, an estimate of randomness is computed for the bags containing the unknown striation pattern with all possible labels for its identity. The only assumption is that the striation patterns of the training set are drawn independently from the same, but unknown probability distribution.

Randomness of the bag is tested in a way analogous to what is done in traditional hypothesis testing (Vovk 2005). The null hypothesis is that unknown striation pattern **x** with assigned identity label *y* [*i.e.* the pair (**x**,*y*)] belongs in the bag and does not significantly decrease the bag's randomness. The alternative hypothesis is that the pair (**x**,*y*) does not belong in the bag and thus *y* must be a different label than the one assigned. *P*-values are computed for randomness estimates. *Thus conformal prediction regions for tool mark pattern identities can be thought of as generalizations of confidence intervals known from textbook hypothesis testing*. Traditionally confidence intervals are computed for population parameters (*e.g.* a sample average) to give an indication of the regions where their true values may fall. Technically, the Neyman-Pearson interpretation of (1-e)×100% confidence interval for an estimated population parameter (here, striation pattern identities) constructed from a random sample of a given sample size, will contain the true population parameter (1-e)×100% of the time (Vovk 2005). The value e is called the level of significance and is the probability that any given confidence interval constructed from a random sample will *not* contain the true population parameter.

Note that the null hypothesis can be accepted for multiple label prediction regions of the striation pattern's identity. In such cases the identity assignment (*i.e.* the prediction region) at the (1-e)×100% confidence level is ambiguous. While multi-label output is not wholly uninformative, hopefully the prediction region will contain only one label with a *p*-value ≤ 0.05. This means that the conformal prediction algorithm has produced a prediction region with only one label and a confidence level of at least 95%.

P-values for striation pattern test identities are found by computing nonconformity scores. The nonconformity score, α_i for the i^{th} striation pattern using one-vs.-one multiclass SVMs was computed as

$$\alpha_i = \frac{1}{k-1} \sum_{j=1}^{k(k-1)/2} \lambda_{i,j}$$

where $\lambda_{i,j}$ is a matrix element of an n row by $k(k\text{-}1)/2$ column matrix of Lagrange multipliers (Vovk 2005). The formula just sums all the columns in this matrix and weights the resulting n-dimensional vector by $1/(k\text{-}1)$.

A p-value for each possible labeling $\text{tlab}_i \in \{1,2,...,k\}$ of a test striation pattern is computed as

$$p_{\text{tlab}_i} = \frac{\#\{j \in \{1,2,...,n\} : \alpha_j^{\text{tlab}_i} \geq \alpha_{\text{test-pattern}}^{\text{tlab}_i}\}}{n}$$

where $\alpha_j^{\text{tlab}_i}$ is the nonconformity score of the j^{th} pattern when the test pattern is labeled as screwdriver tlab_i, $\alpha_{\text{test-pattern}}^{\text{tlab}_i}$ is nonconformity score of the test pattern labeled as screwdriver tlab_i, and p_{tlab_i} means the p-value of the test pattern labeled as screwdriver tlab_i (Vovk 2005). A set of k p-values is computed for each test pattern, one for each possible screwdriver identity. For a chosen significance level e (*i.e.* level of confidence 1-e) a 1-e confidence region of labels is determined by selecting those labels of the test pattern with p-values $\geq$ e. Ideally the output confidence region contains only one label. If a multi-label confidence region is output, it counts as a correct identification if it contains the true label of the striation pattern, though obviously it is less informative. Empty regions can also be output if the CPT algorithm cannot confidently identify the striation pattern. Empties automatically count as errors (Vovk 2005).

III. Results

1. Toolmark impression data collection and database

The collection of toolmark impression data has been successful in amassing a large assortment of fired cartridge casings, "scraping" tools and their corresponding striated toolmarks.

1.1 Cartridge case impression pattern collection

Glock striation pattern collection

The Glock family of firearms has a unique design, which provides a useful impressions/striation when cartridge cases are examined. A head stamp with the regions of interest appears in Figure 24.

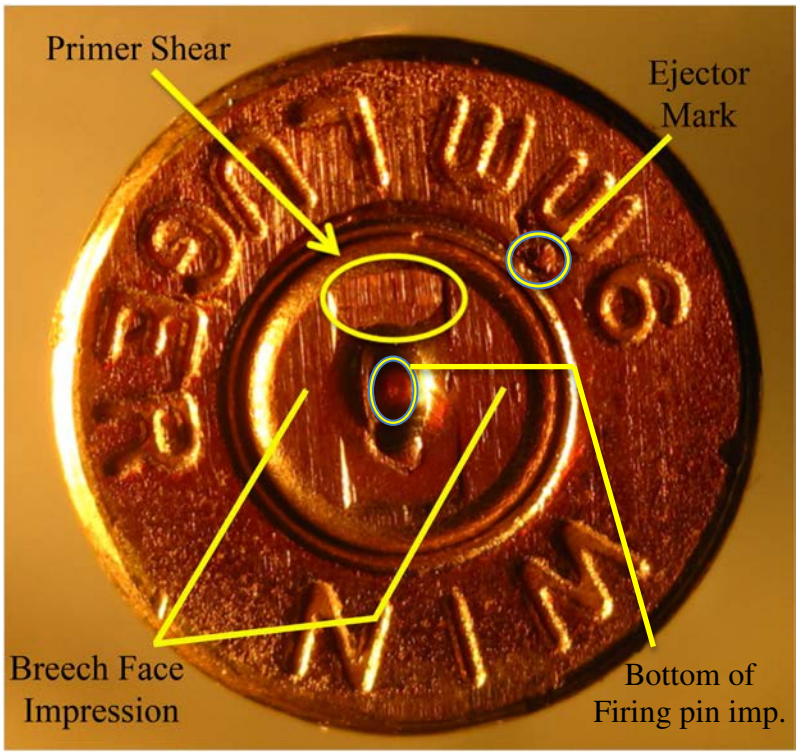


FIGURE 24. Regions of interest being recorded in the database of firearm impression patterns.

Glocks use an elliptical firing pin, which requires a rectangular firing pin aperture that is large enough to accommodate the size and shape of the pin when it emerges from its hidden position upon commencement of the firing sequence. This rectangular aperture is responsible for the primary toolmark striation pattern examined in this study, the primer shear. In figure 19 above, the cartridge case has been rotated 180^{o} from its position when fired and the primer shear appears at 12 o'clock.

The primer shear striation patterns from one hundred sixty-three, 9mm cartridge cases fired from twenty-four Glock 9mm semi-automatic pistols have been collected and scanned into our growing research database. For these primer shears, confocal scans were conducted using 50x magnification and high numerical aperture (0.95 NA) yielding striation patterns of very high detail and relatively low noise. The primer shear marks show pronounced striation patterns and serve as useful test case examples for computational pattern identification methodology.

Cartridge cases fired from Glock pistols will continue to be collected and scanned with our confocal microscope and made available in the database past the official end of this project. Figure 25 shows a 3D confocal image of an entire primer shear region. Note the pronounced striation pattern.

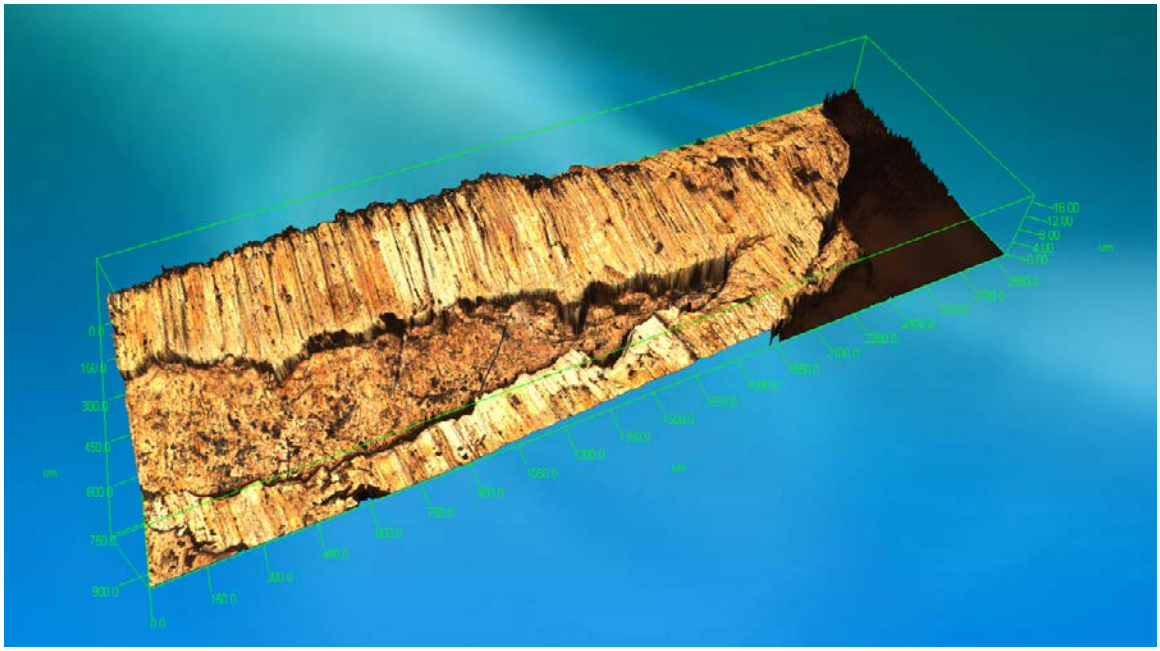


FIGURE 25. Confocal image of an entire primer shear striation pattern. Image acquired with 50x objective (0.95 NA).

The focus of the firearms portion of this project was primer shears, however more portions of the cartridge case surface have been (and will continue to be) scanned and recorded into the database. These surfaces are available for future analysis and are detailed below.

The bottoms of firing pin impressions have been recorded for select cartridges examined and are available in the database for future analysis. Figure 26 shows the bottom portion of a firing pin impression. The left hand image is an all-in-focus 2D confocal image of the area. The pseudo-color surface to the right shows the impression in 3D.

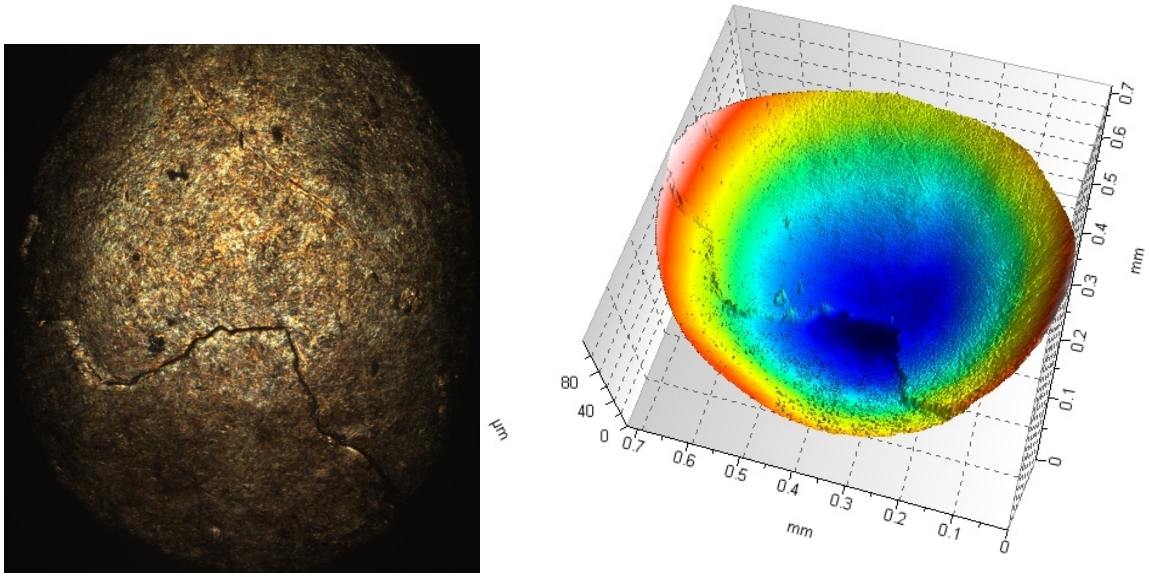


FIGURE 26. Confocal image of bottom of firing pin impression pattern. Left image is all-in-focus 2D. Right image is same area in 3D. Image acquired with 20x objective (0.6 NA).

Because of the steep slopes, unfortunately the entire area of the firing pin impression could not be recorded under high magnification.

Breech face impressions on primers (left and right strips about the firing pin impression, along the diameter of the case) have also been recorded for select cartridges examined and are available in the database for future analysis. Figure 23 shows a 3D confocal image of one of the strip areas of a breech face impression. The “dip” on the right hand side of the top image is the edge of the firing pin impression. The green line across the top shows the area where a profile was determined. The profile appears in the last image in Figure 27.

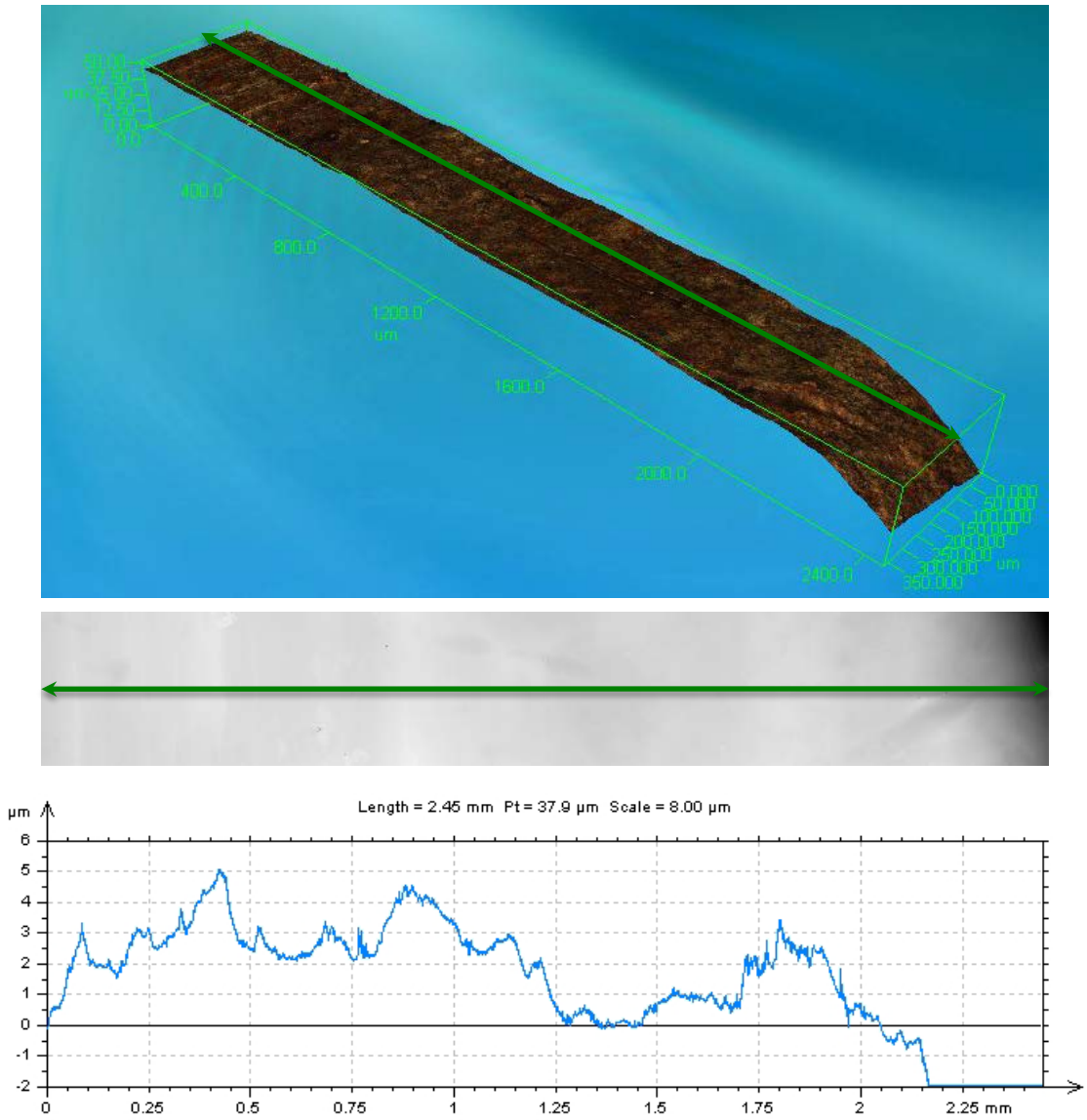


FIGURE 27. Confocal image of left breech face impression pattern on a primer. The green lines show where the profile at the bottom of the figure was taken. Image was acquired with 50x objective (0.95 NA).

Finally, ejector marks have been recorded for select cartridges examined and are available in the database for future analysis. Figure 28 shows an all-in-focus 2D as well as 3D confocal image of an ejector mark.

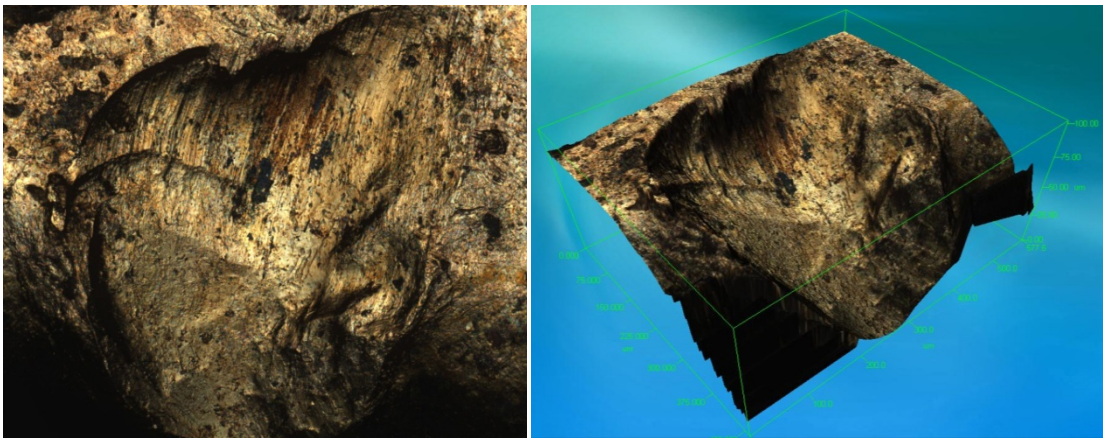


FIGURE 28. Confocal image of ejector mark. Image acquired with 20x objective (0.6 NA).

1.2 Striated toolmark pattern collection

Screwdriver striation pattern collection

Eight Craftsman® brand screwdrivers and ten Iron Bridge® brand screwdrivers have been obtained as exemplars. These screwdrivers were used to make striation patterns in both lead and jewelers wax for inclusion into the database.

Figure 29 shows 3D confocally imaged, 1,500 um strips of striation patterns generated by one of the Craftsman screwdrivers. The top image is an impression made in lead while bottom image is an impression made in wax.

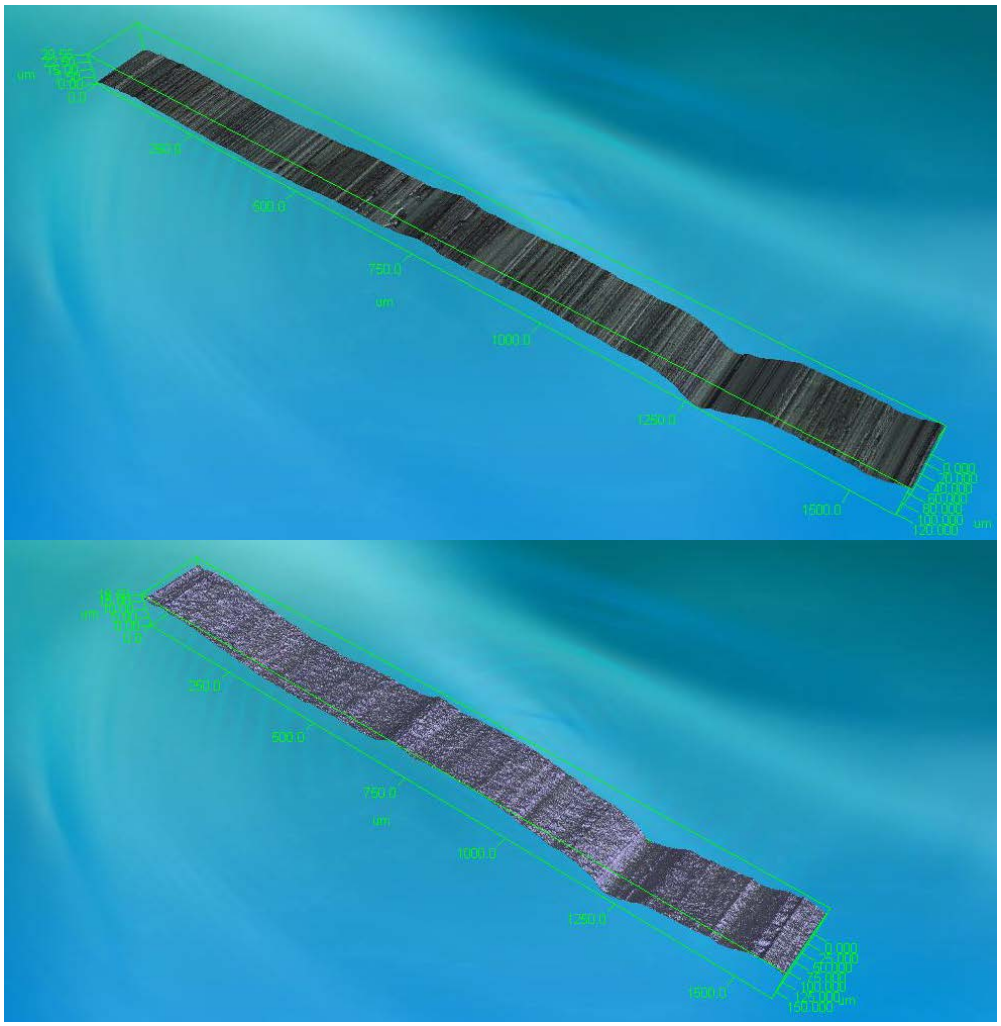


FIGURE 29. Confocal image of Craftsman screwdriver striation patterns. Top image is lead impression media while bottom is jewelers' wax. Images acquired with 50x long working distance objective (0.55 NA).

After the acquisition of several wax striation patterns it was realized that the wax surfaces are pitted with small "nooks and crannies" and thus contain many sharp angles. The recorded surfaces were significantly more noisy than those surfaces generated in lead. Thus it was decided that while wax is a good media for toolmark generation when properly coupled with side lighting and compound microscopy, it would not be prudent to continue with wax toolmarks for confocal microscopy at this juncture. A total of 180 striation patterns in lead media generated by eighteen screwdrivers have been scanned into the database. Scanning of screwdriver striation patterns will continue past the official end of this project.

Chisel striation pattern collection

Five consecutively manufactured chisels (Mayhew® Brand) were obtained from the reference collection of Gerard Petillo (independent firearm/tool mark examiner)(cf. Figure 30). The five chisels have been used to create 50 striation patterns on a lead medium at a 30^{o} angle of attack (5 replicate striation patterns per side of each chisel).

FIGURE 30. Five consecutively manufactured chisels (G. Petillo reference collection.).

In order to better represent the conditions that trained forensic tool mark examiners must deal with, the striation patterns were made by dragging the chisels (in a jig) across the surface of the lead impression media, by hand. This process generally yielded many non-contiguous sub-striation patterns (cf. Figure 31), which range in width, but all are between the left and right edges of the total chisel striation pattern generated.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

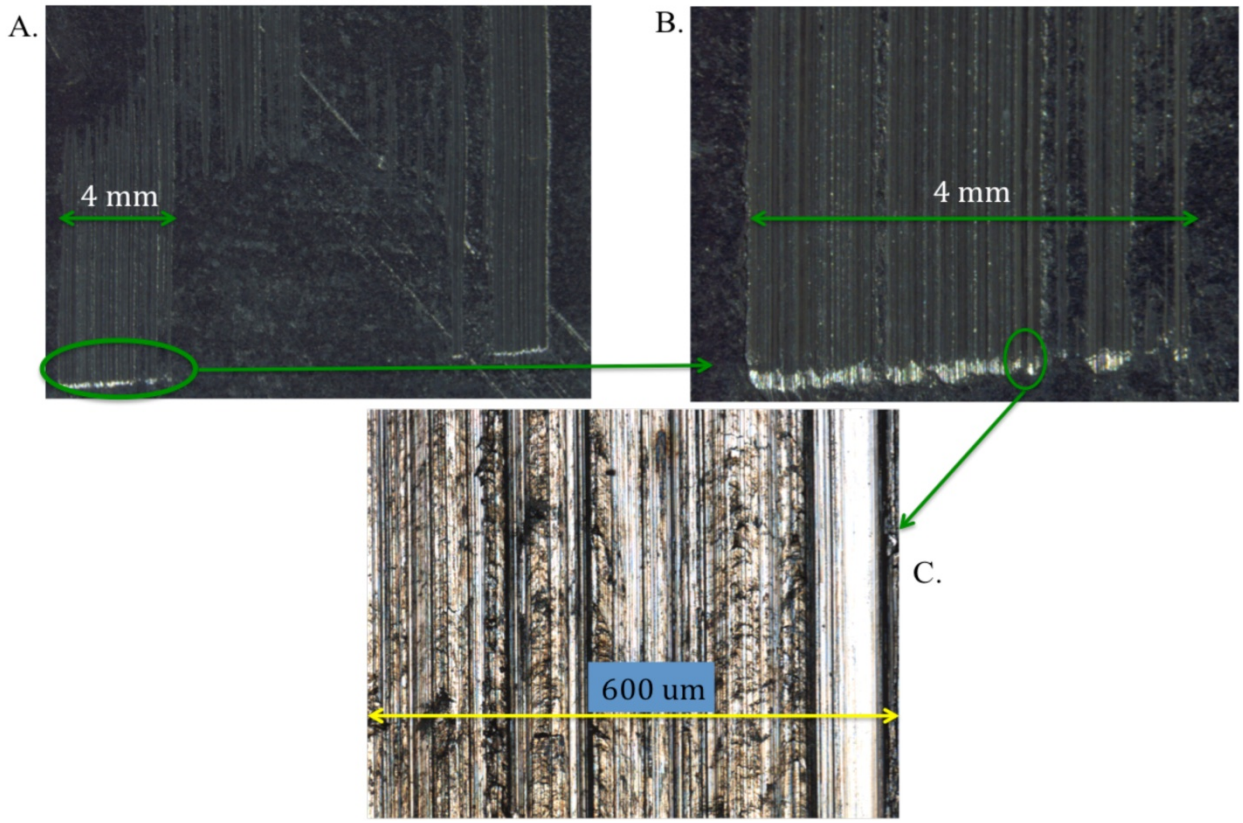


FIGURE 31. Typical striation pattern generated by a chisel. A. shows the entire pattern. Note the many non-contiguous sub striation patterns. B. is a contiguous sub-pattern, i.e. “striation patch”. C. is one of the tiles making up the striation patch.

Within themselves, these sub-striation patterns are contiguous and will be referred to as “striation patches” for brevity. Below is a portion of a striation pattern on lead media generated by chisel 3 side A.

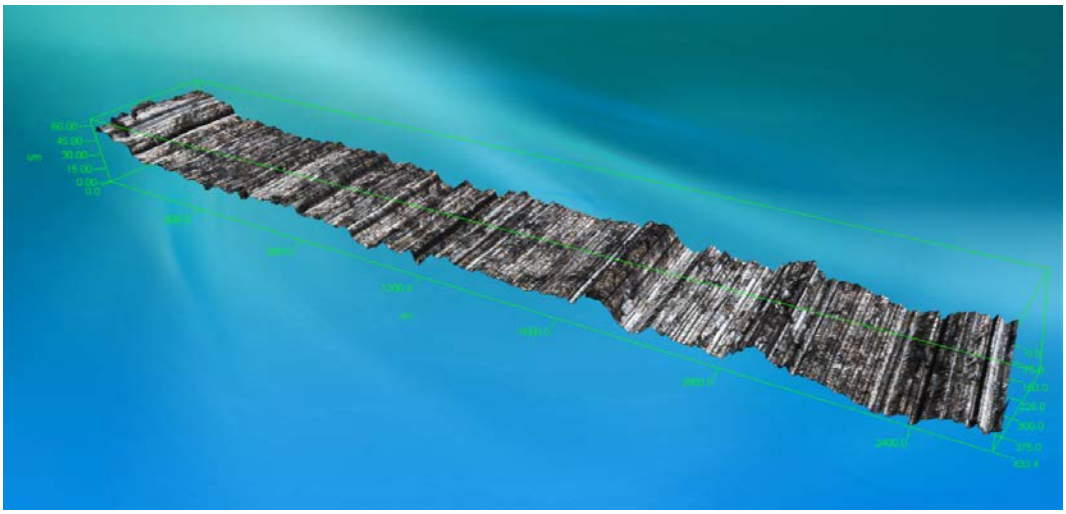


FIGURE 32. Striation patch from striation pattern on lead generated by chisel 3 side A.

Confocal images of these striation patches were obtained with the 20x objective (0.6 NA). Each striation patch generally did not fit within the field of view under 20x magnification and thus consists of a variable number of 20x field of view tiles. One hundred and thirty four striation patches consisting of 1109, 20x tiles total were recorded in our database. The tiles that composed a given striation patch, were not stitched together and will be analyzed separately. This is due to the fact that the striation patches vary widely in length. In order to use the multivariate statistical tools developed for this project, all patterns must be of the same length. All 1109 tiles are available in the database for future analysis.

1.3 Database and web interface

The database of all 3D toolmarks recorded in the process of carrying out this project (striated/impressed surfaces on cartridge case head stamps fired from 9mm Glocks, screwdriver striation patterns and chisel striation patterns) is available to the firearms/toolmark research community and the forensic firearms/toolmark practitioner community at http://toolmarkstatistics.no-ip.org/. Users can sign up to request an account, where after approval, they will have full access to the database. Several pieces of software and statistical analysis scripts were generated in the process of carrying out this project and are also available for download/use by users.

The database is meant for exploration of what 3D microscopy is capable of, research, algorithmic development/testing and for interested practitioners to generate qualitative images of

3D toolmarks for case/court presentation purposes. It is research sized (i.e. small to medium) at the time of writing this report and will continue to grow past the official end of this project.

A web interface has been developed for the database so that the data collected as well as the statistical/visualization software developed for this project can be searched or downloaded by interested users. Figure 33 shows a screenshot of the homepage for the database.

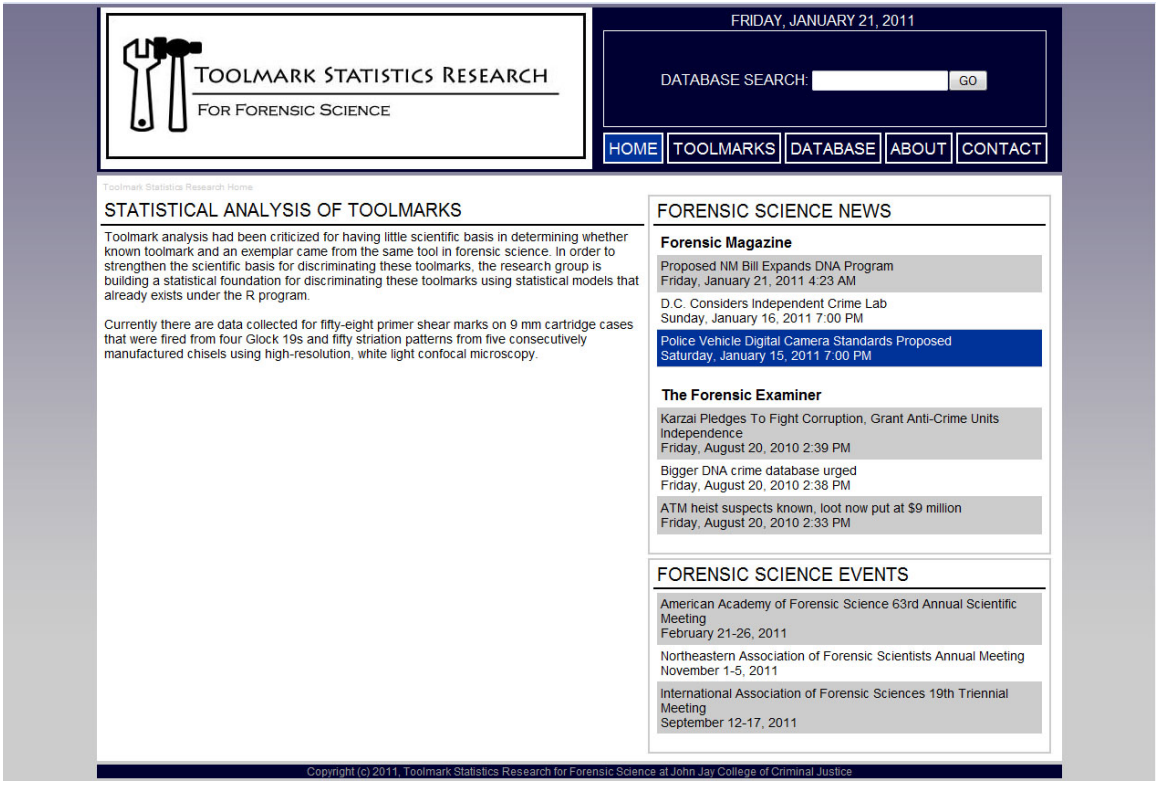


FIGURE 33. Database homepage.

The database software for this project is FileMaker Pro (for prototyping) and MySQL (for production level). Fields in the database are: impression type, subfields: tool/firearm employed, tool/cartridge manufacturer, model, lot number, serial number or other manufacturer's identifying marks, dates of tool production and collection, dates of tool mark production, miscellaneous notes and point cloud data scanned from impression. Figure 34 shows a search page.

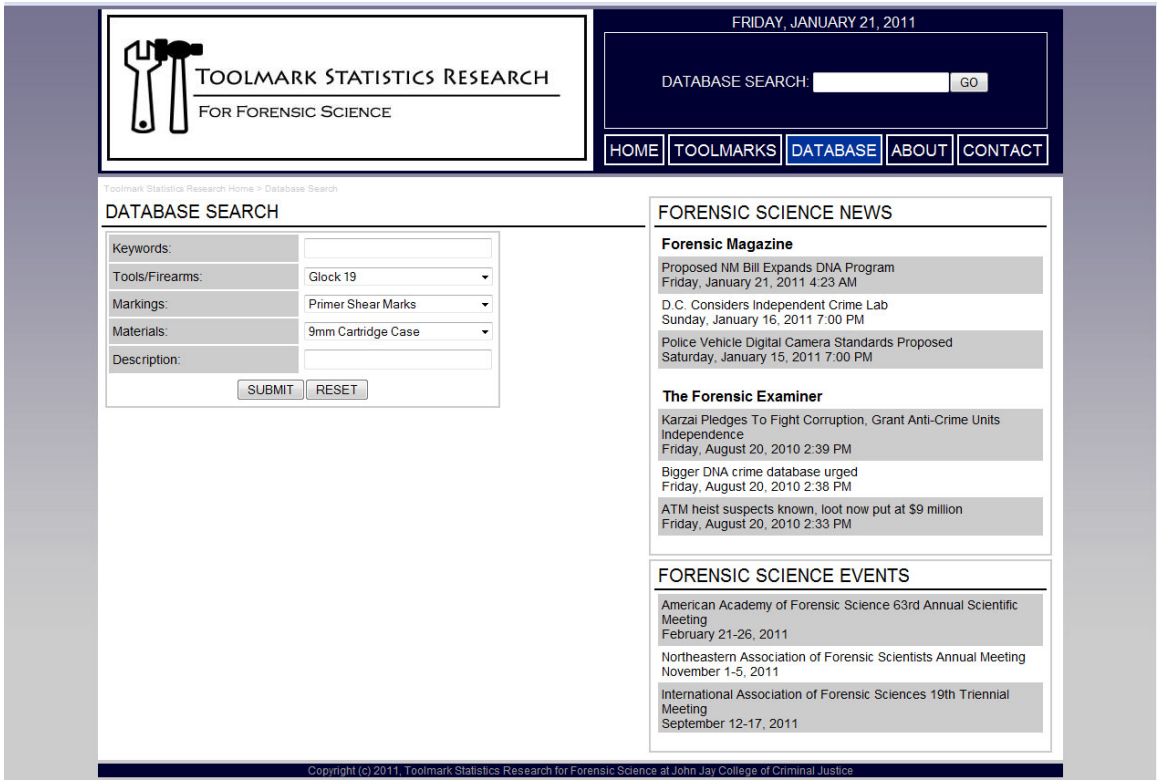


FIGURE 34. Database search page.

Queries are returned as text descriptors that can be clicked to download data files to the user's computer. The 3D surface data is encoded in 16 or 32-bit gray levels and is stored in binary using the format of the Mountains® metrology software system. A screenshot of the search results page is shown in Figure 35.

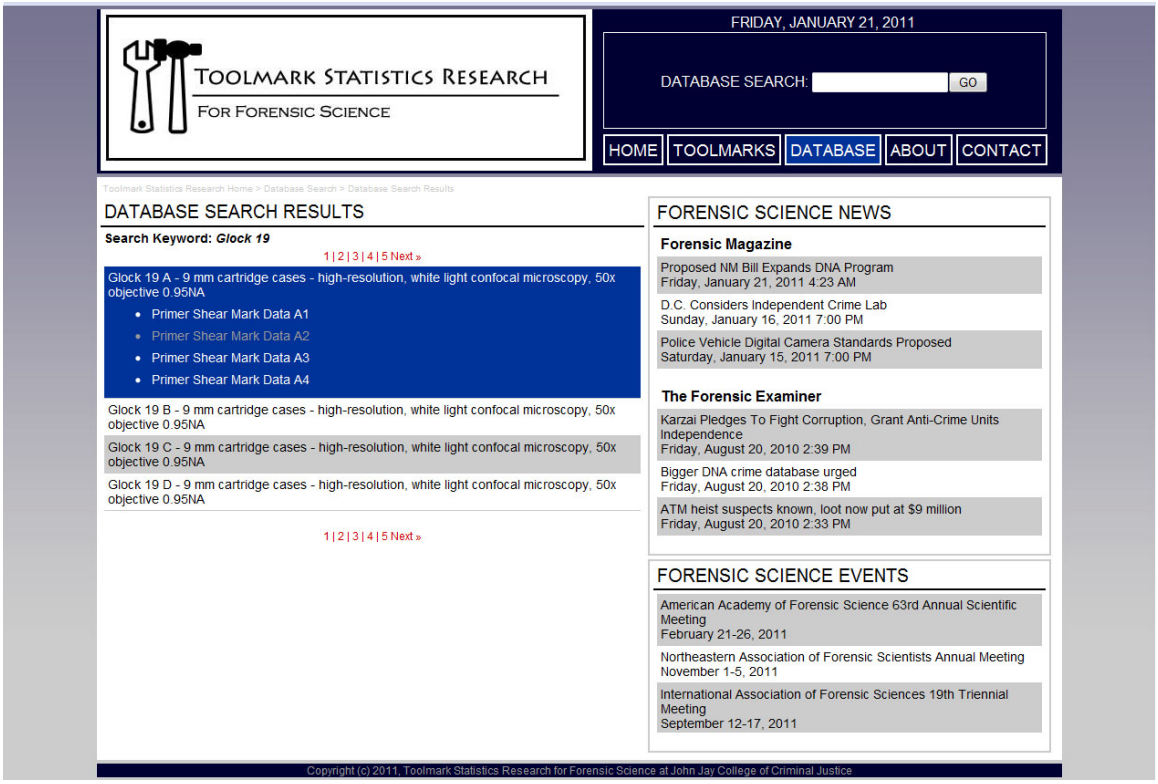


FIGURE 35. Search results page.

The Mountains® binary data format was chosen because it is generally well known in the scientific community (specifically metrology/mechanical engineering) and is published. Users however will not need to have the Mountains® software to open the data files downloaded from the database. A Java language "plug-in" has been written for the open-source digital imaging/analysis software suite ImageJ (developed at the NIH). The plug-in, which is available on the website, allows ImageJ functionality to be used to perform measurement tasks as well as interactive 3D viewing of the tool mark surfaces in the database. Brief description of this functionality is given below. Users of the database can also develop their own analysis software to open and operate on our data files (this was our original intention for making the toolmark data record in this study available to the wider forensic tool mark analysis community).

1.4 Surface visualization and measurement software

Surface and profile format for the files in the database is Mountains® .sur format. Data is saved in binary format to save space. Again, the Mountains® program is not needed to open the files. Instead the open source digital image processing program, ImageJ can be used

(http://rsbweb.nih.gov/ij/). The ImageJ suite was chosen because it is a fully extensible programming paradigm (besides being open source and platform independent). ImageJ can be used to make basic measurements on the surface and view it in an interactive 3D format. Figure 36 shows a primitive fixed 3D image of one of the primer shear striation patterns produced with ImageJ.

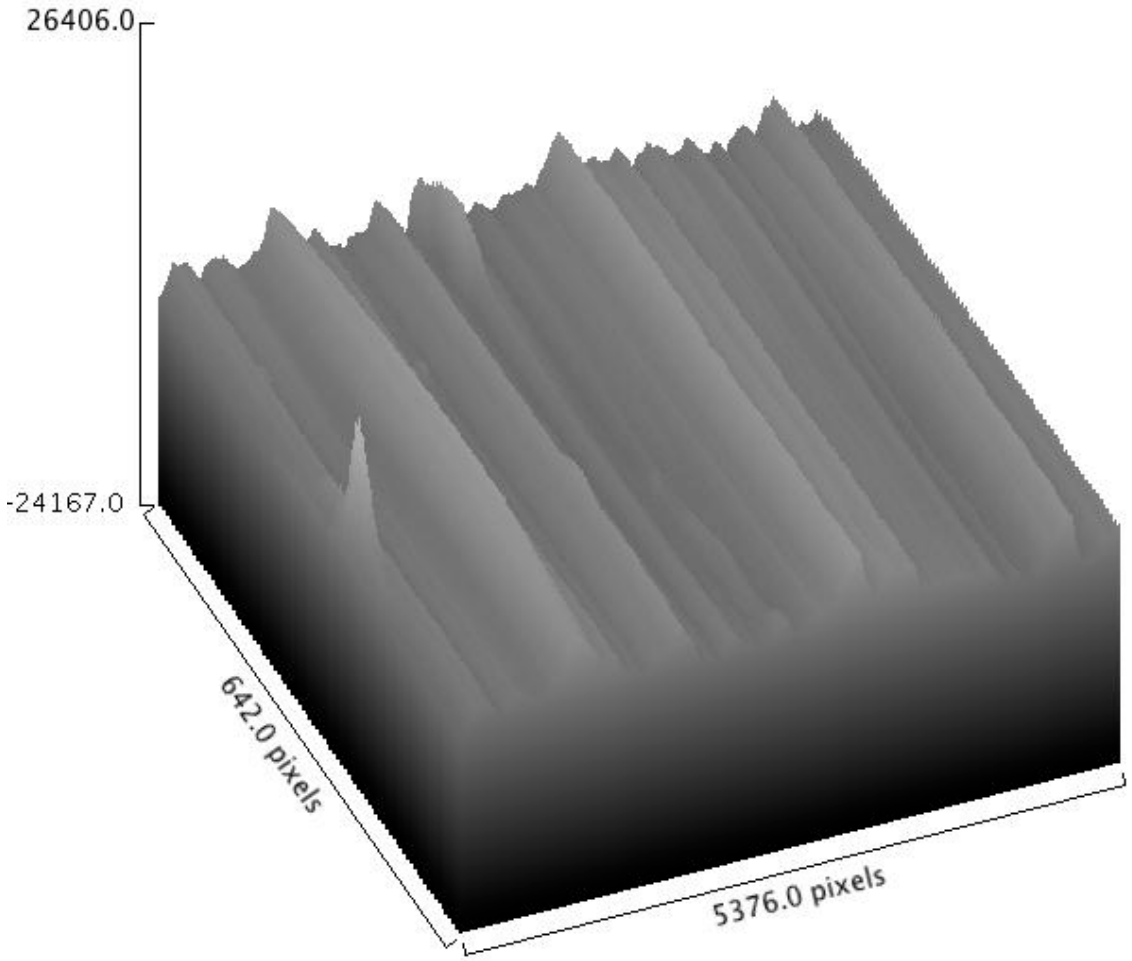


FIGURE 36. Basic 3D primer shear image produced by ImageJ from data in our database. A practitioner/researcher can generate figures such as this by simply downloading data from our database, installing ImageJ and opening the data file. Basic measurements can be made on the fixed 3D images, such as that shown in the figure. The units in the figure are in pixels and 16-bit grey levels, however they can be converted to any set of length units.

When the ImageJ plug-in is installed (which is just a copy and paste operation for the user), a calibrated surface image appears as shown below in Figure 37.

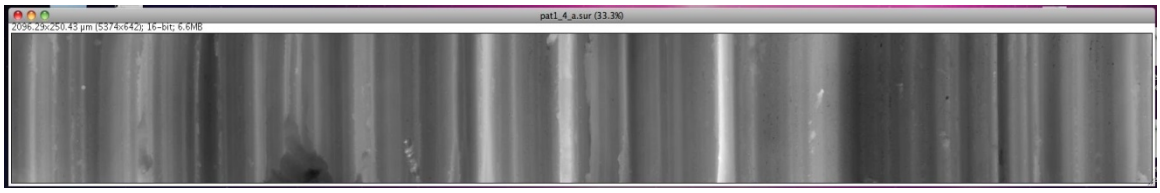


FIGURE 37. Micrometer calibrated 2D image of a Glock 19 primer shear.

The topography files exist in separate windows and can be easily aligned for visual inspection.

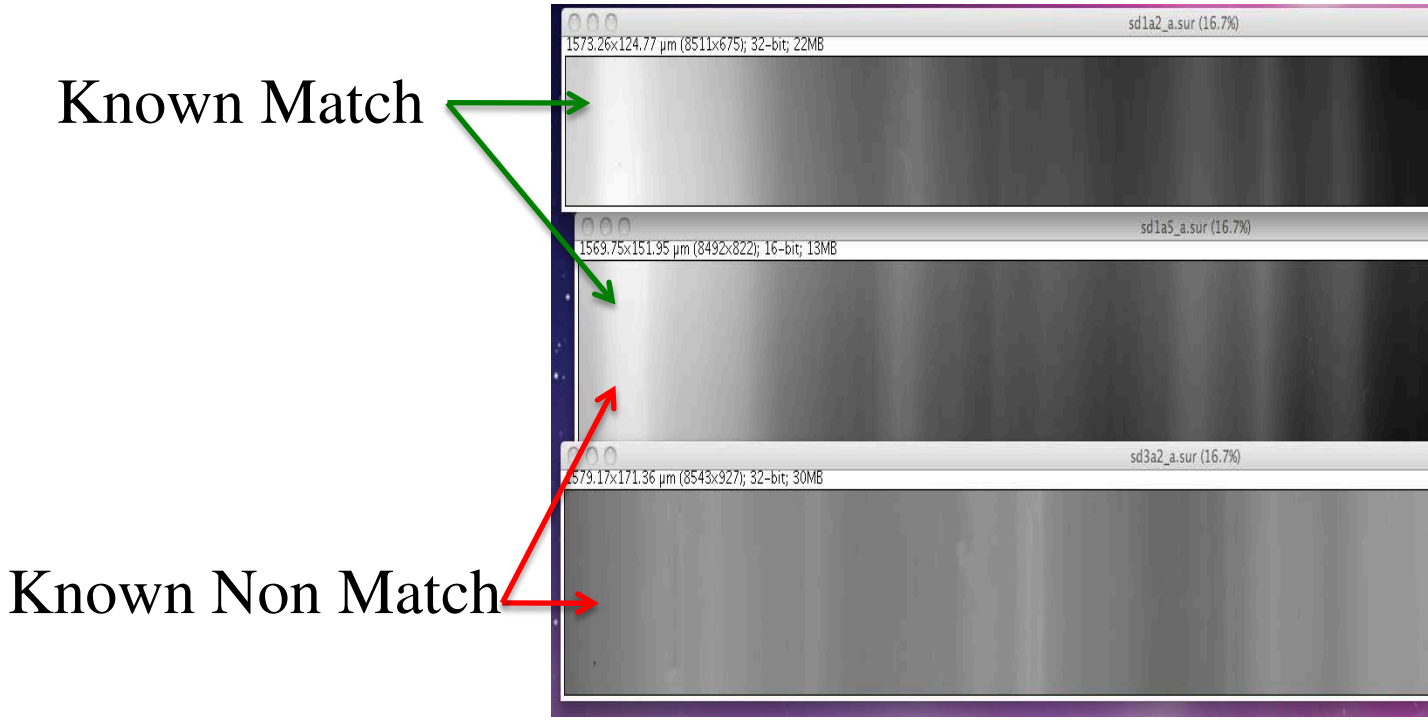


FIGURE 38. Topographies of three screwdriver striation patterns (two screwdrivers), shown in grey levels.

The ImageJ toolbar (Figure 39) allows the user access to numerous measurement and manipulation tools such as rotation, length and angle measurement.

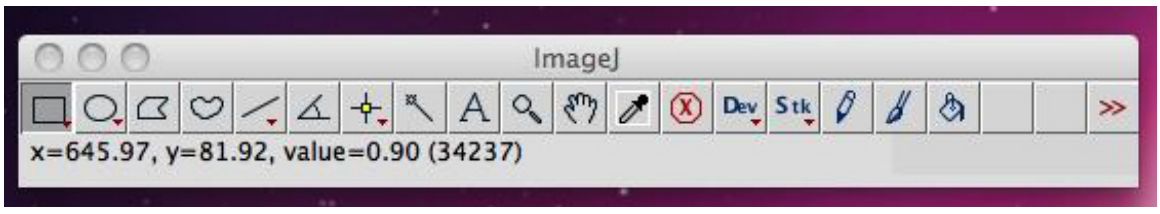


FIGURE 39. ImageJ toolbar. Makes measurement and manipulation of calibrated tool mark images from the database simple and flexible.

The surface files can also be shown as 3D interactive plots (Figure 40). The color encodes the surface height.

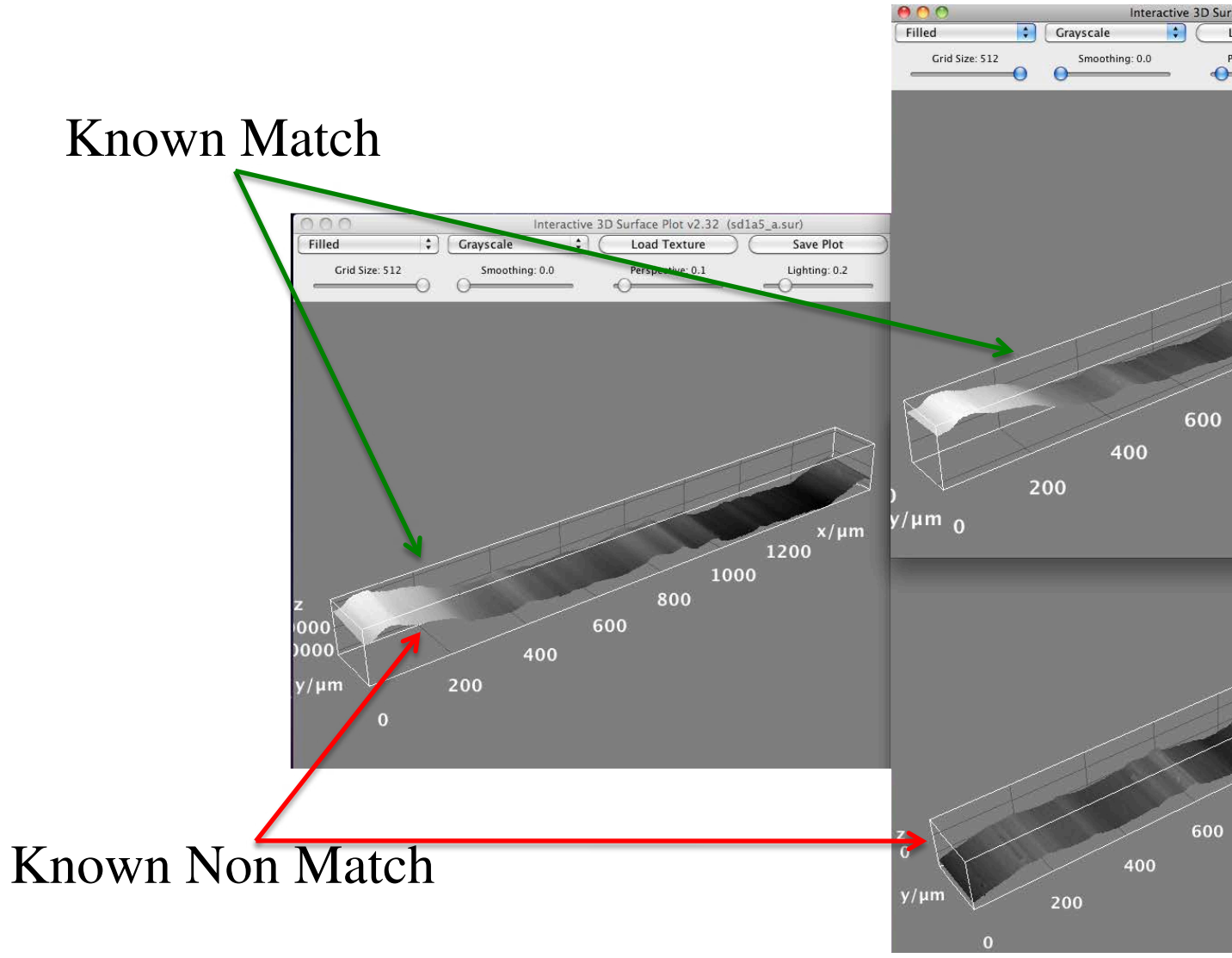


FIGURE 40. Interactive 3D ImageJ images screwdriver striation patterns.

Topography profiles across arbitrary tool marks in the database can also be generated as shown below in Figure 41.

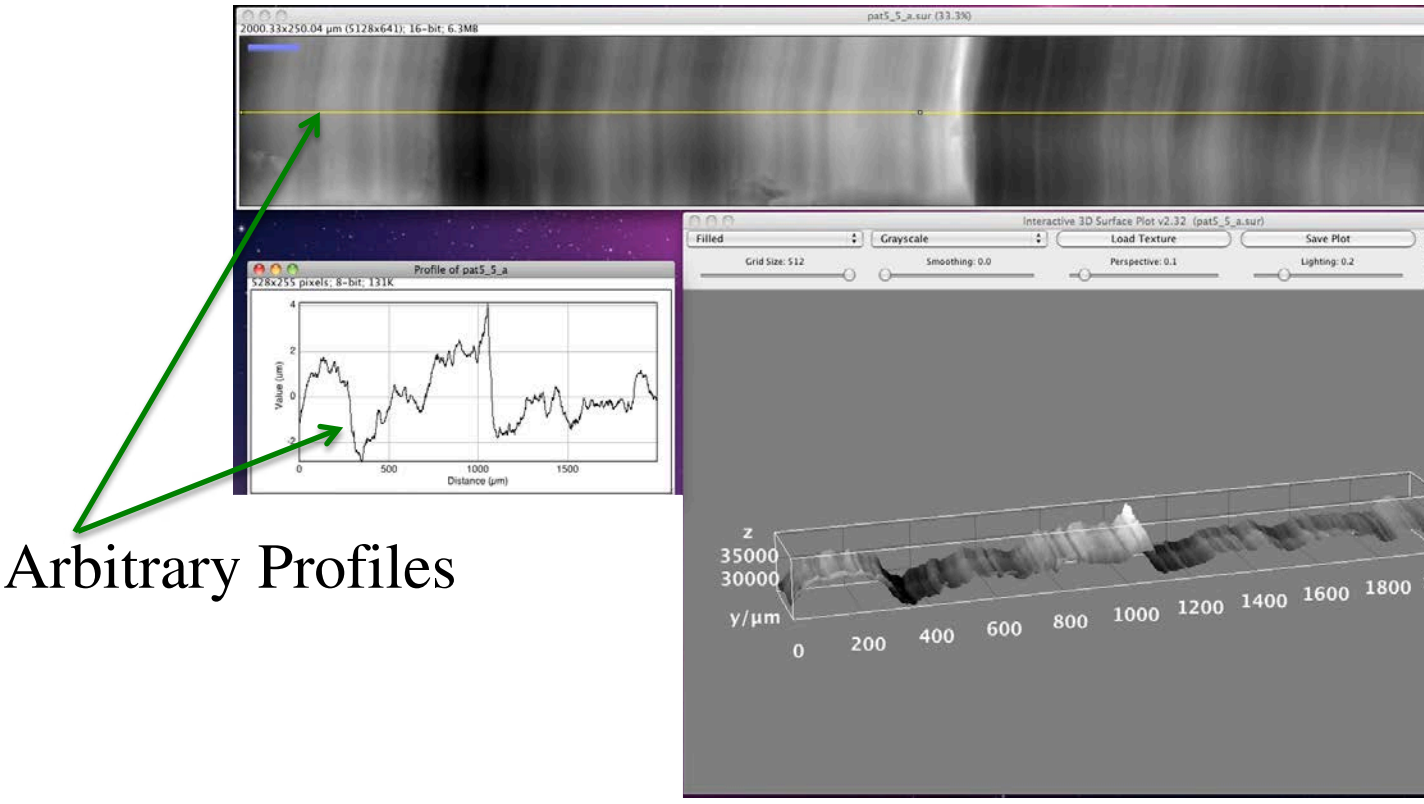


FIGURE 41. Basic primer shear images produced by ImageJ from data in our database. Imaging and measurement tools of ImageJ applicable to any toolmark in the database however.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

All images generated with ImageJ can be saved in any number of standard formats (JPEG, TIF, etc.) and can be opened in other imaging programs (e.g. Photoshop) for further annotation, preparation for court exhibits, lectures and publication presentations.

1.5 Profile simulator software

A portion of the database consists of 9mm cartridges fired for the Hamby-Thorpe study (Hamby 2009b). For original Hamby-Thorp study, two to three cartridges/Glock were available. The statistical analysis techniques used in this project however are numerically more reliable with five or more "replicates" per experimental unit (i.e. cartriges/gun). In order to exploit the Hamby-Thorp benchmark data set, a wavelet decomposition based simulator was written in the Mathematica programming language (Mathematica was chosen in the interests of a speedy prototyping process.). The program takes as input two or more profiles generated by the same tool. Figure 42 shows the collection of three mean profiles from primer shears of the three cartridges fired from Glock G26_SN_CMR289US (database designation).

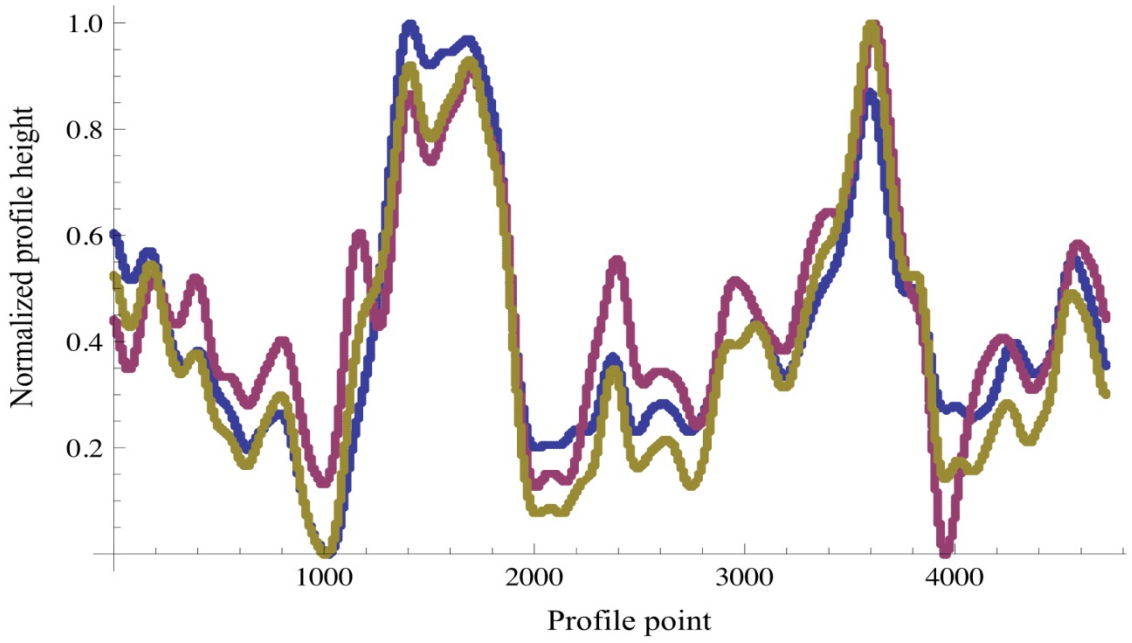


FIGURE 42. Mean profiles (aligned) of three primer shears on the three cartridges fired from Glock G26_SN_CMR289US.

The pyramid algorithm for the discrete wavelet transform (DWT), which is available in Mathematica, was used to perform multi-resolution analysis (MRA) on each profile (Mallat 2008, Percival 2006). At each level of decomposition (scale) a set of wavelet coefficients is determined for each increment of translation. Figure 43 show the wavelet coefficients for the 13 levels (12 detail, 1 mean) of its MRA using a fourth-order Coiflet wavelet basis (Fu 2003).

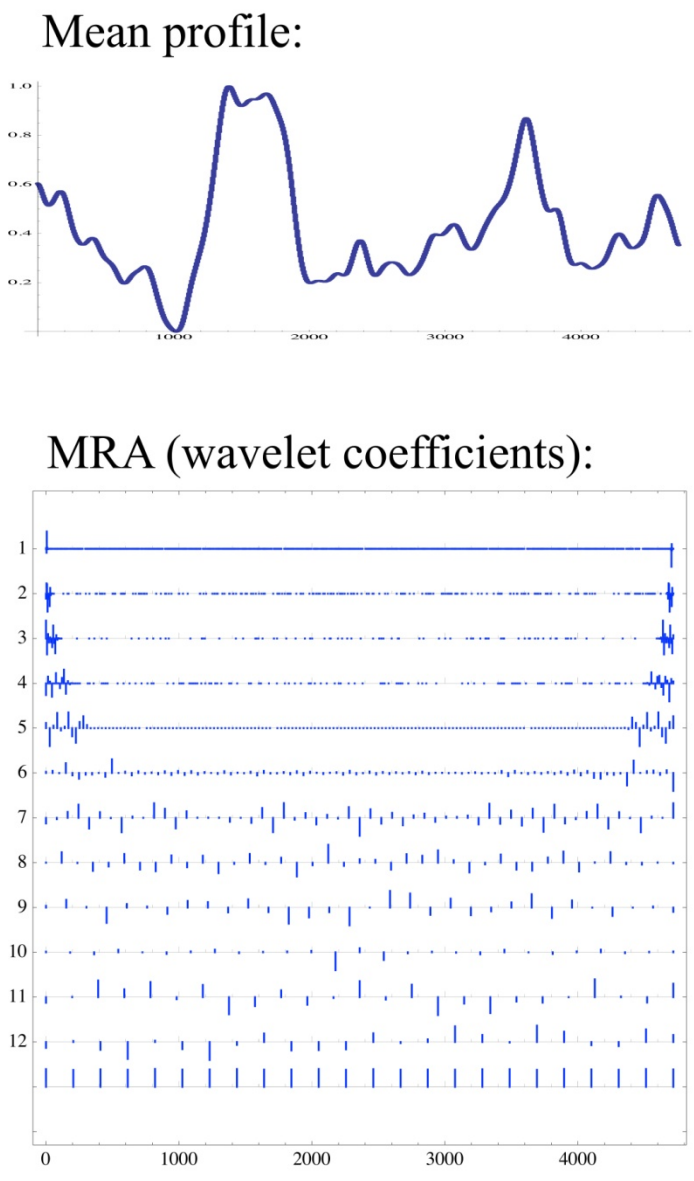


FIGURE 43. Mean profiles 1 of G26_SN_CMR289US along with the wavelet coefficients from

its MRA, Coiflet(4) basis. Extreme left and right boundary coefficients are noisy for the highest detail levels (levels 1-5). Effects from these coefficients are later excluded from the simulated profiles.

Wavelet coefficients from each level are collected together segregated for a group of profiles. Each translational increment at each level for a group of profiles is fit to a rough non-parametric distribution with a Gaussian kernel and a standardized window width. A profile is then simulated by sampling "simulated wavelet coefficients" from each set of distributions at each level. The inverse DWT is then applied to a simulated set of coefficients in order to construct the simulated profile. A particular simulated profile can be kept or thrown away. The "keep criteria" used by the software is a user specified level of correlation. If the correlation coefficient between the simulated profile and any number of real profiles (specified by the user) reaches a user defined level of correlation, the simulated profile is kept. Otherwise it is thrown away. The number of points to drop from the left and right boundaries can be specified, though 5% from either end is recommended.

The Glock in question *could reasonably* have generated kept profiles. Kept simulated profiles can then be fed back into the simulator in order to simulate more profiles. Thus the user has control over the real profiles used as input to the simulator, the wavelet basis, the number of real profiles to compare the simulated profiles to, ability to feed simulated profiles back into the simulation. Hence any dataset of two or more real profiles (from any striated surface, not just primer shears) can be built up to any sized data set that could reasonably have been generated by the source tool of the real profiles. Loosely speaking, a simulated set of profiles can be as "similar" or "different" for a real set of profiles as a user desires. An example set of 30 profiles generated from the three mean profiles of Glock G26_SN_CMR289US is shown in Figure 44. The correlation coefficient for each of these profiles is greater than 0.85 with each of the three real profiles.

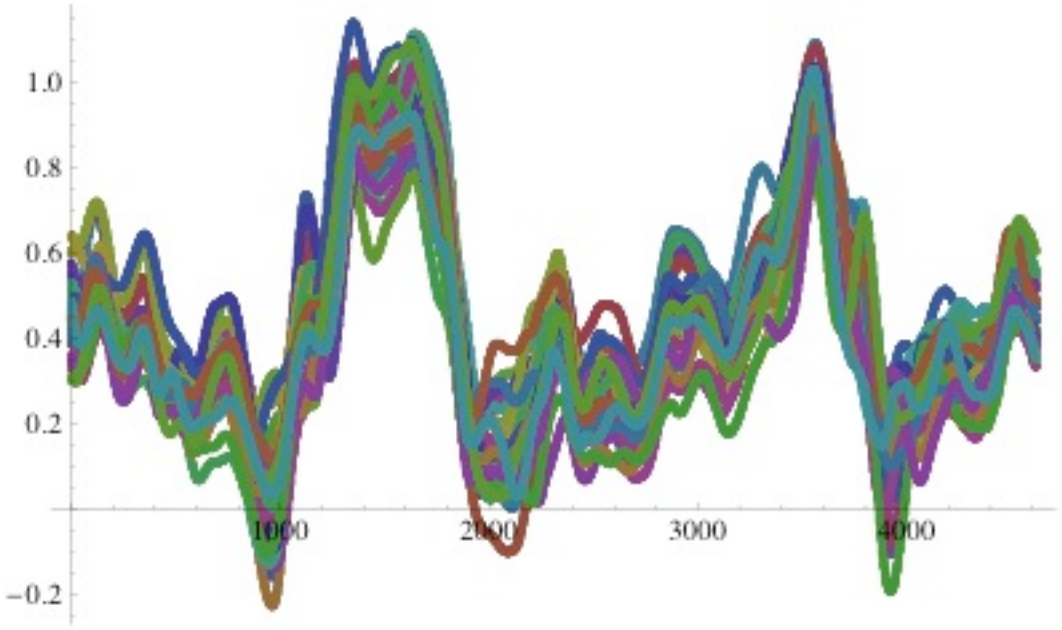


FIGURE 44. Thirty simulated profiles of Glock G26_SN_CMR289US along with the three real profiles used in the simulation. Coiflet(4) wavelet basis used in the simulation.

If the correlation coefficient between the simulated profile and any number of real profiles from the same experimental unit (i.e. primer shear or screwdriver striation pattern) is greater than or equal to a user defined level of correlation, the simulated profile is kept. This cut-off allows the user to set a minimal level of similarity between the simulated patterns and the real patterns used to generate them. The cut-off should be set to a low value. For the studies described in section 2 (below) the "minimal" correlation coefficient between real and simulated profiles was chosen to be 0.5. This cut off was chosen by examining the distributions of "known match" scores (KM) and "known-non match" (KNM) scores for real profiles. For example, Figure 45 shows the KM and KNM distributions for mean profiles of the real cartridge case primer shears. The crossover point is approximately 0.6. To be more conservative and create an even more challenging data set, it was decided to choose a cut-off point well below this, 0.5. As can be seen from the figure, a score of 0.5 is well into the KNM regime.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

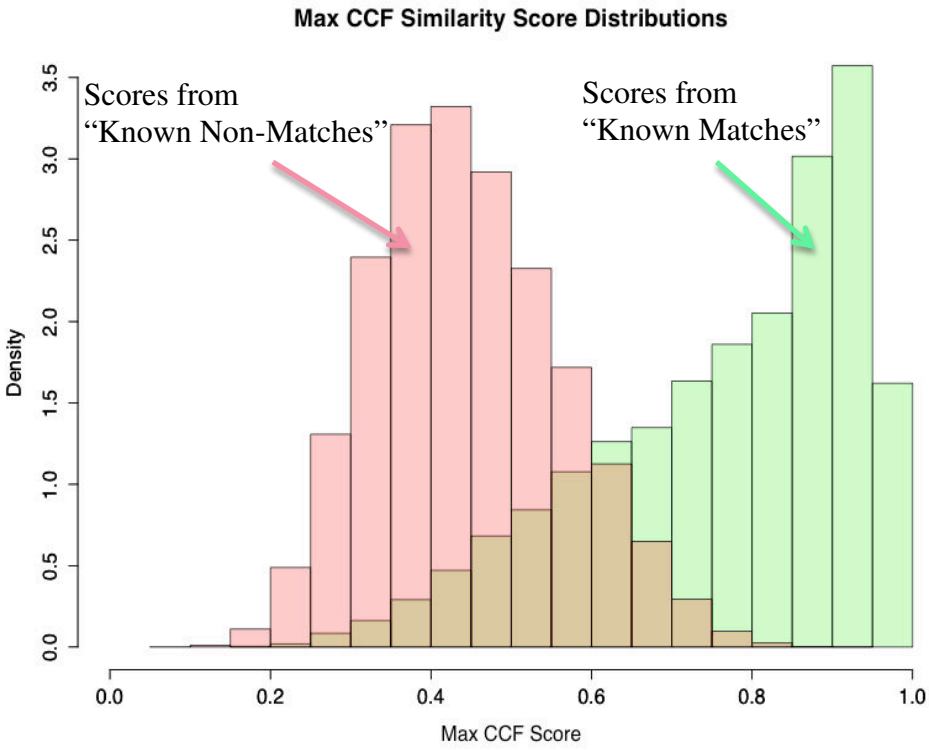


Figure 45. Correlation "similarity" scores between 162 primer shear mean profiles from the database.

Typically the correlation between the simulated profiles and their real counterparts was not as low as the cutoff (though users can set it to be). Choosing the cut-off to be a low level of correlation allows for variation and challenging profiles to be included in the data set. However the mean correlation between the simulated profiles can still float near the mean correlation value between real profiles. Consider the following example. Figure 46 shows a set of mean primer shear profiles from a Glock in the database which were relatively consistent (visually) from cartridge case to cartridge case.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

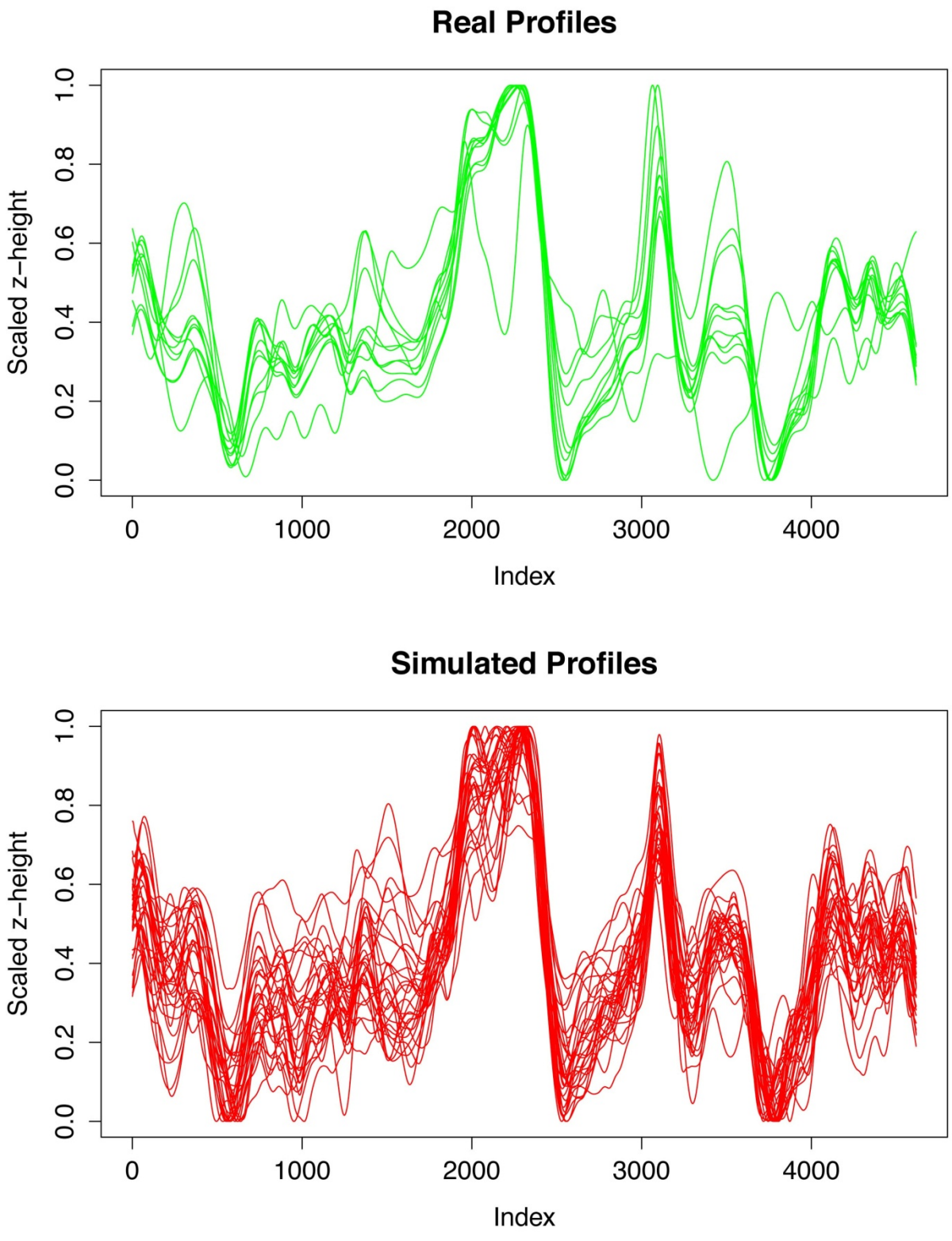


Figure 46. Glock primer shear mean profiles labeled “pat1” in the database.

The eleven profiles shown in green are real. The thirty profiles shown in red are simulated from the real. Overall visual inspection indicates that the simulated profiles are plausible variations of the real profiles. The mean correlation coefficient similarity score between the real profiles was 0.82. The mean correlation coefficient between the simulated profiles is 0.84. The mean correlation coefficient between the simulated and real profiles is commensurate with these, 0.83, numerically indicating their validity as plausible representations of the real profiles.

In somewhat of a contrast to Figure 46, Figure 47 shows a set of mean primer shear profiles from a different Glock in the database. This set is visually less consistent from cartridge case to cartridge case. Again the green profiles are real (twelve) and the red (thirty) are simulated.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

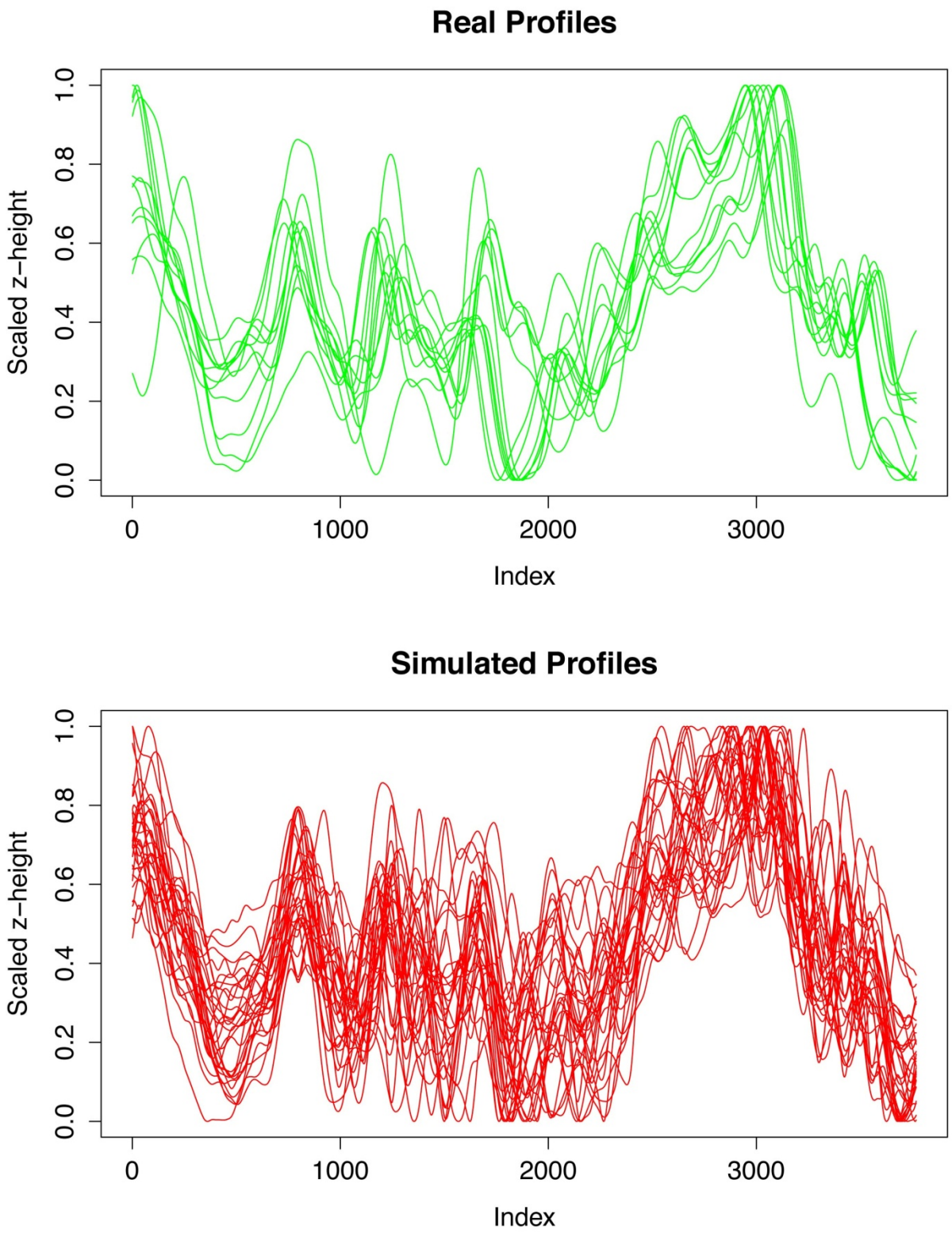


Figure 47. Glock primer shear mean profiles labeled “lauren” from the database.

The mean correlation coefficient similarity score between the real profiles was 0.70. The mean correlation coefficient between the simulated profiles is 0.68. The mean correlation score between the simulated and real profiles is 0.70. Note that the minimum correlation cut-off for the simulated profiles was set to 0.5 for both Glocks. This routine of validation/plausibility checks for all sets of simulated profiles was used for both statistical studies (cartridge case primer shears and screwdriver striation patterns) discussed in section 2.

1.6 R software and statistical analysis scripts

Software has been written to open and analyze the surfaces in the database with the R statistical analysis suite (http://www.r-project.org/). R software has also been written to perform conformal prediction theory computations for support vector machine classifiers and k-nearest neighbor classifiers. The R scripts used for PCA, CVA, SVM and error rate estimations used in this project as well as all of the above R software are available on the project's website discussed above (http://toolmarkstatistics.no-ip.org/). Note that, at the time of writing this report, the analysis scripts can only be used with 2D toolmark *profiles*. These profiles however can be generated from the toolmark surfaces with the R software discussed above. Profiles available for immediate use/exploration by users are mean, median, mode, user selected and random. Spline functions can be fit to the profiles so that first and second derivatives of the profiles can be used in the statistical analysis scripts if the user desires.

Arbitrary profiles of the surface can also be transformed into "barcode"-like signals. The scheme fits a cubic spline function to the selected profile. The fit function is then differentiated to find its critical values. It is at the critical value where the "peaks" and "valleys" of the profile occur. When a critical value is found (to within a user-set tolerance) a vector of 1s is used to represent a peak and a vector of -1s is used to represent a valley. The widths of these motifs (i.e. peaks and valleys) are computed by finding half the distance between a motif (peak or valley) and the motif immediately to its left (valley or peak). Distances from the extreme left or right of the profile, halfway to the closest motif, are recorded as vectors of zeros. An example barcode is shown in Figure 48.

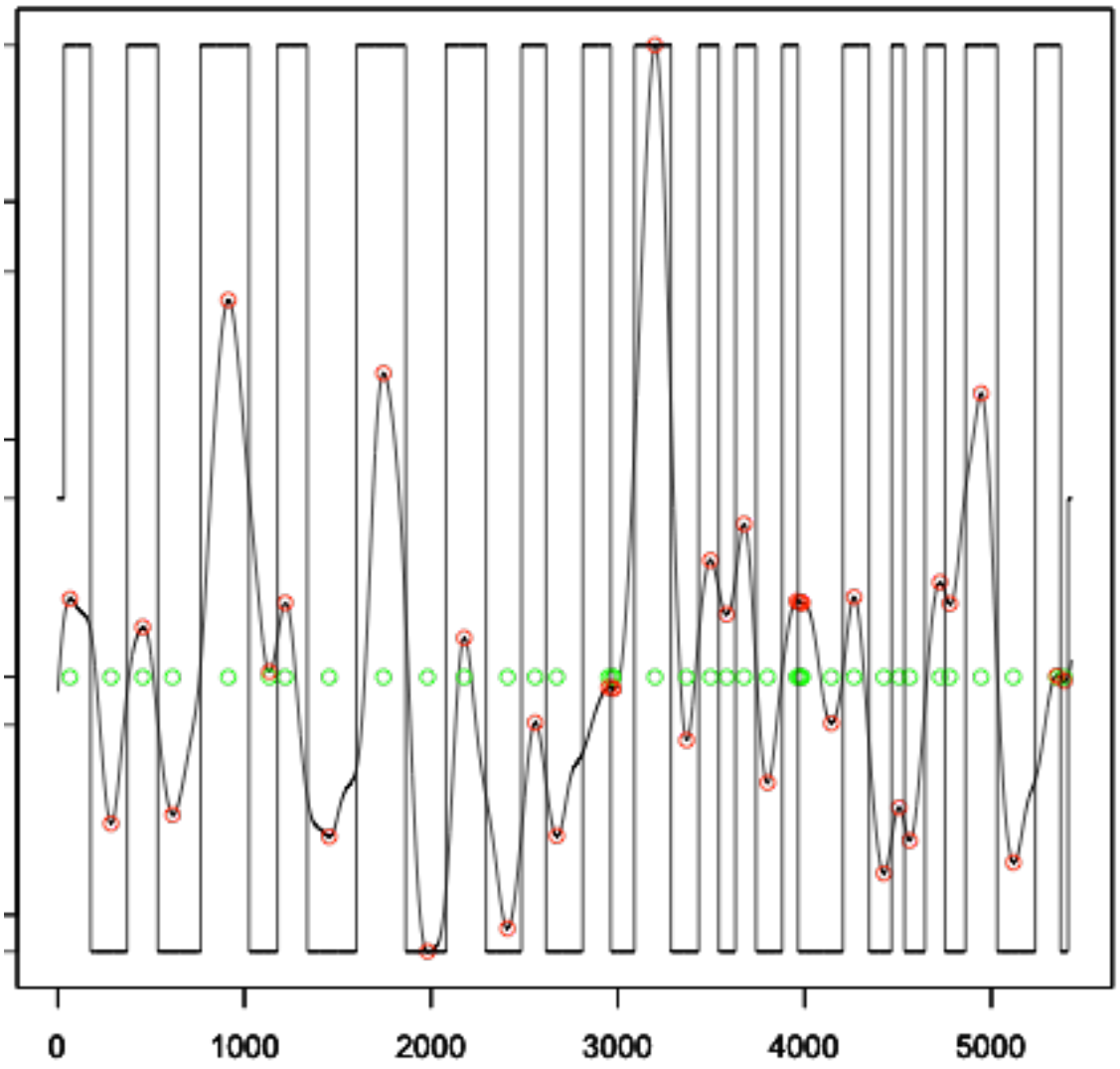


FIGURE 48. Barcode with corresponding surface profile overlaid. The red circles mark the peaks and valleys found by barcode generating algorithm. The green circles mark zeros of the profile's first derivative.

The profile corresponding to the barcode appears superimposed. The red circles are the peaks and valleys picked out by the algorithm. The green circles mark the actual critical points of the profiles derivative, found by the algorithm.

The barcode algorithm basically behaves as an "edge detector". Thus it can be used on all the profiles making up a surface. A composite of the barcodes is a representation of where the major striation pattern appears. An example of the barcode algorithm applied to an entire striation pattern is shown in Figure 49.

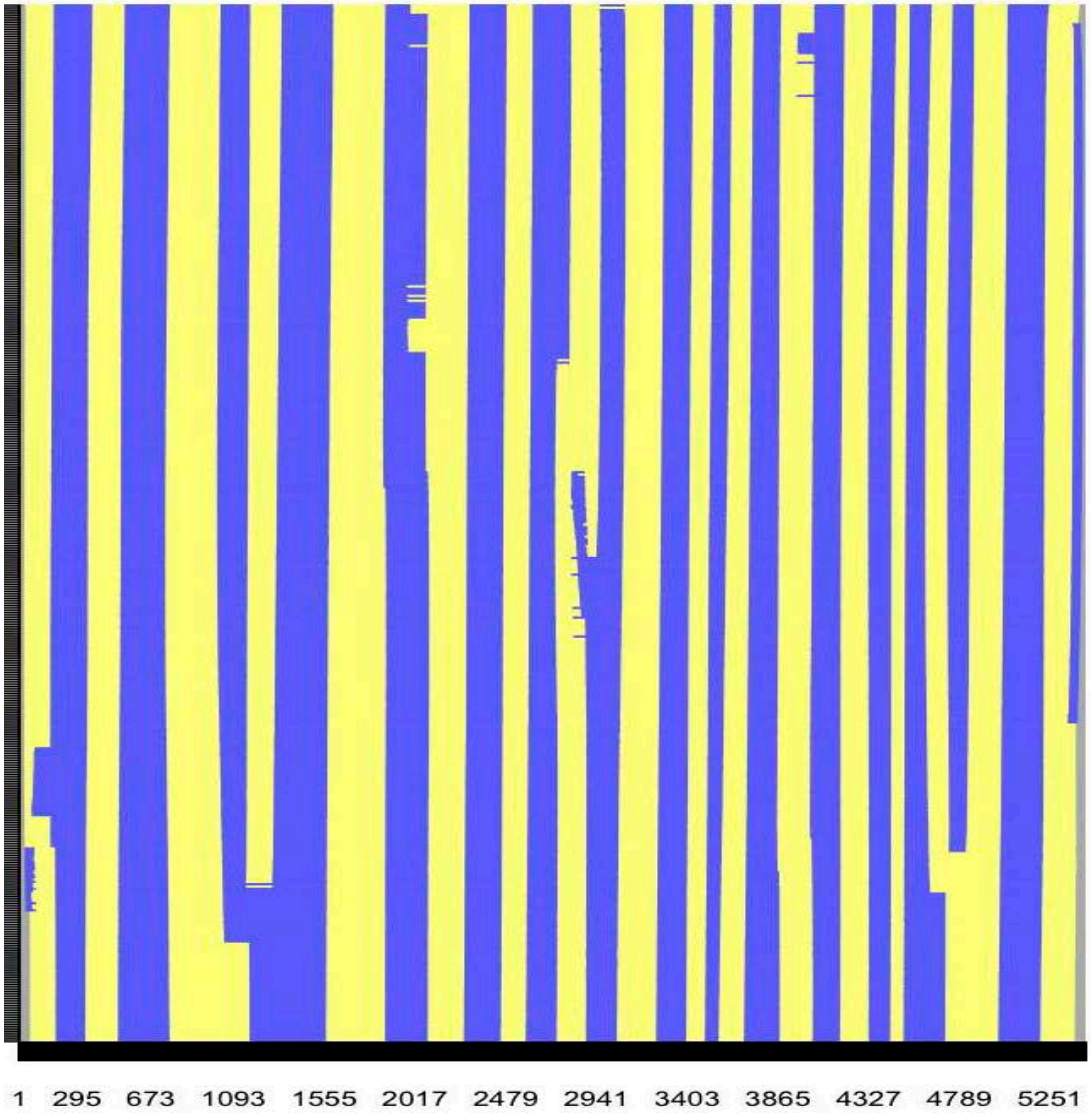


FIGURE 49. Edge representation of a striation pattern surface. Surface is a primer shear. The image is built up form profile barcodes corresponding to profiles making up the surface.

Finally, if the mode is taken down each of the columns of Figure 49 one obtains the "mode-barcode" for the striation pattern (cf. Figure 50). Profile barcodes, surface edges or mode-barcodes can all be input into the statistical analysis software as physical representations of the tool mark.

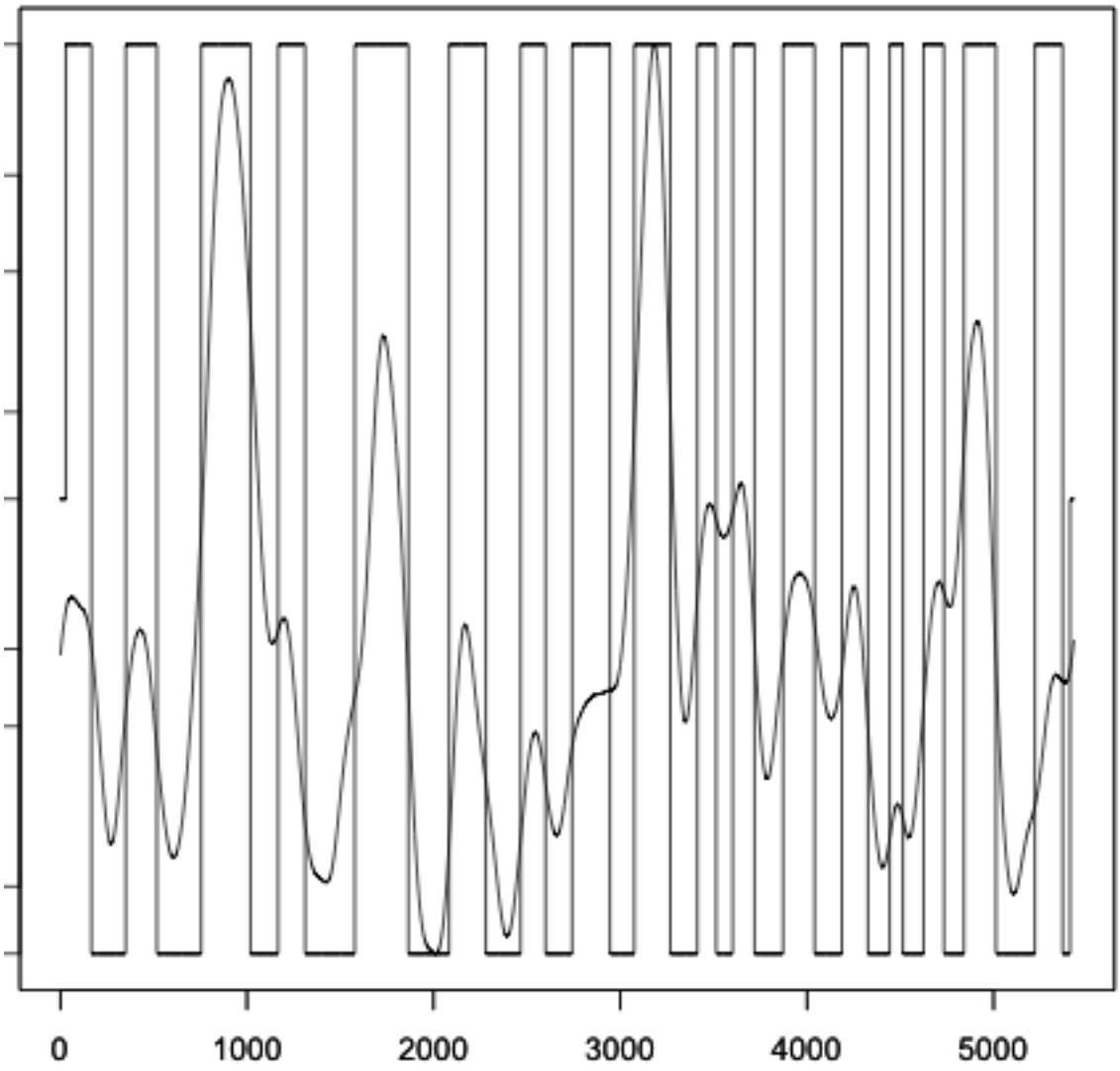


FIGURE 50. Mode-barcode representation of primer shear striation pattern. The mean profile of the primer shear surface is superimposed for reference.

As can be seen from the above discussion users of the R software developed in the course of this project have a great deal of flexibility in the way in which profiles extracted from striation surfaces can be represented. However, as a matter of note, the statistical studies presented in the next section, take a standard approach to striation pattern representation. Only mean waviness profiles are used in the analysis.

2. Statistical Analyses

2.1 Glock primer shear striation patterns

One hundred sixty-two primer shear striation patterns generated by breach face shear on Glock 19 fired cartridge cases (twenty-four guns) were statistically analyzed. Enough field of view "tiles" were scanned so that one end of the striation pattern to the other was scanned. The resulting tiles were stitched together and very noisy portions at the ends were cropped out. Figure 51 shows four (un-aligned), ~2000 μm wide, primer shear striation patterns on cartridge cases fired from Glock #2 after the noisy ends have been cropped.

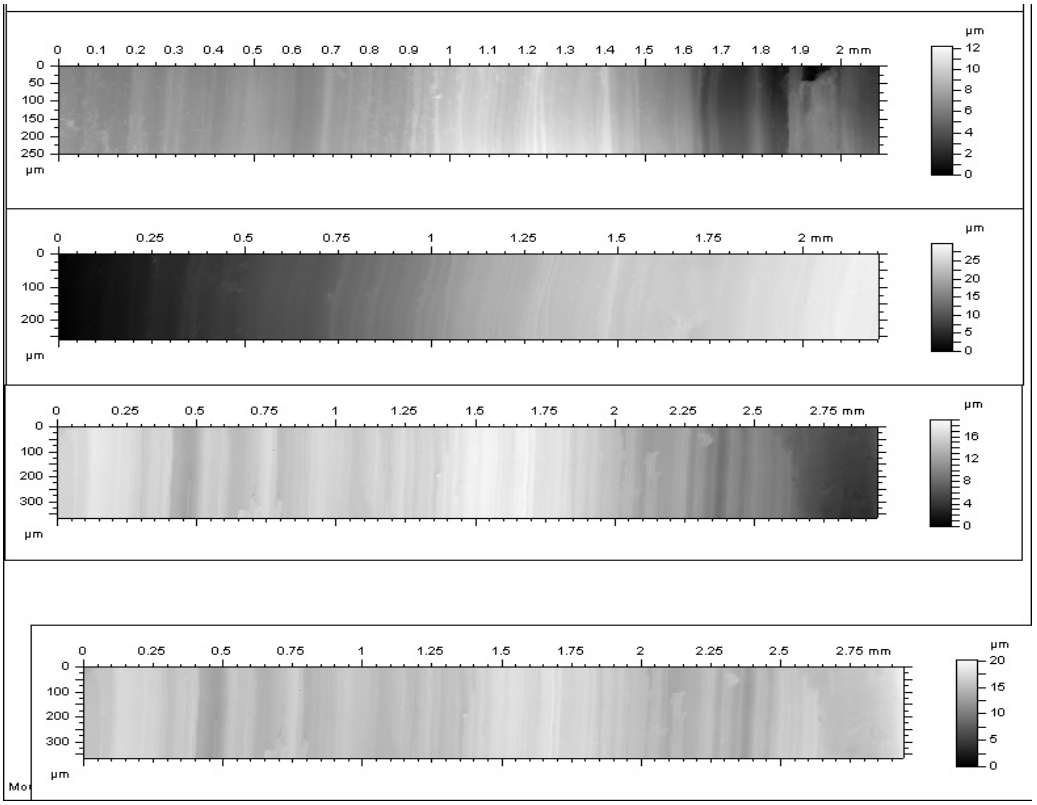


FIGURE 51. Four (unaligned) primer shear striation patterns on cartridge cases fired from the same Glock.

Typically, scanned surfaces require first ("leveling" or linear trend removal), second or third order polynomial removal. It was determined that, all scanned striation patterns would be subjected to third order polynomial "form" removal. Third order was chosen because it removed a good deal of the gross "warping" apparent in all of the primer shear patterns. Generally speaking, third order form removal took out about as much warp as fourth order but more than second order. An example of a form removed primer shear pattern appears below in Figure 52.

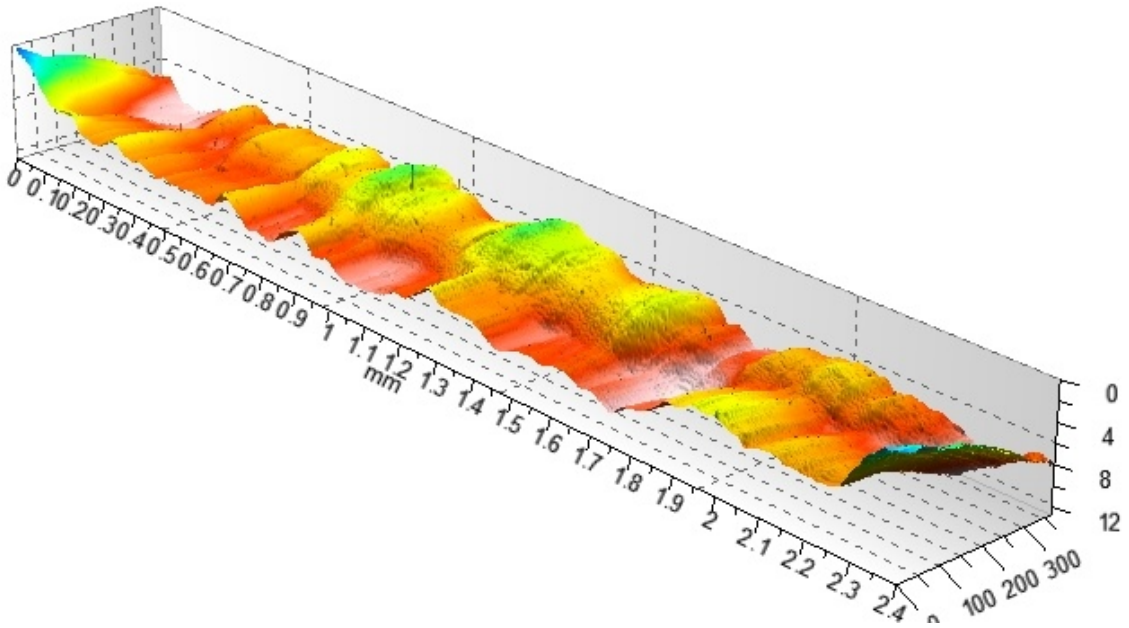
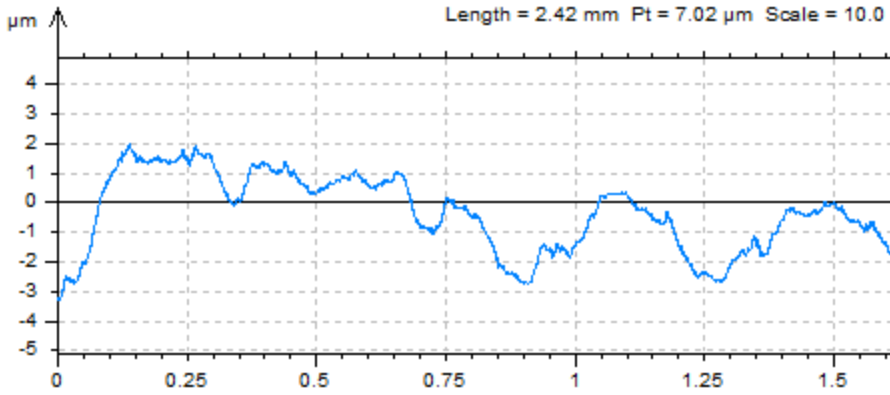


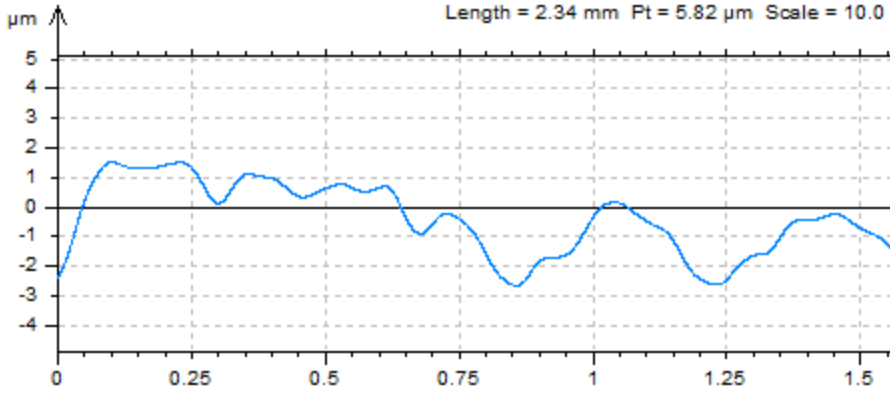
FIGURE 52. Detrended (form removed) primer shear striation pattern on cartridge case fired from Glock #3, cartridge case 2.

The resulting detrended surfaces were then filtered into roughness and waviness components using the Gaussian filter and $\lambda xc = \lambda yc = 0.025$ mm cutoff values (Muralikrishnan 2009, Chu 2010). An example of the output for these computations is shown in Figure 53.

Mean total profile:



Mean "waviness" profile:



Mean "roughness" profile:

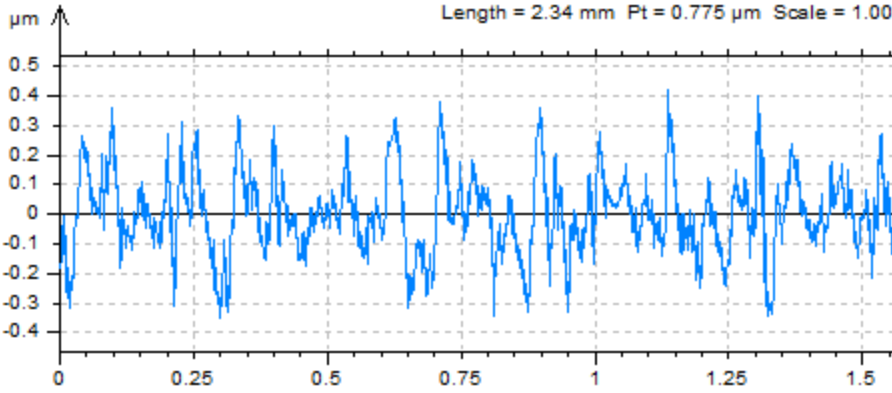


FIGURE 53. Mean total profile, mean waviness profile and mean roughness profile from the detrended breach face shear striation pattern of Glock #3.

For striation patterns where the "lines" run vertically down the surface in the y-direction, there is an obvious high redundancy of information (cf. Figure 54).

FIGURE 54. Form removed primer shear from Glock #3, cartridge case 2 viewed from above. The dimension of the surface is 2.1 mm x 250 μm but scaled to appear square for display purposes. Note the prominent striation pattern running from the top to the bottom of the figure.

Thus, mean profiles were used because they provide good overall summaries while having file sizes that were manageable. It is current standard practice to use the mean profile as input into the statistical discrimination algorithms instead of the entire surface (Vorburger 2007, Bachrach 2002, Chu 2010, Bachrach 2010, Chumbley 2010, Faden 2008).

Next, the roughness and waviness components from each surface were closely examined. From the profile plots it was clear that almost all of the "line structure" in the striation pattern, apparent in a comparison microscope, was contained in the waviness surfaces across all of our samples. Therefore, only extracted waviness was processed through the statistical regime discussed below. The waviness component of each mean profile was loaded into the R statistical program for further processing (R Core Development Team 2009).

The profiles were first normalized such that their z-heights ranged between 0 and 1 (as discussed above). The mean profiles require alignment, i.e. registration, to be processed as maximally similar multivariate feature vectors. Using the max CCF methodology the profiles were aligned. Within a collection of cartridge cases from a particular Glock, the longest profile

was chosen as a reference. The remaining profiles are then maximally aligned with respect to the reference profile. Next, the mean profiles between guns were aligned. This was done via the group-mean-profile procedure.

In order to fully exploit the Hamby-Thorp portion of the data set it was needed to increase the number of mean profile replicates per Glock. Plausible replicates were generated using the wavelet simulator discussed. The wavelet expansion was used because it offers a principled multi-scale description of surface morphology and allows for statistical analysis to be carried out efficiently (Maksumov 2004, Reizer 2010). It was decided to balance the whole data set and simulate enough profiles so that each gun is represented by 30 mean profiles. The real data set consisted of 162 collected profiles taken from a subset of 24 different Glocks in the database. After simulations were carried out the data set size was 720 profiles (30 total for each Glock). Criteria for keeping simulated profiles was a correlation of greater than 0.5. This low bound to "similarity" was chosen to generate a challenging set of profiles to discriminate. Profiles were simulated in blocks of ten. The growing sets of group profiles (both real and simulated) were fed back into the simulator as input until the set reached 30 acceptable profiles ((again, criteria for an acceptable simulated profile was a correlation "similarity score" greater than or equal to 0.5 with the real profiles).

The augmented data set (720 profiles) was renormalized and registered. The profiles were stacked together and trimmed to the same length (4233 points or about 1.7 mm), forming a data matrix. For the profiles of our test set, the data matrix had dimensions 162 waviness profiles by 3618 points per profile. The z-heights of the profiles were renormalized and realigned between groups.

The resulting data matrix of profiles contained 3618 columns. The matrix was first mean centered. Two classification regimes were employed, PCA-SVM and PCA-CVA-SVM. Hold-one-out cross validation (HOO-CV) was used to choose a reasonable sized dimension in which to carry out error rate estimation of the classification methodology.

One-vs.-one multiclass support vector machines (SVMs) (linear kernel, penalty parameter $C = 1$) were then applied to the PCA dimensionally reduced data space (Vapnik, 1998). The lowest HOO-CV error, 4.1%, occurred first at PCA dimension 22 (99.5% variance retained). The relatively high HOO-CV error rate of 4.1% was not surprising as the simulated portion of the data set was constructed to be challenging to classify.

Using 2,000 bootstrap resampling iterations, 22D PCA-SVM produced a refined bootstrap error rate estimate of 2.5%. The 95% confidence interval for the error rate, also determined by bootstrapping, was 1.3% to 3.2%. Figure 52 shows the bootstrapped error rate optimisms for SVM classification, which was used to compute the confidence interval around the error rate.

The classification results with the PCA-CVA-SVM analysis were better. The data's dimension was first reduced to 31-dimensions, using PCA (99.9% variance retained), removing most of the redundancy. HOO-CV was then applied with CVA-SVM to choose a final dimension size. A 2.5% HOO-CV error rate (the lowest HOO-CV error rate) was found first in a 15D CVA space. The refined bootstrap error rate (2000 resampling iterations) on the resulting 15D PCA-CVA-SVM discrimination model was 1.1%, with a 95% confidence interval of 0.6%-1.8%, using 2,000 bootstrap resampling iterations. Thus both classification regimes produced low error identification error rate estimates. PCA-CVA-SVM not surprisingly preformed a bit better as CVA attempts to project the data into a more group clustered data space.

Conformal prediction theory using the SVM classifier was then applied to the PCA-CVA projected data set. A random sample of 25% of the data set, (180 profiles, both real and simulated) was selected as an unknown test set. The identification confidence for the algorithm was set to 95%. As should be the case, the empirical (i.e. finite sized data set) error rate was 5.3%, very close to the theoretical long run limit of 5%. In and of itself, the CPT error rate is not very interesting as the method is guaranteed to an error ≤ 5% of the time in the long run, at the 95% level of confidence (Vovk 2005). The efficiency of the "correct" identifications must be gauged as well. A "correct" answer (confidence region) output by a CPT based classifier need only contain the correct identification label. Technically "correct" confidence regions can contain more than one, or even, all possible tool identifications. Thus the numbers of multi-label and "uninformative" (containing all labels) answers output by the CPT algorithm are important in assessing the classifiers performance. For this dataset no "uninformative" confidence regions were produced. The multi-label confidence region rate was 24%. This rate was a bit high, however the number of labels in these multi-label regions never exceeded two.

2.2 Screwdriver striation patterns

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Each toolmark was scanned using the 50x-long working distance objective (0.6 NA). Under the confocal microscope, the left edge of each toolmark was denoted to provide a point of reference for the section of striations. Once the left edge was marked, we moved in (towards the right) 1,000 µm. The purpose of this was principally for scan time savings. The confocal microscope collects slices of information in the z-direction. Because the left edge of the striation patterns are generally so much higher than the rest of the mark, scanning from the left edge would have increased the scanning time dramatically, with relatively little gain of information. From this point (1,000 µm in from the left edge), seven sections were selected so that there was some overlap for the confocal microscope software to stitch together. The noise-cut method used for all the toolmarks was Z-interpolation. This is the same process that was used for cartridge cases.

Third order polynomial form removal was used to level the images, as with the Glock cartridge casings (Figures 55 and 56).

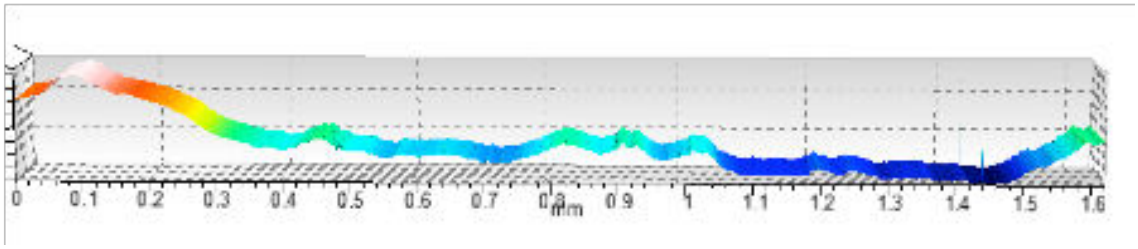


FIGURE 55. Three-dimensional image of a striation toolmark before form removal

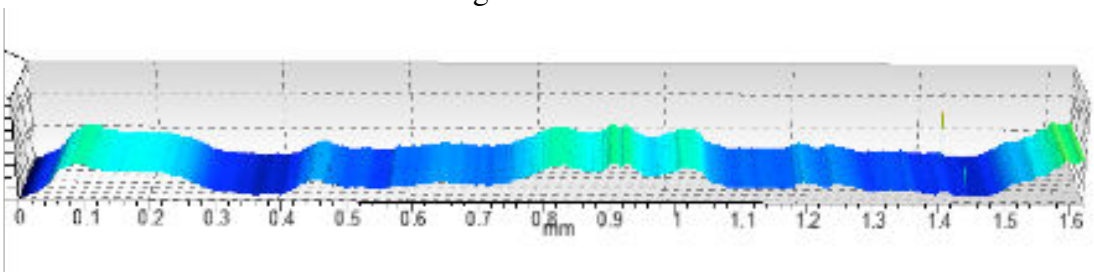


FIGURE 56. Three-dimensional image of a striation toolmark after form removal

The resulting detrended surfaces were then filtered into roughness and waviness components using the same parameters as were used with the primer shear study. An example of the output for these computations is shown in Figures 57 and 58.

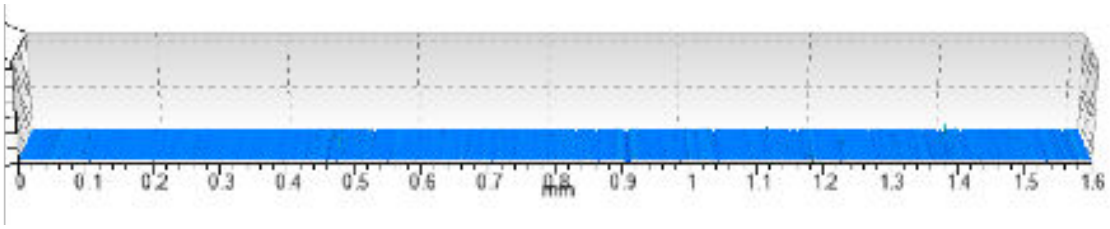


FIGURE 57. Three-dimensional image of a roughness profile.

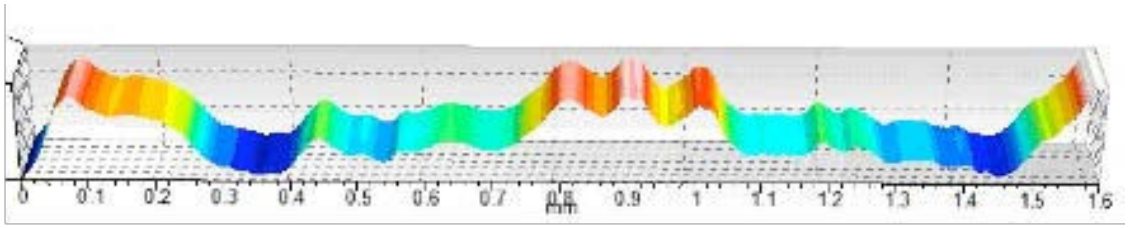


FIGURE 58. Three-dimensional image of waviness profile.

Just as for the primer shear study, mean waviness profile was focused on to obtain information about the general shape of the toolmark. Mean waviness profiles were computed from all waviness surfaces in R. Figure 59 shows one such mean waviness profile.

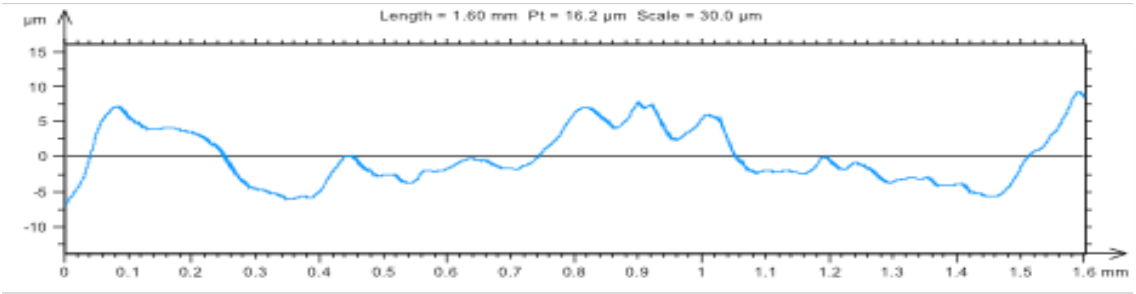


FIGURE 59. Mean profile of waviness profile.

Prior to statistical analysis, the mean profiles required rescaling and registration. The same procedures were followed as for the primer shear study. First, the algorithm rescaled each profile so the height information would be on a scale of 0 to 1. In essence, the lowest point on the profiles would be marked 0, and the highest point in the profiles would be marked 1. The algorithm also aligned the profiles within and between each group as was done in the cartridge case primer shear study.

Eight Craftsman® and ten Iron Bridge® ¼-inch slotted screwdrivers were examined and labeled. Furthermore, each side of the screwdriver tip was labeled either A or B. As a result, the toolmark striation patterns of thirty-six different screwdrivers were analyzed. The toolmark

medium used in this study was lead because, as explained previously, lead is soft enough to make test marks without damaging the tool's working surface and produced less noisy images with the confocal microscope. Five replicate striation patterns for each side of each screwdriver were produced for a data set size of 180 patterns.

As was the case for the primer shear study above, PCA-SVM and PCA-CVA-SVM were exploited to assess the striation pattern identification error rate estimates. Unlike the primer shear study however, no simulated data was used. For the PCA-SVM computations, HOO-CV was used to find a lowest error rate (3.3%) first at 19 PCA dimensions (99.0% variance retained). Using the data set projected into 19D PCA space, the refined bootstrapped error rate estimate was found to be 2.6% with 2,000 resampling iterations. The 95% confidence interval for the error rate, as determined by the bootstrap optimism histogram was 0.6% – 6.1%. This 95% confidence interval for the error rate is wide and was a bit surprising at first. Visual examination of mean profiles that were incorrectly identified in the initial HOO-CV process were relatively straight forward to pair with the screwdriver that created it. We have observed this behavior in past machine learning projects when too few replicate patterns per group are used in the fitting/validation process. Though five replicates per group is generally accepted practice in the chemometrics community (where many of the machine learning algorithms used here are also used), we decided to examine if the confidence interval around the error rate would narrow if more replicate profiles per screwdriver were included. A simulation run was performed on the real screwdriver profiles (180) using the same operating parameters as were used for the cartridge case primer shear profiles. Twenty-five profiles were simulated for each screwdriver brining the data set size up to 1080 patterns (30 profiles per screwdriver). Hold-one-out cross-validation indicated that 13 PCA dimensions would yield a reasonable space in which to carry out bootstrap error rate estimation. The refined bootstrap error rate estimate for 13D PCA-SVM discrimination model was 0.41%. The refined bootstrap 95% confidence interval around the error rate estimate, [0.09%,0.83%] was indeed narrower using the augmented data set with thirty replications per screwdriver.

For the PCA-CVA-SVM computations the data was first reduced to 99.9% variance with PCA (31D). HOO-CV found the lowest error rate (0%) first at 7 CVA dimensions. Again, as was the case for the primer shear study, these results using CVA *combined* with PCA are much better than the PCA based results alone. Also the model is much simpler (i.e. lower in dimension). The

refined bootstrap identification error rate estimate using this model was found to be 0.6% with a corresponding confidence interval of 0% to 2.8%. Again 2000 resampling iterations were used and the confidence interval was determined by the bootstrap optimism histogram. Interestingly CVA did not seem to improve the discrimination ability of the model much when used on the augmented screwdriver dataset. 21D PCA-8D CVA-SVM (PCA dimension was 99.9% variance retained, CVA dimension determined with HOO-CV) refined bootstrap error rate was 0.40% with a 95% confidence interval of [0.09%, 0.74%]. This is basically the same as the PCA-SVM error rate estimate to within small statistical fluctuation.

Conformal prediction theory using the SVM classifier was applied to the PCA-CVA projected data set at the 95% level of confidence. A random sample of 40% of the data set, (72 screwdriver striation pattern mean profiles) was selected as an unknown test set. The empirical error rate was 1.4%, which seems a bit optimistic at the 95% confidence level and probably reflects a statistical fluctuation due to the smaller size of data set. A study with the extra simulated profiles is underway. The rate at which multi-labeled confidence regions were produced was very low at 1.4%. Again, as was the case for the primer shear study, no multi-labeled confidence region had more than two labels (in this study there was only one multi-label region produced). No uninformative confidence regions appeared.

IV. Conclusions

1. Discussion of findings

This research outlines a set of objective and testable methods to associate toolmark impression evidence with the tools and firearms that generated them. Striation patterns are the focus. The results complement previous univariate based toolmark discrimination studies and are consistent with and buttress the qualitative conclusions of the forensic firearms and toolmark examination community.

Three dimensional confocal microscopy, surface metrology and multivariate statistical methods lie at the heart of the approach presented in this project. Through the studies described, practitioners can see how a surface metrological-statistical scheme can provide an investigative aid and estimate algorithmically based identification error rates for firearm and toolmark comparisons.

Striated toolmarks were collected from screwdrivers and chisels. Striated and impressed toolmarks were collected from cartridge cases. Quantitative confocal images of the surface topographies of all toolmarks examined have been included in a database. The forensic research and practitioner community can access information in the database at the website URL: http://toolmarkstatistics.no-ip.org/ . Data from this project is being made available for further research by the academic and practitioner communities and for interested practitioners to construct images for court exhibits. Several pieces of software, including software for visualization of/measurement on the toolmark surfaces in the database, were generated in the course of the project. All software and R statistical analysis scripts used are available on the website.

The reasonably complete striation patterns from screwdrivers and the primer shear from 9mm Glock fired cartridge cases could be summarized as multivariate feature vectors in the form of mean profiles. These mean profiles were used with standard multivariate machine learning methods in order to estimate identification error rates from such an algorithmic regime. A combination of principal component analysis (PCA), canonical variate analysis (CVA) and support vector machines (SVM) proved most effective for accomplishing this task with low identification error rate estimates, generally ~1% with 95% confidence intervals ~[0%,3%]. Bootstrap resampling was used to estimate these identification error rates and confidence intervals. A summary of these results appears in table 1.

Table 1. Summary of statistical analysis results.
[a]Fitted discrimination model dimensions. See text.
[b]Hold-one-out cross-validation error rate estimate. Used to choose model dimensions.

[c]Refined bootstrap error rate estimate. Square brackets are 95% confidence intervals around error rate estimates.

	# Patterns	Dimensions[a]	HOO-CV[b]	Bootstrap[c]
Cartridge Case PCA-SVM	720 (162 real, 558 simulated)	22	4.1%	2.5% [1.3%,3.2%]
Cartridge Case PCA-CVA-SVM	720 (162 real, 558 simulated)	31,15	2.5%	1.1% [0.6%,1.8%]
Screwdriver PCA	180 (180 real)	19	3.3%	2.6% [0.6%,6.1%]
Screwdriver PCA-CVA-SVM	180 (180 real)	31,7	0%	0.6% [0%,2.8%]
Screwdriver aug.[d] PCA	1080 (180 real, 900 simulated)	13	0.56%	0.41% [0.09%,0.83%]
Screwdriver aug.[d] PCA-CVA-SVM	1080 (180 real, 900 simulated)	21,8	0.56%	0.40% [0.09%,0.74%]

[d]Augmented with simulated patterns, based on the real patterns.

Conformal prediction theory (CPT) was used to assign rigorous levels of confidence to all PCA-CVA-SVM toolmark identifications. Such levels of confidence can help a judge or jury assess the quality of an algorithmic association of a tool to a toolmark. The CPT classifiers proved to be reasonably efficient, producing only small multi-label confidence regions and only at relatively low rates. Uninformative confidence regions were not observed. Note that bootstrapping methods, PCA, CVA, SVM and CPT have very few underlying assumptions built in, and this was a major reason why they were chosen. This is a major advantage to their use in a courtroom setting where their results will be far more likely to stand up to adversarial scrutiny and be less open to attack.

Unfortunately the three-dimensional impressed toolmarks and the "patchy" chisel striation patterns proved too complicated for our current suite of developed software to analyze at this time. (This is another reason why we are making the data collected for the project available to the wider research community.) Development of open source software for the machine learning analysis of complete three-dimensional impression patterns and incomplete toolmarks will be the subject of future research.

That said, practitioners could apply the machine learning regime presented here, to any set of reasonably complete striation patterns (i.e. of reasonable quality), and generate tool-toolmark association error rate estimates and identifications at a chosen level of confidence. Given the findings of the studies presented above as well as those of previous univariate based projects, their results will no doubt be consistent with the theory that no two striation patterns derived from different tools are identical.

2. Implications for policy and practice

Impression evidence left at crime scenes is indispensable and cannot be allowed to become inadmissible in court. Access to our database and methodology will provide the law enforcement community with data and standardized methods to make quantitative comparisons. Computational pattern recognition is already widely used in industry, including chemical engineering, audio/visual engineering, mail and product sorting, computer security, marketing, etc. It is absolutely critical that the forensic community take advantage of the enormous potential of pattern recognition and the computing power available today. Adopting these statistical techniques for impression pattern comparison will yield standardized and efficient protocols as well as reproducible, independently verifiable, fair and accurate conclusions. In practice, the methodologies and data developed for this project gives forensic toolmark examiners literature and tools that will preserve this valuable evidence's admissibility in court.

3. Implications for further research

The results of toolmark pattern classification resulting from other 3D microscopy technologies needs to be extensively compared. Do interferometric and focus variation type microscopies provide the same topographies as confocal microscopy (i.e. to within an acceptable level of random error)? Some studies have been done, in particular at NIST (Song 2006). Still, more work is needed.

Work needs to be done to extend the multivariate machine learning methodology developed here to full 3D surface data sets. Our preliminary computational experiments suggest that registration of a large number of 3D surfaces for direct feature vector computation will be (unavoidably) extremely intensive computationally and not apt to run well on current desktop computers. Thus rotation and translation invariant features for use with multivariate classification techniques need to be explored for performance enhancement.

Open source, platform independent interactive imaging/ preprocessing/ measuring programs for 2D and 3D toolmark analysis need to be developed for the forensic toolmark examination community. Software in this project has taken the first steps towards that goal. Still however, work needs to be done to standardized noise removal in acquired 3D topographies. GPUs can be harnessed for heavy numerical computations involved in noise removal and surface filtering operations and wavelet computations. Also, an R port of the wavelet based profile

simulator developed in Mathematica for this project needs to be accomplished so that anyone interested can use it.

Bachrach *et al.* and Chumbley *et al.* have shown that the angle at which a "scraping" tool is used to create a striation pattern is crucial to computing the toolmark's identity. Future research can and will focus on varying the angle of attack when generating striation patterns. Other real world toolmark media such as bone must also be carefully considered.

Recently genomics has been able to leverage empirical Bayes methods due to the large size of the data sets routinely encountered in that field (Benjamini 1995, Efron 2010, Kall 2008a,b). Considering the number of yes/no decisions a machine learning task makes when classifying an unknown against several possible groups, empirical Bayes methods may be applicable to toolmark analysis as well. Specifically, local false discovery rates can provide another "match quality" measure in a way similar to CPT.

V. References

Association of Firearm and Toolmark Examiners. (1998). Theory of identification as it relates to toolmarks. *AFTE Journal*, 30(1), 86 – 88.

Bachrach, B. (2002). Development of a 3D-based automated firearms evidence comparison system. *Journal of Forensic Sciences*, 47(6), 1 – 12.

Bachrach, B., Jain, A., Jung, S., Koons, R. D. (2010). A statistical validation of the individuality and repeatability of striated toolmarks: Screwdrivers and tongue and groove pliers. *Journal of Forensic Sciences*, 55(2), 348 – 357.

Banno A. (2004). Estimation of Bullet Striation Similarity Using Neural Networks. *Journal of Forensic Sciences*, 49(3), 1-5.

Banno, A., Masuda, T., & Ikeuchi, K. (2004). Three dimensional visualization and comparison of impressions on fired bullets. *Forensic Science International*, 140(3), 233 – 240. doi:10.1016/j.forsciint.2003.11.025.

Benjamini, Yoav; Hochberg, Yosef (1995). Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *J Royal Statistical Soc Ser B,* 57(1): 289–300.

Biasotti A., Murdock J., & Moran, B. R. (2008). *Firearms and toolmark identification*. In: Faigman D., Kaye D., Saks M., Sanders J., (Eds.). Modern scientific evidence: the

law and science of expert testimony (Vol. 4). Thomson Reuters/West.

Biasotti, A. (1959). A statistical study of the individual characteristics of fired bullets. *Journal of Forensic Sciences*, 4(1), 34 – 50.

Brinck, T. B. (2008). Comparing the performance of IBIS and BulletTRAX-3D technology using bullets fired through 10 consecutively rifled barrels. *Journal of Forensic Sciences,* 53(3), 677 – 682.

Brundage, J. (1998). The identification of consecutively rifled gun barrels. *AFTE Journal*, 30(1), 438 – 444.

Buckleton J., Nichols R., Triggs C., Wevers G. (2005) An Exploratory Bayesian Model for Firearm and Toolmark Interpretation. *AFTE Journal*, 37(4), 352-61.

Bunch, S. G. (2000). Consecutive matching striation criteria: A general critique. *Journal of Forensic Sciences*, 45(5), 955 – 962.

Burd, D. Q., & Gilmore, A. E. (1968). Individual and class characteristics of tools. *Journal of Forensic Sciences*, 13(3), 390 – 396.

Champod, C., Baldwin, D., Taroni, F., & Buckleton, J. S. (2003). Firearm and toolmarks identification: The Bayesian approach. *AFTE Journal*, 35(3), 307 – 316.

Chu,W., Song, J., Vorburger, T., Yen, J., Ballou, S., & Bachrach, B. (2010). Pilot study of automated bullet signature identification based on topography measurements and correlations. *Journal of Forensic Sciences*, 55(2), 341 – 347.

Chumbley, L. S., Morris, M. D., Kreiser, M. J., Fisher, C., Craft, J., Genalo, L. J., Davis, S., Faden, D., & Kidd, J. (2010). Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm. *Journal of Forensic Sciences*, 55(4), 953 – 961.

Cork D.L., Rolph J.E., Meieran E.S., & Petrie C.V. (2008) Ballistic Imaging. Washington: National Academies Press.

Cowles, D. L., & Dodge, J. K. (1948). A method for comparison of toolmarks. *Journal of Criminal Law and Criminology*, 39(2), 262 – 264.

Daubert v. Merrell Dow Pharmaceuticals., Inc., 509 U.S. 579, 113 S.Ct. 2786 (1993).

De Forest, P. R., Gaensslen, R. E., & Lee, H. C. (1983). *Forensic science: An introduction to criminalistics.* U.S.A.: McGraw-Hill, Inc.

De Kinder, J., & Bonfanti, M. (1999). Automated comparisons of bullet striations based

on 3D topography. *Forensic Science International*, 101(2) 85 – 93.

Duda R.O., Hart P.E., Stork D.G. (2001) *Pattern Classification*. 2 ed. New York: Wiley.

Du Pasquiera, E., Hebrardb, J., Margota, P., & Ineichen, M. (1996). Evaluation and comparison of casting materials in forensic sciences applications to toolmarks and foot/shoe impressions. *Forensic Science International*, 82(1), 33 – 43.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. New York: Cambridge University Press.

Faden, D., Kidd, J., Craft, J., Chumbley, L. S., Morris, M., Genalo, L., Kreiser, J., & Davis, S. (2007). Statistical confirmation of empirical observations concerning toolmark striae. *AFTE Journal*, 39(3), 205 – 214.

Frye v. United States. 293 F. 1013 (D.C. Cir 1923).

Fu, S., Muralikrishnan, B. and Raja, J. (2003). Engineering Surface analysis with different wavelet bases. J. Manuf. Sci. Eng.,125(4), 844-852.

Fukunaga K. (1990) *Statistical Pattern Recognition*. 2 ed. San Diego: Academic Press.

General Electric Company, et al. v. Robert K. Joiner et al. 522 U.S. 136 (1997).

Geradts Z., Bijhold J., Hermsen R., Murtaugh F. (2001). Image Matching Algorithms for Breech Face Marks and Firing Pins in a Database of Spent Cartridges of Firearms. *Forensic Science International*, 119, 97-106.

Geradts, Z., Keijer, J., & Keereweer, I. (1994). A new approach to automatic comparison of striation marks. *Journal of Forensic Sciences*, 39(4), 974 – 980.

Greene, R. S., & Burd, D. Q. (1950). Special techniques useful in toolmark comparisons. *Journal of Criminal Law and Criminology*, 41(4), 523 – 527.

Grodsky, M. (1999). Elmer's glue-all: A low cost toolmark casting medium. *Journal of Forensic Identification*, 49(2), 117 – 121.

Grzybowski, R. A., & Murdock, J. E. (1998). Firearm and toolmark identification – Meeting the Daubert challenge. *AFTE Journal*, 30(1), 3 – 14.

Grzybowski, R. A., Miller, J., Moran, B., Nichols, R., & Thompson, R. (2003). Firearm/toolmark identification: Passing the reliability test under federal and state evidentiary standards. *AFTE Journal*, 35(2), 209 – 241.

Hamby, J. E., Brundage, D. J., & Thorpe, J. W. (2009a). The identification of bullets fired from 10 consecutively rifled 9mm Ruger pistol barrels: A research project involving 507 participants from 20 countries. *AFTE Journal*, 41(2), 99 – 110.

Hamby J & Thorpe J. (2009b) The Examination, Evaluation and Identification of 9mm Cartridge Cases Fired from 617 Different GLOCK Model 17 & 10 Semiautomatic Pistols. *AFTE Journal*, 41(4), 310-324.

Howitt, D., Tulleners, F., Cebra, K., & Chen, S. (2008). A Calculation of the Theoretical Significance of Matched Bullets. *Journal of Forensic Sciences*, 53(4), 868-875.

Forensic Technology. (2001). IBIS User Guide, Version 3.3. Montreal: Forensic Technology WAI Inc.

Jolliffe I.T. (2004). *Principal component analysis*. 2nd ed. New York: Springer.

Kall L., Storey J. D., MacCross M. J. and Noble W. S. (2008a). Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Research, 7(1), 40-44.

Kall L., Storey J. D. and Noble W. S. (2008b). Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics, 24, i42-i48.

Kennedy R.B., Pressmann I.S., Chen S., Petersen P.H., Pressman A.E. (2003) Statistical Analysis of Barefoot Impressions. Journal of Forensic Sciences, 48(1), 55-63.

Kennedy R.B., Chen S., Pressmann I.S., Yamashita A.B., Pressman A.E. (2005) A Large-Scale Statistical Analysis of Barefoot Impressions. Journal of Forensic Sciences, 50(5), 1071-9.

Kumho Tire Company, Ltd., et al. v. Patrick Carmichael, etc., et al., 119 S. Ct. 1167 (1998).

Langville A.N., Meyer C.D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton: Princeton University Press.

Leon F.P. (2006). Automated Comparison of Firearm Bullets. *Forensic Science International*, 156, 40-50.

Li M., & Vitanyi P. (2008). *An introduction to Kolmogorov complexity and its applications*. 3rd ed. New York: Springer.

Mallat S. G. (2008). A Wavelet Tour of Signal Processing, 3rd ed. New York: Academic Press.

Maksumov, A., Vidu, R., Palazoglu, A., & Stroeve, P. (2004). Enhanced Feature Analysis Using

Wavelets for Scanning Probe Microscopy Images of Surfaces. *Journal of Colloid and Interfacial Science*, 272. 365-377.

Miller, J. (1997). Square cap nails: Manufacturing, identification, and value. *AFTE Journal*, 29(3), 282 – 287.

Miller, J. (1998a). Cut nail manufacturing and toolmark identification. *AFTE Journal*, 30(3), 492 – 498.

Miller, J. (1998b). Reproducibility of impressed and striated toolmarks: 4d cut flooring nails. *AFTE Journal*, 30(4), 631 – 638.

Moran B. (2002). A Report on the AFTE Theory of Identification and Range of Conclusions for Tool Mark Identification and Resulting Approaches To Casework. *AFTE Journal*, 34(2), 227-35.

Muralikrishnan B., Raja J. (2009). *Computational Surface and Roundness Metrology*. New York: Springer.

National Academy of Sciences. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, D.C.: The National Academies Press.

Neel, M., & Wells, M. (2007). A comprehensive statistical analysis of striated tool mark examinations part I: Comparing known matches and known non-matches. *AFTE Journal*, 39(4), 176 – 198.

Nichols, R. G. (2003). Consecutive matching striations (CMS): Its definition, study and application in the discipline of firearms and toolmark identification. *AFTE Journal*, 35(3), 298 – 306.

Percival D. B. & Walden A. T. (2006). *Wavelet methods for time series analysis*. New York: Cambridge University Press.

Peterson, J., & Markham, P. (1995). Crime laboratory testing results, 1978 – 1991, II: Resolving questions of common origin. *Journal of Forensic Sciences*, 40(6), 1009 – 1029.

Petraco, N., Petraco, N. D., & Pizzola, P. A. (2005). An ideal material for the preparation of known toolmark test impressions. *Journal of Forensic Sciences*, 50(6), 1407 – 1410.

Petraco, N., Petraco, N. D. K., Faber, L., & Pizzola, P. A. (2009). Preparation of toolmark standards with jewelry modeling waxes. *Journal of Forensic Sciences*, 54(2), 353

– 358.

R Core Development Team. (2009). R: A language and environment for statistical computing [computer program]. 2.9.1th ed. Vienna, Austria: R Foundation for Statistical Computing.

Reizer, R. (2011). Simulation of 3D Gaussian surface topography. *Wear*, 271, 539-543.

Rencher AC. Methods of Multivariate Analysis. 2 ed. Hoboken: Wiley, 2002.

Roberge, D., & Beauchamp, A. (2006). The use of BulletTrax-3D in a study of consecutively manufactured barrels. *AFTE Journal*, 30(2), 166 – 172.

Saunders C., Davis L. & Buscaglia J. (2011). Using Automated Comparisons to Quantify Handwriting Individuality. J Forensic Sci, 56(3), 683 - 689.

Semwogerere, D., & Weeks., E. R. (2005). Confocal microscopy. In *Encyclopedia of Biomaterials and Biomedical Engineering* (Vol. 1, p. 705 – 714). New York: Informa Healthcare U.S.A., Inc.

Senin, N., Groppetti, R., Garofano, L., Fratini, P., & Pierni, M. (2006). Three-dimensional surface topography acquisition and analysis for firearm identification. *Journal of Forensic Sciences*, 51(2), 282 – 295.

Scholkopf B & Smola A. J. (2002). *Learning with Kernels*. Cambridge: MIT Press.

Smith C. A. B., (1947). Some Examples of Discrimination. Ann. Eugenics, 18, 272-283.

Song J., Vorburger T., Renegar T., Rhee H., Zheng A., Ma L., et al. (2006). Correlation of topography measurements of NIST SRM 2460 standard bullets by four techniques. *Measurement Science and Technology*, 17(3), 500-3.

Storey, J. D. & Tibshirani. (2003). Statistical significance for genome wide studies. *Proc. Natl. Acad Sci*, 100(16) 9440-9445.

Taroni F., Bozza S., Biedermann A., Garbolino P. & Aitken C. (2010) Data Analysis in Forensic Science: A Bayesian Decision Perspective. 1 st. ed. New York: Wiley.

Taroni F., Champod C., & Margot P. (1996). Statistics: A future in toolmarks comparison? *AFTE Journal*, 28(4), 222 – 232.

The Committee on the Judiciary House of Representatives. (2009). *Federal rules of evidence*. URL: http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/EV2009.pdf, Last accessed: 7/15/2011.

Theodoridis S., Koutroumbas K. (2006) *Pattern Recognition*. 3 ed. San Diego: Academic

Press.

Tontarski, R. E., & Thompson, R. M. (1998). Automated firearms evidence comparison: A forensic tool for firearms identification–An update. *Journal of Forensic Sciences*, 43(3), 641 – 647.

United States of America v. Amando Montiero, Valdir Fernandes, Angelo Brandao, Brina Wurie, Luis Rodrigues, Manuel Lopes, 407 F. Supp. 2d 351 (2006).

United States of America v. Chaz Glynn, 578 F.Supp.2d 567 (2008).

United States of America v. Darryl Green, 405 F. Supp. 2d 104 (2005).

United States of America v. Donald Scott Taylor, 2009 U.S. Dist. LEXIS 101072 (2009).

United States of America v. Edgar Diaz, Rickey Rollins, Don Johnson, Robert Calloway, Dornell Ellis, Emile Fort, Christopher Byes, Paris Ragland, Ronnie Calloway, Allen Calloway, Terrell Jackson, and Redacted Defendant No. One., 2007 U.S. Dist. LEXIS 13152 (2007).

Vapnik, V.N. (1998). Statistical learning theory. New York: Wiley.

Vorburger T. V., Yen J. H., Bachrach B., Renegar T. B., Filliben J. J., Ma L., Rhee H. G., Zheng A., Song J. F., Riley M., Foreman C. D. & Ballou S. M. (2007). Surface topography analysis for a feasibility assessment of a national ballistics imaging database. NISTIR 7362, URL: www.nist.gov/pml/div683/grp02/upload/nistir2007-7362.pdf, Last accessed: 12/28/2011.

Vovk V., Gammerman A., & Shafer G. (2005). *Algorithmic learning in a random world*. 1st ed. Springer, New York.

Wayman J. L. (editor). National Biometric Test Center Collected Works 1997-2000. URL: http://www.engr.sjsu.edu/biometrics/nbtccw.pdf, Last accessed: 12/28/2011.

Zeiss Axio CSM 700 Confocal Microscope Software Manual.

VI. Dissemination of research findings

The products of the this research have been presented at the following conferences and professional meetings:

1. International Conference on Surface Metrology, Worcester Polytechnic Institute, Worcester, M.A., October 25. 2010. Title: “Forensic Surface Metrology, Firearms and Tool Mark Evidence”.

2. Northeastern Association of Forensic Scientists (NEAFS) 2010 meeting, Manchester Village, V.T., November 12 2010. Title: "Statistical Analysis of Chisel Marks".
3. Federation of Analytical Chemistry and Spectroscopy Societies (FACSS) conference. Rayleigh, N.C., October 19. 2010. Title: "Application of Chemometrics and Advanced Pattern Recognition to Trace Evidence Analysis".
4. Hofstra University, Department of Chemistry, New York, October 13. 2010. Title: "Application of Advanced Computational Pattern Recognition to Trace Evidence Analysis".
5. Olympus Tech Tour, New York, September 21. 2010. Title: "Confocal Microscopy and Tool Mark Analysis: Pushing Out the Frontiers of Forensic Science".
6. National Institute of Justice Pattern and Impression Evidence Symposium, Clearwater, Florida, August, 2. 2010. Title: "Addressing the National Academy of Sciences Challenge: Methods for Statistical Pattern Comparison of Striated Tool Marks".
7. Association of Firearms and Tool mark Examiner (AFTE) Annual Training Seminar, Chicago, I.L., May 31, 2011. Title: "Confocal Microscopy and Striated Tool Marks: A Statistical Study and Potential Software Tools For Practitioners". This presentation was awarded "Best Paper of the 2011 AFTE Seminar".
8. American Academy of Forensic Sciences (AAFS) Annual Conference, Chicago, I.L., February 25, 2011. Title: "Computational Pattern Recognition of Striation Patterns: Fired Cartridge Cases and Chisel Striation Patterns".

The website will be advertised on the AFTE forum website (http://www.afte.org/forum/smf1/index.php) at or near the official end of this project. Some preliminary findings of this research has been published in the following peer-reviewed journal article:

Carol Gambino, Patrick McLaughlin, Loretta Kuo, Frani Kammerman, Peter Shenkin, Peter Diaczuk, Nicholas Petraco, James Hamby and Nicholas D. K. Petraco, "Forensic Surface Metrology: Tool Mark Evidence", Scanning 27(1-3), 1-7 (2011).

Further results will be formatted into additional articles and will be submitted for publication in the Journal of Forensic Sciences, Forensic Science International, or the AFTE journal. In addition, the products of this research will possibly be published in a textbook

covering tool mark examination as well as other forms of impression evidence. Lastly, the A&E Television program "Forensic Files" interviewed members of our research group about research in this project.

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.