Final Technical Report Guidelines

Report Title: Significance of Association in Tool Mark Characterization
Award Number: 2009-DN-R-119
Author(s): L.S Chumbley, M. Morris

Abstract

One weakness that currently exists in the field of comparative examination of evidence is the general failure of current approaches to adequately assess the significance of association through quantitative measures that provide a statistical evaluation of evidence. While various efforts have been made and methodologies employed over the years, such as the measurement of consecutive matching striations, tool mark comparisons remain difficult to quantify in a robust statistically valid sense. While the desire to develop a methodology that allows the examiner to assign confidence levels and predict error rates is universal, the unfortunate truth is that such a description (similar to what is possible in the field of DNA) is an unattainable goal in toolmark analysis since the population will continually increase and the variability cannot be satisfactorily defined. However, this is not to say that statistical relevance cannot be assigned to toolmark examination. Relevance can be assigned if one is careful about the structure of the study attempted.

In a recent study of tool marks produced by sequentially made screwdriver tips the authors developed a computer algorithm that was able to reliably separate matching tool marks from those that do not match using an analysis based on Mann-Whitney U-statistics applied to data files containing 2-dimensional information obtained using an optical profilometer. These successful results indicate that significance of association can be accomplished by statistical evaluation of the data files. The work carried in the present project (and discussed in this report) built upon this success by providing additional statistical information that will increase the relevance of the measurements obtained. Thus, the overall goal of this work was to increase the statistical relevance of toolmark analysis. To achieve this goal two distinct objectives were identified:

1) Extend the previously developed statistical methodology to allow for self-calibration to control rates of false non-matches.

Our previous work (Chumbley et al., 2010) has focused on the use of the Mann-Whitney U-statistic as an index for assessing the similarity of toolmarks. While it has been empirically shown to be useful in sorting mark-pairs made by the same tool from mark-pairs made with different tools, it is also influenced by many other aspects of the toolmark structure, hence a single value cannot be used as objective evidence (with quantifiable risk) for or against a match. The current work has focused on overcoming this difficulty by using multiple test marks made in the laboratory, in a ``self-calibrated'' analysis. In short, comparison values between lab marks (that are known to match) form the basis for comparisons between lab marks and evidence marks, eliminating the need for "universal" critical values (i.e. single sets of constant references values such as those found in commonly used statistical tables) for the comparison index. A formal statistical analysis based on likelihood functions have been developed to allow for control of false non-match calls.

2) Empirically validate the methodology developed by performing experiments using a different type of tool mark.

The first part of the project dealt with objective one. In this task the variability of marks from screwdriver tips by characterized examining multiple marks made by a trained forensic examiner. The data was used to establish the variation in U-statistic values inherent in the system and allowed likelihood analysis to be conducted denoting significance of association between lab-lab comparisons and lab-field comparisons made at the same angle. With the initial model established, modifications were then undertaken to generalize the model to be applicable to marks made at all angles. In this analysis it was shown that the angle at which mark was made could be deduced to a fairly high level of accuracy.

In the second part of the project the second objective was attained by applying the algorithm developed in the previous study to an entirely new system separate and apart from the screwdriver tool marks studied initially. In this case, markings produced by shear cutting metal wire using the shear face on pliers were analyzed and tested to determine the applicability of the approach. Since the marks produced are not regularly striated, this study represented a significant extension concerning the performance of the algorithm. The study found that with adjustment of the analysis parameters used, known matched sets of data from these quasi-striated marks, i.e. marks characterized by groups of striations instead of regular striae, could be successfully differentiated from known non-match sets. Areas for improvement were also identified that will make the system even more reliable. Successful validation of the methodology has created a wide range of possible future applications for the developed statistical algorithm that could revolutionize comparative tool mark analysis.

Table of Contents

Executive Summary

1. Synopsis of the Problem

The problem addressed by this work is one that has received much attention over the past few years, namely, **how can the field of forensic examinations be moved from a simple subject comparison to a more objective analysis involving quantifiable measurements and valid statistical descriptions?** This is a difficult proposition since the majority of comparative forensic analyses are essentially open systems and subject to variability. However, this problem can be effectively addressed in a number of ways. In this project, efforts to increase the objectivity of the analysis involved examining data obtained from multiple marks, then using this data in both a likelihood analysis and to investigate the angle effect, i.e., how a toolmark changes as you vary the angle of attack of the tool associated with making a toolmark.

Also problematic is that the wide variability of toolmark types makes it difficult to develop an algorithm suitable for more complex marks. The earlier algorithm developed in [1] was therefore tested to determine suitability for use in quasi-striated marks.

2. Purpose

The research hypotheses adopted for this study are stated as follows:

Hypothesis #1:. The effectiveness of quantitative toolmark comparison can be improved by including multiple marks made by the suspect tool, and by the use of statistical models that reflect relevant sources of variation and correlation.

Hypothesis #2: Objective analysis of quasi-striated marks is possible. The same (or a similar) algorithm applicable to striated marks can be developed that provides the same level of performance and confidence as seen for striated marks.

The methods used to test these hypotheses varied substantially. Therefore, the work is described in two distinct sections or tasks, each task being related to one of the above stated hypotheses. Task I consisted primarily of a mathematical analysis of existing data to see if means could be found to increase statistical relevance. The models developed were then tested in various ways. Task II involved experimental data acquisition of toolmarks of a type that have been shown to be difficult to analyzed using methods designed for regularly striated marks. The purpose of this study was to determine whether the current algorithm developed at Ames Lab / Iowa State University and used successfully on striated marks was robust enough to be used in evaluating other kinds of tool marks in which parallel striae are not so dominant.

3. Research Design

Task I: Statistical Relevance

The research design for the statistical validation study used data obtained from 50 sequentially manufactured screwdrivers. Toolmarks from these screwdrivers were made in lead at a variety of angles ranging from 30 degrees to 85 degrees, using both sides of the screwdriver. At least four replicates were made of each toolmark, providing a large database of measurable toolmarks. Quantification was carried out using a stylus profilometer.

The analysis undertaken was based on the assumption that a single tool mark was found at the crime scene. A suspect tool is obtained and forensic examiners make several marks using the

suspect tool in the lab under controlled conditions. By comparing multiple marks all known to be made by the same tool in the lab, a sample of matching mark-pairs can be created from that tool. The same lab tool marks can be compared to the field mark to create a smaller sample of mark-pairs with an unknown matching status. Rather than evaluating only one mark-pair, there now exist two samples that can be compared; if there is no apparent systematic difference between these two samples, this supports the argument that the marks were all made by the same tool, i.e. that the crime scene tool mark and lab tool marks "match."

Let x_0 represent the tool mark that was found at the crime scene. Let x_1, ..., x_n represent the n tool marks that were made by the suspect tool in the lab. Note that all comparisons of the marks not including x_0 are known matches all made by the same tool under the same conditions. Let y_{ij} represent the numerical index value of similarity that results from comparing x_i to x_j. If all pairwise comparisons of available tool marks are made, this results in two different types of data. The first set, y_{0j} for $j = 1, ..., n$, includes all comparisons of the field mark to a lab mark. This type of data is designated field-lab comparisons. The second set is y_{ij} for $i, j = 1, ..., n$ and $i < j$. These data values represent indices of similarity for a known match from the suspect tool, which are designated lab-lab comparisons.

The collection of all data values will be denoted by the vector **y,** which is of length N. When all possible comparisons are made there are $\binom{n}{2}$ lab-lab comparisons and n field-lab comparisons, so $N = \binom{n}{2} + n$. For the purposes of this report, the data is discussed in terms of the number of lab marks, such as a data set of size n. Note that this means there are $n + 1$ tool marks under comparison.

The question is whether or not the suspect tool was used to make the tool mark found in the field. If the field-lab comparisons are indistinguishable from the lab-lab comparisons then there is no evidence that the tool marks are different, and the data are therefore consistent with the hypothesis that all marks were created using the same tool. However, if the field-lab comparison values are relatively small compared to the lab-lab comparison values, then there is evidence that the field mark and the lab marks were created using different tools. However, if the data representing field-lab comparisons are generally smaller in value than the data representing lab-lab comparisons, then there is evidence that the field mark and the lab marks were created using different tools.

A simple hypothesis test is used to compare the field-lab and lab-lab comparisons. Let μ_0 be the mean for a field-lab comparison, that is $E(y_{0j}) = \mu_0$ for $j = 1, ..., n$. Let μ_1 be the mean for a lab-lab comparison, so $E(y_{ij}) = \mu_1$ for $i, j = 1, ..., n$ and $i < j$. It is assumed that all data values have a common variance defined as $Var(y_{ij}) = \sigma^2$ for $i, j = 0, 1, ..., n$ and $i < j$, that each y_{ij} is normally distributed. Finally, denote by **y** an N-element vector of all comparison values, and assume that the joint distribution of **y** is multivariate normal. The mean of **y** is a vector of means μ_0 and μ_1 with the form $\mu = \left(\mu_0 \mathbf{1}_n', \mu_1 \mathbf{1}_{N-n}'\right)'$, where $\mathbf{1}$ designates a vector of 1's of size indicated by the subscript, and the variance of each element of **y** is σ^2. To finish defining the joint distribution of **y**, one needs to develop an appropriate dependency structure reflecting the way the data are generated.

Each y_{ij} is the result of comparing two tool marks, specifically x_i with x_j. Thus, at most four physical tool marks (i.e. up to two for each pair-wise comparison) are involved in the consideration of covariance between two data values. Since these four tool marks are not necessarily unique, one can say two data values are correlated with correlation ρ if a common

tool mark is involved in both comparisons. That is, y_{ij} is correlated with y_{kl} if $i = k$, $i = l$, $j = k$ or $j = l$. Two comparisons with no marks in common are uncorrelated. The case of two comparisons made on the same pair of tool marks is not considered because those similarity values would be identical, i.e. there is no measurement-specific "error" in this system, and so no point in replication. Let **R** be the $N \times N$ correlation matrix of **y** defined by the following entries

$$Corr(y_{ij}, y_{kl}) = \begin{cases} 0 \text{ if } i \neq k, i \neq l, j \neq k \text{ and } j \neq l \\ \rho \text{ if } i = k \text{ or } i = l \text{ or } j = k \text{ or } j = l \\ 1 \text{ if } i = k \text{ and } j = l \end{cases} \quad (1)$$

To finish defining the joint distribution of all pairwise comparisons of tool marks, let **y** be the vector of all pairwise comparisons (both field-lab, and lab-lab) ordered so that each y_{ij} is such that $i < j$. Then $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{R})$ where $\boldsymbol{\mu} = \left(\mu_0 \mathbf{1}_n', \mu_1 \mathbf{1}_{N-n}'\right)$ and each element of the matrix **R** is as defined in (1).

The model described suggests there could be separate means for the two available samples, field-lab and lab-lab comparisons. This enables hypothesis testing to be conducted, and using Normal model theory and weighted least squares, maximum likelihood estimates (MLEs) can be easily derived. Details of how this was done are provided in the full text. Likelihood analysis is effective in determining a match of tool marks under carefully controlled test conditions in which all tool marks are produced the same way. However, in order to be useful in practice, it must also perform well when tool marks are made under different conditions. It is known empirically that when tool marks are made at the same angle, and the field mark was not made by the suspect tool, analysis of the field-lab comparisons show no correlation while the lab-lab comparisons show high correlations match. However, for angles that differ by 10° or more, even the lab-lab comparisons show low correlation. In other words, even when marks are made by the same tool, if the tool angles differ by 10° or more, the comparison values resemble those from non-matching tool marks.

Knowing this, it is important to generalize the approach to account for these effects. It is impossible to know the tool angle that was used to make a mark left at the crime scene. However, in a lab, tool marks can be made at any angle to try to better match a crime scene mark. If enough tool marks are made in the lab at angles differing by 10° or less, it should be possible to find the best match of a field mark to lab marks to estimate the angle at which the field mark was made.

Before the basic model is modified to account for angle information, more notation is necessary. Let a_i be the tool angle, in degrees, at which tool mark x_i is made for $i = 0, 1, ..., n$. Tool angles will be incorporated as a function of their absolute difference, $|a_i - a_j|$, which will be denoted by the similarity measure $d(a_i, a_j)$.

The mean response for data values is large when the angles are the same and approaches zero as the difference in angles increases. Comparisons of matching tool marks made at angles differing by 10° or more resemble non-matches, so these facts will be reflected in $d(a_i, a_j)$. The function chosen to represent the difference in angles was $d(a_i, a_j) = \exp[-\theta(a_i - a_j)^2]$ where $\theta = 0.01$. This was chosen so that $d(a_i, a_j) = 1$ when tool angles match, is much smaller when $|a_i - a_j| = 10$ and approaches zero as $|a_i - a_j|$ continues to increase. This functional form has no particular

physical basis, but empirically reflects the trends seen in the data. The data available in this study did not support formal estimation of the parameter θ, but this might be possible – and desirable – for larger data sets. Therefore, a modified data model incorporating tool angle is

$$y_{ij} \sim N(\mu_{ij}, \sigma^2) \text{ where} \tag{2}$$

$$\mu_{0j} = \mu_1 + \alpha_0 d(a_0, a_j) \text{ for } j = 1, ..., n \text{ and} \tag{3}$$

$$\mu_{ij} = \mu_1 + \alpha_1 d(a_i, a_j) \text{ for } i, j = 1, ..., n \text{ and } i < j. \tag{4}$$

where α_0 and α_1 can be thought of as "regression coefficients" represent the extent of the effect of angle similarity on the data. So, where the field and lab marks are not made by the same tool, it would be expected that α_0 would be near zero, since similar angles should not improve the "apparent similarity" of tool marks in this case. That is, the mean of a data value is a linear function of the overall mean of the comparisons and the similarity measure between the tool angles used in making the marks. Similar to the initial model different mean functions for field-lab and lab-lab comparisons are allowed for through different regression slopes. Thus, inferences about whether or not the suspect tool made the crime scene marks can again be made with a likelihood ratio test. However, of interest now is comparing the alternative hypothesis defined by the model described in (2) through (4) with a null hypothesis defined by a simpler model that does not discriminate between field-lab and lab-lab comparisons. That null model can be stated as

$$y_{ij} \sim N(\mu_{ij}, \sigma^2) \text{ where} \tag{5}$$

$$\mu_{ij} = \mu_0 + \alpha d(a_i, a_j) \text{ for } i, j = 0, 1, ..., n \text{ and } i < j. \tag{6}$$

Both the null and alternative models assume that the angles are known for every tool mark made in the lab; that is a_1, ..., a_n are known. The angle of the mark made in the field, a_0, is unknown. Thus, a_0 is a parameter in the model along with α, α_0, α_1, μ, σ^2 and ρ. The correlation structure described in equation (1) remains for this model and ρ will still be chosen using a grid search between 0 and 0.5. Since the tool angle needs to be accurate within 10° to see evidence of a match, the maximum likelihood estimation procedure incorporates a grid search of values for a_0 in increments of 5° between 20° and 90°. These angle bounds were chosen as reasonable angles for which a tool mark could be made and leave behind a viable mark. Maximum likelihood estimates for the remaining parameters can be computed using weighted least squares provided values of a_0 and ρ.

With a model now in hand extensive testing could be conducted. The results of these tests are summarized under Section 4 of the Executive Summary, **Findings and Conclusions**. As stated above under Section 2 of the Executive Summary, **Purpose**, the experimental work of this proposal fell naturally into two Tasks. The experimental methodology of Task II will now be summarized.

Task II: Analysis of Quasi-Striated Marks

For this experiment, 50 pairs of sequentially manufactured slip joint pliers, as nearly identical as possible, were used to produce samples for evaluation. To make the samples, copper wire of 0.1620" diameter was obtained and cut into two-inch lengths with bolt cutters to distinguish the ends from the shear cuts made by the pliers. The cut lengths of wire were placed centered in the plier jaws on the shear surface and shear cuts of the copper were made, taking care to keep the sides of the pliers consistent as the shear cuts were made. The total number of copper samples thus obtained was 1000, with 500 shear cuts in contact with one side of each pair of pliers

(designated Side A) and 500 shear cuts in contact with the other side (i.e. Side B). Each shear cut mark surface was scanned optically with an Alicona Infinite Focus G3 profilometer at 10x magnification to acquire the surface geometry of the mark. Shear cuts made in this manner are quasi-striated, i.e. linear striae do exist but vary considerably across the surface of the shear cut mark due to the shearing process involved. If the process were completely cutting in nature regular striae would be the expected result.

Once the data were acquired, noise spikes around the edges of the mark where no signal was acquired were removed by development of a cleaning routine and the data was de-trended. De-trending was necessary because due to the manner in which the data were collected the line profile of each mark data file had an increasing linear trend in the z direction moving from one side of the mark to the other. Such a trend is common when using profilometers since the surface analyzed is rarely exactly parallel with the direction of scanning. Trending was corrected by subtracting a fitted plane, fit to match that of the original data file using a least squares process, from that of the trended data.

Comparisons between the marks were made using the algorithm developed by the PIs in a previous study and completely described in [1]. This algorithm uses a Mann-Whitney statistical approach to make comparisons between two data sets. Comparisons were conducted at two locations, close to the end of the mark and close to the start of the mark. These mark locations were chosen to examine differences between the beginning of the shear cut, where the mark has short and variable length striae, and the end of the mark, where the striae are longer and appear to be more regular.

Each side of the pliers was considered to be a separate data set, the assumption being, as confirmed by forensic examiners, each side acts as a different surface. Given there are 50 pairs of pliers, with two sides for each pair of pliers and ten replicate shear cuts for each side of each pair of pliers, the total number of samples possible for examination came to 1,000 discrete data sets.

4. Findings and Conclusions

Task I: Statistical Relevance

Since it was known that all the tool marks used in each analysis are made by the same tool, one expects the field-lab and lab-lab comparisons to result in similar data values, so only one regression slope for the similarity measure, $d(a_i, a_j)$, should be needed for an adequate model fit since lab and field marks are actually interchangeable in this case. Thus, the null model described in equations (5) and (6) above should be the best model for these data and the distribution of p-values (i.e. attained significance levels) from these likelihood ratio tests should be distributed approximately uniformly between 0 and 1. This was indeed the case in most instances. However, it was noticed that a large portion of the LRTs resulted in very small p-values, contrary to fact, leading to a close examination of the individual tool marks. Two recurring issues in the data were present in the majority of field marks that resulted in LRTs with small p-values, the first related to the quality of the acquired data, the second related to data acquisition at the very ends of each profilometer scan. Both types of poor data resulted in incorrect / inaccurate analysis by the algorithm. In addition to problems associated with data quality, it was found that some of the very small p-values originated from a more fundamental problem with the algorithm described in [1]. In several instances best matching windows were found that were physically impossible, for example, they occurred at opposite ends of the mark in highly sloped regions. This “opposite end” match results in problems for the algorithm when

running the validation step. After removing obviously poor data, the data sets and the analyses were run again, producing much improved results. This indicates that the performance of the algorithm may be enhanced significantly by the introduction of screening algorithms to test the data quality before matching is attempted.

To further examine the efficiency of the modified models and the LRT, also examined was the accuracy of the estimates of a_0, the tool angle with which the field mark was made. Overall, the null model does fit well to the data in most cases where the data are matches, and the estimation process for the field angle seems reasonably accurate.

To create non-matching data sets, all n = 20 tool marks made from the same tool were considered to be lab marks. Field marks were chosen out of the remaining five tools, one mark from each of the 5 angles available for each tool. The data sets were assembled by comparing the field mark with each of the lab marks and comparing the lab marks pairwise with one another. This process was repeated for two different sets of lab tool marks, resulting in a total of 50 data sets for which the lab tool and field tool are not the same.

For non-matching data, all of the field-lab data values should be close to zero regardless of the tool angles since the marks were made by different tools. However, most lab-lab comparisons will result in a larger comparison value since they are true matches. This discrepancy should show up in the models through the regression slopes. One could expect that the alternative model will be a better fit to these data since the coefficient for the angle similarity function, $d(a_i, a_j)$, will be close to zero for the field-lab comparisons but large for the lab-lab comparisons. Thus the alternative model should be a better fit to these data and the p-values from the LRT should be small, i.e. the distribution of p-values should be skewed with greater frequencies associated with smaller p-values.

When this model was tested on the data sets, as expected, the p-values were mostly small and had an overall skewed shape. This supports the hypothesis that the alternative model which allows for multiple regression slopes is a better fit to these data and provides evidence that the lab tool marks do not match the field mark in most of these tests.

Task II: Analysis of Quasi-Striated Marks

A sampling format was set up to compare three different groups of data: known matches, known non-matches from the same pair of pliers (i.e. different sides), and known non-matches from different pairs of pliers. The same algorithm used in an earlier work for striated marks [1] was applied in this study to examine the quasi-striated marks made by the slip joint pliers, and details can be found in the quoted references and in the full text concerning the operation of the algorithm.

Initial results used the same algorithm parameters that had been successfully employed for striated mark comparisons. However, the success of identifying known matches was relatively low, there being little separation between the returned T1 values (a normalized form of the Mann-Whitney U statistic; see, e.g., Conover [2] for details) of known matches and non-matches. From the minimal success of the first attempt at matching the plier marks, several changes were decided upon for further comparisons. This involved using de-trended data to remove problems associated with non-flat samples and a series of experiments was conducted where the window sizes were varied in a consistent manner to evaluate the effect window size has on the resulting T1 value. Window sizes were varied in two ways, firstly in the absolute size of the search and validation windows employed, and secondly in the ratio of sizes employed.

The results of these experiments showed that the performance of the algorithm increased dramatically. By increasing window size, while known non-matches returned values centered around zero regardless of window size, the T1 value for known matches increases from just slightly over zero to an average of 6.36 and 6.09. However, the data range increased as well, and at the larger window sizes numerous outliers exist and failure of the algorithm occurs in some cases, especially for the short edge comparisons. This problem was addressed by changing the ratio of window sizes used for the search and validation steps. Results of these experiments showed that higher window ratios did have a significant effect in reducing the number of outliers and spread of the known matches. While a slight degradation in the maximum T1 values obtained was seen for the known matches there was still significant separation between the known matches and non-matches. Less change was seen in the results for the known non-matches, whose average values still were centered around zero.

Outliers are seen in all the data sets, both known match and known non-match. Examination of these data files points to the same problem with the algorithm noted above under the Task I: Statistical Relevance section, namely, the "opposite end" match problem. In its current form, the algorithm has maximum flexibility, allowing marks to be compared along a linear direction both forwards and backwards. However, a screening option is being considered that will automatically determine whether an "opposite end" match has occurred and alert the user to this possibility. The user can then examine only those files so flagged and decide whether an incorrect match has occurred.

5. Implications for Policy and Practice

Task I: Statistical Relevance

The implications of this study for policy and practice are as follows. The statistical analyses carried out on toolmarks in this work clearly adds further evidence that the long-held assumptions under which forensic examiners operate, namely, that all toolmarks are unique, does have a sound scientific basis. However, if a truly objective, quantitative analysis is desired such as was carried out in this study, it may require examiners to generate more lab samples for comparison than is typically done now.

Task II: Analysis of Quasi-Striated Marks

Successful application of the algorithm described in [1] to quasi-striated toolmarks has two major implications. Firstly, the work shows once again that there is a scientific basis for objective, quantitative toolmark identification. Secondly, as this represents one of the first successful analyses on a quasi-striated toolmark, it implies that it should be possible to characterize and classify all types of toolmarks, given the right type and quality of data and the appropriate analysis algorithm.

I. Introduction

Research into objective, quantitative, analysis of evidence such as is found in comparative forensic examinations has been a focus of the group at Ames Laboratory / Iowa State University for quite some time. The work conducted under this project, for which this document constitutes the final report, falls into two reasonably distinct tasks. Task I involved continued examination of data acquired under an earlier effort concerning screwdriver toolmarks in an attempt to add further statistical relevance to the characterization and analysis previously undertaken. Task II involved an examination of an entirely new set of marks, namely, shear cut marks produced when wire is mechanically separated by slip-joint pliers. The toolmark produced in this action is not regularly striated, but varies over the length of the mark.

For both tasks the problem statement and literature review is the same when considering the broad scope of the work undertaken. However, the research hypotheses tested vary, as do the experimental methods used and the results. Therefore, the organization of this report is such that the Methods and Results of these two distinct Tasks will be dealt with separately for the sake of clarity.

1. Statement of the Problem

The problem addressed by this work is one that has received much attention over the past few years, namely, **how can the field of forensic examinations be moved from a simple subject comparison to a more objective analysis involving quantifiable measurements and valid statistical descriptions?** Various attempts have been made in recent years to introduce objectivity into toolmark analysis, driven equally by the well-known Daubert decision and the highly publicized success of DNA testing, which can link a suspect to a crime scene with a high degree of statistical reliability. This is possible since the boundary conditions involved in DNA testing are well known. This is not the case in the majority of comparative forensic analyses, which are inherently more open and subject to variability. We believe this problem can be effectively addressed in a number of ways. A first attempt was made in a previous study by the authors [1] involving the objective comparisons of striated marks produced by sequentially made screwdrivers. Much of the work presented in this report builds on this earlier effort.

In this project we have expanded the scope of the investigation to examine the variability between striated marks. We sought to increase the objectivity of the analysis by examining data obtained from multiple marks made from the same tool. This data was used in both a likelihood analysis and to investigate the effect making a toolmark at different angles has on the possibility of identification.

Also problematic is that the wide variability of toolmark types makes it difficult to develop an algorithm suitable for more complex marks. The earlier algorithm developed in [1] was therefore tested to determine suitability for use in quasi-striated marks. This involved examination of irregular, quasi-striated marks that result due to shear cutting. The specific samples used were obtained by shear cutting copper wire using sequentially made pairs of pliers.

2. Literature Review

Historical Background

The history of forensic analysis of toolmarks and firearms stretches back nearly 180 years to the first documented case of firearms identification in 1835 [2]. Early firearms identification relied primarily on the identification of the caliber, any macroscopic imperfections of the bullet, and the shape and type of bullet used in the crime [3]. The case in question occurred in the City of London, England in 1835 when Henry Goddard, a part of the police force, was able to identify the mold mark on a fired lead ball used in committing a crime. He also was able to identify the paper patch used in firing the black powder weapon. From these clues, Goddard was able to deduce the guilty party and bring him to justice [2].

The late 1800s and early 1900s saw an increased interest in firearm identification. This interest included several court cases within the United States, and promoted research conducted throughout the U.S. and Europe. Published works included titles such as, "La Deformation Des Balles de Revolver" (Deformation of Revolver Bullets, 1889), "The Missile and the Weapon" (1900), "Zur Sachverstandign Beurteilung Von Geschossen" (The Expert Examination of Fired Bullets, 1905) written by A. Lacassogne of Lyon, France, Dr. Albert Llewellyn Hall of Buffalo, New York and Dr. R. Kockel of Leipzig, Germany, respectively [2]. Some credit Dr. Kockel with the first use of striation matching of toolmarks, which occurred around 1900. In his first paper, Kockel identified knife cuts made in wood through oblique lighting and photography. In a later notable paper, he described the examination of marks through magnification and measured the relative spacing with calipers. Additionally, this paper noted the change in geometry of the toolmark with different attacking angles of the knife blade [3].

Most early studies and cases largely focused on ballistic toolmarks, with the exception of a few studies including Dr. Kockel's work as previously described. In 1948, Dr. Thomas of the University of Ghent added to the toolmark references by publishing a paper describing the toolmarks left on a skull by an axe. Since then, many different types of toolmarks have been characterized [3].

In 1942 a notable paper was published by Burd and Kirk examining the marks made by screwdrivers. In this study [4] the authors addressed four different points: 1) the effect of varying the angle of application of the screwdriver on a toolmark, 2) establishing the necessary criteria for identification, 3) assessing the similarity between tools with identical appearance and manufacturing process, and 4) classifying the different types of marks that can be encountered. Burd and Kirk pointed out in the study the traditional method of examining toolmarks with oblique lighting and a comparison microscope will only yield a match if, and only if, the marks in question have a similar contour, since this is reflected in the "lines" or striations seen through the microscope. The authors go on to conclude several important points. First, two marks made with the same tool must be made with a difference in vertical angle of no more than 15 degrees if a match is to be obtained. Similarly, two marks made with the same tool must be made with a difference in horizontal angle of no more than 20 degrees if a match is to be determined. The authors also established the maximum percentage of lines that matched in non-match comparisons did not exceed 25% and when match comparisons were performed this percentage jumped to around 80%. Additionally, examination of "identical" tools produced noticeably unique marks that could not be matched to another "identical" tool.

Tongue and groove pliers were evaluated in 1980 by Cassidy [5]. These pliers are often used to pry open door handles and their marks are simple striated marks stemming from a plier tooth

sliding across a surface gripped in the pliers' jaws. For this study Cassidy procured three sets of upper and lower jaws that were sequentially broached with no further manufacturing processes applied to preserve any subclass characteristics present from the broaching process. He observed no subclass characteristics that might be mistaken for individual characteristics. In the study's discussion, Cassidy demonstrates that the pliers' teeth were broached perpendicular to the direction that the marks are made and would not produce any subclass characteristics in the striated marks. Furthermore, actual tongue and groove pliers in production go through many processes after broaching; thus, marks produced by these mass production pliers would produce only marks that have individual characteristics.

The studies cited above in no way constitute the entirety of the research carried out on toolmarks. By way of example, studies related to firearm identification [6-9], tool mark comparisons [10,11], knife markings [12,13], tire and shoe tracks [14,15], bolt cutters [16-18], drill bits [19], rotary glass cutters [20] etc., can all be found in the AFTE journal. The sheer volume of work attests to the effort on the part of examiners and others to define the limits of their field and establish best practices and protocols. All these studies have shown that given the right conditions, comparative studies can be used to correctly relate evidence obtained at a crime scene to evidence obtained either from suspects or to exemplars produced in a laboratory setting. It is equally well documented that important exceptions exist in every field, where identification is difficult, questionable, or impossible. In such cases examiners by training are taught to err on the side of caution. However, as this is a subjective assessment the field will always be open to questions of impropriety unless the status of the evaluation can be made more objective in nature.

Recent Studies

Since the 1993 Daubert vs. State of Florida decision the scientific basis for conducting comparative forensic examinations has been called into question. Attacks concerning the reliability and relevance of many forms of forensic evidence have increased in recent years due to the widespread publicity surrounding a number of unfortunate misidentifications that have occurred.

The lack of studies aimed at determining definitive error rates associated with any particular field that relies upon comparative identification was recognized by a recent National Academy of Science report Strengthening Forensic Science: A Path Forward [21]. As stated in the Executive Summary of the report, "*A body of research is required to establish the limits and measures of performance and to address the impact of sources of variability and potential bias. Such research is sorely needed, but it seems to be lacking in most of the forensic disciplines that rely on subjective assessments of matching characteristics. These disciplines need to develop rigorous protocols to guide these subjective interpretations and pursue equally rigorous research and evaluation programs.*" The development of such protocols is difficult at best, given the complex nature of the problem to be addressed. However, such studies are starting to emerge in an effort to answer these critical assessments.

With the availability of inexpensive computing power and increasingly precise metrology instruments, toolmarks are being reexamined through objective statistical comparison of their 3-D profiles. In 2007 Faden et al. [10] developed a computer algorithm to compare and match surface data taken from a stylus profilometer. In the study, 44 sequentially manufactured screwdriver tips were used to create marks at 30, 60 and 85 degrees from both sides of the screwdriver blade and the profilometer used to record the surface contours of the mark through 9600 data points. A computer program was then used to compare the collected profilometer

traces. Three different comparison data sets were generated: 1) true matches, 2) true non-matches, and 3) comparisons between side A and side B of the screwdriver blades. The Pearson correlation was calculated for all comparisons. Faden et al. determined that while there is a significant separation in the correlation values between true match and true non-match marks at the same angle, the Pearson correlation is not effective at determining when an actual match exists. Moreover, marks made from different sides of the same screwdriver tip produced a separation of data consistent with that of non-matches.

In 2010 Bachrach et al. [22] expanded the research of statistical comparison of toolmarks by evaluating screwdriver marks, and tongue and groove plier marks through confocal microscopy. In this study, Bachrach et al. examined marks made by screwdrivers at different angles in lead and aluminum. In addition, they examined the marks from tongue and groove plier marks in lead, brass and galvanized steel. After scanning the marks with a confocal microscope, the mark data were normalized to level the data, and then put through a signature generation process. This process took the cross sectional profile of the mark and applied a Gaussian band pass filter to eliminate class characteristics within the mark. Then, two signatures were run through a correlation component to evaluate the two signatures' similarity to each other. From this study, several conclusions were drawn. First, striated toolmarks in the same medium and produced under the same conditions are repeatable and sufficiently specific to allow identification. Second, striated toolmarks created with the same conditions, but different media, have a high reproducibility. Third, screwdriver marks depend on the angle at which they are made more than the media in which they are created. Fourth, the probability of two tools displaying similar features is extremely small. Finally, the probability of error originated from a poor toolmark image, not from the tool's failure to create an individual toolmark.

Chumbley et al. [1] continued with the work performed by Faden et al. in 2010. In this study, a statistical algorithm was used to evaluate its effectiveness in comparison to actual toolmark examiners. Again, data were collected by a stylus profilometer for 50 sequentially manufactured screwdriver tips. Marks were made at 30, 60, and 85 degrees for both sides of the screwdriver tip, A and B. The mark profiles collected were then analyzed by a statistical algorithm. These calculated results were then compared to a double blind study where 50 experienced toolmark examiners evaluated a given sample set with which the algorithm had difficulty. The results from this study showed that while the objective algorithm was very effective in discriminating between known matches and known non-matches, it still did not reach the level of performance of experienced examiners.

Objective statistical comparison continued through research done by Petraco et al. in 2012 [23]. In research supported by the U.S. Department of Justice, Petraco evaluated striated marks from screwdrivers and chisels, as well as striated and compressed marks from cartridge cases. Like Bachrach et al., Petraco et al. also used confocal microscopy when collecting the surface profiles of the sample marks. The results of this study showed chisel marks were patchy at best and proved too complicated for the developed software to analyze successfully. Screwdriver and cartridge cases had much more success in comparisons and had very low error rates. With the successes and the difficulties associated with this current software, Petraco et al. have made their marks and software open sourced and accessible to others in the forensic community.

Summary

Examination of past and recent papers on this subject reveals several common themes or findings that should be noted. These are summarized below:

1. A large body of work exists showing that comparative examinations can successfully be used to identify toolmarks from specific tools / firearms. Such comparisons are traditionally qualitative in nature.

2. Quantitative methods can be employed satisfactorily to discriminate between known matches and known non-match comparisons. However, toolmark analysis is inherently different than DNA analysis where all parameters can be adequately described. The diversity of the field precludes statistical validation on the same level as can be obtained for DNA.

3. Quantitative analysis of different types of toolmarks present different challenges. Statistical analyses that have been shown to work well for a striated mark perform less well or fail completely when used on other marks such as impressed marks.

3. Research Hypotheses

The research hypotheses adopted for this study address Summary points #2 and #3 as noted above. Namely, these are stated as follows:

Hypothesis #1: The effectiveness of quantitative toolmark comparison can be improved by including multiple marks made by the suspect tool, and by the use of statistical models that reflect relevant sources of variation and correlation.

Although the effectiveness of these methods cannot be validated to the same level of satisfaction as methods used with DNA data, the results still serve to show that the conclusions reached are objective in nature and possess a sound scientific basis.

Hypothesis #2: Objective analysis of quasi-striated marks is possible. The same (or a similar) algorithms applicable to striated marks can be developed that provide the same level of performance and confidence as seen for striated marks.

By showing that all types of toolmarks can be analyzed objectively the long-held assumption that all toolmarks are unique is strengthened. Objective, quantitative analysis also opens the door to further development of advanced characterization algorithms.

The methods used to test these hypotheses and the results obtained are presented below. The methods employed vary quite substantially for the two hypotheses, separating the work quite naturally into two distinct tasks. Task I, whose goal was to test Hypothesis #1, consists primarily of a mathematical description of the approach taken to increase statistical relevance and the testing of that approach. Task II, whose goal was to examine Hypothesis #2, involved more of an experimental approach to using the current algorithm developed at Ames Lab / Iowa State University.

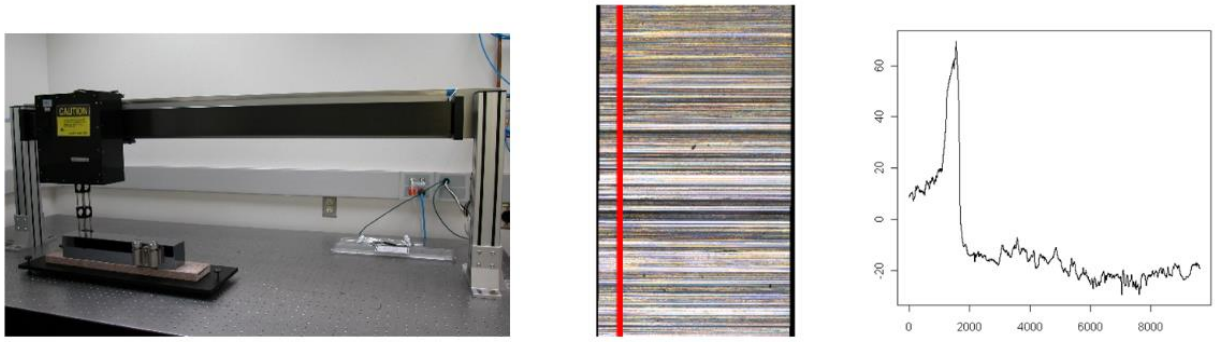
As this work constituted the theses of two graduate students at Iowa State University, the two Tasks as outlined above will be discussed separately, with appropriate sections related to Methods, Results, and Conclusions contained entirely within each Task heading. Much of the text below comes directly from what will be submitted in partial fulfillment of the graduation requirements for these students. Summaries of the work will also be submitted to internationally refereed journals for evaluation by the PIs peers. Plans at this time include submitting results

associated with Task I to Technometrics and those associated with Task II to the Journal of Forensic Science.

TASK I: Statistical Relevance

II. Methods

As previously mentioned, several authors have proposed algorithms to quantify the matching process for striated tool marks. This study focused on the algorithm proposed by Chumbley et al. [1]. For their algorithm, the first step in quantifying the comparison of tool marks is to digitize the marks. Data obtained in that study was acquired using a Hommelwerk surface profilometer, Figure 1(a), with a vertical resolution of 0.005 microns over a 7 mm trace that contained 9600 discrete points or pixels. The data obtained consists of depths of the grooves recorded at a function of "pixel" location. When the numeric depths are plotted by pixel location, the result is a digital tool mark like that shown in Figure 1(c) (N.B. The actual mark left by the screwdrive starts at ≈ pixel location 2000. The large spike on the left is typical of the edge of the toolmark.) Because the striated surface is essentially a set of parallel ridges, most of the useful information about the individual characteristics of the tool can be characterized by this single index data series.



a. b. c.

Figure 1: Using a profilometer to digitize a tool mark. (a) Stylus profilometer. (b) Magnified tool mark showing the location of a profilometer scan, approximately 5 mm in length. (c) Digitized tool mark. Vertical axis is surface height in microns, horizontal axis is pixel number.

The process of matching tool marks generally used in crime laboratories today does not utilize the digitized profilometer data just described, but involves an expert comparing the tool mark found at the crime scene to one or more made in the lab with the suspect tool. A comparison microscope allows the two marked surfaces to be independently moved on different stages while the examiner views both of them in a "split field". The examiner locks in the "best" matching microscopic subset of striae from both tool marks and then looks for similarities in the surrounding striae. In an analogous strategy, the algorithm proposed by Chumbley et al. [1] uses numerical optimization to determine the "best" matching subset and returns a single numerical index value of similarity based on comparing other segments of the data series representing the two tool marks. For details on the process, reference Chumbley et al. [1].

1. Data

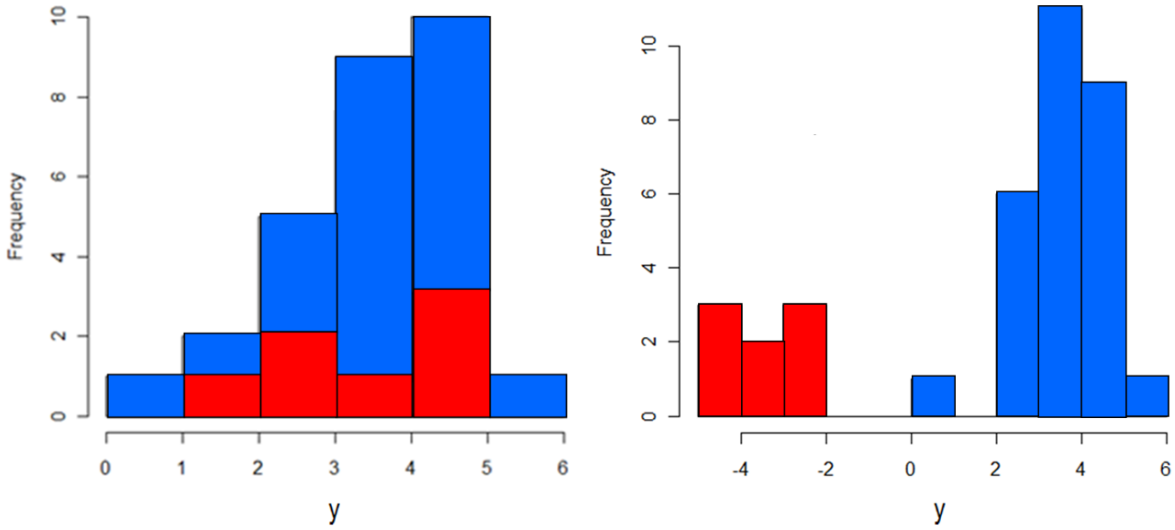
A single data value, as we will refer to it in this paper, is the numerical index of similarity that results from comparing two tool marks. The particular data used in this study, denoted as *y* below, are centered-and-scaled Mann-Whitney U-statistics, which are approximately distributed as normal variates with mean zero and variance one for marks made by different tools (see, for

example, Conover [2]). The extent of data is determined by the collection of tool marks we have to compare. Suppose a single tool mark was found at the crime scene for which we wish to find a match. A suspect tool is obtained and forensic examiners make several marks using the suspect tool in the lab under controlled conditions. Comparing multiple marks all known to be made by the same tool in the lab, we can create a sample of matching mark-pairs from that tool. The same lab tool marks can be compared to the field mark to create a smaller sample of mark-pairs with an unknown matching status. Rather than evaluating only one mark-pair, we now have two samples that we can compare; if there is no apparent systematic difference between these two samples, this supports the argument that the marks were all made by the same tool, i.e. that the crime scene tool mark and lab tool marks "match."

Let x_0 represent the tool mark that was found at the crime scene. Let x_1, ..., x_n represent the n tool marks that were made by the suspect tool in the lab. Note that all comparisons of the marks not including x_0 are known matches all made by the same tool under the same conditions. Let y_{ij} represent the numerical index value of similarity that results from comparing x_i to x_j. If all pairwise comparisons of available tool marks are made, this leaves us with two different types of data. The first set, y_{0j} for j = 1, ..., n, includes all comparisons of the field mark to a lab mark. We will call this type of data field-lab comparisons. The second set is y_{ij} for i, j = 1, ..., n and $i < j$. These data values represent indices of similarity for a known match from the suspect tool which we will call lab-lab comparisons.

The collection of all data values will be denoted by the vector **y** which is of length N. When all possible comparisons are made there are $\binom{n}{2}$ lab-lab comparisons and n field-lab comparisons, so $N = \binom{n}{2} + n$. For the purposes of this paper, we will discuss data in terms of the number of lab marks, such as a data set of size n. Note that this means there are $n + 1$ tool marks under comparison.

The overall question of interest in these analyses is whether or not the suspect tool was used to make the tool mark found in the field. Comparing the two groups of data described previously will help in answering this question. An illustration of possible results can be seen in Figure 2 with the red portions of the bars representing the field-lab comparisons, and blue portions of bars representing lab-lab comparisons. If the field-lab comparisons are indistinguishable from the lab-lab comparisons, such as shown in Figure 2(a), then there is no evidence that the tool marks are different, and the data are therefore consistent with the hypothesis that all marks were created using the same tool. However, if the field-lab comparisons are relatively small compared to the lab-lab comparisons, as in Figure 2(b), then there is evidence that the field mark and the lab marks were created using different tools.



(a) Known match (b) Known non-match

Figure 2: Histograms for datasets showing field-lab comparisons (Red) with lab-lab comparisons (Blue) under known match and non-match conditions.

2. Basic Model

Since we would like to determine whether or not tool marks were made by the same tool, a simple hypothesis test could be used to compare the two samples of field-lab comparisons and lab-lab comparisons. Toward development of such a test, let μ_0 be the mean for a field-lab comparison, that is $E(y_{0j}) = \mu_0$ for $j = 1, ..., n$. Let μ_1 be the mean for a lab-lab comparison, so $E(y_{ij}) = \mu_1$ for $i, j = 1, ..., n$ and $i < j$. We will assume that all data values have a common variance defined as $Var(y_{ij}) = \sigma^2$ for $i, j = 0, 1, ..., n$ and $i < j$. (More general models that include different variances for lab-lab comparisons and field-lab comparisons were investigated in preliminary work, but did not appear to improve performance of the method.) It is further assumed that each y_{ij} is normally distributed. The similarity index of Chumbley et al. [1] is essentially a U-statistic (e.g. Conover [2]), for which an assumption of approximate normality is justifiable. The assumption may also be reasonable for other similarity measures.

Although this is sufficient to fully define the distribution for a single data value, we also need to address the joint distribution of all pairwise comparisons. Each data value is normally distributed, and we will further assume that the joint distribution of **y** is multivariate normal. The mean of **y** is a vector of means μ_0 and μ_1 with the form $\mu = (\mu_0 1'_n, \mu_1 1'_{N-n})'$ and the variance of each element of **y** is σ^2. To finish defining the joint distribution of **y**, we need to develop an appropriate dependency structure reflecting the way the data are generated.

Correlation

Each y_{ij} is the result of comparing two tool marks, specifically x_i with x_j. Thus, at most four physical tool marks are involved in the consideration of covariance between two data values. Since these four tool marks are not necessarily unique, we will say two data values are correlated with correlation ρ if a common tool mark is involved in both comparisons. That is, y_{ij} is correlated with y_{kl} if $i = k$, $i = l$, $j = k$ or $j = l$. Two comparisons with no marks in common are uncorrelated. We do not consider the case of two comparisons made on the same pair of tool

marks because those similarity values would be identical, i.e. there is no measurement-specific "error" in this system, and so no point in replication. Let **R** be the $N \times N$ correlation matrix of **y** defined by the following entries

$$Corr(y_{ij}, y_{kl}) = \begin{cases} 0 \text{ if } i \neq k, i \neq l, j \neq k \text{ and } j \neq l \\ \rho \text{ if } i = k \text{ or } i = l \text{ or } j = k \text{ or } j = l \\ 1 \text{ if } i = k \text{ and } j = l \end{cases} \quad (1)$$

With this correlation structure in place, we can finish defining the joint distribution of all pairwise comparisons of tool marks. Let **y** be the vector of all pairwise comparisons ordered so that each y_{ij} is such that $i < j$. Then $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{R})$ where $\boldsymbol{\mu} = (\mu_0 \mathbf{1}_n', \mu_1 \mathbf{1}_{N-n}')$ and **R** is as defined in (1). The complete model for a data set of size $n = 4$ is shown in (2) through (4).

$$\mathbf{y} = (y_{01}, y_{02}, y_{03}, y_{04}, y_{12}, y_{13}, y_{14}, y_{23}, y_{24}, y_{34})' \quad (2)$$

$$E(\mathbf{y}) = \boldsymbol{\mu} = (\mu_0, \mu_0, \mu_0, \mu_0, \mu_1, \mu_1, \mu_1, \mu_1, \mu_1, \mu_1)' \quad (3)$$

$$Var(\mathbf{y}) = \sigma^2\mathbf{R} \text{ with } \mathbf{R} = \begin{pmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & \rho & 0 & 0 & \rho & \rho & 0 \\ \rho & \rho & 1 & \rho & 0 & \rho & 0 & \rho & 0 & \rho \\ \rho & \rho & \rho & 1 & 0 & 0 & \rho & 0 & \rho & \rho \\ \rho & \rho & 0 & 0 & 1 & \rho & \rho & \rho & \rho & 0 \\ \rho & 0 & \rho & 0 & \rho & 1 & \rho & \rho & 0 & \rho \\ \rho & 0 & 0 & \rho & \rho & \rho & 1 & 0 & \rho & \rho \\ 0 & \rho & \rho & 0 & \rho & \rho & 0 & 1 & \rho & \rho \\ 0 & \rho & 0 & \rho & \rho & 0 & \rho & \rho & 1 & \rho \\ 0 & 0 & \rho & \rho & 0 & \rho & \rho & \rho & \rho & 1 \end{pmatrix} \quad (4)$$

Likelihood Analysis

The model described suggests there could be separate means for the two available samples, field-lab comparisons and lab-lab comparisons. To test whether the same tool made the field and lab tool marks, we can set up the hypothesis test

$$H_0 : \mu_0 = \mu_1 \text{ vs } H_A : \mu_0 \neq \mu_1. \quad (5)$$

Using Normal model theory and weighted least squares, maximum likelihood estimates (MLEs) can be easily derived for $\boldsymbol{\mu}$ and σ^2 given a value for ρ as follows

$$\hat{\boldsymbol{\mu}}|\rho = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \quad (6)$$

$$\hat{\sigma}^2|\hat{\boldsymbol{\mu}}, \rho = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\mu}})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\mu}})/N \quad (7)$$

where $\mathbf{X} = 1_N$ under H_0 and $\mathbf{X}= \begin{bmatrix} 1_N & 0 \\ 0 & 1_{N-n} \end{bmatrix}$ under H_A. Using a grid search for ρ, we can compute parameter estimates as the values that maximize the normal likelihood. However, we must first establish viable values for ρ.

Based on physical considerations, we require that ρ be non-negative. This is reasonable because correlation is used to model the effect of a tool mark common to two pairs. The structure of this correlation matrix forces stricter boundaries on the range of values for ρ. In particular it can be shown that $\rho < 0.5$. The proof of this relies on the fact that **R** must be positive definite. Specifically, it can be shown that the unique eigenvalues for **R** are

$$\lambda_1 = 1 - 2\rho,\ \lambda_2 = 1 + \rho(n-3) \text{ and } \lambda_3 = 1 + 6\rho + 2\rho(n-4)$$

By definition, a matrix is positive definite only if all of its eigenvalues are positive. Applying this definition to the three eigenvalues results in the following implications and inequalities

$$\lambda_1 > 0 \implies \rho < 1/2 \tag{8}$$

$$\lambda_2 > 0 \implies \rho > \frac{-1}{n-3} \text{ or } \rho < \frac{1}{3-n} \tag{9}$$

$$\lambda_3 > 0 \implies \rho > \frac{-1}{2n-2} \text{ or } \rho < \frac{1}{2-2n} \tag{10}$$

Using (8) we can see that $\rho < 0.5$. Looking at (9), if n = 1, 2 or 3 we get the inequalities $\rho > 1/2$, $\rho > 1$ and $\rho > 1/0$ respectively, which are all impossible given (8) or beyond the known bounds for the Pearson correlation coefficient. Thus we know we need a sample size of 4 or more. For both (9) and (10), the larger the sample size becomes the closer the bound gets to zero. Combining these results with our requirement that ρ not be negative, we can say $\rho \in [0, 0.5)$.

Using a likelihood ratio test (LRT) for the null and alternative models described in (5), the resulting p-value will determine whether or not there is evidence of a match. To demonstrate this process, we return to the data that provided the histograms in Figure 2. In the first example, Figure 2(a), there was a data set of size n = 7, that is one field mark and seven lab marks, resulting in a total of 28 data values all of which were known to come from the same tool. Seven of the values are shown in red and represent the field-lab comparisons. The remaining data values are shown in blue and represent known matches from lab-lab comparisons. From the histogram we can see that the field comparisons are indistinguishable from the lab comparisons which was to be expected. Table 1 shows the parameter values and maximized likelihood for both the null and alternative models. The likelihood ratio statistic,

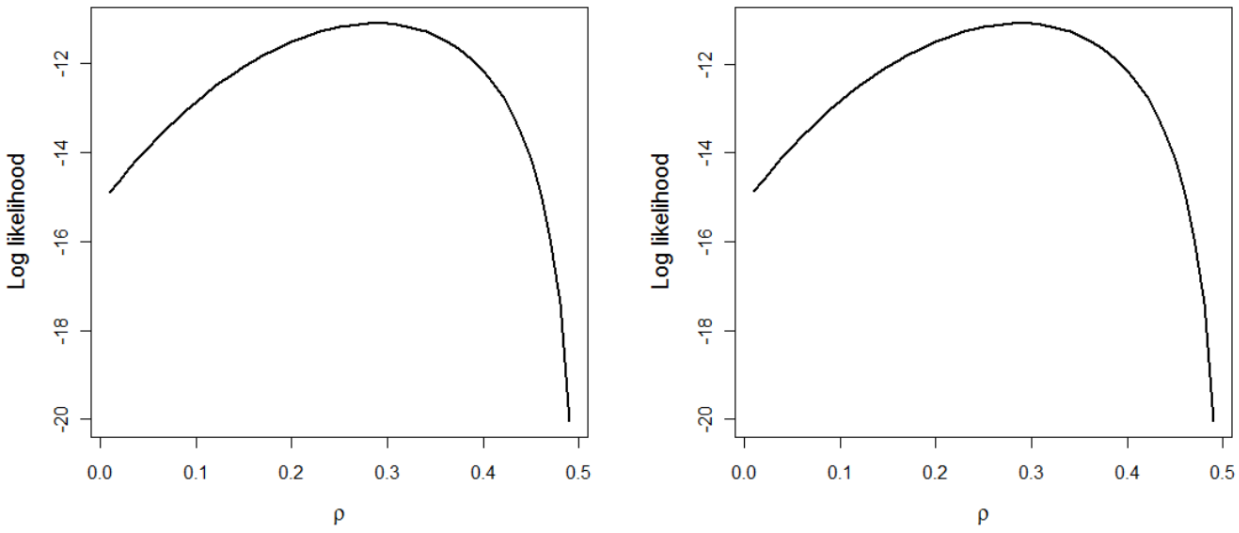
$$\lambda = \frac{\ell(\hat{\mu}_0, \hat{\sigma}^2, \hat{\rho})}{\ell(\hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}^2, \hat{\rho})}$$

is such that -2 ln (λ) approximately follows a χ^2 distribution with one degree of freedom. For these data -2 ln (λ) = 0.030 resulting in a large p-value of 0.8875. (Wilks [3] is a historical reference for statistical tests constructed in this way.) The null hypothesis that the means are equal is rejected and conclude there is no evidence that the tools are different.

	Null Model	Alternative Model
$\ln \ell(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2, \hat{\rho})$	-11.100	-11.086
$\hat{\rho}$	0.289	0.288
$\hat{\mu_0}$	3.558	3.468
$\hat{\mu_1}$		3.588
$\hat{\sigma}^2$	1.176	1.173
$-2\ln(\lambda) = 0.030$, p-value $= 0.8875$		

Table 1: MLEs for known matching data displayed in Figure 2(a)

Figure 3 shows the profile likelihoods for ρ for the null model (a) and the alternative model (b). The plots are nearly identical for the two models suggesting that the models are the same, and only one mean is necessary. This supports the conclusion from our hypothesis test since we failed to reject the hypothesis that μ_0 and μ_1 are equal. Furthermore, both plots are maximized around 0.3 which reinforces the need for correlation in the model.



(a) Null Model (b) Alternative Model

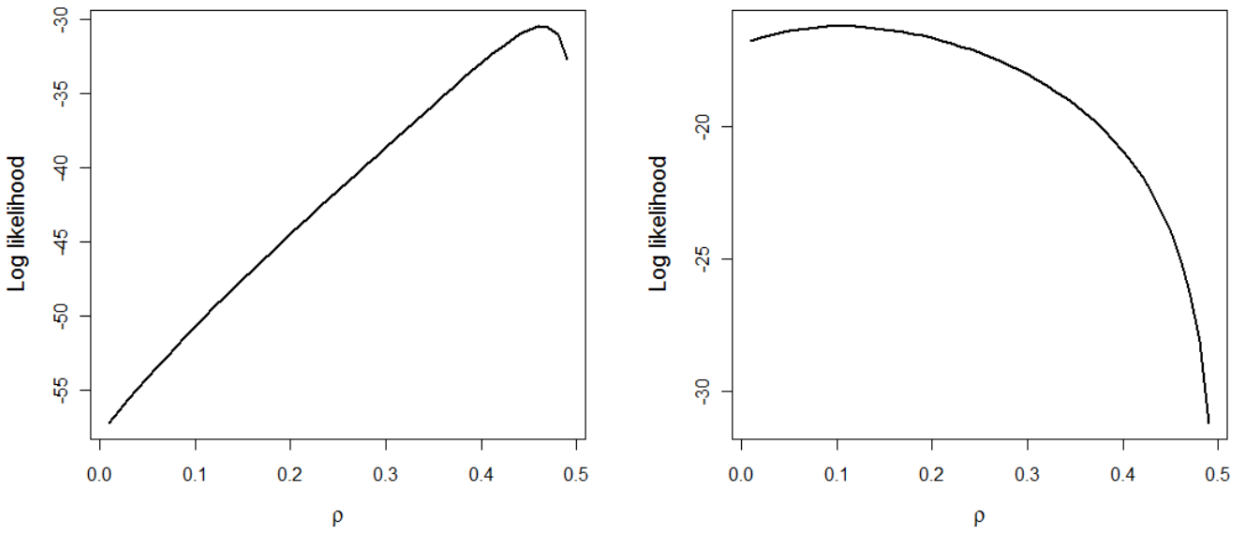
Figure 3: Profile likelihoods for ρ for the Null and Alternative Models

In the second example, Figure 2(b), the field tool mark was known to originate from a different screwdriver than the lab marks it was tested against. Here $n = 8$ so there was one field mark available and eight lab tool marks for a total of 36 data values. Again, the histogram shows the field-lab comparisons in red and the lab-lab comparisons in blue, however, now we see a clear separation in the two types of samples with the field-lab comparisons being much smaller than the lab-lab comparisons. Table 2 shows the parameter estimates from maximizing the likelihood for these data under the null and alternative hypotheses and the likelihood ratio statistic, $-2\ln(\lambda) = 28.417$. The small p-value of $9.78 \times 10{-}8$ indicates we should reject the null hypothesis that the means of the two samples are equal. From this we would conclude there is strong evidence that the two samples were created using different tools.

	Null Model	Alternative Model
$\ln \ell(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2, \hat{\rho})$	-30.4099	-16.2016
$\hat{\rho}$	0.464	0.105
$\hat{\mu_0}$	2.025	-3.357
$\hat{\mu_1}$		3.563
$\hat{\sigma}^2$	10.416	0.958
$-2\ln(\lambda) = 28.417$, p-value $= 9.78 \text{ x } 10^{-8}$		

Table 2: MLEs for known non-matching data displayed in Figure 2(b)

When we examine the profile likelihoods for ρ from these non-matching data, Figure 4, we can see the two plots are very different from one another. The plot from the null model, (a), shows that the correlation coefficient reaches a maximum near the upper bound around 0.45, but the plot from the alternative model, (b), reaches its maximum near the lower bound around 0.1. This supports the results that the two models are different and a second mean is needed to model these data. However, the MLE of ρ is non-zero in each model reinforcing the argument that the correlation structure in (1) is appropriate.



(a) Null Model (b) Alternative Model

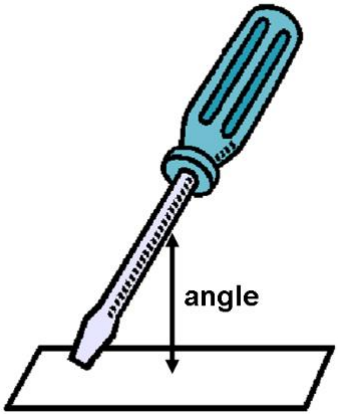
Figure 4: Profile likelihood for ρ for Null and Alternative Model

3. Angle

Angle Influence

As we saw from the examples in Section 2, the likelihood analysis is effective in determining a match of tool marks under carefully controlled test conditions in which all tool marks are produced the same way. However, in order to be useful in practice, it must also perform well

when tool marks are made under different conditions. When a screwdriver is scraped against a metal surface, specific circumstances such as the pressure exerted on the tool and the angle between the tool and the surface, can affect the appearance of the tool mark. Here we will consider the angle at which a mark is made, a measurable quantity that can be analyzed and accounted for to enhance the model and analysis described in Section 2. For clarification, the angle at which a mark is made is measured as the smallest angle the tool makes with respect to the marked surface, illustrated in Figure 5.

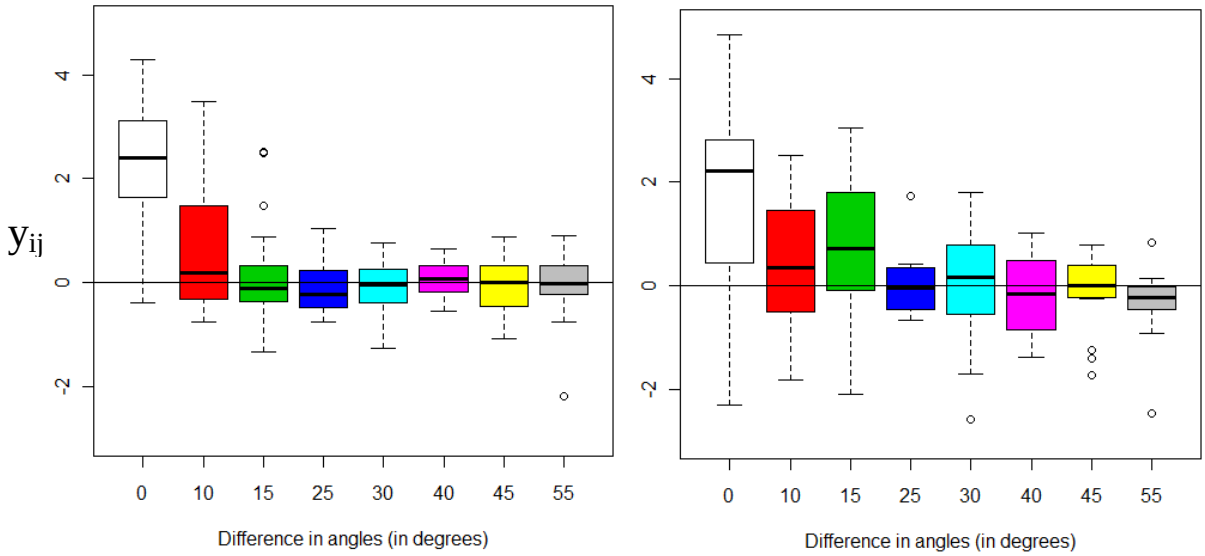


a.

b.

Figure 5: a.) Depiction of the angle between a screwdriver and a marked surface. b.) Jig used for making markings

Chumbley et al. [1] demonstrated that when two tool marks are made by the same tool, similarity indices are generally much larger when the tool angle is the same in each case. To demonstrate this effect with our own data, we produced tool marks with a single screwdriver at 30°, 45°, 60°, 75° and 85°. At each angle, four separate tool marks were made. Although they are made at different angles, they all represent what we have defined until now as marks that match. Pairwise comparisons of the twenty available marks were made resulting in 190 data values. This process was repeated for two different screwdrivers. Figure 6 displays boxplots of the data for each tool grouped by the difference in the tool angles of the marks being compared.

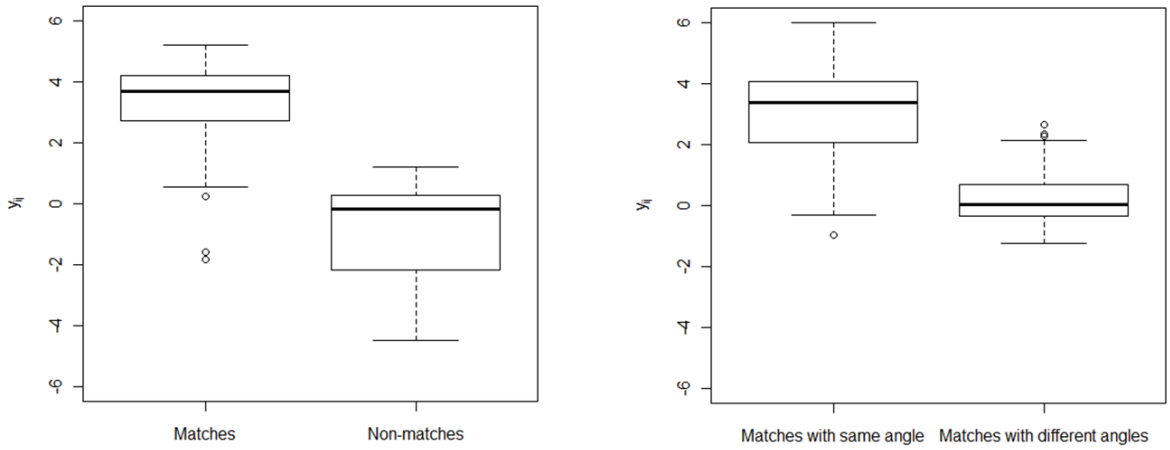


(a) Tool 1 (b) Tool 2

Figure 6: Boxplots for known match comparisons y_{ij} at different angles.

We know empirically that when tool marks are made at the same angle, and the field mark was not made by the suspect tool, the field-lab comparisons are centered around zero, while the lab-lab comparisons that match have a mean around two or three. From the boxplots in Figure 6, we can see that when the angles are the same, the data are centered around two or three like a true match. However, for angles that differ by 25°, the data are smaller and centered around values closer to zero. In other words, even when marks are made by the same tool, if the tool angles differ by 25°, the comparison values resemble those from non-matching tool marks. This is in agreement with statement by toolmark examiners who state that when the angle varies by 10˚ or more the marks start to appear distinct from each other and positive identifications can not be made.

To better demonstrate the similarities between comparing matches to non-matches and comparing matches made at the same angle to those made at different angles, we refer to Figure 7. The data that were used to make Figure 2 were grouped into matches and non-matches and are shown in the boxplots in Figure 7(a). We then used the same data but included only the comparisons that were made by the same tool, and collected the tool angle information for each mark. The matching data was then grouped into matches made at the same angle and those made at different angles, which is shown in Figure 7(b). For these data, the only available angles were 30°, 60° and 85° so all angles differed by at least 10°. Side-by-side, the similarities between comparing matches to non-matches, and comparing matches made at the same and differing tool angles is more apparent. This further shows that for tool angles that differ by 10◦ or more, the data are no longer identifiable as a match.



(a) Boxplots of matches and non-matches.

(b) Boxplots of marks made at the same and different angles.

Figure 7: Boxplots showing the similarities between data from different tools made at the same angle and data from the same tool made at different angles.

Knowing that tool angle has a significant effect on the data, it is important to generalize the approach described in Section 2 to account for these effects. One difficulty is that it is impossible to know the tool angle that was used to make a mark left at the crime scene. However, in a lab, tool marks can be made at any angle to try to better match a crime scene mark. Note here that with current procedures, tool mark examiners often do make marks at multiple angles for this purpose. If enough tool marks are made in the lab at angles differing by 10° or less, perhaps we can find the best match of a field mark to lab marks to estimate the angle at which the field mark was made and determine if the suspect tool was consistent or inconsistent with the crime scene mark.

Model with Angle

Before we modify the basic model of Section 2 to account for angle information, we need to introduce more notations. Let a_i be the tool angle, in degrees, at which tool mark x_i is made for i = 0, 1, ..., n. Since we know the closer the angles are, the better the match will be, we will incorporate tool angles as a function of their absolute difference, $|a_i - a_j|$, which we will denote by the similarity measure $d(a_i, a_j)$.

We saw from Figure 6, that the mean response for data values is large when the angles are the same and approaches zero as the difference in angles increases. We also observed that comparisons of matching tool marks made at angles differing by 10° or more resemble non-matches, so these facts will be reflected in $d(a_i, a_j)$. The function we chose to represent the difference in angles was $d(a_i, a_j) = \exp[-\theta(a_i - a_j)^2]$, where $\theta = 0.01$. This was chosen so that $d(a_i, a_j) = 1$ when tool angles match, is much smaller when $|a_i - a_j| = 10$ and approaches zero as $|a_i - a_j|$ continues to increase. Figure 8 shows this behavior of $d(a_i, a_j)$ as a function of $|a_i - a_j|$.

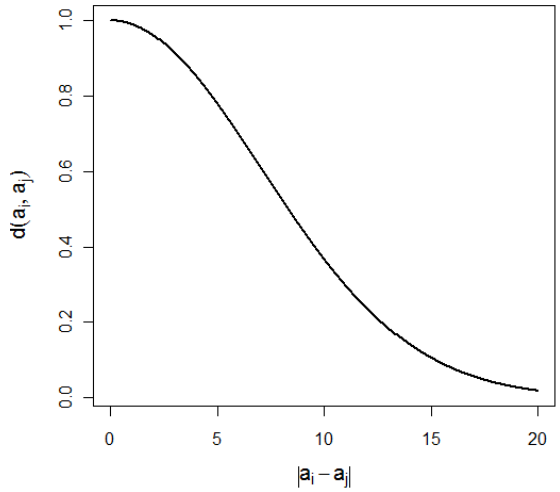


Figure 8: Graph showing $d(a_i, a_j)$ as a function of the absolute distance between angles, $|a_i - a_j|$.

A modified data model incorporating tool angle is:

$$y_{ij} \sim N(\mu_{ij}, \sigma^2) \text{ where} \tag{11}$$

$$\mu_{0j} = \mu_1 + \alpha_0 d(a_0, a_j) \text{ for } j = 1, ..., n \text{ and} \tag{12}$$

$$\mu_{ij} = \mu_1 + \alpha_1 d(a_i, a_j) \text{ for } i, j = 1, ..., n \text{ and } i < j. \tag{13}$$

where α_0 and α_1 can be thought of as regression coefficients representing the extent of the effect of angle similarity on the data. So, where the field and lab marks are not made by the same tool, it would be expected that α_0 would be near zero, since similar angles should not improve the "apparent similarity" of tool marks in this case. That is, the mean of a data value is a linear function of the overall mean of the comparisons and the similarity measure between the tool angles used in making the marks. Similar to the previous model we have allowed for different mean functions for field-lab comparisons and lab-lab comparisons through different regression slopes. Thus we can again make inference about whether or not the suspect tool made the crime scene marks with a likelihood ratio test. However, we are now interested in comparing the alternative hypothesis defined by the model described in (11) through (13) with a null hypothesis defined by a simpler model that does not discriminate between field-lab and lab-lab comparisons. That null model can be stated as

$$y_{ij} \sim N(\mu_{ij}, \sigma^2) \text{ where} \tag{14}$$

$$\mu_{ij} = \mu_0 + \alpha d(a_i, a_j) \text{ for } i, j = 0, 1, ..., n \text{ and } i < j. \tag{15}$$

Both the null and alternative models assume that we have known angles for every tool mark made in the lab; that is a_1, ..., a_n are known. The angle of the mark made in the field, a_0, is unknown. Thus, a_0 is a parameter in the model along with α, α_0, α_1, μ, σ^2 and ρ. The correlation structure described in Section 2.1 remains for this model and ρ will still be chosen using a grid search between 0 and 0.5. Since we know the tool angle needs to be accurate within 10° to see evidence of a match, we will perform a grid search for a_0 in increments of 5° between 20° and 90°. These angle bounds were chosen as reasonable angles for which a tool mark could be made and leave behind a viable mark. Maximum likelihood estimates for the remaining

parameters can be computed using weighted least squares provided values of a_0 and ρ. Let $\boldsymbol{d_0} = (d(a_0, a_1), d(a_0, a_2), ..., d(a_0, a_n))'$ and $\boldsymbol{d_1} = (d(a_1, a_2), d(a_1, a_3), ..., d(a_{n-1}, a_n))'$. Then the MLEs for this modified model can be computed as

$$\hat{\boldsymbol{\beta}}|a_0, \rho = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \tag{16}$$

$$\hat{\sigma}^2|\hat{\boldsymbol{\beta}}, a_0, \rho = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/N \tag{17}$$

where $\beta = \begin{bmatrix}\mu_0 \\ \alpha\end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix}\mathbf{1}_N & (\mathbf{d_0'}, \mathbf{d_1'})'\end{bmatrix}$ under the modified null model and $\beta = \begin{bmatrix}\mu_1 \\ \alpha_0 \\ \alpha_1\end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix}1 & \mathbf{d_0} & \mathbf{0_n} \\ \mathbf{1}_{N-1} & \mathbf{0}_{N-n} & \mathbf{d_1}\end{bmatrix}$ under the modified alternative model.

III. Results

1. Introduction

To test the modified models, tool marks were made from all the available jigs in our lab, corresponding to tool angles of 30°, 45°, 60°, 75° and 85°. At each of the available tool angles, four unique tool marks were made using each of six screwdrivers, resulting in a total of 5 × 4 × 6 = 120 tool marks. To fully test the modified model, we will consider two scenarios, one where the suspect tool made the field mark and one where it did not. In the first set, all the data from a single tool are used. This reflects a situation where the analysis should indicate a "match" and the results can be seen in Section 1.1. For the second scenario, we consider the case where the field mark was made by a different tool than the lab marks. The non-matching results are discussed in Section 1.2.

Results for Matches

To collect the data for known matches, only tool marks made by the same tool were compared to one another. Data sets were compiled using all 20 tool marks for a given tool; four marks from each of the five available angles. Each tool mark within a set was chosen one-at-a-time to be the field mark leaving the remaining n = 19 lab marks for comparison, three of which are made at the same tool angle as the field mark. This process was done for every available tool mark from all six tools and the likelihood ratio test described in Section 3.2 was performed for each (i.e. 120 tests in all).

Since we know all the tool marks used in each analysis are made by the same tool, we would expect the field-lab comparisons and lab-lab comparisons to result in similar data values so only one regression slope for the similarity measure, $d(a_i, a_j)$, should be needed for an adequate model fit. Thus, the null model described in (14) and (15) should be the best model for these data. Based on this assumption, we would expect that the distribution of p-values from these likelihood ratio tests should be distributed approximately uniformly between 0 and 1. The actual resulting p-values for all six tools are shown in Figure 9.

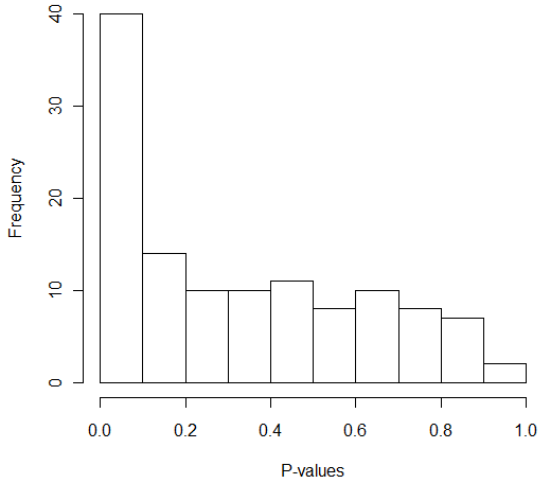
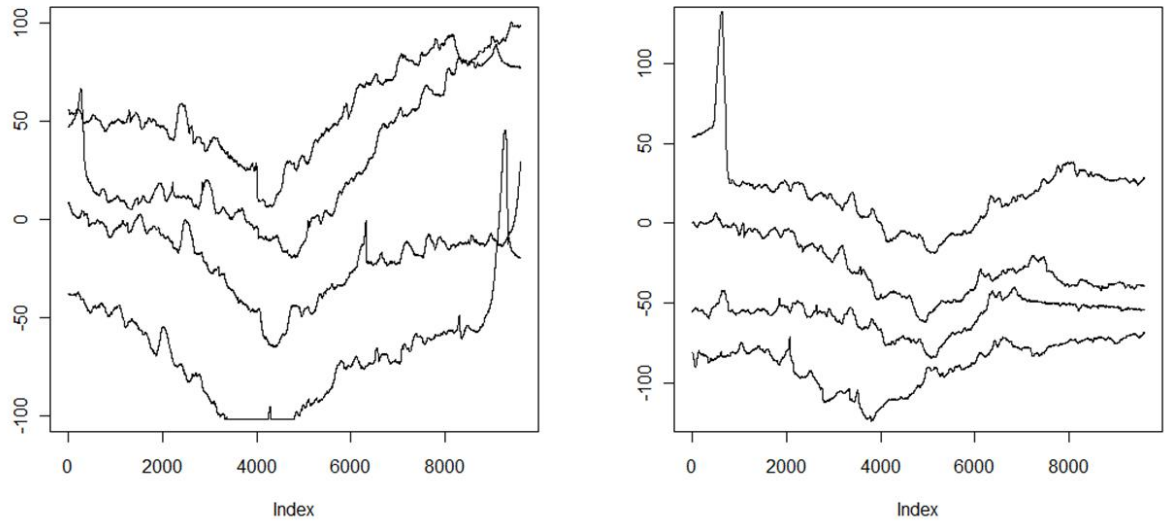


Figure 9: Histogram of p-values for all matching data.

With the exception of the first bin from 0.0 to 0.1, the p-values do appear to be close to uniform in distribution. However, there is a significant peak where a large portion of the LRTs result in very small p-values which is evidence of non-matching data, contrary to the fact. To determine the source of these unexpected results, we looked for trends in the specific field tool marks whose LRTs had p-values less than 0.1. In doing so, we noticed that when a likelihood ratio test returned a small p-value for a field mark, it was often the case that one or more of the other three tests from marks made at the same angle by the same tool also led to small p-values. This suggested the possibility that a single tool mark might be responsible for several unexpected test results, leading us to a further look into the individual tool marks.

We noticed two recurring issues in the data were present in the majority of field marks that resulted in LRTs with small p-values. The first was an issue in the quality of data, in particular, a flaw in the individual tool marks which can occur in the form of a "flatline" or a "sharp peak". During the digitizing process for a tool mark, the stylus profilometer reads the depths of the grooves of the striae in the tool mark, as described in Section 1. In some of the tool marks, the stylus reached the minimum or maximum value it can record. As a result, areas of the digitized tool mark are perfectly flat over a consecutive set of pixel locations. For consistency, we define a flatline in a tool mark to be any region of 200 or more consecutive pixels with constant response, i.e. a variance of zero. In a few circumstances, a flatline occurred for one of the tool marks but not the others made at the same angle with the same tool. These were responsible for some of the apparent "non-matches" and small p-values.

An example of a matching set of tool marks where one mark has a flatline flaw is shown in Figure 10(a). The four lines represent all four digitized tool marks made by the same tool at the same angle. Note that the actual value of the depth on the y-axis is not meaningful since these lines have been shifted vertically to make them easier to see. The top three digitized marks appear very similar, but the fourth tool mark contains a flatline beginning in the vicinity of pixel 3000. As a result of this flaw, the validation step of the algorithm of Chumbley et al. [1] is misled when comparing this flatlined tool mark to its true matches since areas of the tool marks that should be similar are not.



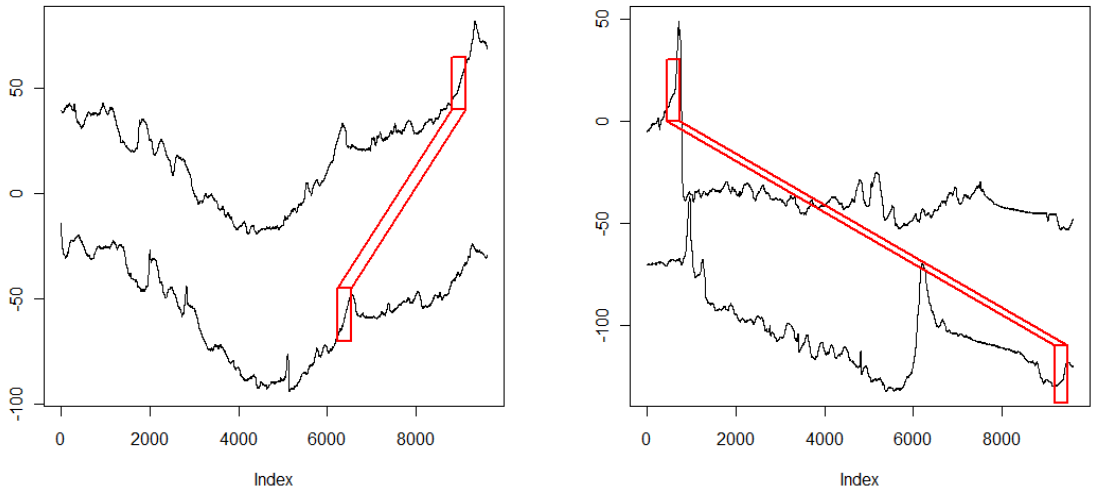
(a) A set of tool marks demonstrating a flatline. (b) A set of tool marks demonstrating a sharp peak.

Figure 10: Examples of flaws in the tool marks.

The second type of data quality flaw is a "sharp peak". This is a phenomenon that tends to occur at the end of a tool mark when the stylus goes from recording the tool mark to recording the flat lead plate on which the tool mark was made, resulting in a sharp jump in the digitized mark. Similarly to the flatline flaws, these peaks cause the validation process in the matching algorithm to fail since those areas cannot be aligned with the other tool marks made at the same angle. These peaks occur at varying degrees of severity in the tool marks; we chose to define a tool mark as having a sharp peak if a consecutive set of 200 pixel locations have a variance of 750 or more. An example of a sharp peak is shown in Figure 10(b). Even though the four tool marks appear to have the same pattern along most of the pixel locations, there is a sharp peak in the top mark near the left end of the mark. As a result, when chosen as the field mark, comparisons to this tool mark resulted in a small p-value since the spike ruins the similarity that should be apparent in corresponding segments of the tool marks during the validation step.

In addition to problems associated with data quality, we found that some of the very small p-values originated from a more fundamental problem with the algorithm of Chumbley et al. [1]. In the first step of the algorithm, "best match" windows are chosen from each of the tool marks being compared, within which the correlation between the two digitized marks is greatest. The validation step then compares the marks in nearby "validation windows" based on the locations of these best match windows. If the best match windows are chosen incorrectly, i.e. the algorithm selects a pair of segments that do not physically correspond, the rest of the algorithm will fail since the validation windows will not line up properly. After reviewing the tool marks that returned small p-values, we noticed there were several instances where data values using that mark had best matching windows that were incorrectly chosen. This is a failure of the algorithm that we will call a "bad match". We define a bad match as a situation where the best match windows are two windows that do not actually correspond, even though a good pair of matching windows can be seen in the tool marks.

An example of a bad match is shown in Figure 11(a). The red boxes represent the two windows that were chosen by the algorithm to be the best match. However, we can see that they do not really match, but match can occur when the best matching windows are chosen on the opposite outer edges of the tool marks. Figure 11(b) shows an example where the best match windows occur on the extremes of the tool mark, but opposite one another so they do not represent a true match.



(a) An example of a bad match. (b) An example of a bad match.

Figure 11: Examples of flaws in the matching process of the algorithm.

After examining all of the digitized tool marks and data values, we found that there were four tool marks that have sharp peaks, seven tool marks that have a flatline, and one that has both a flatline and a sharp peak. In addition, there were 32 data values, out of the 180 that were from true matches in both tool and tool angle that have a poorly chosen best match window and were deemed bad matches. Because all of these situations were caused by either a failure in the algorithm or poor quality of tool marks, they were removed from their respective data sets and the analyses were run again. The p-values of the remaining108 likelihood ratio tests are shown in Figure 12. The initial peak of p-values between 0 and 0.1 is substantially reduced in Figure 12 relative to Figure 9. The p-values appear closer to uniform in distribution as we had expected since all data values represent matches in this case.

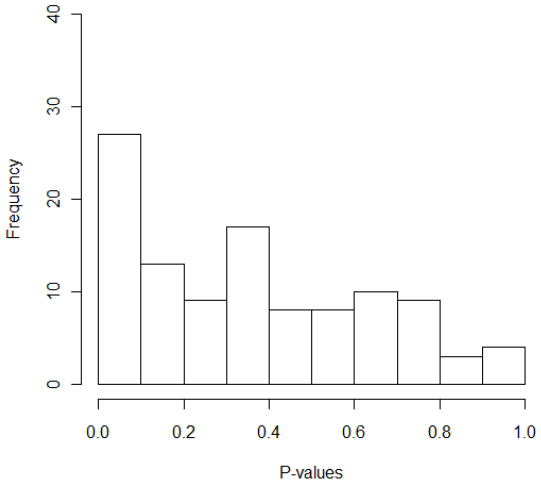
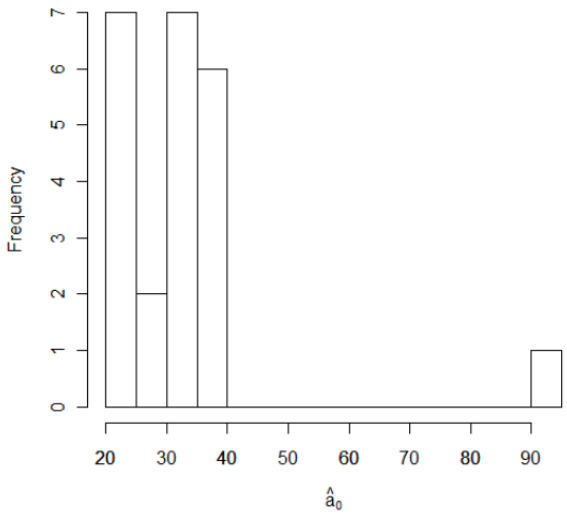


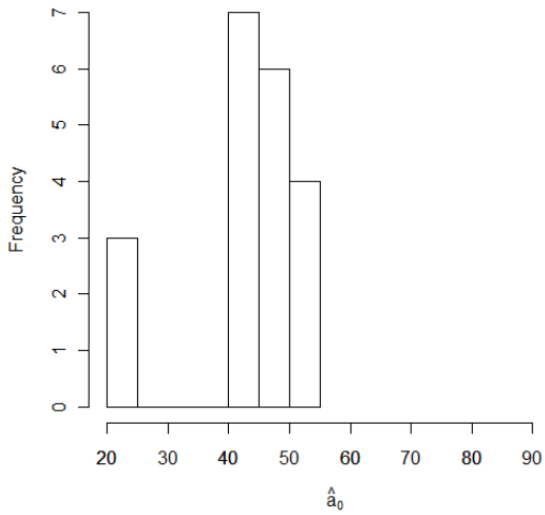
Figure 12: Histogram of p-values after removing bad matches and tool marks with flatlines or sharp peaks.

To further examine the efficiency of the modified models and the LRT once the problematic data were removed, we also examined the accuracy of the estimates of a_0, the tool angle with which

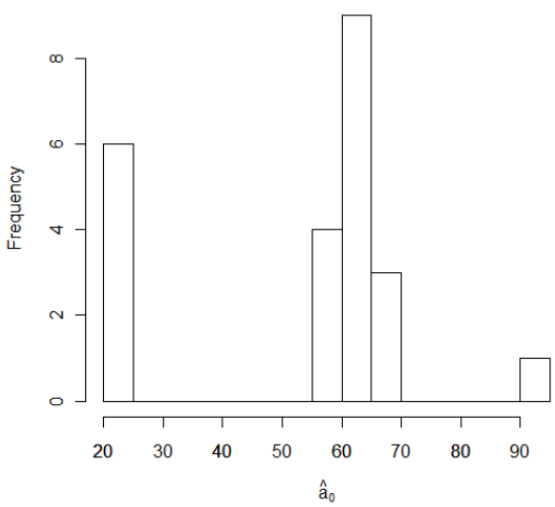
the field mark was made. The histograms in Figure 13 show the estimates, $\hat{a}_0$, based on the null model described in (14) and (15) grouped by the true value of a_0. Each of the histograms in Figure 13 shows that the estimation does accurately predict the angle within 5 degrees of the true angle in most cases. Occasionally an estimate is very different from the true value, as shown, for example, by the small bar at 90° when the actual angle was 30° in Figure 13(a). Since we know there are issues with both the quality of the data and the matching algorithm some of these inaccurate estimates may be associated with more subtle problems in the data generation process. Overall, the null model does fit well to the data in most cases where the data are matches, and the estimation process for the field angle seems reasonably accurate.



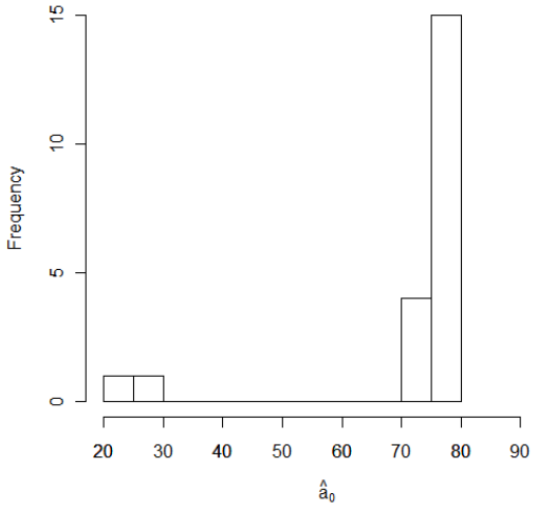
(a) Values of $\hat{a}_0$ when $a_0 = 30$



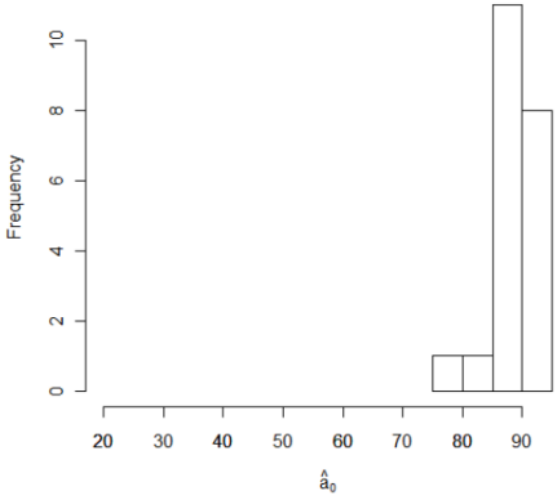
(b) Values of $\hat{a}_0$ when $a_0 = 45$



(c) Values of $\hat{a}_0$ when $a_0 = 60$



(d) Values of $\hat{a}_0$ when $a_0 = 75$



(e) Values of $\hat{a}_0$ when $a_0 = 85$

Figure 13: Estimated values of a0 grouped by the true value of a0 for matching data.

Results for Non-Matches

To create non-matching data sets, we began by using all n = 20 tool marks made from the same tool and considered these the lab marks. Field marks were chosen out of the remaining five tools, one mark from each of the 5 angles available for each tool. The data sets were assembled by comparing the field mark with each of the lab marks and comparing the lab marks pairwise with one another. This process was repeated for two different sets of lab tool marks, resulting in a total of 50 data sets for which the lab tool and field tool are not the same.

For non-matching data, all of the field-lab data values should be close to zero regardless of the tool angles since the marks were made by different tools. However, most lab-lab comparisons will result in a larger comparison value since they are true matches. This discrepancy should show up in the models through the regression slopes. We expect that the alternative model will be a better fit to these data since the coefficient for the angle similarity function, $d(a_i, a_j)$, will likely be close to zero for the field-lab comparisons but should still be large for the lab-lab comparisons. Thus we would expect the alternative model to be a better fit to these data and so the p-values from the likelihood ratio tests should be small, i.e. the distribution of p-values should be skewed with greater frequencies associated with smaller p-values. The p-values that resulted from these tests are shown in the histogram in Figure 14. As expected, the p-values are mostly small and have an overall right skewed shape. This supports the hypothesis that the alternative model which allows for multiple regression slopes is a better fit to these data and provides evidence that the lab tool marks do not match the field mark in most of these tests.

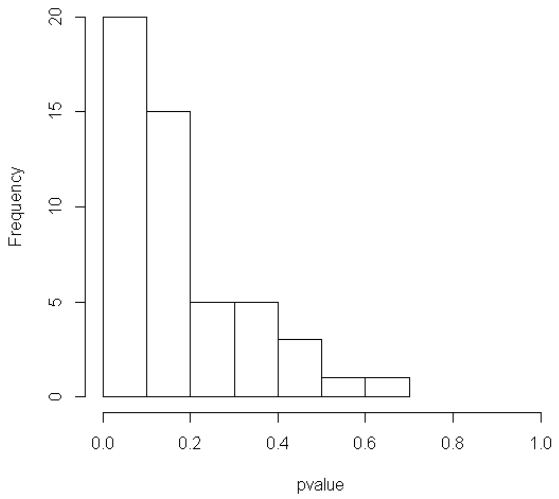


Figure 14: Histogram of p-values for non-matching data.

IV. Conclusions

1. Discussion of Findings

This work confirms that the angle at which a tool is used can have a significant effect on the similarity of the tool marks. Since it is these similarities that are crucial to the tool mark matching process, it is necessary to account for these effects in the models and analyses used to examine tool marks. Including the tool angle in the model as a function of the difference in angles of marks being compared does seem to yield positive results, both in the ability of the likelihood ratio tests to choose an appropriate model, reflecting a match or non-match, and in estimating the unknown field angle. As long as lab tool marks are made at angles within 15° of the field angle, the estimation process is reasonably accurate when the tool marks match.

Due to the constraints of available resources, we were only able to collect tool marks from a few angles, those being 30°, 45°, 60°, 75°, and 85°. As a result we were forced to choose a parameter value for θ in the similarity function of the modified models. Having tool marks made at more angles would allow us to include θ in the estimation process. The lack of angles available also limited the precision with which we could evaluate the model since there were few angles close to one another. Having more angles available would allow us to see how well we can estimate the field angle when there are many marks made at similar angles in the lab set.

Another area where future work is necessary is in studying the quality of tool marks. In the case of matching data, we saw there are flaws in the tool marks as well as in the algorithm used to compute the match score that need to be addressed. In the tool marks, any unusual area of the digitized mark, such as a flatline or sharp peak, can lead to incorrect estimation and conclusions for the matching status. Both of these flaws are caused by failure of the measurement device. Similarly, if the algorithm fails to accurately choose appropriate best matching windows in the tool marks, the analyses will have incorrect conclusions. Although further work may be necessary to account for the effects of varying data quality in the model, we saw that once the bad matches and flawed data had been removed, the modified models and likelihood ratio tests based on them did perform as was expected.

2. Implications for Policy and Practice.

The rigorous analysis conducted, using replicate marks made from a series of screwdrivers, was able to show that replicate-to-replicate variation is much less than tool-to-tool variation, allowing comparisons of field-lab and lab-lab comparisons to be conducted successfully to match the field mark to the correct lab produced mark. The study shows that, if given enough lab replicate marks for comparison, the analysis is so reliable that even the angle at which the field mark was made can be determined to a high degree.

The implications of this study for policy and practice are as follows. The statistical analyses carried out on toolmarks in this work clearly adds further evidence that the long-held assumptions under which forensic examiners operate, namely, that all toolmarks are unique, does have a sound scientific basis. However, if a truly objective, quantitative analysis is desired such as was carried out in this study, it may require examiners to generate more lab samples for comparison than is typically done now. While it is know anecdotally that examiners do currently prepare replicate marks, it is not known how many are made, or how rigorously all of those marks are examined. The need to generate additional samples to replicate the results of this study in the forensic lab could increase the work load of examiners. However, that drawback might be offset by the increased reliability and objectivity of the data produced.

3. Implications for Further Research.

The results found in this study suggest a number of possible areas for further research. While the work described here effectively provides for control over the rate of false non-match calls, it does not address the more important issue of false match calls. Current preliminary research indicates that this may also be possible through a different statistical testing structure. Since it is clear that risk-control may heavily depend on making more lab marks for comparison to single evidence mark, quality control processes will also be particularly important for determining which, if any, of the replicate lab marks are of poor quality and so should not be included. We will propose that this can be done with indices of the sort we have been using, but applied to pairs of lab-generated marks so as to identify any marks that consistently show weak similarity to others.

V. References

1. Chumbley, L.S., Morris, M.D., Kreiser, J., Fisher, C., Craft, J. Genalo, L.J., Davis, S., Faden, D., and Kidd, J., "Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm," Journal of Forensic Sciences, Vol. 55, No. 4, Jul. 2010, pp. 953-961.

2. Conover, W.J. (1999), *Practical Nonparametric Statistics,3rd Edition,* ISBN-13: 978-0471160687, John Wiley & Sons, New York.

3. Wilkes, S.S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypothese," *The Annals of Mathematical Statistics* **9**: 60-62.

VI. Dissemination of Research Findings

As stated at the beginning of this report, this project was conducted as part of the thesis work of Ms. Amy Hoeksema, a graduate student at Iowa State University, and much of the text comes from what will be include in her Ph.D. thesis. The entire thesis will be available from Iowa State University. In addition, a paper entitled "Significance of Angle in the Statistical Comparison of Toolmarks" is being prepared for submission to the refereed technical journal Technometrics. .

Dissemination has also occurred by oral presentations given by Prof. Morris and Ms. Hoeksema including the following:

M. Morris, A. Hoeksema, S. Chumbley, "Statistical Modeling of Tool Mark Comparisons Incorporating Material Correlation and Tool Angle," Pattern and Impression Evidence Symposium, Clearwater Beach, FL, August, 2012.

A.Hoeksema, "Statistical Comparison of Toolmarks in Forensics," Joint Statistical Meetings, San Diego, CA, August 2012.

A.Hoeksema, "Statistical Comparison of Toolmarks in Forensics" Conference on Data Analysis, Santa Fe, NM, February 2012.

TASK II: Analysis of Quasi-striated Marks

II. Methods

For this experiment, 50 pairs of sequentially manufactured slip joint pliers were purchased from Wilde Tool Co., Inc. so as to be as nearly identical as possible. It is well known the manufacturing process greatly affects the resulting toolmarks a tool makes due to the surface features imparted on the tool during manufacturing [24, 25]. For this reason, a detailed description of the way the pliers used in this study were manufactured is in order.

All of the plier-half blanks examined in this study were hot forged from the same die, followed by cold forging from the same forging die. Following forging, holes were punched to seat the fastener, i.e. the bolt that will hold the two halves of the pliers together. At this point a difference is introduced in the blanks. On slip joint pliers, one half of the pliers has a small hole, while the other half has a larger, double hole allowing the user to gain a better grip when using the pliers (see Figure 15). Once the plier holes were punched the teeth and shear cutting surfaces were created using a broaching process. It is this machining method that creates the scratch minutiae on the surface of the plier halves responsible for producing the characteristic toolmark that is of interest in forensic examinations.

Figure 15: Slip joint pliers in their unfinished and finished states. From left to right: plier halves (single and double hole) before broaching; an example flat side of pliers that will be polished; finished and labeled pliers (sides A and B).

The plier halves for this study were cut on two separate broaching machines; halves with the smaller hole were all broached on one machine, while the halves with the double hole were broached on a second. At this point in the process the manufacturer stamped numbers 1-50 on each plier half as they were finished being broached. Thus, the 50 pairs could be assembled with

confidence that they were actually made sequentially. After broaching, both halves were given the same heat treatment and shot peened to surface harden the metal. The long, flat surface was then polished and the pliers were assembled and gripped. As a final step the company branded the double hole side of each pair of pliers. For the purposes of this study each half of the pliers was assigned as either A or B, with Side B being the branded half of the pliers (see Figure 15).

To make the samples, copper wire of 0.1620" diameter and lead wire of 0.1875" diameter were obtained and cut into two-inch lengths with bolt cutters to distinguish the ends from the shear cuts made by the pliers. Next, the cut lengths of wire were placed centered in the plier jaws on the shear cutting surface with pliers side B facing down. Shear cutters are defined by AFTE as "opposed jawed cutters whose cutting blades are offset to pass by each other in the cutting process" [26]. Since the marks were created using the shear surface and both sides perform cutting action, by definition they are shear cutting marks. Alternating shear cuts of lead and copper were made with each pair of pliers for a total of 20 shear cuts. All odd numbered shear cuts were lead samples; all even numbered shear cuts were copper. The total number of copper samples thus obtained was 1000, with 500 shear cuts in contact with Side A, 500 shear cuts with side B.

When the wire is mechanically separated, the two surfaces of the shear edges move past each other. The resultant action is therefore a combination of both cutting the surfaces and a shearing action of the edges as they move through the material. The result is two surfaces being created on each half of the separated wire sample, comprising both shear cut and impression markings, roughly at 90˚ to each other with both being ≈ 45˚ to the long axis of the wire. Only the shear cut surface on the 'A' and 'B' sides of the sample were scanned and analyzed. A schematic showing the process is shown in Figure 16.

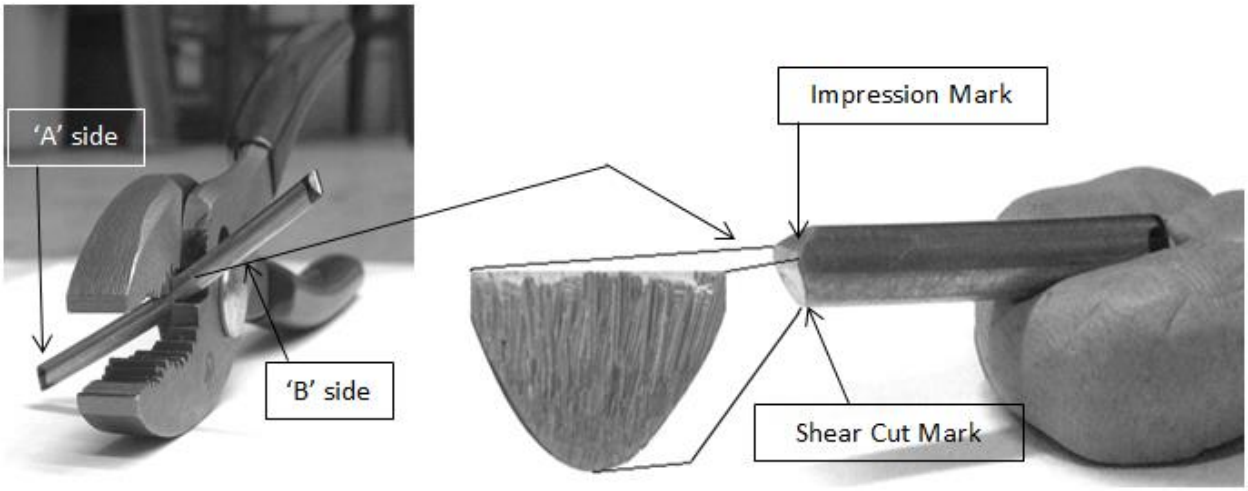


Figure 16: The left photo shows an example wire sample mid-shear cut, revealing how the toolmark gains its angle. The right photo shows the B side sample. Note that the analyzed mark is not completely circular. The A side sample (not shown) is similar in shape.

For the purpose of this study, only the copper samples were evaluated. Each shear cut mark surface was scanned optically with an Alicona G3 Infinite Focus microscope at 10x magnification and a two micron vertical resolution to acquire the surface geometry of the mark. An example of a typical scan is shown in Figure 17. The tool mark is seen to be quasi-striated, i.e. parallel linear striae do exist but it clearly varies across the surface of the shear cut mark.

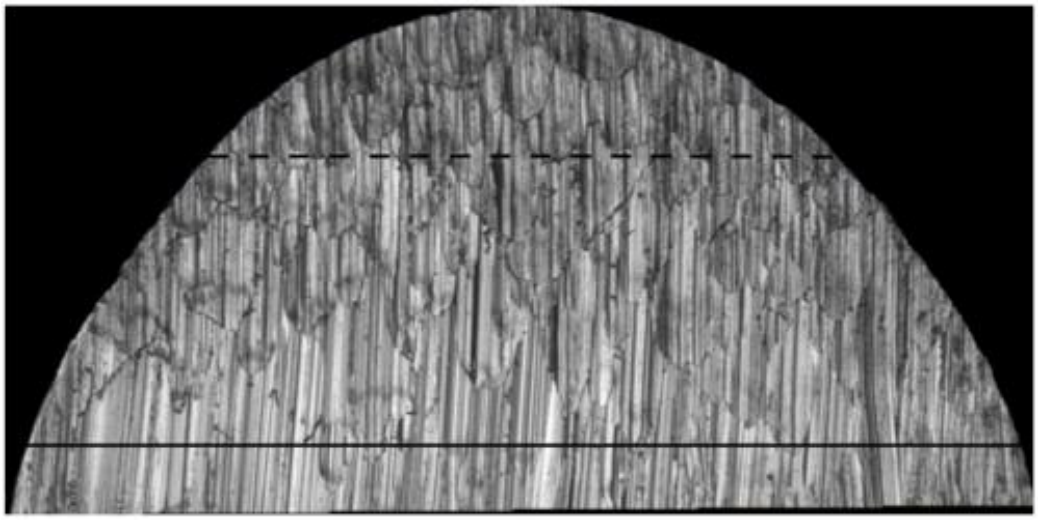


Figure 17: Areas examined during comparisons. Dashed line is referred to as the "short edge," the solid line is referred to as the "long edge." Width of the sample varies slightly from sample to sample but is roughly the diameter of the wire, i.e. 0.1620".

When the data are acquired, noise spikes occur around the edges of the mark where the shear cut surface drops off because there is no surface here for the profilometer to scan. This noise is non-informative for the matching process, and is not desirable in the data file. Therefore, the raw data are processed using a computer routine to remove the extraneous noise spikes. This process is referred to as a cleaning routine and does not affect the data that characterizes the shear cut surface. An example of a clean and uncleaned data file can be seen in Figure 18.

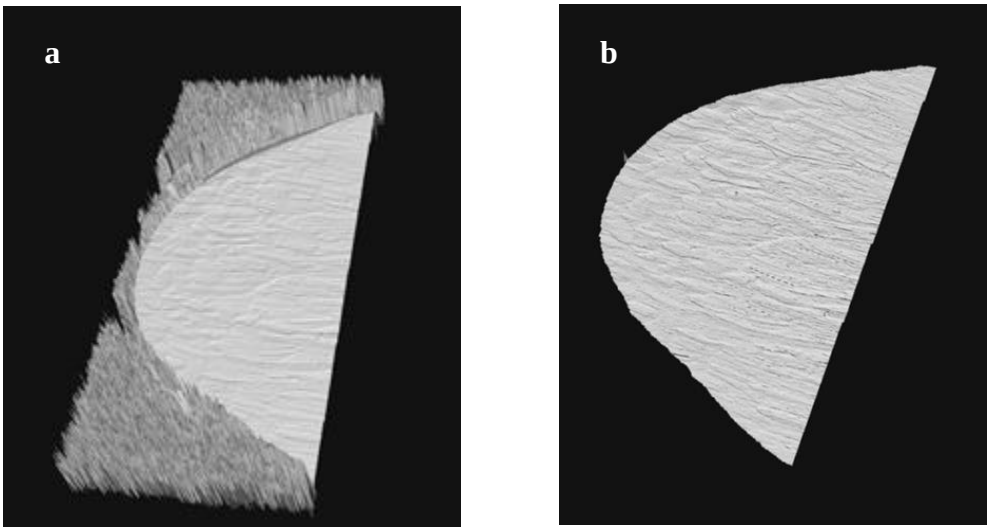


Figure 18: a) Raw data; b) cleaned data with noise spikes removed

All raw data files contained trended data. Simply put, due to the manner in which the data were collected the line profile of a mark data file had an increasing linear trend in the z direction moving from one side of the mark to the other. Such a trend is common when using profilometers since the surface analyzed is rarely exactly parallel with the direction of scanning. Because the files were a rectangular collection of 3D data (shown in the uncleaned data of Figure 18a), trending was corrected by subtracting a plane matching that of the trended data from the file. To accomplish this, the detrending routine selects left and right diagonal points from the data (approximately 40 on each side, 80 in total) and uses a linear least squares method to fit the

appropriate plane for the data. It then subtracts the fitted plane from the data to achieve an appropiately leveled data file for comparison. As a reference, these final data files are roughly 2200 by 4500 pixels, each pixel being 0.8 microns in height and width.

Comparisons between the marks were made using the previously described algorithm [1]. This algorithm uses a Mann-Whitney statistics approach for the comparisons, where the degree of relevance that exists between each comparison is given by the returned value of the statistical analysis known as a T1 value. The comparisons were divided into two different groups, those made close to the end of the mark, as designated by the solid line in Figure 16, and those made close to the start of the mark, shown by the dashed line in Figure 16. From this point on, the dashed line data will be referred to as the short edge and the solid line data as the long edge. These mark locations were chosen to examine differences between the beginning of the shear cut, where the mark has short and variable length striae, and the end of the mark, where the striae are longer and appear to be more regular.

Each side of the pliers was considered to be a separate data set, the assumption being, as confirmed by forensic examiners, each side acts as a different surface. Given there are 50 pairs of pliers, with two sides for each pair of pliers and ten replicate shear cuts for each side of each pair of pliers, the total number of samples possible for examination came to 1,000 discrete data sets.

III. Results

1. Experimental Results

A sampling format was set up to compare three different groups of data: known matches, known non-matches from the same pair of pliers (i.e. different sides), and known non-matches from different pairs of pliers. The comparison setups are as follows:

> *Set 1*: Compare known matches. These should be marks from the same side of pliers. Comparisons were made between marks 2 and 4 and between marks 6 and 8 for each side of the pliers, side A and side B.
>
> *Set 2*: Compare known non-matches from the same pair of pliers. Comparisons were made between side A and side B for marks 10, 12 and 14.
>
> *Set 3*: Compare known non-matches from different pairs of pliers. The samples were divided into 12 groups of four, each numbered consecutively, e.g. tools 1-4, 5-8, etc.

Comparisons were made for both side A and side B. Table 3 shows an example comparison setup for the first group of pliers.

Table 3: Comparisons for Set 3, Group 1

Comparison	Plier number	Side	Mark number	Plier number	Side	Mark number
A	1	A	16	2	A	16
B	3	A	16	4	A	16
C	1	A	18	4	A	18
D	2	A	18	3	A	18
E	1	A	20	3	A	20
F	2	A	20	4	A	20

The same algorithm used in an earlier work for striated marks [1] was applied in this study to examine the quasi-striated marks made by the slip joint pliers. The algorithm has two primary steps: Optimization and Validation. During the Optimization step, the regions of best agreement between the two marks are determined by the maximum correlation statistic, or "R-value." The size of the region is assigned by the user and is hereafter referred to as the "Search Window." The second step of the algorithm, Validation, uses both rigid and random window shifts to verify the regions chosen in the Optimization step indeed correspond to a true match. These windows are hereafter referred to as the "Valid Windows" and their width is also user determined. The R-values in this step must clearly be lower than the R-value in the Optimization step, as the highest R-value has already been calculated. However, in the instance where a true match exists, the R-values associated with the rigid shift valid windows should be larger than those associated with the random shift valid windows, the assumption being, if an excellent match exists at one location then very good matches should exist at any number of corresponding locations. If true, this is indicative a true match does exist. Conversely, rigid window shifts do not produce systematically larger R-values than random shifts in the case of a true non-match, since the high values found during the Optimization step exists due to random chance rather than any physical relationship between the items being compared. Further discussion of this algorithm can be found in the literature [1].

Initial Results

Originally, the size of the search and valid windows were set at the comparison software's default 200 and 100 pixels, respectively, and the comparisons were conducted with samples from the first 20 pairs of pliers. This setup produced 400 different comparisons for the long and short edge comparisons. When a comparison is made, indication of a true match is found when the T1 value of the statistic returned is relatively high. Little or no relationship between the marks results in T1 values centered near 0.

Results of these early comparisons can be found in Figure 19. In these box plots, the bold line in the middle of the box represents the median, the lower quartile by the bottom line of the box, and the upper quartile by the top line of the box. The whiskers are within one and a half times the difference between the upper and lower quartiles. Any outliers outside the whiskers are denoted by dots. In these plots, known matches are in the comparisons designated Set 1, while Sets 2 and 3 show comparisons between known non-matches from different sides of a pair of pliers and non-matches between different pairs of pliers, respectively. It is evident that with these window sizes, the success of identifying known matches was relatively low, there being little separation between the returned T1 values of known matches and non-matches.

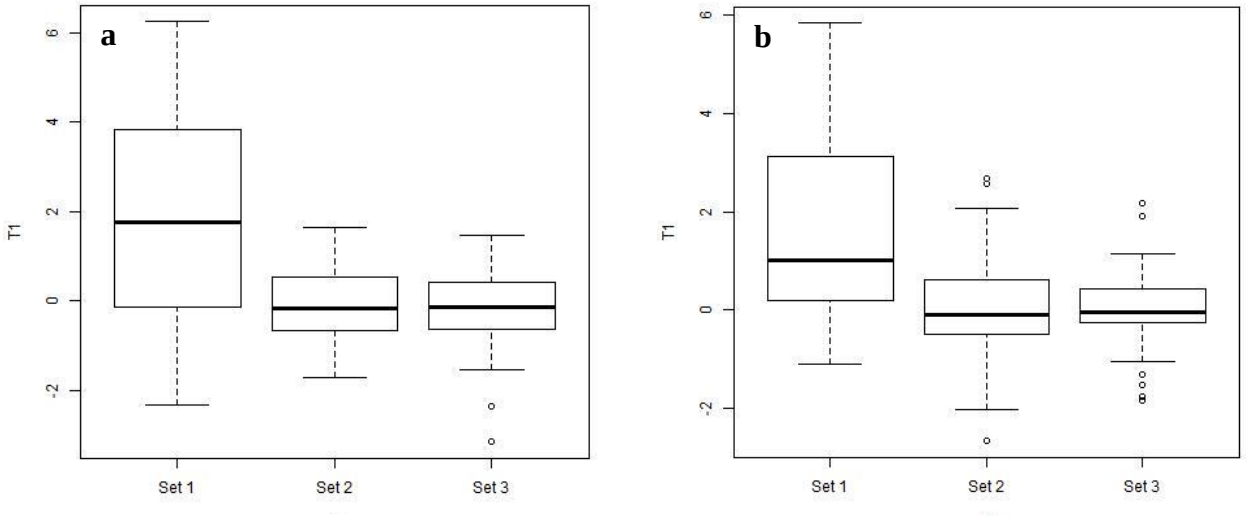


Figure 19: Original data comparisons for (a) short edge, (b) long edge.

Size Effect

From the minimal success of the first attempt at matching the plier marks, several changes were decided upon for further comparisons. Firstly, the original data shown in Figure 19 compared trended data, which was corrected in subsequent comparisons as described above. Secondly, it was decided to vary the window size for all plier mark samples. The initial values used were chosen simply because they had proven effective for comparison of fully striated marks. A series of experiments was conducted within each plier comparison set where the window sizes were varied to evaluate the effect window size has on the resulting T1 value. In other words, the question asked was: does the size of the window play a large role in the discrimination between known matches or known non-matches? In this series of experiments Search and Valid windows were assigned four different values. The Valid window was always half the size of the Search window. Search windows were set at values 100, 200, 500, and 1000 pixels, respectively, to examine the effects of one smaller Search window and two larger Search windows. These new settings were extended to all 50 pairs of pliers and their corresponding toolmarks in the copper wire, bringing the total number of comparisons to 3,952.

The results of these comparisons can be found in Figures 20 and 21. Observation shows that the T1 value increases dramatically with increasing window size. While known non-matches return values centered around zero regardless of window size, the T1 value for known matches increases from just slightly over zero to an average of 6.36 and 6.09 for the largest window size for the long and short comparisons, respectively. However, the data range increases as well. At the larger window sizes, numerous outliers exist and failure of the algorithm occurs in some cases, especially for the short edge comparisons.

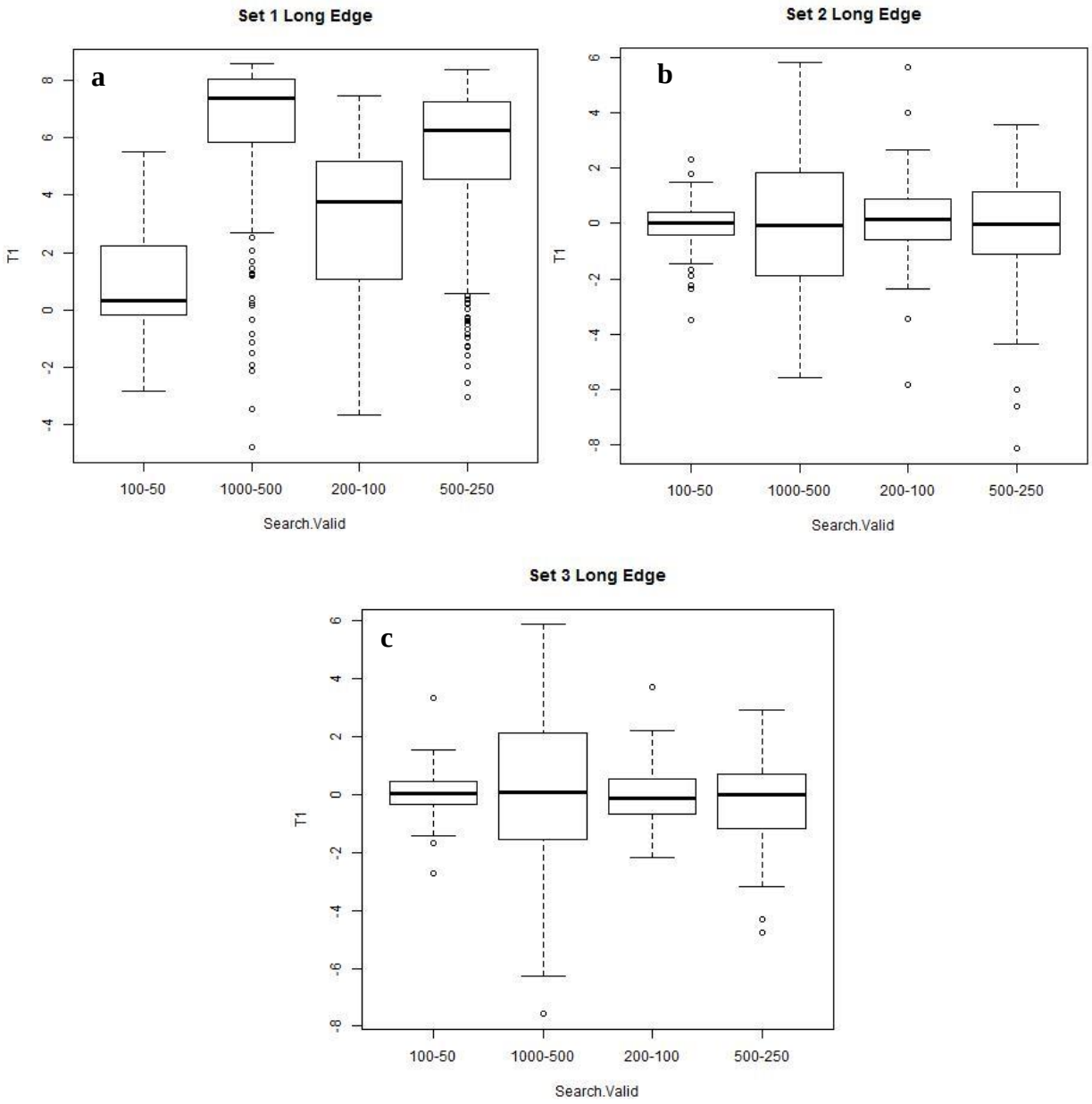


Figure 20: Long edge comparisons. a) Known matches from the same set of pliers. b) Known non-matches from the same set of pliers. c) Known non-matches from different sets of pliers.

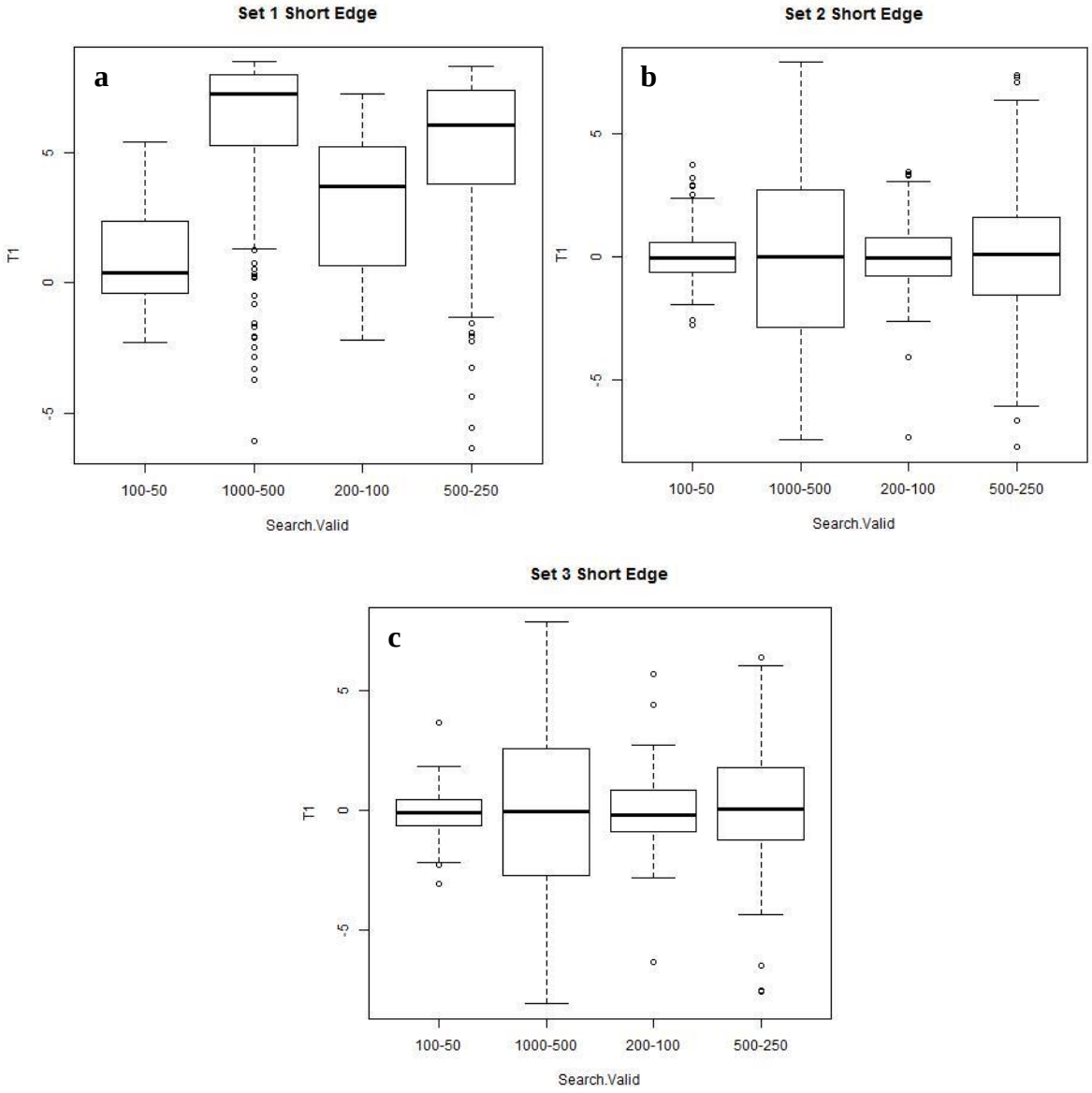


Figure 21: Short edge comparisons. a) Set 1: Known matches from the same set of pliers. b) Set 2: Known non-matches from the same pair of pliers. c) Set 3: Known non-matches from different pairs of pliers.

The large number of observed failures and the increased number of outliers at the larger window sizes for known matches directly results from the constraints placed on the way the Search and Valid windows are chosen and compared. One of the standard conditions under which the algorithm operates is the Search and Valid windows are never allowed to overlap. In some cases, especially with the short edge comparisons, the shorter length of line from which data can be selected and compared results in far fewer data points for comparison. This problem is exacerbated as the window sizes increases. For larger sizes, there simply is not enough data available to meet these conditions in all instances. Thus, this stipulation can cause the algorithm to return no T value. Note that the problem is most apparent for known matches rather than non-matches. This is because the validation process is more critical when an actual correlation exists between the marks being compared than if no correlation exists.

Table 4 summarizes the instances in which the algorithm failed to return values. It can be clearly seen that the return rate decreases with the shorter line profiles as the window size increases. As

a reference, set 1 has a total of 200 comparisons, set 2 has 150 comparisons and set 3 has 144 comparisons.

Table 4: Cases in which the algorithm returned no T values for each window size

Long edge comparisons				
Set	100-50	200-100	500-250	1000-500
1	0	1	1	1
2	0	1	3	3
3	0	2	3	5
Short edge comparisons				
Set	100-50	200-100	500-250	1000-500
1	0	0	1	9
2	1	0	3	19
3	1	0	3	24

Ratio Effect

As a first attempt at a solution, two additional window size ratios (i.e., the ratio of the size of the search window to the size of the validation window) were examined: 4 to 1 and 6 to 1. It was hoped that by limiting the size of the Valid windows less spread in the data would be seen. For each new ratio, four different window sizes were chosen and the algorithm was run again following sets 1, 2 and 3 at both the long and short edge locations on the mark. For these exploratory tests the data were limited to pliers 1-25, the assumption being the abbreviated data set would be representative of the full 1-50 pliers data. Results of this examination can be found in Figures 22 and 23. This set of parameters does indeed appear to have a significant effect in reducing the number of outliers and spread of the known matches (i.e. Set 1) as compared to the 2:1 ratio data. A slight degradation in the maximum values obtained was seen for the known matches. Less change is seen in the results for the known non-matches (Sets 2, 3). Average values still were centered around zero and spread seemed to increase somewhat in some cases for the known non-matches.

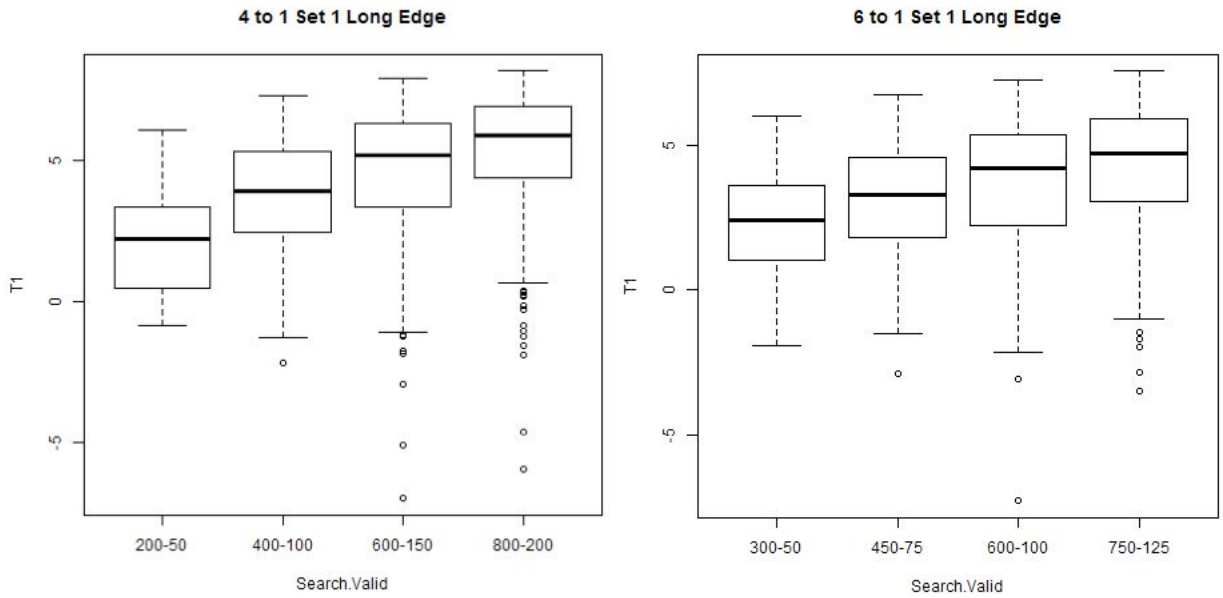


Figure 22a: Effect of changing window size ratio on Set 1, known matches, long edge.

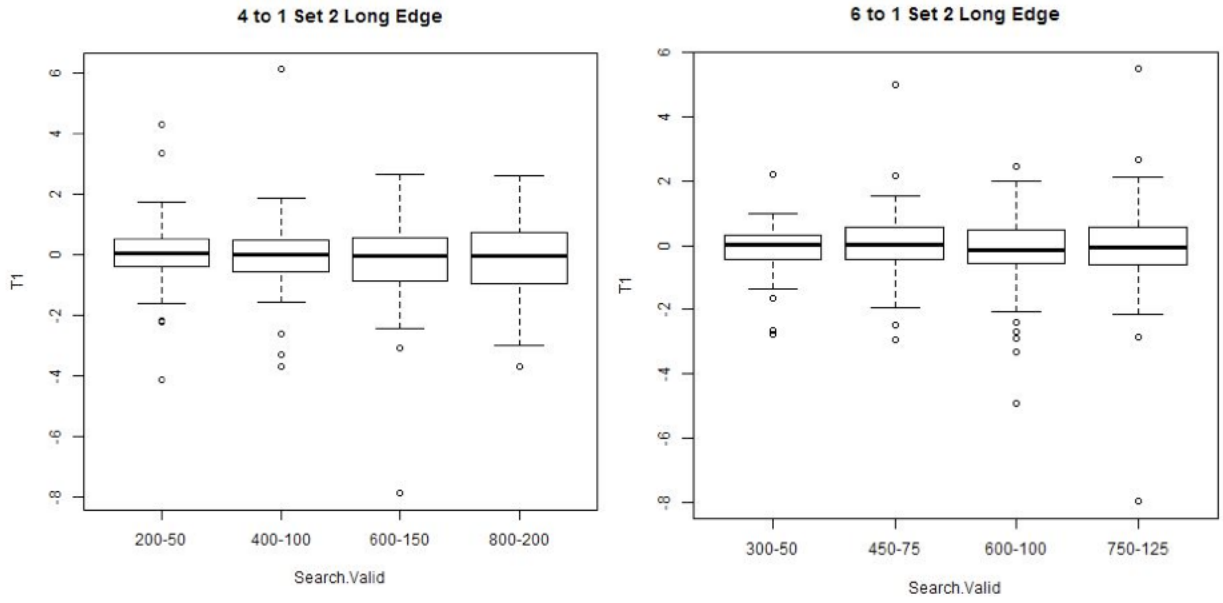


Figure 22b: Effect of changing window size ratio on Set 2, known non-matches, different sides of the same pliers, long edge.

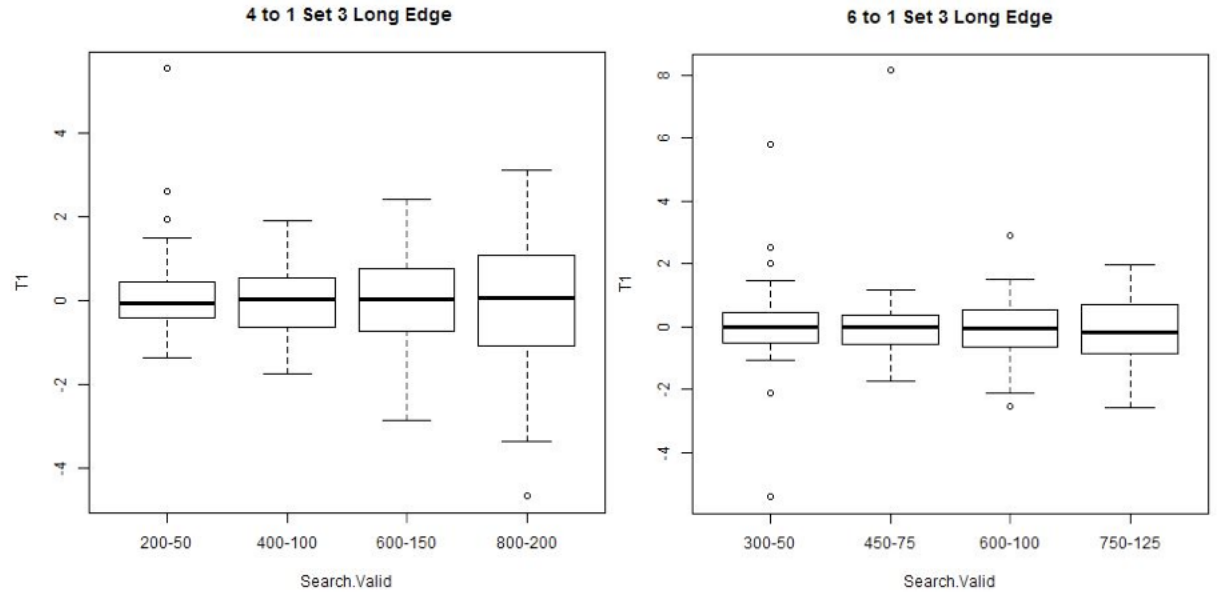


Figure 22c: Effect of changing window size ratio on Set 3, known non-matches, different pliers, long edge.

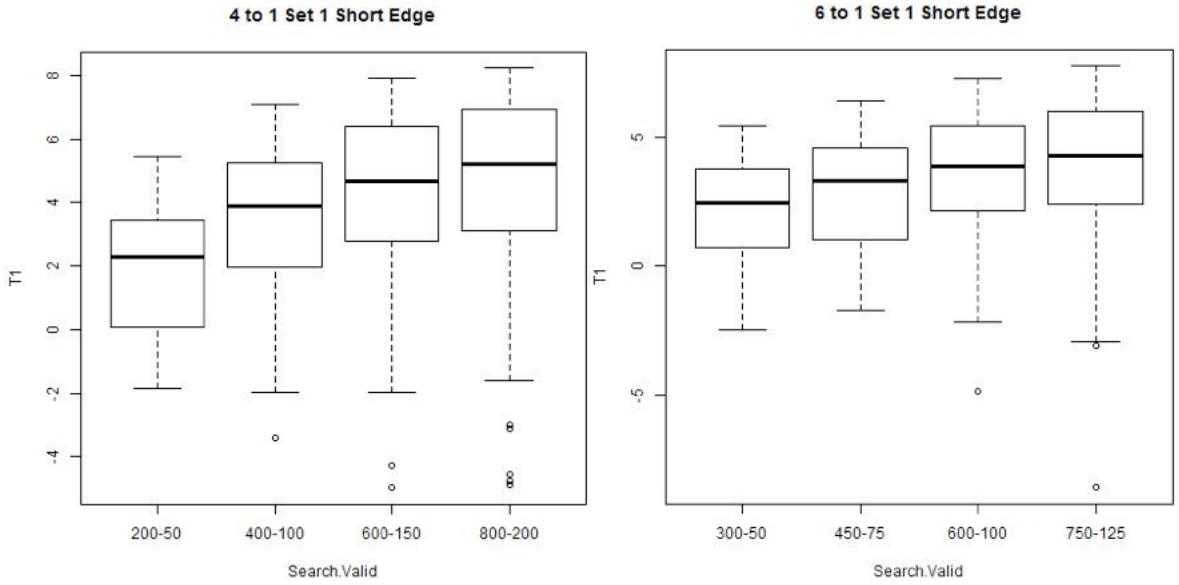


Figure 23a: Effect of changing window size ratio on Set 1, known matches, short edge.

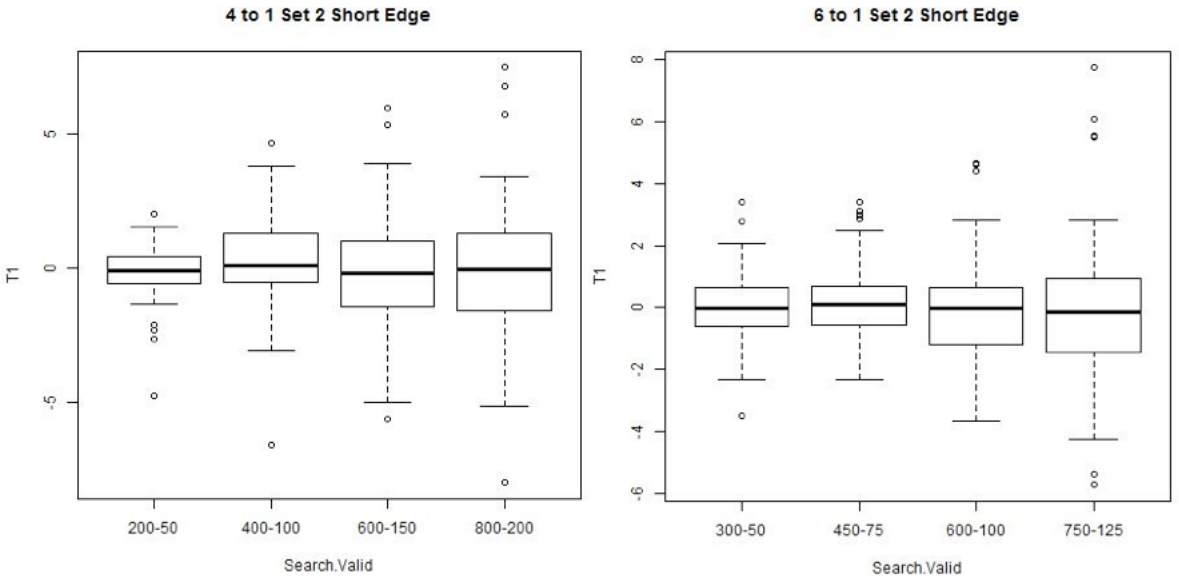


Figure 23b: Effect of changing window size ratio on Set 2, known non-matches, different sides of the same pliers, short edge.

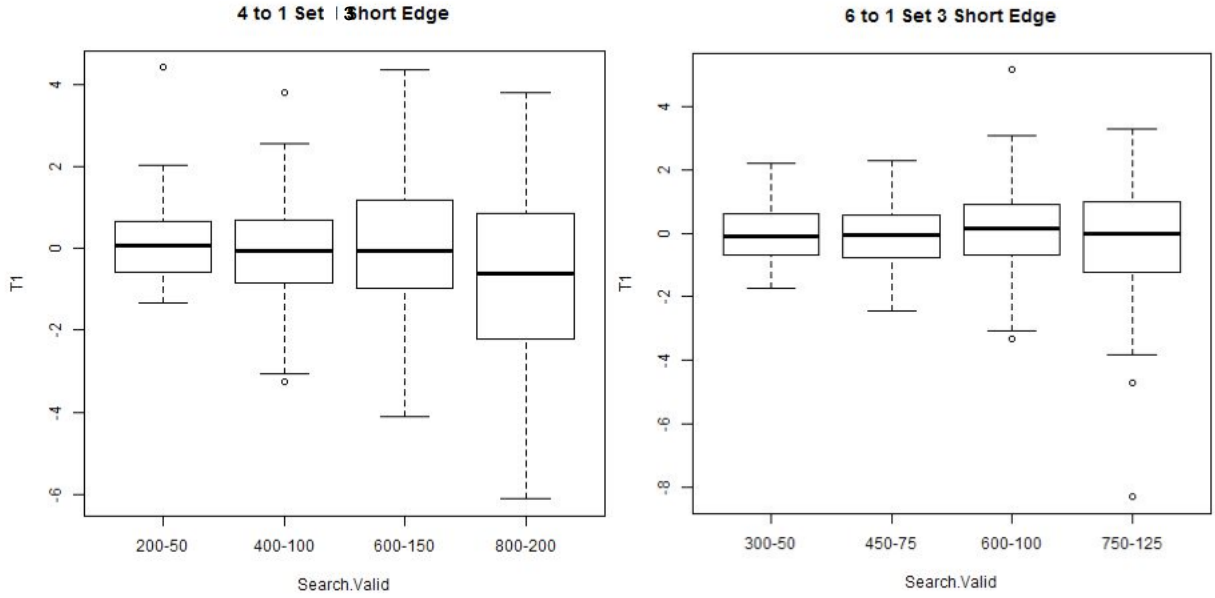


Figure 23c: Effect of changing window size ratio on Set 3, known non-matches, different pliers, short edge.

2. Discussion

When using the developed algorithm, ideally the data should show a clear separation between T1 values for known matches as opposed to known non-matches, with no overlap occurring, even when considering outliers. While elimination of overlap in the outliers has not been achieved it is clear that a high degree of separation is seen in the majority of cases when the search parameters are adjusted from the defaults used for the striated screwdriver marks. This suggests that the current algorithm is more robust than it initially appeared, and could be suitable for discrimination if performance can be enhanced and the spread in the data can be decreased to produce complete separation between known matches and non-matches. These tests also indicate the size of the Search and Validation windows can have a critical role in determining when a match can be discriminated from a non-match. Since the size and number of Valid windows is user defined, future work must involve a series of experiments to determine what operation parameters are best suited for each individual class of marks. For example, the relatively small Search and Valid window sizes that worked well for screwdriver marks were inadequate for the plier marks. However, increasing the Search and Valid window size proved effective in increasing separation between known matches and non-matches for slip joint pliers and changing the size ratio has an effect on the spread of the data.

Outliers are seen in all the data sets, both known match and known non-match. Examination of these data files points to a consistent problem with the current state of the algorithm, which the authors refer to as the "opposite end" match problem. This seems to be an area where further improvements can be made. In earlier work involving screwdriver comparisons [1], it was noted the algorithm often returned false match values, incorrectly identifying the match areas on opposite ends of the mark's cross-sectional profile. "Opposite end" matches appear to occur most often in known non-matches, however non-match values have been returned for known matches as well with similar opposite end match problems. In detrending the data, many of these problems have been eliminated; however a few opposite end match problems still exist. One such example can be seen in Figure 24 for a plier comparison datafile, which consists of detrended data. One data set is shown at the top while the second is shown at the bottom.

Simple chance where the opposite ends of the mark have a very similar profile over the small area of the search window, as denoted by the box, has resulted in the computer declaring an excellent match. Obviously, such a match is physically impossible, no matter how good the numbers. Due to the matches occurring at the ends, the validation routine is thwarted in its efforts to discriminate this incorrect result since few rigid translations are possible. In the example shown, translations to the left of the match are prevented by the bottom scan and translations to the right are restricted to a few pixels right next to the match region by the top scan. Thus, the error goes undetected unless the files are visually examined.

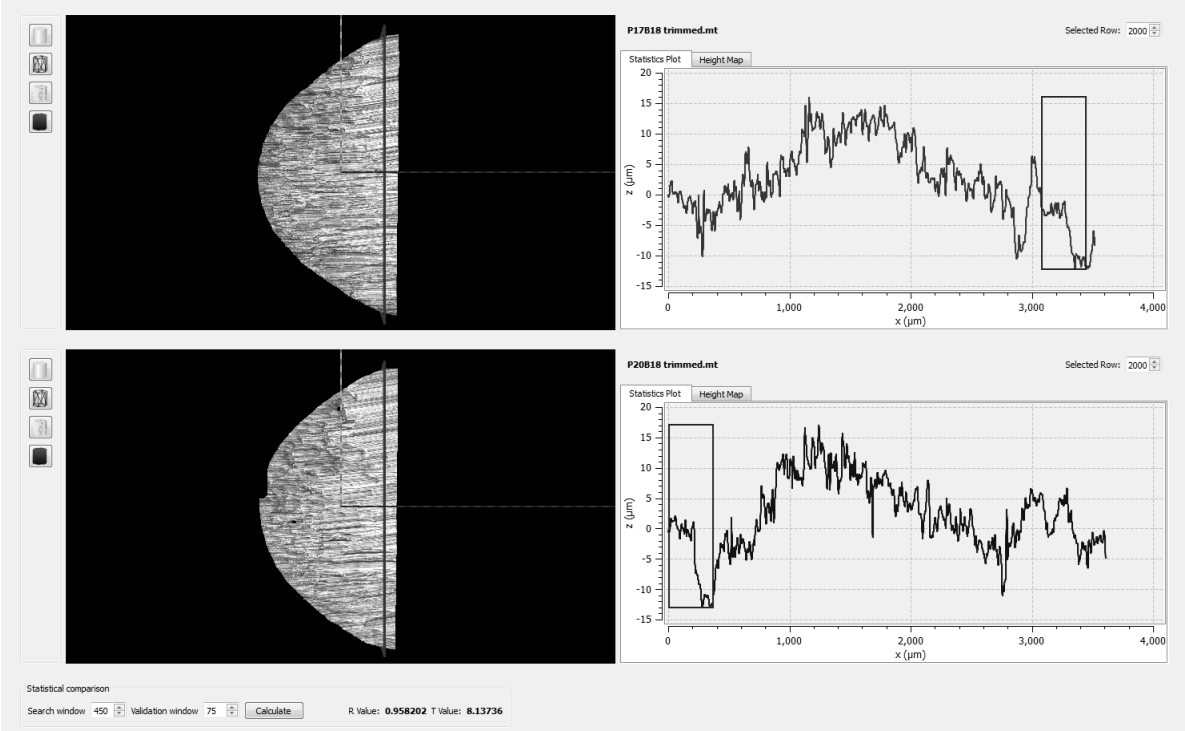


Figure 24: Incorrect opposite ends match for long edge comparison of known non-matches from different pairs of pliers. The search and valid windows were 450 and 75. T1 value is 8.137.

In its current form, the algorithm has maximum flexibility, allowing marks to be compared along a linear direction both forwards and backwards. Such a methodology requires no contextual information to be known about the mark. A fully striated mark may leave few clues as to what is the "left" side of the mark vs. the "right" side, as determined by how one holds the screwdriver, Figure 25. As shown by the bold arrows, pulling the screwdriver across the surface in opposite directions leaves the same mark, but it is rotated 180 degrees. While this situation is usually easily recognized by a trained examiner making a test mark, it is more of a problem for an automated system. To the machine, both situations result in a series of parallel lines. If the scan is constrained to run comparisons in only 1 direction (dotted line), this match may be missed since "left" could be viewed as "right" and vice versa. For this reason currently the algorithm is written to be as flexible as possible with comparisons run in both directions so it is not necessary to know which side of the mark was on the left and which was on the right as it was being made.

Determining the correct scanning direction is less of a problem for a shear cut wire, where contextual information such as "left" and "right" can be easily assigned due to the macroscopic shape of the object itself, Figure 24b. In this instance the situation is somewhat similar to distinguishing between class characteristics in a firearm examination.

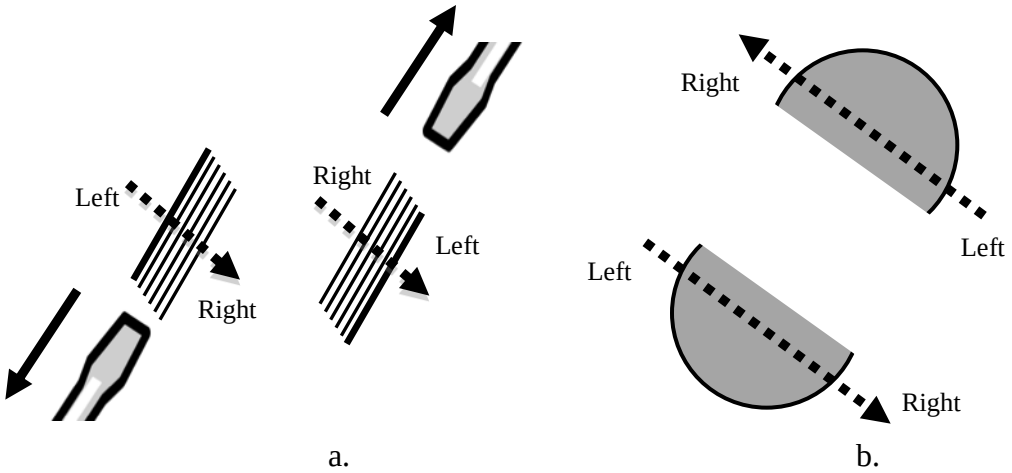


Figure 25: a) Fully striated marks hold few clues to "left" vs. "right" for the automated scan as denoted by the dashed line. b) Shear cut wire sample scan directions are easily distinguishable by the macroscopic shape.

Currently each data file needs to be examined separately in order to determine whether an "opposite end" match has occurred. A screening option is being considered that will automatically determine whether an "opposite end" match has occurred and alert the user to this possibility. The user can then examine only those files so flagged and decide whether an incorrect match has occurred. Clearly, in this instance the examiner will have to use their contextual knowledge of the marks being compared to make this determination.

IV Conclusions

1. Discussion of Findings

A study of 1000 shear cut copper wire samples produced using 50 sequentially manufactured pliers were characterized using an Alicona G3 Infinite Focus Microscope operated as an optical profilometer. An objective analysis of the data was carried out using a computer-based algorithm that had previously been employed to successfully compare striated marks produced by screwdrivers. Initial analyses results produced inconclusive when using the same parameters employed successfully for the screwdriver marks. However, further experiments showed that changing the comparison parameters, specifically the sizes of the search and validation windows, and the ratio of the sizes, the algorithm could produce better statistical separation of known match/non-match comparisons. Reasons for error were identified and future improvements to the algorithm are planned. The major improvement anticipated is a screening option for the identified matched search windows to eliminate the possibility of clearly incorrect "opposite end" matches.

2. Implications for Policy and Practice

Successful application of the algorithm described in [1] to quasi-striated toolmarks has two major implications. Firstly, the work shows once again that there is a scientific basis for objective, quantitative toolmark identification. Secondly, as this represents one of the first successful analyses on a quasi-striated toolmark, it implies that it should be possible to characterize and classify all types of toolmarks, given the right type and quality of data and the appropriate analysis algorithm.

This work, coupled with the PI's previous results on screwdriver toolmarks, imply that an objective, semi-automated or automated system for characterization can be developed. While the technical details to make this a reality are numerous, there seems to be no valid theoretical reason that would prevent this. This is not to say that a system could be developed with the same level of statistical relevance as is seen for DNA samples. This clearly is not the case. However, it is clear that comparative analysis is not solely a subjective matter.

3. Implications for Further Research

The algorithm developed at Ames Lab / Iowa State has been shown to be more robust than originally thought. Having looked at fully striated and quasi-striated shear marks, the next logical step is to examine full impression marks to see if the algorithm can be successful applied, or modified in such as manner as to be successful, to these types of marks. Development of a more user-friendly software package that incorporates all of the lessons learned as regards the analysis of toolmarks is also desirable and should yield a useful product for forensic examiners.

V. References

1. Chumbley, L.S., Morris, M.D., Kreiser, J., Fisher, C., Craft, J. Genalo, L.J., Davis, S., Faden, D., and Kidd, J., "Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm," Journal of Forensic Sciences, Vol. 55, No. 4, Jul. 2010, pp. 953-961.

2. Hamby, J.E., and Thorpe, J.W., "The History of Firearm and Toolmark Identification," AFTE Journal, Vol. 31, No. 3, 1999, pp. 266-283.

3. Meyers, C.R., "Firearms and Toolmark Identification: An Introduction," AFTE Journal, Vol. 25, No. 4, Oct. 1993, pp. 281-285.

4. Burd, D.Q., and Kirk, P.L., "Toolmarks: Factors Involved in Their Comparison and Use as Evidence," Journal of Criminal Law and Criminology, Vol. 32, No. 6, 1942, pp. 679-686.

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

6. Bisotti, A., "A Statistical Study of the Individual Characteristics of Fired Bullets," Journal of Forensic Science 4(1), 34-50, 1959.

7. Bonfanti, M.S., and DeKinder, J., "The Influence of the Use of Firearms on their Characteristic Marks," Association of Firearm and Tool mark Examiners, 31(3), 318-323 1999.

8. Bonafanti, M.S., and De Kinder, J., "The Influence of Manufacturing Processes on the Identification of Bullets and cartridge cases – A Review of the Literature," Science and Justice 39, 3-10, 1999.

9. Bisotti, A., and Murdock, J., "Criteria for Identification or State of the Art of Firearm and Tool mark Identification," Association of Firearm and Tool mark Examiners, 4, 16-24, 1984.

10. Faden, D., Kidd, J., Craft, J., Chumbley, L.S., Morris, M., Genalo, L., Kreiser, J., and Davis, S., "Statistical Confirmation of Empirical Observations Concerning Toolmark Striae," AFTE Journal, Vol. 39, No. 3, 2007, pp. 205-214.

11. Baldwin, D., Morris, M., Bajic, S., Zhou, Z., and Kreiser, M.J., "Statistical Tools for Forensic Analysis of Tool marks Ames (IA)," Ames Laboratory Technical Report IS-5160, 2004.

12. Tuira, Y.J., "Tire Stabbing With Consecutively Manufactured Knives", AFTE Journal, Vol. 14, No. 1, January, 1982.

13. Cilwa, R.B., Townshend, D.G., "Tool Mark Identification: Knife to Cut Wire", AFTE Journal, Vol. 8, No. 4, Dec., 1976, pps. 66-67.

14. Hamby, J.E., "Matching of Tool Marks Made in Rubber," U.S. Army Crime Lab, AFTE Newsletter No. 20, June, 10, 1972, pp. 29.

15. Warren, G., "Shoeprint - Wax Casting Material Information," AFTE Journal, Vol. 15, No. 2, April, 1983, pps. 77-78.

16. Butcher, S., and Pugh, D., "A study of marks made by bolt cutters," Journal of the Forensic Science Society, Vol. 15, No. 2, 1975, pp. 115-126.

17. Hornsby, B., "MCC bolt cutters," AFTE Journal, Vol. 21, No. 3, 1989, pp. 508.

18. Hall, J., "Consecutive cuts by bolt cutters and their effect on identification," AFTE Journal, Vol. 24, No. 3, 1992, pp. 260-272.

19. Reitz, J., "An unusual toolmark identification case," AFTE Journal, 1975, Vol. 7, No. 3, 1975, pp.40-43.

20. Warren, G., "Glass cutter impression identification," AFTE Journal, Vol. 23, No. 4, 1991, pp.925-927.

21. Report by a committee for the National Research Council for the National Academy of Sciences, "Strengthening Forensic Science: A Path Forward", National Academies Press, March, 2009.

22. Bachrach, B., Jain, A., Jung, S., and Koons, R.D, "A Statistical Validation of the Individuality and Repeatablility of Striated Tool Marks: Screwdrivers and Tongue and Groove Pliers," Journal of Forensic Sciences, Vol. 55, No. 2, Mar. 2010, pp. 348-357.

23. Petraco, N. et al., "Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons," Document 239048, NCRJS, July 2012.

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

25. Monturo, C., "The Effect of the Machining Process as it Relates to Toolmarks on Surfaces," AFTE Journal, Vol. 42, No. 3, 2010, pp. 264-266.

26. AFTE. "Shear Cutters." Glossary of the Association of Firearm and Toolmark Examiners. 5th ed. Montreal: Forensic Technology, Incorporated, 2007. 182. CD.

VI. Dissemination of Research Findings

As stated at the beginning of this report, this project was conducted as part of the thesis work of Ms. Taylor Grieve, a graduate student at Iowa State University, and much of the text comes from what will be included in her Ph.D. thesis. The entire thesis will be available from Iowa State University. In addition, the following paper has been submitted:

Taylor Grieve, L.S. Chumbley, J. Kreiser, Max Morris, and Laura Ekstrand, "Objective Comparison of Marks from Slip-Joint Pliers," Association of Firearms and Toolmark Examiners (AFTE).

Dissemination has also occurred by an oral presentation given by Ms. Grieve:

T. Grieve, L.S. Chumbley, J. Kreiser, M. Morris, S. Zhang, L. Ekstrand, " Comparison of Striated Marks From Slip Joint Pliers," AAFS, Washington, February, 2013.

A second paper is currently being written for submission to the Journal of Forensic Science. This paper will contain a further analysis of the data presented by Ms. Grieve in the AFTE journal. It is planned that this paper will be ready for submission by Summer, 2013.