Surface Topography Analysis for a Feasibility Assessment of a National Ballistics Imaging Database

T.V. Vorburger
J.H. Yen
B. Bachrach
T.B. Renegar
J.J. Filliben
L. Ma
H.G. Rhee
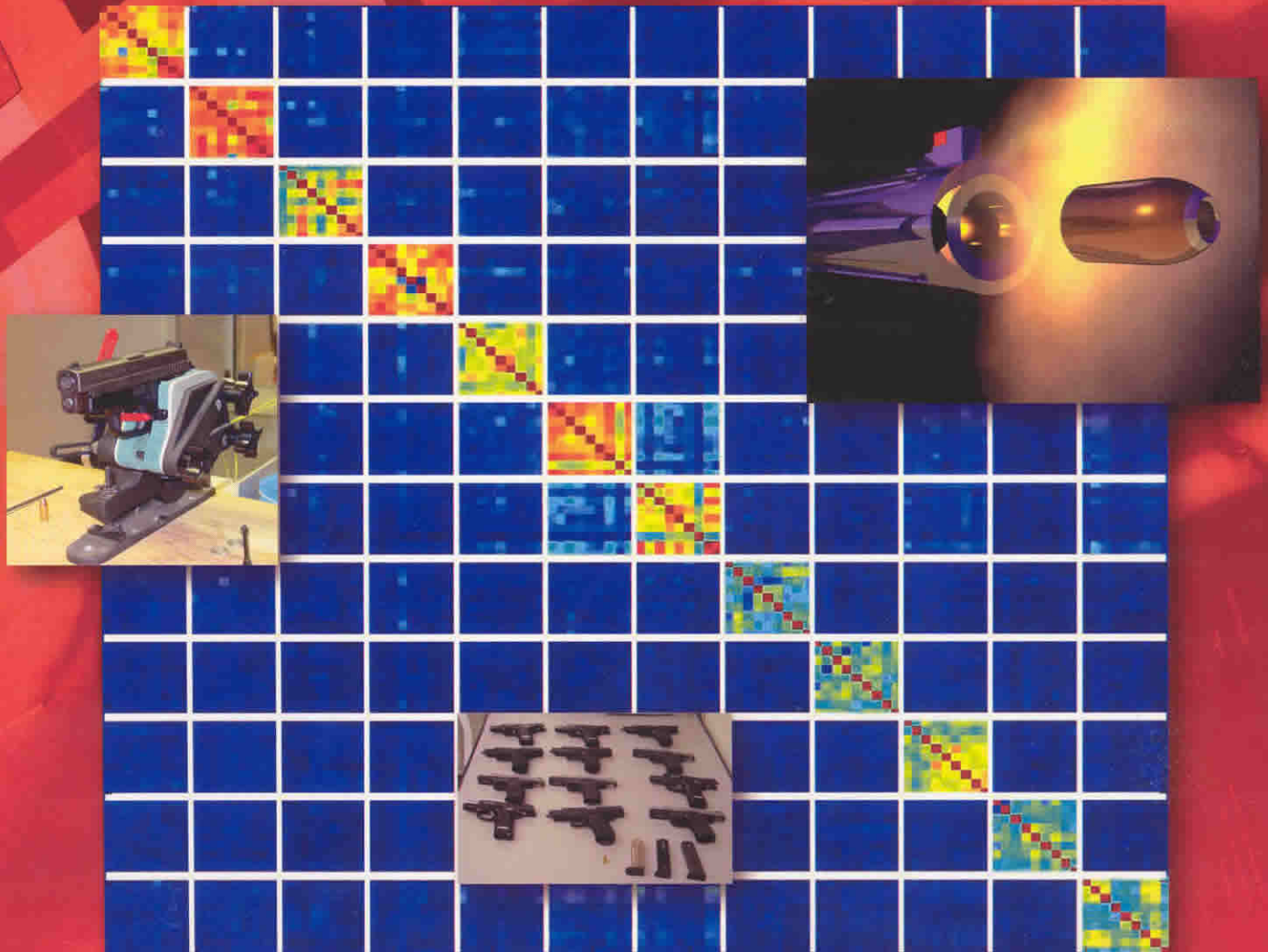A. Zheng
J.F. Song
M. Riley
C.D. Foreman
S.M. Ballou

NISTIR 7362

Surface Topography Analysis for a Feasibility Assessment of a National Ballistics Imaging Database

T.V. Vorburger, J.H. Yen, B. Bachrach[1]
T.B. Renegar, J.J. Filliben, L. Ma
H.-G. Rhee, A. Zheng, J.-F. Song
M. Riley, C.D. Foreman, S.M. Ballou

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

[1]Intelligent Automation, Inc.
15400 Calhoun Drive
Rockville, MD 20855

U.S. Department of Commerce
Carlos M. Gutierrez, Secretary

Technology Administration
Robert Cresanti, Under Secretary of Commerce for Technology

National Institute of Standards and Technology
William Jeffrey, Director

Table of Contents

Glossary of Abbreviations

ACCF	Areal Cross Correlation Function, a statistical function of three dimensional surface topography
ANOVA	Analysis of Variance
ATF	Bureau of Alcohol, Tobacco, Firearms, and Explosives
BF	Breech face
CCF	Cross Correlation Function, a statistical function of two dimensional surface topography
DAS	Data Acquisition Station, a component of IBIS (below)
EEEL	Electronics and Electrical Engineering Laboratory, an organizational unit of NIST
FP	Firing pin
FBI	Federal Bureau of Investigation
FTI	Forensic Technology, Inc., Montreal, Canada
IAI	Intelligent Automation, Inc., Rockville, MD
IBIS	Integrated Ballistics Identification System
ITL	Information Technology Laboratory, an organizational unit of NIST
I-2D	IBIS 2D system using BrassCatcher software Version 3.4.5.
I2	I-2D in some charts
MEL	Manufacturing Engineering Laboratory, an organizational unit of NIST
NA	The National Academies, Washington, DC
NBID	National Ballistics Imaging Database
NBIDE	NIST Ballistics Imaging Database Evaluation
NIBIN	National Integrated Ballistics Identification Network
NIJ	National Institute of Justice
NIST	National Institute of Standards and Technology, Gaithersburg, MD
NLC	National Laboratory Center, ATF, Ammendale, MD
NRC	National Research Council, Washington, DC
NTC	The ATF's National Tracing Center,
N-3D	Experimental topography measurement and correlation system developed at NIST
N3	N-3D in some charts
OLES	Office of Law Enforcement Standards, a division-level unit of NIST
Ra	Roughness Average, a statistical parameter of a surface profile
RBID	Reference Ballistics Imaging Database
Rq, rms	Root mean square roughness, a statistical parameter of a surface profile
SAS	Signature Analysis Station, a component of IBIS
Sa, Sq	Parameters analogous to *Ra* and *Rq* for areal (3D) surface topography

Executive Summary

Introduction

This document reports on a study to determine the feasibility and utility of a national ballistics database of casing and bullet images. The purpose of such a proposed database would be to provide a reference collection of ballistic images against which casings or bullets found at the scene of a crime may be compared, with the intent of uniquely identifying the weapon that generated the spent casing or bullet. In the same fashion that the national fingerprint database serves as a forensics tool for person identification for law enforcement officers, this ballistics database would serve as an important tool for weapon identification. To construct the database, the proposed plan would be to conduct test firings for every weapon sold over the counter, with the hope that if any such weapon is subsequently used in a crime, it may be identified and traced back to the original owner.

The study was conducted in support of a National Academies (NA) Committee to Assess the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database. The study was sponsored by the Department of Justice's National Institute of Justice (NIJ), through the Office of Law Enforcement Standards (OLES) under the Electronics and Electrical Engineering Laboratory (EEEL) of the National Institute of Standards and Technology (NIST). The study was conducted by members of OLES, NIST's Manufacturing Engineering Laboratory (MEL), NIST's Information Technology Laboratory (ITL), and the private company, Intelligent Automation, Inc. (IAI). The purpose of this report is to inform the NA Committee about the results and observations of the NIST study. The NA Committee issues its own report of its investigations and findings, which may include results from this report.

Project Questions

The critical question for this feasibility study is

1. Can we distinguish and identify guns based on casings information sufficiently well to support a national ballistics identification system? That is, are the markings (firing pin, breech face, ejector markings), which an individual gun leaves on its fired casings, unique enough to distinguish it from other guns of the same type?

    Note: Due to time limitations, most of the work so far has been limited to casings. The issue of bullets will be addressed in another report.

Other important questions are

2. Is our ability to distinguish guns and gun types affected by different ammunition types?

3. Are guns from some manufacturers easier to discriminate than guns from other manufacturers?

4. What is the relative usefulness of each casing region (firing pin impression, breech face impression, ejector marks) for gun discrimination?

5. How do three-dimensional (3D) surface topography imaging methods compare with the optical imaging technologies currently in wide use?

Our project goal is to carry out a rigorous study to provide unambiguous answers to the above questions and related questions of interest.

Casing Collections

To date, this study investigated two collections of fired casings, identified here as the De Kinder collection and the NBIDE (NIST Ballistics Imaging Database Evaluation) collection. The De Kinder collection was created several years ago by De Kinder et al. [7] to test the performance of the widely used Integrated Ballistics Identification System[*] (IBIS) [3] in a large database. The NBIDE collection was created in May 2005 by NIST personnel as a part of the current feasibility study. These collections are described in more detail below.

De Kinder
The De Kinder experiment used 600 autoloading pistols, all the same Model P226 Sig Sauer, to make 4200 test fires, seven for each gun. We used the test fired casings from ten of these guns, providing us with 70 casings for analysis as part of this feasibility study. The 70 casings were also among those originally studied by De Kinder. As part of the present study, these 70 De Kinder casings were imaged by both reflectance based (2D) and surface topography (3D) methods and analyzed via a cross-correlation method.

The advantage of the De Kinder database is that it allows us to observe the variability in the surface markings caused by different ammunition and to assess the implications for gun distinguishability. The data set also allows for the assessment of how distinguishable nominally identical guns can be. One limitation of the De Kinder database is that all conclusions are limited in scope to the single gun manufacturer and model employed, the Sig Sauer P226. Another limitation is that almost all the cartridges are of different manufacture (brand), making it difficult to quantify variability of markings within the same brand as opposed to different brands.

NBIDE
The NIST study, also referred to as NBIDE, attempted to answer questions left unanswered in the De Kinder study. A central component in the NIST study was to determine the effect of gun type (manufacturer) on gun identifiability. As noted above, the De Kinder study was limited to a single gun type (Sig Sauer P226). To go beyond this, the statistical design included three gun types (Sig Sauer, Ruger, and Smith&Wesson), four guns of each type, three ammo types, and three firing repetitions for each ammo type taking place over three days. This design helped us to ascertain the existence and magnitude of gun type and ammo type on gun distinguishability. However, the NBIDE study had several limitations. First, the conclusions are, strictly speaking, limited in scope to the three gun types and the three ammo types studied. Second, the total

---

[*] Certain commercial equipment are identified in this paper. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the equipment are necessarily the best available for the purpose.

number of firings (144) and analyzed casings (108) are small compared to the large sizes envisioned for a national ballistics database. Third, all the guns used for NBIDE are brand new.

Image Analysis Methods

We compared two image analysis methods used to collect reflectance (2D) and topography images (3D), respectively:

- An IBIS 2D system [3], based on reflection optical microscopy and running the BrassCatcher software application, Version 3.4.5. This system has been widely used for forensic ballistic examinations. The software has the capability for acquiring images of three marks on spent cartridge cases, namely firing pin impressions, breech face impressions, and ejector marks. We will refer to this system as I-2D.
- An experimental system, not intended for commercialization, based on imaging of 3D surface topography using confocal microscopy and having correlation software developed by the NIST/IAI analysis team. We will refer to this system as N-3D.

These two systems were applied to both of the above casing collections and compared.

The purpose of the NBIDE Study was not restricted to ascertaining the feasibility of a national ballistics database based on the existing I-2D technology used in the NIBIN. Constraining the scope of this feasibility study to existing technology would have limited our efforts to testing whether some previous disappointing results [6-8] from research with I-2D would be supported or refuted with another set of test firings. The National Academies' charge to NIST was broader than that. The broader charge was to assess the feasibility of a high-success ballistics national database by any means—not just I-2D. NIST was thus free to explore other imaging/analysis systems, and decided to emphasize the topography imaging approach as the most feasible alternative or enhancement to existing technology.

The experimental N-3D system consists of five components:

1. An optical topography imaging system that produces a computer-resident digitized 3D surface image,
2. A filtering/outlier-rejection system for pre-processing the 3D image data,
3. A registration process by which the optimal alignment of the individual characteristics of two compared casing images is determined,
4. A statistical calculation to provide a best-alignment correlation metric, which we call $ACCF_{max}$, quantifying how well the two images match,
5. A statistical data analysis methodology that provides a ranked list of best-matched guns as well as other information, such as gun-type and ammunition effects.

Statistical Analysis

Top Ten Analysis
One way to evaluate and compare the I-2D and N-3D technologies is by examining their performances on Top Ten list exercises. This is a list of the ten other entries in the database that

have the highest correlation scores with respect to a reference entry. We first summarize the results for the De Kinder set. For each of the 70 De Kinder casings, there are six correct matches. For a technology to be feasible for a very large database, the Top Ten lists should include all six correct matches nearly all of the time in this database of 70 entries. After all, if a casing that should match does not make the Top Ten list here, then that implies that there are at least five non-matching casings that had higher correlations. Such a low-correlating match is unlikely to make a Top Ten list for a very large database.

Out of a maximum six matches for each casing, the I-2D system produced an average of 3.1 matches using firing pin impressions and an average of 1.0 matches using breech face impressions. The experimental N-3D technology using cross-correlations was somewhat better on firing pin impressions, producing an average of 3.3 matches per Top Ten list, and substantially better on breech face impressions, with an average of 2.8 matches per Top Ten list. Further refinements may well lead to even better results. Using both the breech face and firing pin Top Ten lists for each casing leads to an average of 3.4 correct matches for I-2D and 4.8 correct matches for N-3D.

The De Kinder results by themselves would imply that firing pin impressions produce somewhat better results than breech face impressions, but that neither area of the casing and neither technology yet produces results that are accurate enough for a large database. However, the results from the NBIDE casings make any conclusions drawn from both sets more complicated. Each of the 108 NBIDE casings has eight casings among the other 107 that should match with it; hence, if a casing that should match does not make the Top Ten list, then there are at least three non-matching cases that correlated more highly with the reference casing. The I-2D technology produced an average of 3.7 out of eight correct matches for the firing pin impressions and 5.6 for the breech face impressions. The N-3D technology produced an average of 5.6 correct matches out of eight for the firing pin impressions and 7.9 for the breech face impressions. Using both the breech face and firing pin Top Ten lists for each casing leads to an average of 6.2 correct matches for I-2D and 7.99 correct matches for N-3D.

The NBIDE breech face impressions analyzed with 3D methods stand out from the other datasets here. Not only did most of the NBIDE breech face impressions have all eight correct matches in the Top Ten, most had the correct matches in the top eight ranks. However, while there was considerable separation between matching and non-matching distributions for most of the reference NBIDE breech face impressions, especially those fired from Rugers, some had much less margin for error in that the correlation metrics for matches were only slightly larger than the largest of the correlation metrics for non-matches. Those matching correlation metrics would be in danger of being overtaken by some non-matching correlation metrics in a very large database.

The results of the Top Ten list analysis are summarized in Table 1. The N-3D results for the NBIDE breech face impressions are much better than any other results seen. They suggest that topographic imaging of breech face impressions is good enough to distinguish most of the guns in the study. However, any resulting claims for topographic imaging of breech face impressions have to be reconciled with the much less impressive performance of the same technology on the De Kinder casings. Gun-brand differences might have been a main cause of the differences, as the De Kinder casings all were fired from Sig Sauers; however, the NBIDE study also included

casings fired from Sig Sauers, and the subset of topographic results from those Sig Sauer-fired casings are still much better than the topographic De Kinder breech face results. Alternatively, these differences may arise from differences between new and used Sig Sauer guns or from differences in the ammunition used in the two studies. Just how promising this technology is for very large databases depends on whether the NBIDE or De Kinder results are more representative of the challenges faced by a national database.

In the NBIDE study, only 108 casings fired by only 12 guns covering 3 different brands were studied. Presumably, a larger population of guns would increase the likelihood of "rogue" large correlations from non-matching guns. Also, gun brands and models not covered in the NBIDE study (e.g., low-cost guns high on the list of crime gun usage developed by the Bureau of Alcohol, Tobacco, and Firearms, and Explosives (ATF) [37] but not available as new purchases) may well be more difficult to distinguish and identify than those included. There will likely be guns for which any system will not perform well. We have considered only a minute fraction of the possible gun/ammo combinations, and it is likely that some of these combinations will not be conducive to an easy identification, even by a human examiner.

Table 1. Summary of Top Ten list performance results for I-2D vs. N-3D for De Kinder and NBIDE data sets.

Data Set		Feature under Comparison					
		Firing Pin		Breech Face		Combined	
De Kinder	I-2D	3.1/6	51 %	1.0/6	17 %	3.4/6	57 %
	N-3D	3.3/6	54 %	2.8/6	47 %	4.8/6	80 %
NBIDE	I-2D	3.7/8	46 %	5.6/8	70 %	6.2/8	78 %
	N-3D	5.6/8	70 %	7.9/8	99 %	7.99/8	99.9 %

Direct comparison between the I-2D results and the N-3D results shows that the N-3D research system was more accurate for both breech face and firing pin impressions for both the NBIDE and De Kinder casing collections studied here. The results also show that both I-2D and N-3D perform much better on the NBIDE breech face impressions than on the De Kinder breech face impressions. This further suggests that the differences between NBIDE and De Kinder results are due to physical differences between the two sets of casings. Of course, any firm conclusions require further analysis and data.

Database Feasibility Issues

A typical scenario in which ballistics imaging technology would be used is one where a casing found at a crime scene is then correlated with all the casings in a database that have the same class characteristics. The casings in the database that are chosen for closer scrutiny by a ballistics examiner are those that correlate most highly with the crime scene casing, which we will call the reference casing.

Suppose that there is actually a casing from the same gun in the database, so that it should be a match for the reference casing. Let there be N other casings in the database, where N is a suitably large number. For the real match to make a Top Ten list like those produced by the I-2D

system, only nine or fewer of the N cross correlations with non-matching casings may be greater than the correlation with the real match.

For a first pass model, given several simplifying assumptions, the number of casings in the database that yield a higher cross correlation with the reference casing than does the real match can be modeled as a binomial distribution, Binomial (N,p), where p is the relevant overlap metric (see Sec. 9.5). In this model, the average number of non-matching correlations higher than the true matching correlation increases linearly with N.

This crude probability model enables some observations to be made on how good the correlation methods have to be in order to be successful. For instance, for a given value of N, how small does p have to be in order to have the correct casing in the top 10 at least 90 % of the time? If N is very large, then p has to be very small. In fact, p needs to be around $6.2/N$ to get the right match in the Top Ten 90 % of the time. From this we can make statements of the sort, "If your database is *that* big, then your imaging and correlation techniques better be *that* good to have a reasonable probability of finding a match in it." For instance, if the database has 100 000 guns, then p needs to be on the order of 6.2×10^{-5} to have a 90 % chance of getting the correct match in the Top Ten.

Note that all of the above have been applied to the chances for a single casing. Producing a model that describes the performance of a group of casings or a group of guns is more complicated. Grouping materials with the same class characteristics is a fundamental method used to distinguish groups. Beyond that, for materials with the same class characteristics, there are several levels to which the model can be refined.

Single p – One Group
Suppose we can use the same matching and non-matching distribution for all casings and guns, then a single probability model can be used without modification to refer to all casings.

Grouped by Guns
Results showed there is variability between guns. When we have multiple firings from each gun, we can form separate matching and non-matching distributions for each gun, resulting in a different p for each gun. Thus if a certain gun has an overlap metric p, its casings would tend to make the Top Ten list with Probability $P(N,p)$. For a set of guns, each with its own p and $P(N,p)$, then the success rate of the group of guns is the average of the gun success rates, i.e. the mean of the $P(N,p)$ values. For a given target success rate, say 90 %, it is also useful to see what proportion of the guns have a success rate as high as that target rate.

Grouped by Casings
There can also be variability between casings fired from the same gun. When we have multiple firings from each gun, we can form separate matching and non-matching distributions for each casing, resulting in a different p for each casing. For a set of casings, each with its own p and $P(N,p)$, the success rate of the group of casings is the average of the casing success rates, i.e. the mean of the $P(N, p)$ values. Again, for a given target success rate, say 90 %, it is also useful to see what proportion of the casings have a success rate as high as that target rate.

Note the requirement for multiple firings for guns. If each gun fired $m + 1$ casings, then each casing has only m correct matches. It may be difficult to get good estimates of p using pair-wise comparisons because of the relatively small number of comparisons that can be made. This can be especially problematic because we are most interested in very small values of p, and the pair-wise comparisons can yield estimates of p only as multiples of $1/(mn)$, where n is the number of non-matching correlations per casing. This may lead to too many estimates of 0 for the value of p.

One solution to this problem is to fit continuous distributions to the matching and non-matching samples. These distributions can yield estimates of p that are non-zero but very small. Of course there would remain the issues of whether the fitted distribution is an appropriate fit and how good the fit is, given the limited sample size. In this report, we fit normal distributions using the estimated means and variances of each sample. When compared with the non-symmetrical histogram distributions shown in Section 9, the fitted normal distributions may yield only approximate estimations.

We have considered the above types of groupings. Other types are also possible, such as groupings by casing brand or gun brand or a combination of those.

Discussion of Groups
Suppose that there are differences between groups. Then one can draw different conclusions depending on the level of grouping. In general, having more numerous and more refined groups will lead to more optimistic conclusions, while having fewer groups that are more pooled will lead to more pessimistic conclusions. That is because success in a very large database demands a very small p. Thus casings that have bad distinguishability will tend to increase the estimated p value of their member group to unacceptably high levels. Having a smaller group limits the damage done by a single casing. To use a golfing analogy, playing extremely poorly on one hole is much more harmful in stroke play (where every stroke counts) than in match play (where only holes won or lost count).

To address these questions, we use the N-3D results, which were more accurate than the I-2D results for the two collections studied here. For all except the NBIDE breech faces, the calculated values of p are orders of magnitude too large to be consistent with the requirements of a large database. For a database of size $N = 100\,000$, at most 24 % of the firing pins, and a much lower proportion of the De Kinder breech faces, satisfy a 90 % success rate. The NBIDE breech face impressions are drastically different from the other types of impressions. Under the most optimistic scenario (grouping by casing), for a database of size $N=100\,000$, around 90 % of the paired comparison estimates of p would lead to success rates of 90 % or over , but that proportion is significantly lower if the normal model estimates of p are used. If instead, there is grouping by guns, then the average success rate of the guns is only about 50 %. Under the pessimistic scenario of a single group, then p needs to be on the order of 6.2×10^{-5}, so that the estimate of $p = 0.002$ remains over 30 times too large, despite being orders of magnitude smaller than anything else seen.

Discussion

In summary, our responses to the Project Questions are as follows:

Distinguishability (Question 1)
For a technology to be feasible for a very large database, its Top Ten lists should have obtained close to all possible correct matches in a relatively low sample size experiment like those described in this report. Nothing we have seen here comes close to achieving such high performance standards except for the N-3D performance on the NBIDE breech faces, which suggests that 3D topographic methods are a significant advance for breech face analysis. How promising this technology is for very large databases depends on whether the NBIDE or De Kinder results are more representative of the challenges faced by a national database. In the NBIDE study only 108 casings were fired by only 12 guns covering 3 different brands. A larger population of guns would make more likely the presence of 'rogue' large correlations from non-matching guns. Also, gun brands and models not covered in the NBIDE study (e.g., low-cost guns high on the ATF's list of crime gun usage [37] but not available as new purchases by us) may well be more difficult to distinguish and identify than those included.

Also, in order to perform at levels necessary for very large databases (say around 100 000), the error rates must be very low—so low in fact that for experiments with only 70 or 108 casings as in the report, there can be no overlap between the matching and non-matching sample distributions. While there was considerable separation between matching and non-matching distributions for many of the reference casings, especially those fired from Rugers, others had much less margin for error. Those matching correlations would be in danger of being overtaken by non-matching correlations in a very large database with a much larger population of non-matching correlations. For each individual casing in the NBIDE set, there were only 8 correlations of casings in the matching sample and 99 correlations of casings in the non-matching sample. Thus, one can try to estimate the distributions by pooling the matching and non-matching samples for each gun; however, this likely makes the estimated distributions wider than they should be and in fact would estimate that only half the guns would be successfully matched using the NBIDE breech face data. Estimating very low probabilities with moderately low sample sizes continues to be a challenging problem. We used normal distribution models for the correlation scores themselves in an attempt to ameliorate the problem. Use of the normal models lowered the success rate for the optimistic scenario of grouping by casing (see Sec. 9.10.4).

Gun and Ammunition Factors (Questions 2 and 3)
There seemed to be only marginally significant differences between ammunition types. For the De Kinder breech face impressions, Remington ammunition yielded the largest number of correct gun matches. However, two Remington cartridges were fired from each gun versus one cartridge for each of the other brands. Therefore when a Remington cartridge was compared to the rest of the cartridges of the same gun, there was at least one cartridge of the same brand in that group. This was not true for the other ammunition brands. Therefore, Remington had an "advantage" with respect to the rest. Likewise, the De Kinder firing pin data did not show significant differences among ammunition types. For the NBIDE collection, ammunition type

was marginally significant for both the firing pin and the breech face impressions, with PMC ammunition yielding the highest number of correct gun matches for both regions.

There was a statistically significant difference between the three gun types for the NBIDE firing pin impressions, with Ruger yielding the largest number of correct matches, and Sig Sauer the fewest. However, the gun types did not show significant differences for the NBIDE breech face impressions.

Region Differences (Question 4)
Much work remains to be done on this question. Because the breech face impression is larger than the firing pin impression and ejector mark, it generally contains more peaks and valleys that in turn figure to carry more information than the other two regions. In addition, the flat base surface of the breech face impression should make the individual characteristics easy to recognize. However, the firing pin impression and the ejector mark may carry strong individual characteristics that support positive identifications. The breech face region was clearly superior to the firing pin only for the NBIDE casings. The firing pin data were slightly superior for the De Kinder casings.

A further helpful metric might come from the topography of the ejector marks, which would be particularly useful for exclusion filtering of nonmatches. However, we have not developed a technique for correlating different ejector marks because of their widely varying outer boundaries. It is difficult to develop automated software to correlate the shapes of such regions, particularly when many ejector marks are partially obliterated by the manufacturer's headstamp. Common practice for the I-2D is a manual operation whereby the users draw the ejector mark boundaries themselves when making entries. One of our tasks for future work is to develop a similar analysis program for the existing ejector mark data.

As with ejector marks, class characteristics were not used to differentiate between cartridge cases fired by guns of different manufacture. Class characteristics can provide a valuable filter of non-matching pairs. The overall shape of the firing pin impression, for example, could be used as a class characteristic. Although such a shape will most likely be the same for all guns of a given brand, it may differ for guns of different brands.

Topography (3D) Imaging and Current Optical Imaging Technology (Question 5)
The experimental N-3D results were more accurate than the I-2D results for four experiments. Topography (3D) methods have three advantages:

1. Ballistics signatures are mainly geometrical topographies and N-3D is based on the direct measurement of surface topography.
2. Results for conventional optical images are more sensitive to illumination conditions and surface appearances than topography images.
3. Topography measurements are traceable to dimensional metrology standards.

In addition, the N-3Danalysis scheme of outlier removal, filtering, registration, matching, and statistics is non-proprietary, and this openness should facilitate development of improved algorithms by the technical community. For example, standard topography analysis methods

[30] may be adapted to separate micro- from macro-topography and extract individual characteristics of the surfaces for correlation and identification.

In short, an "open-box" ballistic imaging/analysis methodology has been developed and investigated with accuracy comparable with (and at times higher than) the widely used I-2D methodology. Aside from the question as to whether a large database is feasible, topographv measurement methods can help the ballistics imaging community to improve both ballistics matching systems and the feasibility of such a database. Being of relatively recent vintage, these techniques are still being refined and improved. Questions remain on the best way to handle data processing and optimal correlation metrics. Improvements may also be made using other approaches, such as feature extraction techniques and alternative modes of decomposition (wavelets, principal component analysis, neural nets, etc.). It is possible that refinements will result in great improvements some time from now. However, the requirements of a very large database will likely require substantial improvements.

A disadvantage of the current prototype of the topography approach is that the time required to record the data is much longer than the current I-2D technology. The data gathering and perhaps the correlation algorithms will need to become much more efficient for practical application to a large database.

1. Background

Guns leave telltale markings on fired casings and bullets that enable firearms examiners to link weapons to crimes or one crime to another by comparing fired casings or bullets in an optical microscope. The inspection technology used for this is generally a comparison optical microscope [1] (see for example Fig. 1-1). Evidence evaluated in this way may be used in criminal court to help convict perpetrators of crimes where guns were involved.

Figure 1-1. Photo of a forensic comparison microscope used to inspect pairs of objects and match images of similar objects. Inspection of the bases of a pair of cartridge casings is illustrated here (photo from Leitz* product literature).

However, the confirmation of a link using a comparison microscope is a manual and relatively time consuming process, so automated optical systems have been developed to aid crime lab investigators by processing large volumes of evidence more efficiently. Such work can produce a small set of likely matches, which the firearms examiner can positively confirm using the comparison microscope. A widely used technology [2,3] for this is the Integrated Ballistics Identification System* (Fig. 1-2), sometimes referred to as IBIS 2D, where the surface image is acquired by reflection optical microscopy. We will use the abbreviation I-2D here. This system was developed by Forensic Technology, Inc. (FTI). I-2D workstations generally comprise:

- A data acquisition station (DAS), consisting of a computerized optical microscope capable of generating digitized images of small areas of casings or bullets with the individualized surface markings that help identify a weapon,

---

* Certain commercial equipment are identified in this report. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the equipment are necessarily the best available for the purpose.

- A signature analysis station (SAS) providing access to a large database of the digitized images so that one entry can be compared with many others,
- Correlation software to provide a quantitative estimate of the most likely matches.

Figure 1-2. Photo of an I-2D system (photo from Forensic Technology Inc. product literature).

Such a system reduces the time for initial inspection of ballistics evidence and increases enormously the number of comparisons that can be performed, allowing lab personnel to sift through relatively large amounts of evidence. Using the I-2D technology, the Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF) has led the development of a nationwide set of regional databases called the National Integrated Ballistics Information Network (NIBIN) [4,5] to facilitate the sharing of information among crime labs in different jurisdictions and to increase the number of confirmed matches, called "hits." Over the last decade, NIBIN has been collecting casings and bullets associated with crimes nationwide. This database is accessible to approximately 230 forensic laboratories distributed across the United States. The cornerstone of the NIBIN is the I-2D, which first acquires a reflection microscopy image of the casing or bullet and then applies proprietary analysis code to determine if an incoming crime-related casing or bullet image matches another casing or bullet image entered into the NIBIN database.

For several years I-2D has been the dominant ballistics imaging/analysis technology available commercially [5]. In collaboration with NIBIN, the system has evolved over the past decade and the coordination, placement, and networking of the I-2D devices has been supported by the ATF. For participating forensic laboratories with the I-2D equipment installed, there is a well-defined procedure for acquiring images of given crime-scene casings or bullets, searching roughly 860 000 images in the NIBIN database, and receiving a ranked list of the Top Ten matches from the database that best match the crime-scene casings or bullets in hand.

One issue of the I-2D implementation is that the correlation algorithm for finding hits is proprietary. If the algorithm were openly available, it might receive beneficial enhancements from the technical community at large, in addition to the advances made by the manufacturer with the support of its clients. Another issue is that several experimental reports raise concerns about the performance of the I-2D system for large databases. For example, in a California Department of Forensic Services study [6] consisting of an image database of 792 Smith&Wesson semiautomatic pistols and a reference subset of 50 of the same model pistols, Tulleners observed that the I-2D placed the correct weapon in its top 15 list only 62 % of the

16

time–even though the ammunition were all the same type. More sobering was the fact that the success rate fell to 38 % if the ammunition were different. Studies by De Kinder [7] and George [8] obtained similar results. By comparison, for a practical imaging/searching system, one might envision a success rate criterion of about 90 %.

In contrast to the above articles, a study by Beauchamp and Roberge from FTI studied 2D methods and drew favorable conclusions about the feasibility of a large database having up to millions of entries [9]. Another study of the parameters of correlation databases and experiences with the I-2D methods was performed by Nennstiel and Rahm [10].

The overall success of the NIBIN with the law enforcement community is a contributing influence to the interest in developing a much larger National Ballistics Imaging Database, which would include entries from casings or bullets taken from all newly manufactured handguns in the country. Instead of the approximately 860 000 entries from crime guns estimated in NIBIN [4], this new database would need to handle entries for roughly four million new handguns [11] sold per year. The potential gain would be that any casing or bullet connected with a crime could be matched back through its ballistics signature to the identity of the gun and the original owner, a potentially powerful crime-solving method. The governments of New York [12] and Maryland [13] took the lead in developing state-level Reference Ballistic Imaging Databases (RBIDs) of their own, but the state of California after researching the question [6,14] has not created an RBID. The question of whether to do so or not depends in large part on whether the gains in the number of matches and the aid to investigators will be sufficient to warrant the cost of developing and maintaining the huge database.

At the request of the Office of Science and Technology Policy, the National Academies (NA) have formed the Committee to Assess the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database. The Office of Law Enforcement Standards (OLES) at the National Institute of Standards and Technology (NIST) was selected to manage a research project to assist the NA Committee's work. The research utilized members of OLES, NIST's Manufacturing Engineering Laboratory (MEL), and NIST's Information Technology Laboratory (ITL). To take advantage of the latest research, this team contracted with a private company, Intelligent Automation, Inc. (IAI) of Rockville, Maryland, to assist in the project. In preparation for the collaborative research, we made use of materials and results developed in a previous IAI repetitive fire study [15], the De Kinder et al. study [7], and an existing NIST project on standard bullets and casings [16,17]. Each of these studies had information applicable to the database evaluation. The IAI repetitive-fire study examined bullets fired from gun barrels of different manufacture and brands to assess the validity of firearm identification using topographical data. The NIST project developed information on standard bullets and casings using surface topography techniques and described statistical methods to compare repetitive measurements. The De Kinder project [7], along with the earlier Tulleners project [6], were relevant studies of the application of casing inspection techniques to large databases.

For the NBIDE experimental design, we considered parameters that would produce a typical case scenario in regard to a ballistic database where the first test fires are used and a range of weapons are included. This scenario would reflect the type of cartridge cases that would be collected from the weapons manufacturers for inclusion in a national database. To capture these attributes,

ammunition and weapon types were discussed with the NA Committee, the ATF, and representatives from the Association of Firearm and Toolmark Examiners. The aim was to select weapons that cover a range of quality and precision tooling. Given the limited time available for the analysis of the fired casings, the number of weapons used was restricted.

Ammunition also was chosen carefully in the NBIDE and the selection was guided by casing data gathered by De Kinder (see Attachment D of Ref. 12), which provided information on the primer, hardness, and primer type. The final selection of ammunition was refined by the NA Committee and the ATF.

We used a NIST Standard Reference Material (SRM) 2460, Standard Bullet, and prototypes for SRM 2461, Standard Casings, as check standards. The Precision Engineering Division at NIST, in collaboration with Intelligent Automation, Inc. (IAI) of Rockville MD, initiated testing of three optical methods for imaging topography: the interferometric microscope [18], the Nipkow-disk confocal microscope [19], and the scanning laser confocal microscope [19]. The obtained results were correlated with those measured with a stylus instrument [20], which uses a physically contacting diamond probe to scan the surface and obtain a surface profile. As discussed later in Sec. 2.1.2, the Nipkow-disk instrument was selected to support further research. The research involved the measurements and analysis of the existing casing and bullet collections from the previous studies by De Kinder et al. [7] and Bachrach of IAI [15], respectively. NIST then supplemented these collections with test fires and data developed according to an experimental design procedure. According to the procedure, NIST purchased the specific weapons and ammunition, under the guidelines described earlier, and performed test fires and analyses using double blind methods [21].

It was anticipated that the test results would provide answers to several questions raised by the NA Committee and the research team. These questions focused on such issues as: Are the individual firearms self-consistent? Are they distinguishable? What are the contributing factors to the topographic markings: the firearm, the selected image region, the ammunition type? What are the preferred response metrics: parameters from the I-2D, the topography cross-correlation function (*CCF*), surface roughness? The research team was aware that the choices of imaged region would play a major role in evaluating answers to the offered questions. For example, how much does each imaged region contribute to making an identification of ballistics evidence and what is the cumulative effect of these regions on attaining the highest "hit" rate in a numerical association? Discussion between the NA Committee and the NIST research team also touched on the topics of weapon wear and environmental factors. These topics, although taken under consideration by forensic examiners, were not considered in this portion of the NIST research. The main question to be addressed by this research is the feasibility of a National Ballistics Imaging Database (NBID) of firearms manufactured for sale in the United States. Test fires taken just after firearm manufacturing would constitute the casings or bullets that would be imaged in this database. The information in the database would therefore be from new weapons.

The remainder of this report is organized as follows. In Sec. 2, we describe initial observations and the overall goal of the NIST research. In Sec. 3, we describe the work plan in detail. In Sec. 4, we describe the instrument and methods used for surface topography measurement. In Sec. 5, we describe the overall goals and detailed plans of the designed experiment of test fires. In Secs.

6 and 7, respectively, we describe topography images taken with the Nipkow-disk confocal microscope and measurements taken with I-2D. The data processing and cross correlation of the topography images is described in Sec. 8, and the statistical results are described in Secs. 9 and 10. In Sec. 11, we make several concluding observations from the work and emphasize questions to answer in future work.

2. NIST Participation in the National Academies' Study

A primary goal of the NA Committee study is to determine whether the marks on casings or bullets fired from guns are sufficiently gun-specific and reproducible to make it feasible to build a national database of test fires of new guns to help solve crimes. In support of the Committee's work, NIST and IAI have been performing surface measurements and experiments. The emphasis of this work has been the measurement and analysis of surface topography because ballistic signatures are mainly geometric topographies and test fires produce obvious changes in the surface topography of casings and bullets. Direct measurement of surface topography seems to be the most likely way to enhance the surface information from fired casings and bullets in order to increase the accuracy of ballistics identification to a level consistent with the requirements of an NBID.

The use of surface topography data of casings is the main focus of the test fires, measurements, analysis, and statistical studies we have performed at NIST and IAI. NIST has also collaborated with the ATF and compared the results of topographic measurements with the I-2D acquisitions of the same casings. Topography measurement of bullets is planned for a future report, although some considerations and results are also discussed here.

2.1 Initial Observations

2.1.1 Surface Topography Measurement

We begin, therefore, with the observation that surface topography changes are likely the important, reproducible effects produced by firearms on casings and bullets. Surface topography for these small objects is considered to include both form deviations, with a size close to the scale of the casing or bullet, and fine roughness structures observable down to the resolution of the measuring instrument, usually an optical microscope. Direct measurement and analysis of these topographic structures on fired ballistics evidence should be important for identifying the weapon that fired them. By comparison, the most widely used method, optical reflectance microscopy, produces images that are not so closely related to the topography itself. That is, optical image contrast, which may be represented by the symbol $I(x,y)$, is mainly a function of:

- lighting conditions,
- surface slope,
- shadowing effects,
- multiple reflections, and
- changes in the optical properties (index of refraction and extinction coefficient) over the surface, which produce changes of observed reflectivity over the surface.

Surface topography, therefore, affects the image contrast in a complex fashion through its slope variations, through shadowing effects, and through multiple reflections. The image contrast is also affected by changes in reflectivity over the surface, a non-topographic property that might be related to changes caused by firing but to other causes not related to firing, such as inhomogeneities in the casing or bullet material itself.

Topographic microscopy, on the other hand, is directly sensitive to surface height and spacing variations $z(x,y)$ and can measure these directly, independent of illumination and shadowing effects. This makes it advantageous for studying the marks on bullets and casings left by the gun-firing process. A disadvantage of topographic microscopy, as we will discuss later, is the noise arising during the signal processing required to create a topography image, resulting in dropouts and outliers in the data points. A complication is the need to develop a data separation method whereby the fine roughness signature can be separated from form deviations of the casings or bullets.

We choose an optical topography measurement method here because the traditional method of topography measurement by stylus profiling is a direct mechanical contact technique, which would not be acceptable for use on ballistics evidence. Another technique, optical scattering [22], does not provide enough discrimination of individual peak-valley characteristics to be effective as an identification tool, and scanning electron microscopy [23] is comparatively difficult and exotic to apply to the millions of objects expected for a national database of new-gun test fires.

2.1.2 Primary Measurement Method

The experimental NIST Topography Evaluation System (N-3D) consists of five components:

1. An optical topography measurement system that produces a computer-resident digitized 3D surface topography image,
2. A filtering/outlier-rejection system for pre-processing the topography image data,
3. A registration process by which the optimal alignment of the individual characteristics of two compared casing images is determined,
4. A statistical calculation to provide a best-alignment correlation coefficient, a metric quantifying how well the two images match,
5. A statistical data analysis methodology that provides a ranked list of best-matched guns as well as other information, such as gun-type and ammunition effects.

We chose Nipkow-disk confocal microscopy for measurement of the surface topography after brief tests of its accuracy and a comparison with two other optical techniques. A description of this technique is given in Sec. 3. The other techniques we tested were white light interferometric microscopy and laser scanning confocal microscopy.

Figure 2-1 shows surface profiles measured with these three methods compared with a profile measured with a stylus instrument [24]. The stylus method is taken as the reference method, mainly because of its demonstrated accuracy and because it is less prone to produce data outliers and dropouts than the three optical methods.

The measurements were taken on a land engraved area (LEA) of a standard bullet, Standard Reference Material (SRM) 2460 [16,17]. The illustration shows qualitatively that all four methods agree fairly well. The three optical profiles were stitched together [25] from several profiles measured over about a 1.4 mm length of profile. The peaks and valleys of all four profiles are similar and match up well in position with one another. Pair wise quantitative

comparisons between each optical profile and the stylus profile are indicated by the maximum values of the cross-correlation function (CCF_{max}) also shown. The CCF_{max} value, described later in Sec. 8, is a metric to indicate the agreement between two time series, in this case two surface profiles, when the two correlated profiles are matched in phase as well as possible. If the two profiles were exactly the same, the CCF_{max} value would be unity or 100 %.

Figure 2-1. The profile of LEA 1 of a standard bullet measured by four techniques [24]. 1) stylus instrument; 2) interferometric microscope, CCF_{max} = 92.1 % with respect to the stylus profile; 3) Nipkow-disk confocal microscope, CCF_{max} = 98.9 %; 4) laser scanning confocal microscope, CCF_{max} = 95.3 %. The vertical unit is µm, the horizontal unit is mm.

The CCF_{max} values are given for comparison between each of the three optical profiles and the stylus profile. Initial studies showed that all three optically measured profiles agreed about the same with the stylus profile because the CCF_{max} value representing agreement with the stylus profile was approximately 92 % for all three. There was little to choose between the techniques from an accuracy point of view. We chose the Nipkow-disk approach to pursue further because this technology is already being used commercially by FTI for topography measurement of bullets [27], and a Nipkow-disk system was already in place at IAI. In addition, the laser scanning confocal microscope available to us was more difficult to use than the other microscopes because its capabilities did not include automated stitching software. With the interferometric microscope, we have also observed significant profiling anomalies [28, 29] for roughness measurement of smooth surfaces, with roughness average (Ra) [30] on the order of 100 nm. Surfaces that might be encountered later could have roughness in the 100 nm range.

We then refined the measurement procedure with the Nipkow-disk confocal microscope. Profiles 1 and 3 for Fig. 2-1 are shown again in Fig. 2-2. The top profile is the stylus profile and the second is the profile measured with the Nipkow-disk confocal microscope. Both profiles had curvature removed with a high-pass Gaussian filter [30] with 0.25 mm long-wavelength cutoff and were smoothed with a moving average Gaussian filter with a 2.5 µm short-wavelength cutoff. The sampling interval (point spacing) of the stylus profile is approximately 0.25 µm, and

the sampling interval of the profile obtained with the confocal microscope is approximately 0.625 µm. This profile was then interpolated so that the sampling interval matched that of the stylus profile. The cross correlation function is the third profile. Its maximum value, CCF_{max}, occurring when the two upper profiles are in phase, is 98.9 %. This metric will be discussed in Sec. 8.

Figure 2-2. Computer screen showing a comparison of stylus (top) and Nipkow-disk (second) confocal profiles of LEA 1 of a standard bullet.

Three dimensional topography data taken with the Nipkow-disk confocal microscope are shown in Fig. 2-3. These are topography images of firing pin impressions on two prototype NIST standard casings [31] (P21 and P31), which were nickel replicas of the same master casing. The replicas were fabricated by electroforming [32]. The evident similarity between the data shows qualitatively that both the electroforming process and the measurement process are reproducible. A preliminary calculation of the cross correlation function yields a value of the CCF_{max} of 97 %. However, the calculation of this function depends sensitively on the cutoff of the high-pass filter and hence on how much the overall form of the firing pin impression is attenuated in the analysis.
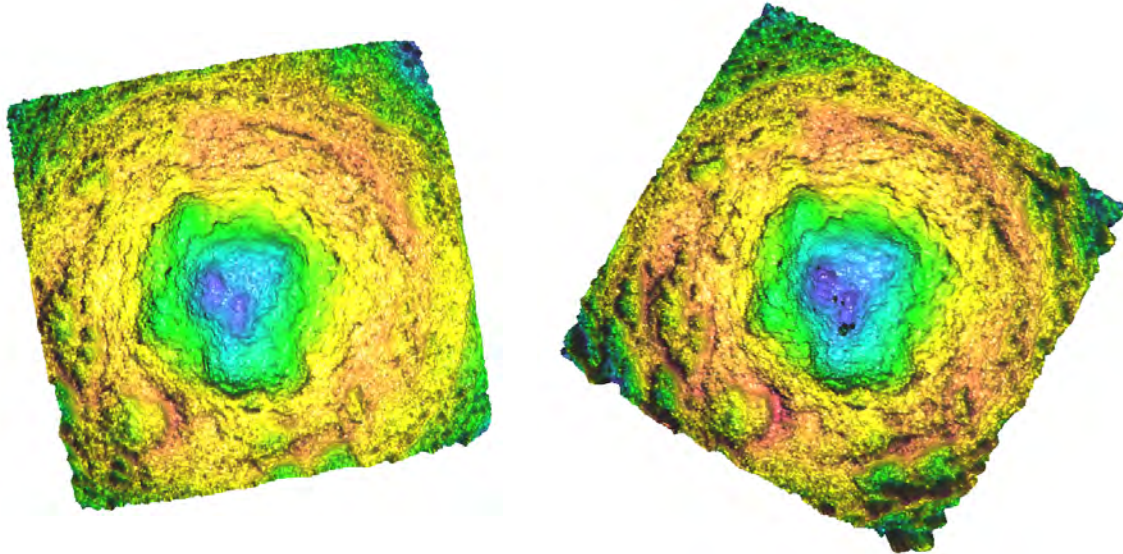
Figure 2-3. Topography images of two prototype standard casings measured with Nipkow-disk confocal microscopy. The casings are electroformed replicas of the same master. Left–casing P21, right–casing P31.

2.2 Further Investigation of Topography Measurement Technology for Ballistics Identification

2.2.1 Two Methods for Representing Casing and Bullet Signatures: Optical Imaging and Topography Imaging

From the point of view of ballistic identification, both the optical images and topography images could be used for representing casing and bullet signatures. The optical image represents the surface appearance of the casing or bullet signatures when observed under an optical microscope. It is well known that tool-marks, including casing and bullet signatures, are mainly geometrical features by nature. When these geometrical features are observed under a microscope, the appearance of the optical image is significantly affected by the optical properties of the surface, such as color, refractive index, and extinction coefficient, as well as by multiple reflections and the lighting conditions that cause the shadowing effects. As a result, the optical image does not precisely represent the surface geometry, especially the micro geometry, which is an important part of the casing and bullet signatures.

On the other hand, a topography map can more precisely represent the surface geometric and micro-geometric characteristics of the casing or bullet signatures, independent of the effects caused by optical properties of the surface and lighting conditions. Furthermore, calibration and check standards, measurement methods and uncertainty procedures for topography measurements have been well established for surface metrology and standardized nationally and internationally. As a result, topography measurements could be a better alternative for ballistics identification, and make the ballistics measurements traceable to national and international standards.

In the following discussions, we use the names "optical imaging" and "topography imaging" (or "topography" for short) to specify these two different methods for representing casing and bullet signatures.

2.2.2   Two Techniques for Ballistics Identification: Optical Image Comparison and Geometric Topography Measurement

The comparison of tool-marks was admitted as evidence in a criminal case as early as 1879 [33]. The first binocular comparison microscope using optical image comparisons for ballistic identification was invented by A. Inostranzeff [33] in 1925. The later versions of the comparison microscopes, such as that developed by Leitz (Fig.1-1) and the I-2D developed by FTI (Fig.1-2), were all based on optical image comparisons. However, with the development of several types of topography measurement techniques [29,34] having high resolution, measurement traceability, large measurement area, and automation software, optical topography measurements of casing and bullet signatures could yield excellent results. Recently, IAI [35] developed a research system and FTI [27,36] developed a commercial one for bullet topography measurement based on confocal microscopy.

2.2.3.   NIST Participation: Investigate and Develop Topography Measurement for Ballistic Identification

It is therefore important to investigate the use of topography measurement technology, including its associated hardware and software issues, as an alternative approach for ballistic identifications. That includes the following questions:
- Is it possible to use topography measurement techniques as an alternative approach for ballistics identification?
- Compared with the I-2D technology currently deployed in the NIBIN, which one is better for ballistics identification?
- What parameter, algorithm and measurement program must be developed for quantitative comparisons of the 2D and 3D surface topographies of casing and bullet signatures?
- How to improve upon existing correlation results [6-10]?

3. Workplan

In order to test the usefulness of topography images as ballistics evidence, we have been developing three data sets of topography images and have been testing how reliably ballistics matchups can be identified from these data. These were

- 70 casings chosen at random from the collection of De Kinder test fires, used previously in the study by them [7],
- 108 casings recently test-fired at NIST from twelve new guns as part of a designed experiment,
- 176 bullets fired from eight guns in the experiment previously conducted by Bachrach [15].

The primary goal of measuring these materials and performing the subsequent correlation analysis was to test how reliably casings or bullets fired from the same guns could be identified. Details of these materials and data sets are given below.

3.1 De Kinder Casings

In the DeKinder study, 4200 test fires [7] were performed by the California Department of Justice Sacramento Laboratory to assess the feasibility of developing a state-wide database of new test fires in California using I-2D technology. Approximately 600 Sig Sauer 9 mm pistols were used for the test fires, and seven rounds of ammunition were fired in each pistol. The ammunition used was:

1. CCI
2. Winchester (Win)
3. Remington Peters (R-P), also referred to as Remington (Rem) here and elsewhere
4. Speer
5. Wolf
6. Federal (Fed)
7. Rem (a repeat).

I-2D optical imaging technology was used to record the surfaces of the fired casings. One purpose of that study was to test whether the individual identities of highly similar pistols could be distinguished sufficiently accurately by the imaging technology. The results of that study [7] and a previous, similar one [6,14] were not promising. We remeasured 70 of the De Kinder casings with topographic imaging to assess whether the new technology can augment or improve upon the conventional optical imaging technology. For each of the 70 casings, we measured the topography of the firing pin impression, breech face impression, and ejector mark. We then used standard correlation algorithms to generate metrics intended to distinguish casings fired by the same pistol. We also compared the accuracy of these results with those obtained with an I-2D system located at the ATF National Laboratory Center (NLC), Ammendale, MD.

3.2 BIDE Casings

Although a good deal of knowledge has been gained from work on the De Kinder casings, aspects of the experimental design there were not optimized for learning about certain noise factors in the identification process for test fires. We therefore carried out a set of 108 new test-fires under a different experimental design to test the relative effects of gun identity, gun type, ammunition type, and drift on the topographic markings. The experimental design included twelve new 9 mm handguns (four each from Sig Sauer, Ruger, and Smith&Wesson (S&W)), three types of ammunition (PMC, Remington, and Winchester), and three repetitions for each combination, for a total of 108 test fires carried out over three different days. An additional 36 rounds of Speer ammunition were also mixed into the test-fire design and test fired in case the analysis of the 108 casings indicated a need for more entries in the experimental design. This design enables us to discern similarities and differences in topography when

- the same gun and the same ammo are fired at different times,
- different ammos are used in the same gun,
- different guns of the same brand are fired, with the same or different ammos, and
- guns of different brands are fired.

The test fires were carried out in the NIST firing range. After the test fires, the topography of the firing pin impressions, breech face impressions, and ejector marks were measured with Nipkow-disk confocal microscopy [19]. Using procedures of a double-blind study [21], the identities of the fired casings were remixed and kept secret from the people performing the measurements and data analysis so that the weapons were not known to them *a priori* before a set of predictions were made by them about which casings were fired from the same gun. These predictions were based on correlations performed on the topography images. After that, the identities of the remixed casings were revealed, and the predictions were compared with the true matches of casings actually fired from the same guns.

The serial numbers and IDs of the BIDE guns are as follows:

Brand	Model	Serial Number	Short ID
Ruger	P95D	31545341	Ruger 41
Ruger	P95D	31545342	Ruger 42
Ruger	P95D	31545346	Ruger 46
Ruger	P95D	31545348	Ruger 48
Sig Sauer	P226	U702930	Sig Sauer 30
Sig Sauer	P226	U702931	Sig Sauer 31
Sig Sauer	P226	U702932	Sig Sauer 32
Sig Sauer	P226	U703333	Sig Sauer 33
Smith&Wesson	9VE	PBV8305	S&W 305
Smith&Wesson	9VE	PBV8306	S&W 306
Smith&Wesson	9VE	PBV8314	S&W 314
Smith&Wesson	9VE	PBV8401	S&W 401

3.3 IAI Bullets

Since 1997, Bachrach et al. [15,35] of IAI have been researching the use of topography data for describing both firearms and toolmark forensic evidence with support from NIJ, the National Science Foundation, and the Federal Bureau of Investigation (FBI). These projects have provided IAI personnel with expertise in the acquisition, processing, and comparison of firearms-related forensic evidence.

As part of the ongoing NIJ study [35], IAI created a substantial collection of sample bullets fired by a variety of guns and barrels of various levels of quality. Table 3-1 summarizes the barrel brands comprising this collection and includes information regarding the manufacturing technique used in the rifling of these barrels, the number of land impressions found on the bullets, and their basic geometrical characteristics. Table 3-1 also includes information regarding the origin of these barrels. This information is coded in the "Notes" column as follows:

a) Consecutive: barrels that were consecutively manufactured (a higher control of ordering), as stated by the manufacturer,
b) Sequential: barrels that were sequentially manufactured (a less stringent control of ordering), as stated by the manufacturer,
c) Standard: barrels that were purchased through a supplier.

Table 3-1. Selected gun/barrel manufacturers and manufacturing techniques.

	Manufacturer	Number of Barrels	Manufacturing Technique	Number of Land Impressions	Width of Impressions [mm]	Barrel Length [cm]	Notes
1	Beretta	11	Gang Broach	6	2.0	10.8	Consecutive
2	Ruger	11	Gang Broach	6	2.0	9.5	Consecutive
3	S&W	11	Elect.-chem mach.	5	2.5	9.5	Standard
4	Taurus	6	Gang Broach	6	1.3	6.4	Consecutive
5	Browning	15	Hammer Forged	6	1.8	10.2	Sequential
6	HiPoint	11	Button Rifling	6/9	1.3/1.6	7.0	Consecutive
7	Sig Sauer	12	Hammer Forged	6	1.7	9.3	Consecutive
8	Bryco	11	Gan Broach	6	1.3	7.9	Consecutive
9	Glock	11	Hammer Forged	6	--	9.5	Standard

These gun types were selected based on the following criteria: a) frequency of association with crime scenes, b) availability of barrels for purchase, c) availability of reliable information regarding their manufacture, and d) degree to which the overall group of selected guns spans the spectrum of commonly used manufacturing techniques. The selection process began with a study of gun models most commonly associated with crimes, using resources such as the ATF's National Tracing Center (NTC) [37]. The NTC maintains statistics of every gun that is "traced" as part of a crime investigation. At IAI's request, NTC provided the statistics of the 25 gun models most often associated with crimes over the last twelve years. That list of candidate guns was narrowed based on the availability of guns for purchase and the availability of information regarding the manufacturing techniques involved in the rifling of their barrels. The nine brands selected were a compromise between all these factors.

The NIJ study includes three main components, and bullet samples were created to address the needs of each of these elements of that study:

a)   The first element of the study is referred to as the "Barrel Wear Study." The purpose of this portion of the study is to assess the effect of barrel wear on the features transferred between the barrel and the bullets fired through it.

b)   The second element of the study is referred to as the "Pristine Bullet Study." The purpose of this portion of the study is to validate the degree to which a one-to-one association between a barrel and a bullet is possible, including an assessment of the probability of erroneous classification (false positive match, false negative match).

c)   The third element of the study is referred to as the "Damaged Bullet Study." The purpose of this portion of the study is to validate the degree to which the conclusions reached for pristine bullets apply to damaged bullets.

Bullet samples were created for each of these three portions of the NIJ study as follows:

a)   Barrel Wear Study: For each barrel brand listed in Table 3-1, one barrel was selected for the Barrel Wear Study. A total of 220 bullets were fired with each barrel, and 80 of these bullets were retrieved using a water tank. These 80 bullets correspond to the bullets 01 through 50, bullets 101-110, bullets 201-210, and bullets 211-220. The order in which these bullets were fired was preserved so that any possible wear effect could be assessed. All bullets fired for the Barrel Wear Study were of Winchester manufacture. The total number of bullets test fired and retrieved under this portion of the project was 720 bullets.

b)   Pristine Bullet Study and Damaged Bullet Study: For each barrel brand listed in Table 3-1, ten barrels (except in the case of Taurus, where only five were available) were used for these two studies. Twenty four bullets were test fired with each of the barrels. Of the 24 bullets, 12 were of Winchester manufacture and 12 were of Remington manufacture. For each of these sets of 12 bullets, ten were retrieved in pristine condition and two bullets were fired in such a manner as to "damage" them. The total number of bullets test fired and retrieved under this portion of the project was 2 040 bullets.

Details of the test firing process for the bullets used in these studies can be found in [15]. As part of the present feasibility assessment for a national ballistic imaging database, it was decided that a small portion of the bullets already available from the NIJ project could be used as sample bullets for this project.

Overall eight of the nine gun barrel brands and 176 sample bullets are selected to be used as part of the present feasibility study. Bullets fired through Glock barrels are excluded because of the nature of the rifling of the Glock barrels. These barrels have polygonal rifling, which makes them extremely challenging for any image acquisition system.

The bullets selected for the present feasibility study are:

a)   Bullets 1, 2, 5, 10, 31, 32, 209, 210, 211, 212 fired through barrel No. 1 of each brand. These bullets were associated with the Barrel Wear Study. All these bullets are of

Winchester ammunition. Altogether, in this group there are 10 bullets from each gun brand and eight brands for a total of 80 bullets.

b)   Bullets 1, 2, (of Winchester ammunition), 11, 12 (of Remington ammunition) fired through barrel No. 2 and barrel No. 6 of each brand. These bullets are associated with the Pristine Bullet Study. In this group, there are 4 bullets $\times$ 2 barrels $\times$ 8 gun brands for a total of 64 bullets.

c)   Bullets 21 (of Winchester ammunition), 23 (of Remington ammunition) fired through barrel No. 2 and barrel No. 6 of each brand. These bullets are associated with the Damaged Bullet Study. In this group there are 2 bullets $\times$ 2 barrels $\times$ 8 brands for a total of 32 bullets.

The above bullets are selected for measurement and analysis. We plan to measure these and report on the results in a subsequent report.

4. Topography Measurements

A Nipkow-disk confocal microscope [38] was used to perform the surface topography measurements. This measurement technology has been observed to be reasonably accurate over a wide range of roughness average (*Ra*) from 3 nm to 500 nm [29] and to correlate well with stylus profiling for measurement of curved bullet surfaces as discussed in the previous section.
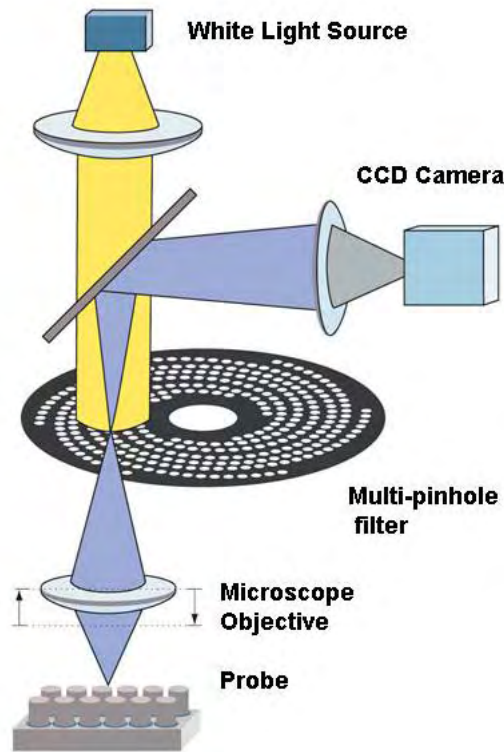

Figure 4-1. Schematic diagram of a Nipkow-disk confocal microscope [38].

A schematic diagram of this measurement approach is shown in Fig. 4-1. Illumination from a white light source is reflected from the surface under investigation and is focused onto a pinhole aperture. If the surface is at the correct height, the reflected light will be focused through the pinhole and a strong optical signal will pass onto the detector. If the surface is not at the correct height, the light will be defocused on the aperture and little or no signal arrives at the detector. Scanning the surface vertically enables one to determine the surface height at a single lateral location by looking for a maximum in the light transmitted to the detector. The Nipkow disk, which spins and consists of multiple apertures, enables the rapid collection of surface heights over all lateral positions in the field of view of the microscope.

4.1 Measurement Conditions

As is customary in ballistics examination, three areas (Fig. 4-2) were measured on the casings—the firing pin impression on the primer, the breech face impression on the primer, and the ejector

mark on the edge of the base. A 20× objective was used, but different fields of view were needed for the three areas.
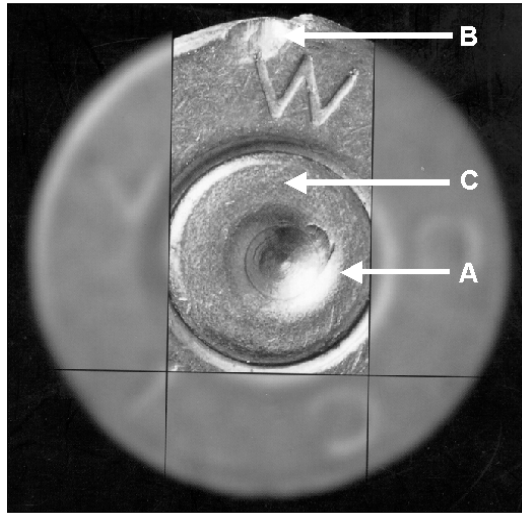

Figure 4-2. Casing signatures include firing pin impression (A), ejector mark (B), and breech face impression (C). Photo by Forensic Technology Inc.

In addition, two microscopes were used for the measurements, with cameras having slightly different fields of view. For the De Kinder firing pin impressions and ejector marks, a single field-of-view of the 20× objective was rectangular with approximate dimensions 0.8 mm × 0.772 mm. For the De Kinder breech face and all NBIDE impressions, a single field-of-view of the 20× objective was square with approximate dimensions 0.8 mm × 0.8 mm. Both cameras used 512 × 512 pixels. Therefore, the nominal pixel size was 1.5625 μm × 1.5078 μm for the rectangular image and 1.5625 μm × 1.5625 μm for the square image.

The firing pin impressions were measured with a single field of view of the objective. The vertical slice (z-slice) interval was approximately 300 nm, and about 400 image slices were measured for a total vertical measurement range of 120 μm. Topography images of four of these impressions from four different Sig Sauer 9 mm semi-automatic pistols are shown in Fig. 4-3.

The breech face impressions required a much larger field of view. The topography images consisted of usually 6 × 6 images stitched into one large image using the commercial stitching software of the instrument. These large images were approximately 4 mm × 4 mm in size and consisted of approximately 2600 × 2600 pixels (see Fig. 4-4). The z-slice interval was about 1 μm, and about 300 slices were measured for a total vertical measurement range of about 300 μm.

The ejector marks were small but elongated, and required either 1 × 2 or 1 × 3 stitched images of 0.8 mm × 0.8 mm patches (see Fig. 4-5). The z-slice interval was about 300 nm and about 300 image slices were measured for a total vertical travel range of about 90 μm.

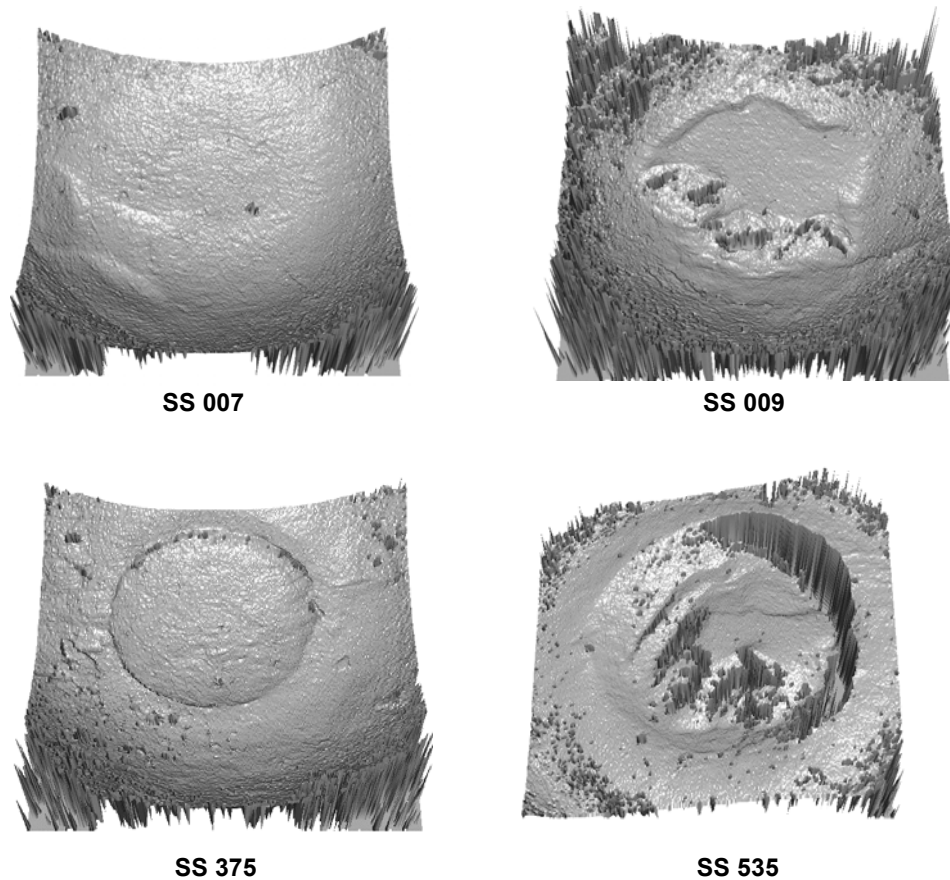**SS 007**

SS 009

SS 375

SS 535

Figure 4-3. Topography images of firing pin impressions using the same CCI ammunition from four different Sig Sauer 9 mm guns with IDs shown.

At this writing, the 9 mm bullets have not yet been measured. The areas to measure on the bullets are the 6 LEAs (see Fig. 4-6). The geometry of the bullet LEAs is quite different from that of the casings described above. The LEA markings are nearly uniaxial, but the areas are curved with a radius of curvature of approximately 4.5 mm. For this geometry, we plan to measure rectangular areas having 1×3 stitched images with size approximately 0.8 mm $\times$ 2 mm and consisting of about 506×1370 pixels. The z-slice interval will be about 0.5 μm and about 500 image slices will be measured for a total vertical range of 0.25 mm. The total amount of data for each bullet will consist of six topography images, one for each LEA.
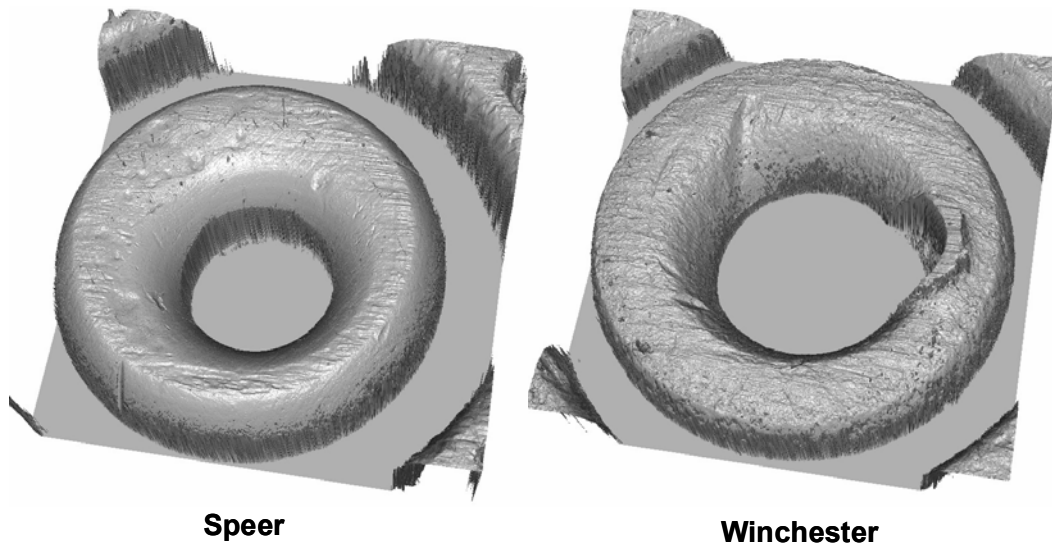
Speer

Winchester

Figure 4-4. Topography images of two breech face impressions fired by the same Sig Sauer #117, 9 mm pistol but with different ammunition, Speer and Winchester.

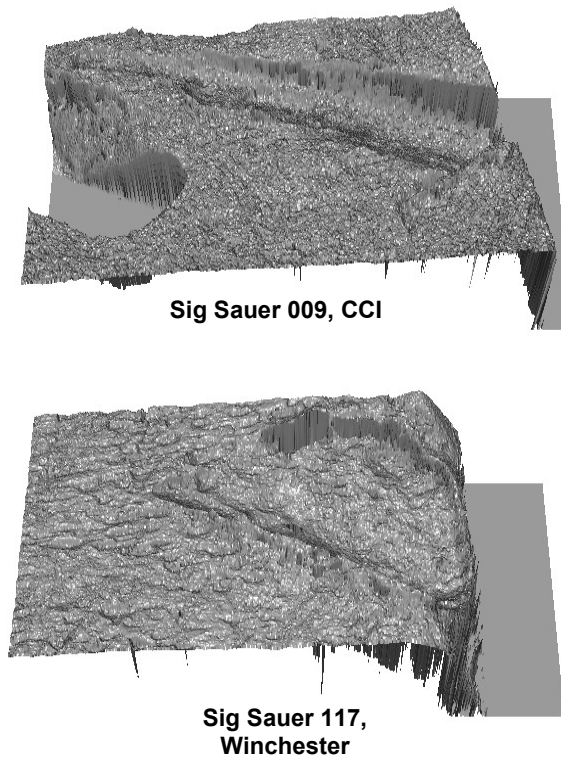**Sig Sauer 009, CCI**

**Sig Sauer 117,
Winchester**

Figure 4-5. Topography images of ejector marks of casings fired by different Sig Sauer 9 mm pistols, SS 009 and SS 117, used in the De Kinder study. The hole on the left side of the upper image is part of the manufacturer's headstamp.

4.2   Calibration and Quality Control

We also checked aspects of the microscope calibration by measuring several types of calibration and check standards. These included:

- A smooth, optical flat reference standard for eliminating any wavefront distortion of the microscope and for checking the vertical resolution limit for measuring roughness. The procedure for eliminating wavefront distortion is a standard procedure from the microscope manufacturer. The rms roughness of the sample is less than 1 nm.
- An internal NIST step height master, one with step height equal to 10.52 μm, or another with step height equal to 12.668 μm, for checking the overall z-scale calibration of the microscope.
- Two sinusoidal roughness standards, a Rubert Model 528 having a roughness average (Ra) of about 0.5 μm and spatial period of about 50 μm, and a NIST SRM 2073a having Ra of about 3 μm and spatial period of about 100 μm, for testing the accuracy of the microscope for straightforward measurements of roughness.
- A prototype NIST Standard Reference Material (SRM) standard casing was used as a check standard.

Wavefront calibration with the optical flat was performed each day that the instrument was used. The SRM 2073a was also measured each day to provide a general control check of the vertical scale and roughness measurement calibration. The step height and 0.5 μm roughness samples were measured at the beginning and end of each major series of data, such as before and after the entire set of firing pin impressions, to provide additional checks on the z-scale calibration. The SRM standard bullet or casing was measured several times as a general control sample with geometry similar to the samples being measured. For example, the firing pin impression of the control casing was measured about five times over several days during the series of measurements of the NBIDE or De Kinder firing pin impressions.

Figure 4-6. One of several land engraved areas (LEAS) on a fired bullet. Most common 9 mm bullets have 6 LEAS. Photo by Forensic Technology, Inc.

5. NBIDE – Experiment

The NIST study attempted to go beyond a re-analysis of the De Kinder casing collection and address questions not addressed by the De Kinder design. A central component in the NIST study was to determine the effect of gun type (manufacturer) on gun identifiability. As noted above, the De Kinder study was limited to a single gun type (Sig Sauer 9 mm). To go beyond this, the NBIDE statistical experimental design included the following specifications:

1. Number of gun types: 3 (Smith&Wesson [S&W or often Sm-W], Ruger, Sig Sauer)
2. Number of guns (total): 12
3. Number of ammo types: 4 (Remington, Winchester, PMC, Speer (extra))
4. Number of days: 3 (48 firings per day × 3 days)
5. Total number of test firings: 144 = 12 guns × 4 ammos × 3 days
6. Total number of casings subsequently analyzed: 108 (Remington, Winchester, and PMC ammos only)

The experiment was designed and conducted in accord with rigorous statistical design principles and techniques. Within a given day, there were 4 sets of 12 firings. Within each set of 12 firings, a given gun was fired once. Across the 3 days, there were a total of 12 sets of 12 gun firings. In a latin square fashion, for each time position (1 to 12) within a set, each gun was fired once, and only once, across the 12 sets. The twelve guns are shown in Fig. 5-1.

Figure 5-1. Photograph of the 12 guns used for the NBIDE designed experiment.

36

All 4 ammos were used every day. The ordering of the ammos for the 4 sets of 12 firings within a day was balanced in an incomplete latin square fashion. Each firing was done remotely using a magnetically actuated trigger pull of the gun held in a fixture (Fig. 5-2). One cartridge at a time was loaded into the fixtured gun according to the order of the experimental design, then fired. The casing was ejected into a bag, not shown, then manually removed, and an entry of the test fire was recorded. Finally, after the 144 test fired casings were collected and recorded, they were re-randomized for the topography acquisitions and analysis so as to assure that the analysis was done in a double blind fashion.

Figure 5-2 Photograph of one of the BIDE guns held in the fixture for remote test firing.

The advantage of the resulting NBIDE database is that it allows one to ascertain the existence and magnitude of gun type and ammo type on gun distinguishability. Because a given gun-ammo combination occurs three times across the experiment, this NBIDE data set also allows for an assessment of how distinguishable (or indistinguishable) a given gun can be across the three firings. As with all experiments, the limitations of the NBIDE database are that 1) the conclusions are limited in scope to the three gun types utilized and the three ammo types and 2) the total number of firings (144) and analyzed casings (108) is still relatively small compared to the large sizes envisioned for a national ballistics database.

6. Topography Data

Topography images were obtained for firing pin impressions, breech face impressions, and ejector marks for 70 De Kinder casings and 108 NBIDE casings. Topography images will also be obtained for the IAI bullets in future studies. Examples of the data are discussed in the following sections.

6.1 De Kinder Casings

6.1.1 De Kinder Firing Pin Impressions

Figure 4-3 shows four topography images of firing pin impressions from the De Kinder casings. All of them may be characterized as roughly spherical depressions with superimposed form deviations and roughness deviations. These images were selected to show the topographic differences that may be obtained among different guns of the same model using the same CCI brand ammunition. The image from SS 007 is smooth and has few striking form deviations. The image for SS 375 has a distinctive crater that the other images lack. The images from SS 009 and SS 535 have somewhat similar winged patterns, but for SS 009 this pattern is imposed on a concave form whereas for SS 535 the pattern is imposed on a slightly convex form.

Figure 6-1 shows a contrary case, significantly different images obtained from the same gun. All four of these images resulted from different ammunition fired by SS 139. The upper left image is rather smooth and spherical with very low form deviations. The upper right image clearly shows a crater, not seen in the upper left hand image. The lower left shows a pattern of uniaxial marks that in fact are associated with Wolf ammunition, not with marks from the gun itself. The lower right image is rather featureless except for the convex form deviation overall. These differences suggest that this gun, SS 139, would be difficult to identify from its firing pin impressions.

6.1.2 De Kinder Breech Face Impressions

The topography images of the breech face impressions were significantly larger in size and contained about eight times as many data points as those for the firing pin impressions. Two of these images are shown in Fig. 4-4. Both of these were obtained from casings fired by gun SS 117. Here again, there seem to be significant differences between these two breech face impressions for both roughness striations and form. However, this is only an initial, qualitative impression. Quantitative tests of similarity are described below.
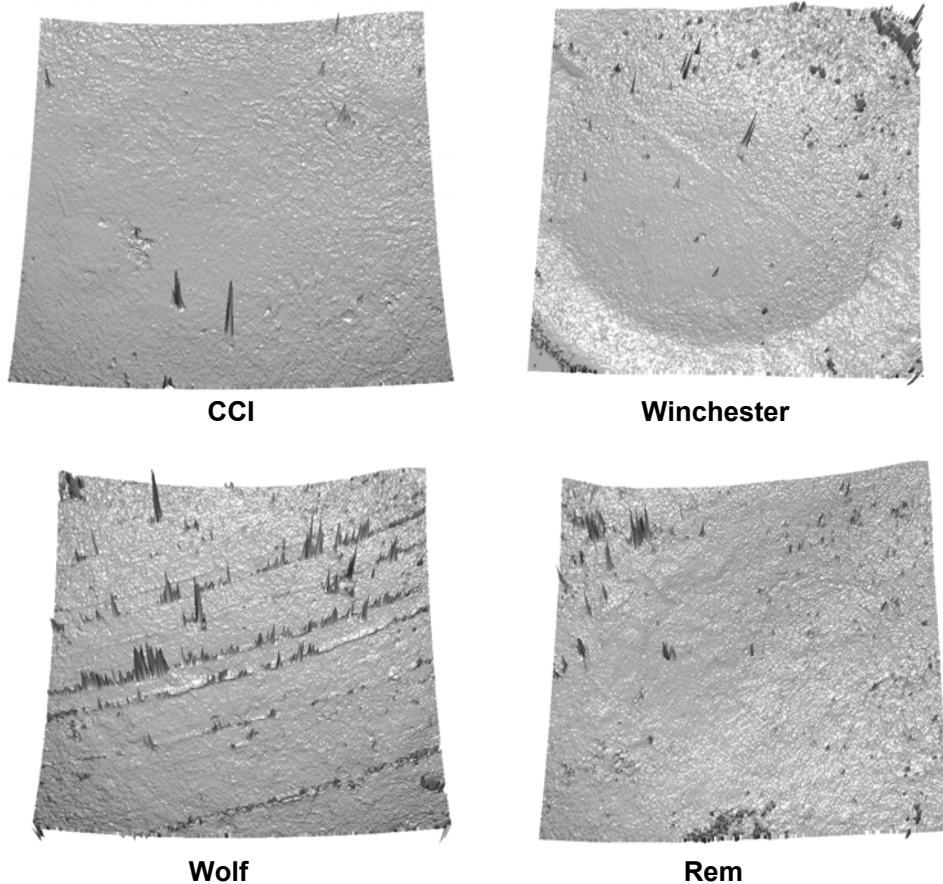
Figure 6-1. Topography images of casing firing pin impressions on different ammunition from the same gun, SS 139.

6.1.3 De Kinder Ejector Marks

Significant differences are apparent in the images of ejector marks for the De Kinder casings. Two of these are shown in Fig. 4-5 for Sig Sauer guns SS 009 and SS 117. SS 009 produces an elongated pattern on the casing, whereas SS 117 produces a clearly triangular pattern. One of the complicating factors of the ejector mark images is the presence of the manufacturer's headstamp lettering on the base of the casing, which obliterates sections of the ejector marks for many of the images. The hole on the left side of the upper image from SS 009 is part of the headstamp, which in this case does not overlap and obliterate the ejector mark.

6.2   NBIDE Casings

6.2.1 NBIDE Firing Pin Impressions

The firing pin impressions on the NBIDE casings were measured with the confocal microscope having a 0.8 mm × 0.8 mm field of view, slightly different from the field of view used for the De Kinder casings, but the other measurement parameters (objective, vertical slice interval) were the

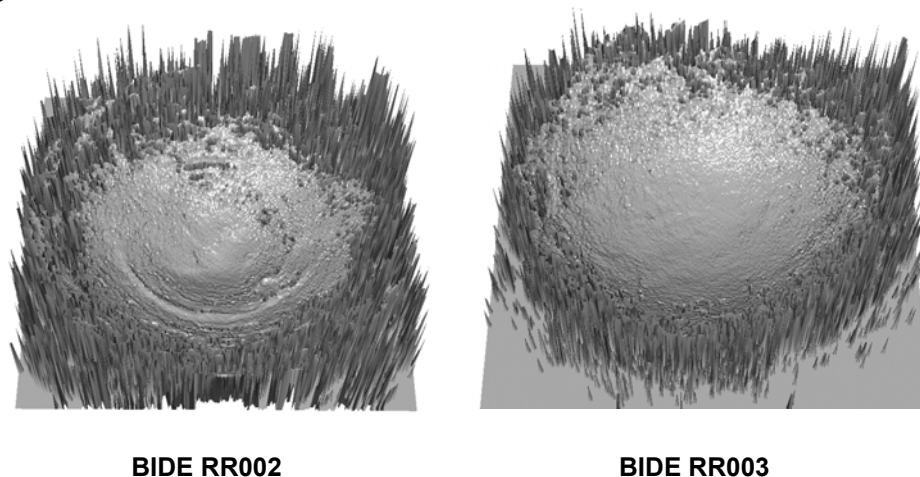same. The topography images are similar to those of the De Kinder casings. Two of these are shown in Fig. 6-2.

BIDE RR002   **BIDE RR003**

Figure 6-2. Topography images of two NBIDE firing pin impressions. The field of view is approximately 0.8 mm × 0.8 mm. The vertical range of the central bowl is approximately 60 μm.
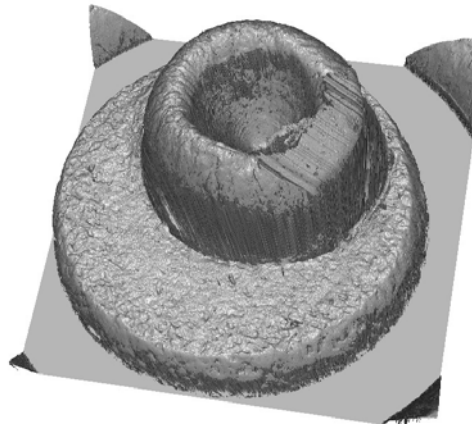
In general the patterns in the center of these images have less structure than those of the De Kinder casings, perhaps because the pistols were new. There are more outliers on the outer areas of these impressions as well. We do not yet know the reason for this. One new type of feature is the coliseum type of circular structure, shown in the image on the left, which was not apparent for any of the De Kinder casings. Also, two of the NBIDE firing pin impressions clearly had holes in the bottom, a characteristic that was easily picked up in the topography images.

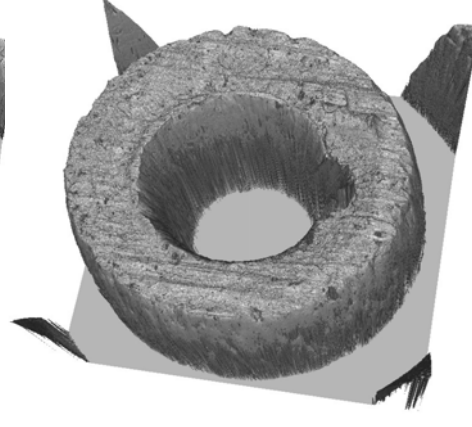## 6.2.2 NBIDE Breech Face Impressions

The NBIDE casing breech face impressions were measured with the same microscope and the same parameters as the De Kinder casing breech face impressions. A number of differences between the individual casings can be discerned. The most remarkable is the large bump near the center of the impression on some of the casings, as shown on the left in Fig. 6-3. This characteristic was present on all of the casings fired by Rugers and on several casings fired by Smith&Wessons.

6.2.3 NBIDE Ejector Marks

The NBIDE casing ejector marks have similar characteristics to the De Kinder casing ejector marks. They vary significantly, primarily because of the unpredictable position of the manufacturers' headstamps. For about half of the images, the ejector mark is partially obliterated by the presence of lettering from the headstamp. Beyond that, ejector marks seem fairly consistent and mostly fall into two categories, triangular marks and long rectangular marks. Two examples are shown in Fig. 6-4. The upper image shows a triangular ejector mark cleanly imprinted on the casing edge. The lower image shows a long rectangular mark cut in two places by the manufacturer's headstamp.

BIDE RR001   **BIDE RR008**

Figure 6-3. Topography images of two NBIDE breech face impressions, showing a large protrusion on the left image and a drag mark on the right image. The lateral field of view is approximately 4 mm × 4 mm; the vertical range is about 0.4 mm for the left image and about 0.2 mm for the right image.

EJE RR005

EJE RR020

Figure 6-4. Topography images of ejector marks of two NBIDE casings used in the designed experiment of test fires.

7. 2D Image Acquisitions and Results

The measurements and correlations of the topography images were compared with acquisitions and correlations using I-2D. As stated earlier, I-2D consists of a DAS using optical microscopy; a database for storage, retrieval, and cataloguing of the images; and a SAS with correlation software to compare the acquired images and select potential matches. The I-2D was used here to acquire images of the De Kinder casings, NBIDE casings, and IAI bullets. The accuracy of the predicted matches from the I-2D method may then be compared with the N-3D method for these materials. The image acquisitions were performed by M. Ols and R. Simmers using the I-2D at the ATF NLC.

Typical I-2D optical microscope images are contained in Figs. 7-1 and 7-2. They emphasize the variation of reflectivity, shadows, and highlights over the surface, rather than the topography directly. Therefore, 2D images and 3D images could be used in a complementary way to enhance the information acquired from casings and bullets, and hence the accuracy of ballistics databases. Fusing the 3D and 2D images into a single image containing both methods of contrast is one approach to this. Such a data fusion approach to surface imaging has recently been implemented by FTI in a bullet imaging system [27]. Similar techniques for surface microscopy have also been developed elsewhere [39].
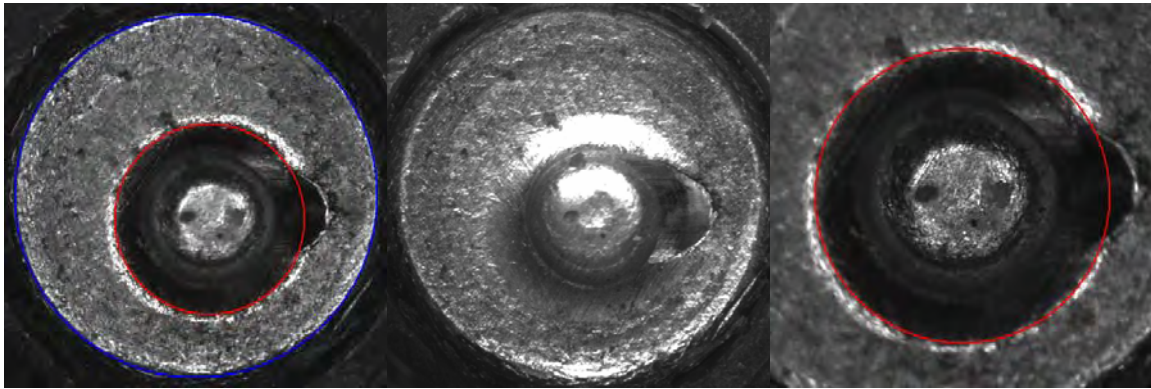

Figure 7-1. I-2D images of the primer section of a Wolf casing fired by Sig Sauer 215 in the De Kinder et al. study. Center−image of both the firing pin and breech face impressions; left−circles selected to demarcate the breech face impression; right−circle selected to demarcate the firing pin impression.

I-2D images were acquired for all three sample sets, and correlations were performed using each entry in the set as a reference. Typical results of I-2D correlations are shown in Figs. 7-3 to 7-5. Figure 7-3 shows the I-2D correlation scores of a Remington casing, fired from Sig Sauer 117 in the De Kinder collection, compared with other casings in the collection. The Top Ten correlation rankings for both the breech face impression and the firing pin impression are shown. The ejector marks (EM) were not used in this correlation analysis. As shown in the top table in Fig. 7-3, the highest correlation score matching the breech face impression of the reference exhibit was 39, obtained for the other Remington casing fired by Sig Sauer 117. As shown in the lower table, the highest correlation score matching the firing pin impression of the reference exhibit was 78, obtained for the Speer casing fired by the Sig Sauer 117. Altogether, two casings, fired by the same gun (SS 117), out of a possible six casings, appear in the Top Ten list

for breech face impressions, and four casings appear in the Top Ten list for firing pin impressions.
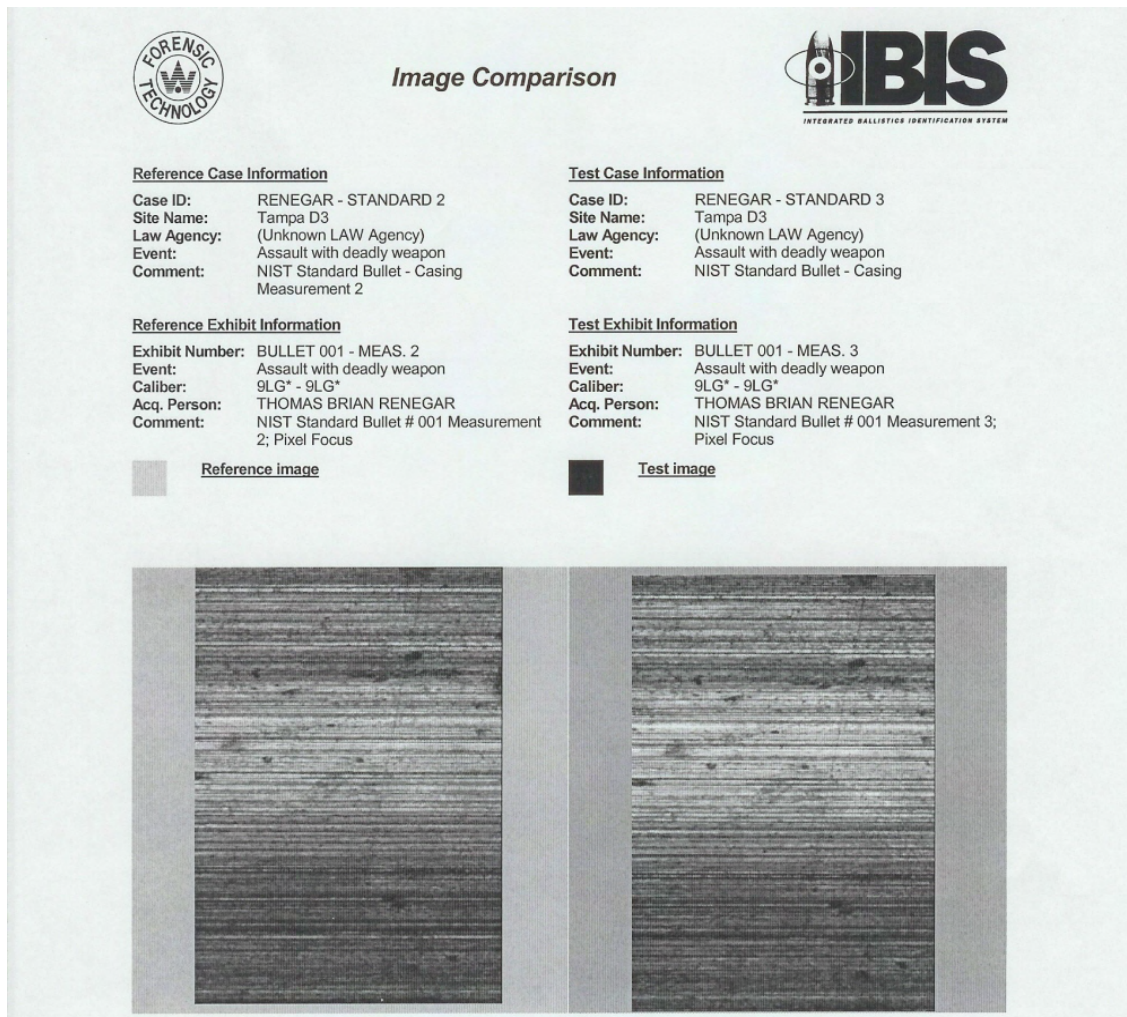

Figure 7-2. Two I-2D acquisition images for LEA 2 of NIST standard bullet #001, measured at two different times by the same operator.

A similar study was also performed for the NBIDE casings. Because of the double blind nature of the NBIDE experiment, the initial results are expressed in terms of RR numbers. An example of one of the sets of correlation results is shown in Fig. 7-4

Simmers and Ols [40] also performed I-2D correlations for the IAI bullets. An example of their results is shown in Fig. 7-5. The correlation results for bullets are more complex than for casings because six different LEAs on the reference exhibit may be compared with six LEAs on each of the other bullet exhibits in the data set. Therefore each pair of bullets are compared six times as the LEAs on the reference bullet are clocked with respect to the LEAs on the compared bullet. Three types of correlation metrics are recorded, as shown in Fig 7-5. The "Max Phase" score is the sum of the scores for all six lands when the compared bullet is rotated into the most favorable orientation with respect to the reference bullet. The "Peak Phase" score is the result for the

highest scoring pair of LEAs in that "Max Phase" orientation. The "Max LEA" is the highest scoring pair of LEAs (out of a possible 36 pairings) when the two bullets are rotated into any position. The three tables in Fig. 7-5 show the Top Ten correlation scores for each of those metrics.

IBIS Correlation Results

Reference Case Information

Case ID: NIST RANDOM CASINGS
Site Name: MD-WAATF1
Law Agency:
Event: Demo, Test and QA
Comment:

Reference Exhibit Information

Exhibit Number: EV18-REM (W/SS117GRP
Event: Demo, Test and QA
Caliber: 9LUG
Acq. Person: REBEKAH GLASS
Comment: THIS IS A NIST TEST CARTRIDGE

Sample Size 69
Tests ordered by *Breech Face*

Rank	Case Id	Exhibit Number	Site Name	Breech Face	Firing Pin	EM or RF
1	NIST RANDOM CASINGS	SS117-REM (ALONE)	MD-WAATF1	39	53	0
2	NIST RANDOM CASINGS	SS535-SPEER	MD-WAATF1	28	32	0
3	NIST RANDOM CASINGS	SS314-SPEER	MD-WAATF1	26	29	0
4	NIST RANDOM CASINGS	SS213-REM (ALONE)	MD-WAATF1	25	13	0
5	NIST RANDOM CASINGS	SS314-FED	MD-WAATF1	25	70	0
6	NIST RANDOM CASINGS	SS139-REM (FROM GRP)	MD-WAATF1	23	9	0
7	NIST RANDOM CASINGS	SS317-CCI	MD-WAATF1	23	42	0
8	NIST RANDOM CASINGS	SS009-REM (FROM GRP)	MD-WAATF1	22	12	0
9	NIST RANDOM CASINGS	EV41-WIN (W/SS215GRP	MD-WAATF1	22	15	0
10	NIST RANDOM CASINGS	SS139-CCI	MD-WAATF1	21	15	0

Tests ordered by *Firing Pin*

Rank	Case Id	Exhibit Number	Site Name	Breech Face	Firing Pin	EM or RF
1	NIST RANDOM CASINGS	SS117-SPEER	MD-WAATF1	13	78	0
2	NIST RANDOM CASINGS	SS314-FED	MD-WAATF1	25	70	0
3	NIST RANDOM CASINGS	SS117-WOLF	MD-WAATF1	2	65	0
4	NIST RANDOM CASINGS	SS117-REM (ALONE)	MD-WAATF1	39	53	0
5	NIST RANDOM CASINGS	SS213-REM (FROM GRP)	MD-WAATF1	19	44	0
6	NIST RANDOM CASINGS	SS430-FED	MD-WAATF1	6	43	0
7	NIST RANDOM CASINGS	SS317-CCI	MD-WAATF1	23	42	0
8	NIST RANDOM CASINGS	SS213-WIN	MD-WAATF1	4	39	0
9	NIST RANDOM CASINGS	SS430-WOLF	MD-WAATF1	2	38	0
10	NIST RANDOM CASINGS	SS535-CCI	MD-WAATF1	17	37	0

Figure 7-3. I-2D Top Ten correlation scores for a Remington casing fired from Sig Sauer (SS) 117. Two other casings fired from SS 117 appear on the Top Ten list for breech face impressions and four appear in the Top Ten list for firing pin impressions. Casing SS117-CCI is labeled here as SS317-CCI.

IBIS Correlation Results

Reference Case Information

Case ID: NIJ/BIDE RE-RANDM CC
Site Name: MD-WAATF1
Law Agency:
Event: Demo, Test and QA
Comment:

Reference Exhibit Information

Exhibit Number: RR# 001
Event: Demo, Test and QA
Caliber: 9LUG
Acq. Person: REBEKAH SIMMERS
Comment:

Sample Size 216
Tests ordered by *Breech Face*

Rank	Case Id	Exhibit Number	Site Name	Breech Face	Firing Pin	EM or RF
1	NIJ/BIDE RE-RANDM CC	RR# 067	MD-WAATF1	191	65	0
2	NIJ/BIDE RE-RANDM CC	RR# 082	MD-WAATF1	165	72	0
3	NIJ/BIDE RE-RANDM CC	RR# 045	MD-WAATF1	131	76	0
4	NIJ/BIDE RE-RANDM CC	RR# 134	MD-WAATF1	88	56	0
5	NIJ/BIDE RE-RANDM CC	RR# 033	MD-WAATF1	84	72	0
6	NIJ/BIDE RE-RANDM CC	RR# 073	MD-WAATF1	69	54	0
7	NIJ/BIDE RE-RANDM CC	RR# 053	MD-WAATF1	57	26	0
8	NIJ/BIDE RE-RANDM CC	RR# 095	MD-WAATF1	57	36	0
9	NIJ/BIDE RE-RANDM CC	RR# 140	MD-WAATF1	54	63	0
10	NIJ/BIDE RE-RANDM CC	RR# 123	MD-WAATF1	52	42	0

Tests ordered by *Firing Pin*

Rank	Case Id	Exhibit Number	Site Name	Breech Face	Firing Pin	EM or RF
1	NIJ/BIDE RE-RANDM CC	RR# 014	MD-WAATF1	34	78	0
2	NIJ/BIDE RE-RANDM CC	RR# 045	MD-WAATF1	131	76	0
3	NIJ/BIDE RE-RANDM CC	RR# 101	MD-WAATF1	7	75	0
4	NIJ/BIDE RE-RANDM CC	RR# 033	MD-WAATF1	84	72	0
5	NIJ/BIDE RE-RANDM CC	RR# 082	MD-WAATF1	165	72	0
6	NIJ/BIDE RE-RANDM CC	RR# 026	MD-WAATF1	11	69	0
7	NIJ/BIDE RE-RANDM CC	RR# 067	MD-WAATF1	191	65	0
8	NIJ/BIDE RE-RANDM CC	RR# 022	MD-WAATF1	44	65	0
9	NIJ/BIDE RE-RANDM CC	RR# 133	MD-WAATF1	11	64	0
10	NIJ/BIDE RE-RANDM CC	RR# 140	MD-WAATF1	54	63	0

Figure 7-4. I-2D correlation scores for NBIDE casing RR #001 showing the Top Ten correlation scores for the breech face impression and the firing pin impression.

Reference Case Information

Case ID: NIJ BULLETS
Site Name: MD-WAATF1
Law Agency:
Event: Demo, Test and QA
Comment:

Reference Exhibit Information

Exhibit Number: E11-11
Event: Demo, Test and QA
Caliber: 9LUG
Acq. Person: REBEKAH SIMMERS
Comment:

Sample Size 131
Tests ordered by **Max LEA/GEA**

Rank	Case Id	Exhibit Number	Site Name	Max LEA/GEA	Peak Phase	Max Phase
1	NIJ BULLETS	E11-12	MD-WAATF1	693	693	2331
2	NIJ BULLETS	E11-1	MD-WAATF1	521	521	1639
3	NIJ BULLETS	E11-2	MD-WAATF1	432	432	1546
4	NIJ BULLETS	E11-23	MD-WAATF1	389	389	1029
5	NIJ BULLETS	E1-10	MD-WAATF1	256	195	844
6	NIJ BULLETS	E1-32	MD-WAATF1	243	166	825
7	NIJ BULLETS	E11-21	MD-WAATF1	240	240	677
8	NIJ BULLETS	E1-2	MD-WAATF1	235	235	787
9	NIJ BULLETS	I2-11	MD-WAATF1	228	228	891
10	NIJ BULLETS	I1-212	MD-WAATF1	224	224	789

Tests ordered by **Peak Phase**

Rank	Case Id	Exhibit Number	Site Name	Max LEA/GEA	Peak Phase	Max Phase
1	NIJ BULLETS	E11-12	MD-WAATF1	693	693	2331
2	NIJ BULLETS	E11-1	MD-WAATF1	521	521	1639
3	NIJ BULLETS	E11-2	MD-WAATF1	432	432	1546
4	NIJ BULLETS	E11-23	MD-WAATF1	389	389	1029
5	NIJ BULLETS	E11-21	MD-WAATF1	240	240	677
6	NIJ BULLETS	E1-2	MD-WAATF1	235	235	787
7	NIJ BULLETS	I2-11	MD-WAATF1	228	228	891
8	NIJ BULLETS	I1-212	MD-WAATF1	224	224	789
9	NIJ BULLETS	I1-211	MD-WAATF1	215	215	821
10	NIJ BULLETS	E1-31	MD-WAATF1	211	211	942

Tests ordered by **Max Phase**

Rank	Case Id	Exhibit Number	Site Name	Max LEA/GEA	Peak Phase	Max Phase
1	NIJ BULLETS	E11-12	MD-WAATF1	693	693	2331
2	NIJ BULLETS	E11-1	MD-WAATF1	521	521	1639
3	NIJ BULLETS	E11-2	MD-WAATF1	432	432	1546
4	NIJ BULLETS	E11-23	MD-WAATF1	389	389	1029
5	NIJ BULLETS	E1-31	MD-WAATF1	211	211	942
6	NIJ BULLETS	I2-11	MD-WAATF1	228	228	891
7	NIJ BULLETS	E2-12	MD-WAATF1	172	172	868
8	NIJ BULLETS	E1-10	MD-WAATF1	256	195	844
9	NIJ BULLETS	I11-1	MD-WAATF1	197	197	838
10	NIJ BULLETS	E2-11	MD-WAATF1	173	170	825

Figure 7-5. I-2D correlation scores for IAI bullet E11-11 as the reference, showing Top Ten rankings by three different metrics: Max Phase, Peak Phase, and Max LEA.

8. Data Processing for Topography Measurements

The topography images are compared in pairs in order to determine quantitatively how well one image matches another. A large number of these pair-wise correlations is required to characterize a modest size data set. For example, for the 70 De Kinder casings, 70 x 69 (= 4830) pair-wise matches are required to correlate all the casings with all the others in the set.

In order to perform the correlations, the topography images undergo several processing steps. These steps are:

- data trimming and decimation,
- removal of dropout data points,
- removal of outlier points,
- filtering, and
- registration and cross-correlation.

These steps are discussed in Secs. 8.1-8.5 and are depicted in Fig. 8-1. Here we show the topographies in terms of top-down, color-coded images instead of the monochromatic isometric views shown before. Briefly, for each topography image, after thinning the data and identifying and interpolating the dropouts and outliers, a long-wavelength cutoff filter is used to remove the low-spatial frequency structures, including the curvature and waviness, and a short-wavelength cutoff filter is used to reduce instrument noise. Then a pair of images is registered and correlated iteratively to find the optimum overlap position between them and a quantitative metric of the similarity between the two images.

In order to accomplish these calculations, two codes with slightly different algorithms and parameters were developed in parallel. One was developed at NIST and was used for calculating exploratory results and for developing the illustrative graphs shown in this section. The other code was developed at IAI and was refined from a code previously developed there. The most significant difference between these codes is that the code developed by IAI has the ability to identify and ignore dropout and outlier points during correlation as well as during the earlier processing steps. We call these codes the NIST code and the IAI code. The IAI code was used for generating the bulk of the results shown in Secs. 9 and 10.

8.1 Data Trimming and Decimation

Pair-wise cross-correlations of large numbers of large data sets can take a long time. For the breech-face impressions of the NBIDE data sets, the cross-correlation analysis running on three 3 GHz PCs took approximately two days. In order to minimize this running time, a number of data points may be removed from the images to reduce the size of the data sets being analyzed. Two data removal approaches are used, trimming of unneeded areas and decimation (or thinning) of points in the areas that are used. These procedures are summarized below.
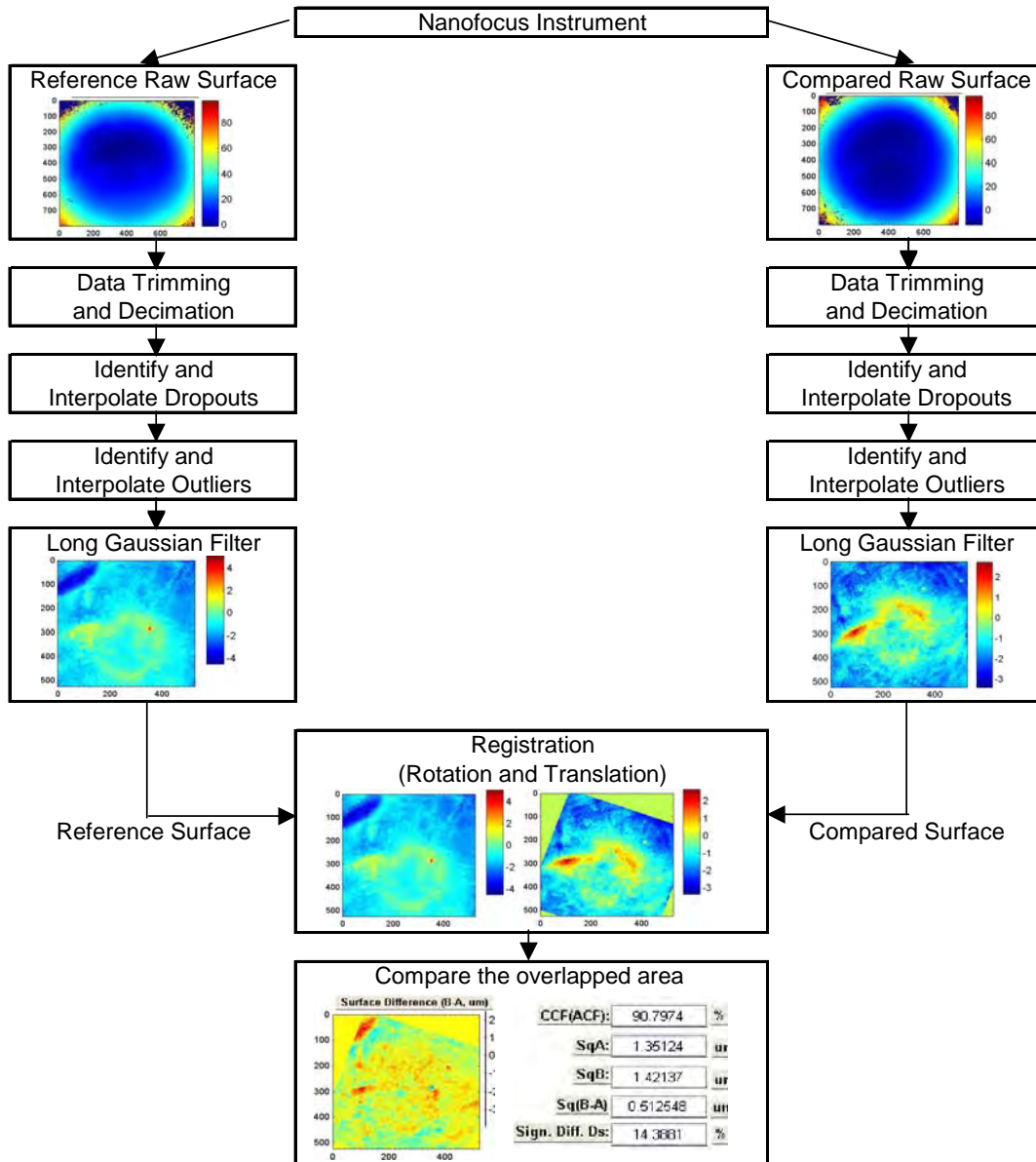
Figure 8-1. Generic data flow of NIST and IAI topography comparison programs.

8.1.1 Trimming

When the breech face images are acquired, they contain extraneous information, which is not useful in the correlation process. Trimming is applied systematically to the breech face impression data to eliminate the extraneous areas. This includes any information captured outside the primer area, and inside the firing pin impression area.

To trim these unnecessary data points in the breech face impression images, the image is loaded into a software trimming program. The user selects and adjusts a circle that inscribes the inner rim of the breech face image. Any data points inside this circle will be excluded (Fig. 8-2). The user then selects and adjusts a circle that circumscribes the outer rim of the breech face image.

Any data points outside of this circle will also be excluded. Once the user has finished these operations, the program removes the unnecessary data points from the 3D topography and saves the data to a new electronic file. This file can then be used for later processing and correlations. Because the trimming is done manually, it is a subjective process, which relies on the user's expertise to trim the edges of the data consistently. However, results shown in Sec. 8.6 indicate a high correlation between similar data sets trimmed by a skilled person.
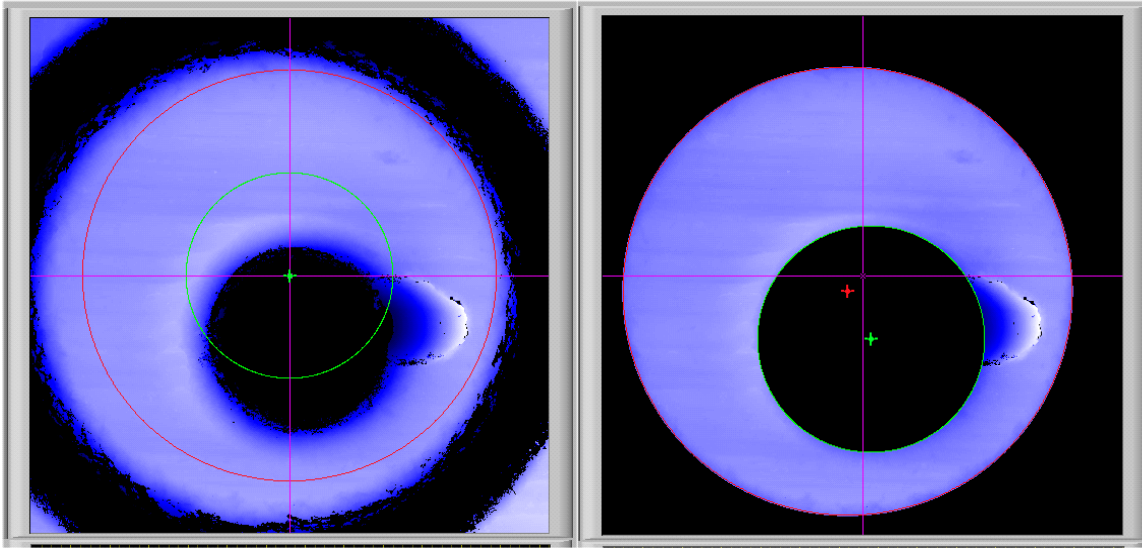

Figure 8-2. Breech face data before trimming (left) and after trimming (right).

8.1.2 Decimation

In order to decrease the computation requirements of the system, it is often useful to work with data sets of lateral resolution lower than that available in the raw data set. Decimating the amount of data to be used for comparison saves both storage and computational requirements. For this reason, the processing of the breech face impressions includes a decimation component wherein the number of data points in the breech face images is reduced significantly. The images are decimated from approximately 2600×2600 data points to approximately 650×650 points. Data decimation is not a necessary step, but often a convenient one. Decimation was not performed on the firing pin impressions.

8.2   Handling Dropouts and Outliers

The purpose of the data pre-processing module is to "clean" the data from unreliable data points, which include dropouts (points that the sensor was not able to acquire) and outliers (points that the sensor managed to acquire, but which are inaccurate or noisy). For this reason, the data pre-processing module consists of a four step process: 1) identification of dropouts, 2) identification of outliers, 3) recording of both types of unreliable points, and 4) interpolation of the unreliable data points. These steps are described below.

1) Identification of dropouts: Most 3D imaging systems provide the user with a "level of confidence" value associated with each data point taken (for optical systems, the level of

confidence usually corresponds to the percentage of light reflected by the target). If the level of confidence is too low, the point is deemed "unreliable," or in other words a dropout.

2) Identification of outliers: As opposed to dropouts, "outliers" are data points inaccurately measured but not reported as inaccurate to the user by the acquisition hardware. For this reason, they are much more challenging to identify. We use two approaches to identify such outliers. The first approach is by estimating the local slope between a point and its neighbors. If the slope is above a certain threshold, the adjacent points will be identified as outliers. The second approach is by considering the statistical distribution of the data. If a particular point is excessively far from the local median in terms of standard deviations, it is identified as an outlier.

3) Recording of unreliable points: Once all dropped and outlier points are identified, a "mask" is created to store this information for use during the comparison stages so that the unreliable points can be excluded from the comparison. In the current software implementation, the mask is an array of the same dimensions as the data, and its entries are "1' for those points deemed to be reliable, and "0" for those points identified as dropouts or outliers. The left side of Fig. 8-3 shows an example of raw firing pin data, where the third dimension (z-axis) is color coded. The right side of Figure 8-3. shows the corresponding mask, where the points identified as dropouts or outliers have been colored blue, while the points deemed "reliable" have been colored red.
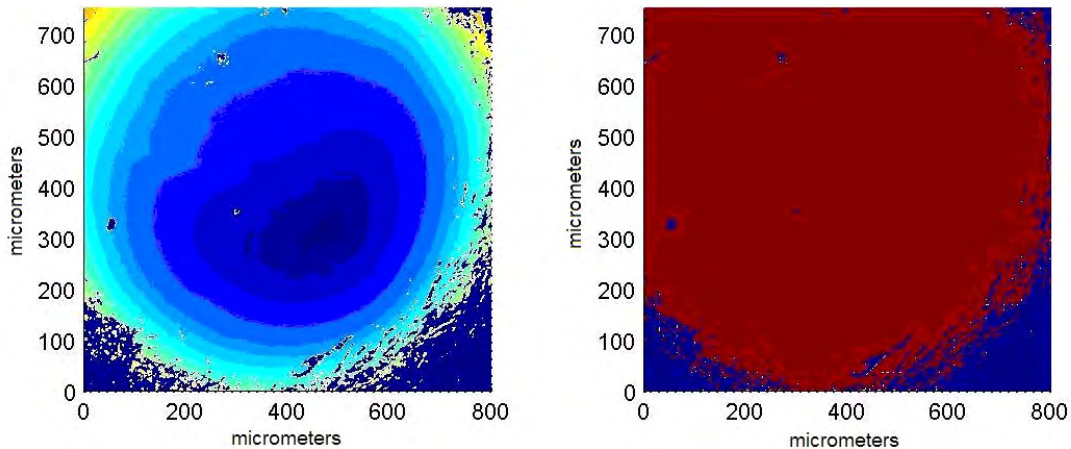

Figure 8-3. Example of raw firing pin topography data (left) and masked data (right) after dropout and outlier identification.

4) Interpolation: For display purposes and in order to accommodate the digital filtering performed at a subsequent stage, the values of dropouts and outliers are replaced by interpolated values based on the neighboring points.

At this stage the IAI code allows for additional data trimming by implementing, at the user's discretion, an algorithm that selects the most promising portion of the available data. This process takes place as follows. Having automatically identified the unreliable points (both dropouts and outliers), the system identifies a section of pre-defined dimensions within the acquired data that: a) shows the least number of unreliable points, and b) satisfies some desirable

constraint, such as being closest to the center of the region of interest, or being to the left of the region of interest, etc. The region thus selected will be isolated and used for the remainder of the process. This step is taken as an aid to the user because often the boundaries of the region of interest include a relatively large number of unreliable points.

8.3   Filtering

The main purpose of the filtering module is to separate those features that are unique to the marking under consideration (individual characteristics) from those that are common to all markings of the same type (class characteristics). Consider, for example, the case of a batch of firing pins manufactured for a given gun model. As all these pins are meant to be used by the same gun model, they are manufactured to the same specifications. The overall geometric shape of these components is therefore very similar. On the other hand, as no two manufactured parts are ever identical, there are microscopic variations unique to each firing pin. It is these unique features that may be of use to identify those cartridge cases struck by the same firing pin. The challenge associated with the development of an effective automated firearm identification system is to separate class characteristics from individual characteristics, and to treat them in the appropriate manner. Filtering helps to isolate those individual characteristics.

We should add that although an assessment of individuality can only be made by considering individual characteristics, class characteristics can also have significant value in forensic identification. For example, a pair of impressions, for which the class characteristics do not match, could not have been created by the same weapon. Therefore, class characteristics can be a powerful tool to exclude a match and hence can be taken as the first step of ballistics identification. Class characteristics were not considered in this study.

In order to emphasize the individual characteristics of the casings, which are mostly represented by the roughness structures, the longest and shortest spatial wavelengths in the topography images are attenuated by applying high-pass and low-pass digital filters, respectively. The longest wavelengths are associated with waviness and shape deviations, which tend to dominate over roughness in topographic comparisons but which may not be as closely related to individual gun characteristics as roughness. The shortest spatial wavelengths are usually associated with system noise.

In the NIST code, the filter algorithm is a close approximation of a Gaussian digital filter [30] developed for high speed computation [41]. The short wavelength cutoff is 2.5 μm and the long wavelength cutoff is 0.25 mm for all topography images, values that coincide with recommended values in standards [30]. In the IAI code, the filter is essentially a moving average series with a finite number of terms tapered by a Gaussian weighting function. The effective cutoff is limited either by the number of terms in the series or the sharpness of the taper. The net effect is that the short wavelength cutoff is about 4 μm and the long wavelength cutoff is about 40 μm for the firing pin impression data, and the short wavelength cutoff is about 37 μm and the long wavelength cutoff is about 150 μm for the breech face impression data.
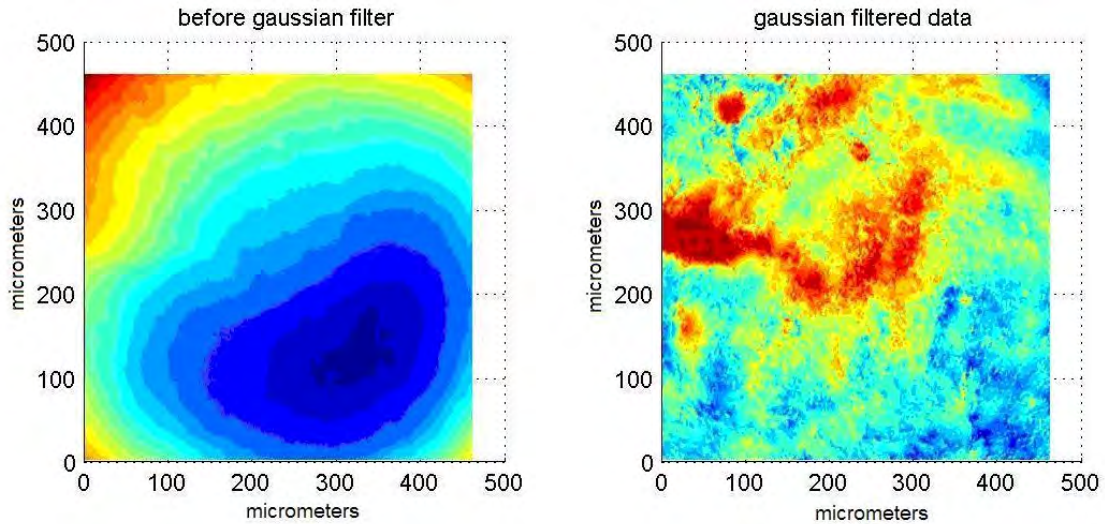
Figure 8-4. Example of a firing pin impression before (left) and after (right) applying filters with long wavelength cutoff λc = 40 μm and short wavelength cutoff λs = 4 μm.

Figure 8-4 shows an example of a firing pin impression before (left) and after (right) the application of a high-pass Gaussian filter with long wavelength cutoff of 40 μm. A low-pass Gaussian filter with cutoff of 4 μm is also applied in order to produce a mild smoothing effect on the image. The image to the left has a non-distinct shape, which may be similar for many other firing pins manufactured for a weapon of the same model. This image is dominated by class characteristics. The firing pin in this example corresponds to a Sig Sauer P226 gun. The image to the right, on the other hand, captures unique features, which in this case did transfer between the firing pin and the primer of the cartridge under consideration. This image is dominated by individual characteristics.
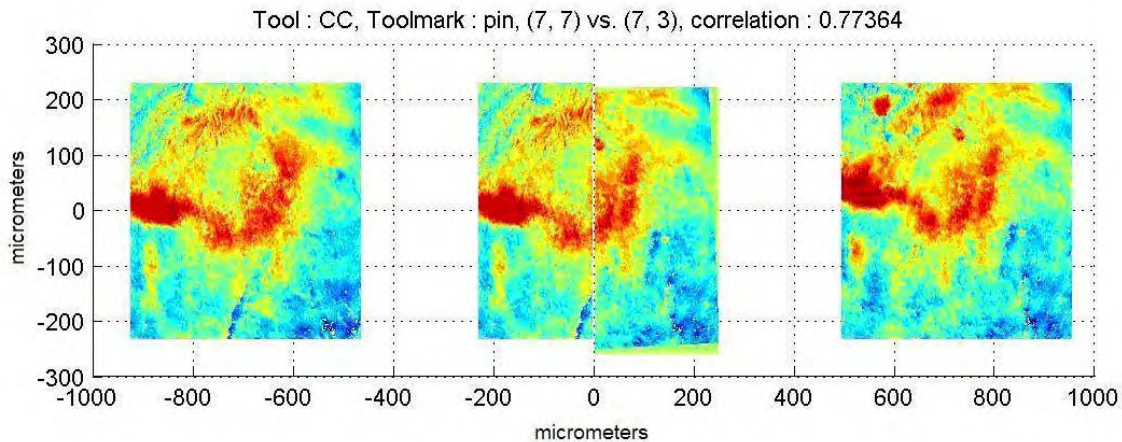

Figure 8-5. Matching unique features found on two different firing pin impressions. Left–firing pin signature of a Remington cartridge case (No. 7) fired by the Sig Sauer 007 gun; right–firing pin signature of a Remington cartridge case (No. 3) fired by the same gun; middle–the two signatures aligned in their optimal correlation position and orientation.

52

Figure 8-5 shows an example of the alignment of the individual markings (often referred to as a signature) of the firing pin impression shown in Fig. 8-4 with those of a firing pin impression on another cartridge casing fired by the same firearm. As can be seen, the degree of similarity in this case is quite high.

8.4   Registration and Correlation

To find the maximum correlation position, a registration program is used to compare two surface topographies while shifting along the *x*- and *y*-directions and rotating around the *z*-axis. For this process the two software codes use different procedures, which seem to function about equally well with the important exception that the IAI code is able to ignore dropout and outlier points, making it advantageous in the case of breech face impressions where the large area of the firing pin impression needs to be excluded from the comparison. For the comparison of firing pin impressions and breech face impressions, both codes use a correlation metric based on the areal autocorrelation and cross-correlation function. Details of the procedures are described below.

8.4.1 Details of the NIST Code

The NIST code uses an affine and rigid rotation and translation image registration method, as described by Bergen, et al [42]. The multi-scale registration scheme is described further by Heeger [43] and implemented by Heeger as a suite of MATLAB functions [44], which were integrated into our 3D-topography comparison software codes.

The areal cross-correlation function, *ACCF* is given by

$$ACCF(A,B,\tau_x,\tau_y) = \frac{ACCV(A,B,\tau_x,\tau_y)}{Sq(A)Sq(B)},\tag{1}$$

where *ACCV* represents the areal cross-covariance function:

$$ACCV(A,B,\tau_x,\tau_y) = \lim_{L_x,L_y \to \infty} \left(\frac{1}{L_xL_y} \int_{-L_y/2}^{L_y/2} \int_{-L_x/2}^{L_x/2} Z_A(x,y)\,Z_B(x+\tau_x,y+\tau_y)\,dx\,dy \right),\tag{2}$$

where $Z(x,y)$ is the measured topography, and τ_x and τ_y are the shift distances in the *x*- and *y*-directions, respectively. For a finite digitized profile, the analytical formula is approximated in the NIST code by its biased estimator [45]:

$$ACCV(A,B,\tau_x,\tau_y) \approx \frac{1}{MN} \sum_{k=1}^{M-k'} \sum_{j=1}^{N-j'} Z_A(j,k)\,Z_B(j+j',k+k'),\tag{3}$$

where $\tau_x = j'd_x$, $\tau_y = k'd_y$, and d_x and d_y are the sampling intervals of the data points in the *x*- and *y*-directions. The areal root-mean-square roughness *Sq* of topography A and B is given by [30]

$$Sq = \left[\left[\frac{1}{L_x L_y} \int_{-L_x/2}^{L_x/2} \int_{-L_y/2}^{L_y/2} Z^2(x,y)\, dxdy\right]^{1/2}\right] \approx \left[\left[\frac{1}{MN} \sum_{k=1}^{M} \sum_{j=1}^{N} Z^2(j,k)\right]^{1/2}\right], \qquad (4)$$

where $Z(x,y)$ is sampled by a set of M points over the evaluation length L_x in the x direction and N points over the evaluation length L_y in the y direction.

From the cross-correlation function we calculate a parameter called $ACCF_{max}$, which is the maximum value of the cross correlation function when the two topographies are shifted to provide the best match. When two casing surface topographies are compared, one is taken as the reference topography A and the other is the compared topography B. In that case:

- If the two topographies are exactly the same (A = B), their *ACCF* function achieves the maximum value (1.00 or 100 %) when the shift distance and rotation angle are zero.
- If two compared topographies A and B have essentially the same pattern but small differences in the topography details, then A ≈ B. For example, when two casings are fired from the same gun, their topographies may have strong correlation. When these two topographies are shifted, their *ACCF* function will have a clear maximum value but not as large as 1.00, because there are some differences between the two topographies.
- If two compared topographies A and B are not correlated, for example, when two casings are fired from different guns, then A ≠ B. Their *ACCF* function will have only random variations without a significant correlation peak.

8.4.2 Details of the IAI Code

The variety of approaches that can be followed to develop a similarity metric for firing pin impressions or breech face impressions (or impressed tool marks in general) is virtually unlimited. In making a selection, it is important to remember that the registration problem will in all likelihood become an optimization problem. From this perspective, practical considerations play an important role in the selection of the similarity metric. For example, the amount of data required to describe impressed tool marks is usually quite significant, so it is important to choose a similarity metric that lends itself to fast computation even for large amounts of data. As an example, it is advantageous to make use of frequency domain methods to speed up the computation of the similarity metric. In addition, it is advantageous to select a similarity metric amenable to multi-resolution computation. The benefit of this approach is that low-resolution versions of the data sets can be used to estimate the neighborhood of the solution to the optimization problem. This neighborhood can be more thoroughly evaluated using sequentially higher resolution versions of the data sets until an optimal solution is obtained. Both these approaches are used in the IAI code.

In the current implementation of the IAI code, a 2D extension of the statistical correlation coefficient is used. However, as mentioned earlier, the code is designed to ignore both dropouts and outliers. Given a data set Z_A, we create a mask M_A defined as follows (see Section 8.2):

$$M_A(i,j) = \begin{cases} 1, \text{ if } (i,j) \text{ is not an outlier or dropout,} \\ 0, \text{ if } (i,j) \text{ is an outlier or dropout.} \end{cases} \qquad (5)$$

Further, we define $Z_A(i,j,\theta)$ and $M_A(i,j,\theta)$ as the rotated versions of Z_A and M_A respectively. Finally, we define the index set $I_A(\theta)$ as follows:

$$I_A(\theta) = \{(i,j) \mid M_A(i,j,\theta) = 1\}. \tag{6}$$

In other words, for a given rotation angle θ, $I_A(\theta)$ corresponds to the set of all the points on the data set $Z_A(i,j,\theta)$ that are considered valid (i.e. which are not dropouts nor outliers). We can now define the similarity metric:

$$ACCF(A,M_A,B,M_B,\tau_x,\tau_y,\theta) = \frac{\displaystyle\sum_{(i,j)\in I_M} Z_A(i,j)Z_B(i-\tau_x,j-\tau_y,\theta)}{\left[\displaystyle\sum_{(i,j)\in I_M}(Z_A(i,j))^2\right]^{1/2}\left[\displaystyle\sum_{(i,j)\in I_M}(Z_B(i,j,\theta))^2\right]^{1/2}}, \tag{7}$$

where $I_{M\theta} = I_A \cap I_B(\theta)$. In other words, all summations are computed only for those points that are valid (i.e., excluding dropout or outlier points) for both Z_A and $Z_B(\theta)$. The ability to ignore unreliable points during correlation is crucial when comparing breech face impressions because of the large number of invalid points inside the firing pin impression and outside the primer boundary. Further, this function correlates only with respect to overlapping points. The set of overlapping points will vary due to the rotation of the data set Z_B, even if no dropouts or outliers are present.

It is important to mention that the similarity metric $ACCF$ is implemented using frequency domain techniques in order to increase the speed of the calculation.

The optimization of the right hand slide of Eqn. (7) with respect to τ_x, τ_y, and θ is performed in a sequential manner. The first step is to estimate the neighborhood of the optimal value of θ. Given a pair of data sets, we denote the optimal correlation value for a given relative orientation as:

$$\text{corr}(\theta_i) = \max_{(\tau_x,\tau_y)\in I_\Delta} ACCF(A,M_A,B,M_B,\tau_x,\tau_y,\theta_i), \tag{8}$$

where the index set I_Δ corresponds to the maximum lateral translations allowed, and is determined *a priori* in IAI's implementation as a percentage of the dimensional sizes of the data. The estimate of the optimal τ_x and τ_y is computed using a frequency domain approach.

As part of the first step of the optimization approach, low resolution versions of Z_A and Z_B are used to solve Eqn. (8) at a discrete number of relative orientations defined by the set $I_\theta = \{-180°, -175°, ..., 175°\}$. The results of this evaluation are used to identify a neighborhood of the optimal θ.

Once the neighborhood of optimal θ has been identified, the process is repeated within this neighborhood using full resolution versions of Z_A and Z_B. The peak correlation value, to be denoted as $ACCF_{max}$, is defined as:

$$ACCF_{max} = \max(\mathrm{corr}(\theta_i)), \forall\, \theta_i = I_{\theta 2}, \tag{9}$$

where the set $I_{\theta 2}$ corresponds to a set of angles within the neighborhood identified in the first step, and their resolution is 1 degree.

We now discuss an example of the registration process used to find the best matching relative position between a pair of images. Figure 8-6 shows the topographies obtained by applying the pre-processing and filtering algorithms to a pair of breech face impressions found on two different cartridge cases fired by the same firearm. The striated marks seen on these breech face impressions seem to be unique to the gun that fired these cartridge cases.

The results of comparing these breech face signatures can be seen in Fig. 8-7. The top image in Fig. 8-6 and the left image in Fig. 8-7 show the breech face for casing #3 fired by Sig Sauer #30 in the original De Kinder collection. The bottom image in Fig. 8-6 and the right image in Fig. 8-7 show the breech face for casing #7 fired by the same gun. The middle image in Fig. 8-7 is a "split screen" image derived from the other two. The right side of this image is the right half of the optimally aligned (rotated and translated) breech face signature of cartridge #7. The two halves match very well, which indicates a high degree of similarity between the two images.

Figure 8-8 shows the corresponding plot of $\mathrm{corr}(\theta_i)$. As seen in this image, the $\mathrm{corr}(\theta_i)$ plot peaks at a relative orientation close to zero degrees, achieving a correlation value $ACCF_{max}$ of 0.65.
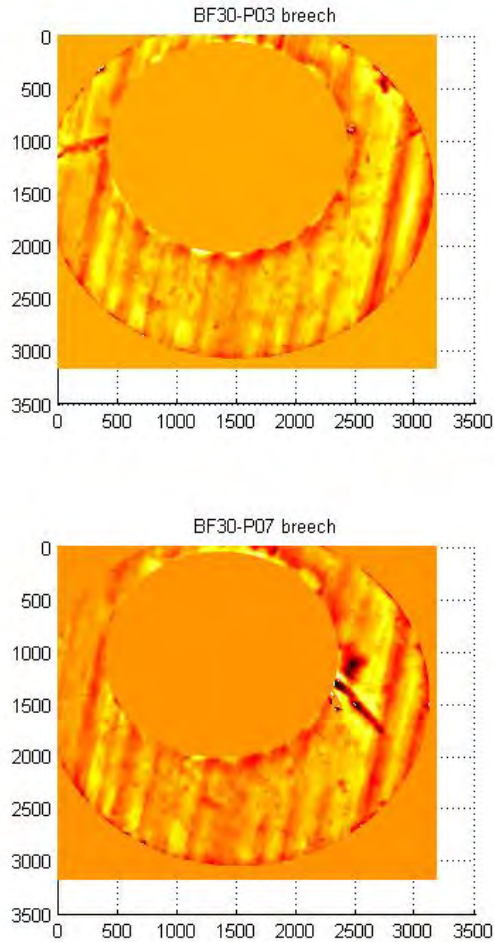
Figure 8-6. Topography images of breech face impressions of two Remington casings fired by the same Sig Sauer #30 gun, one of the original guns studied by De Kinder et al.[7]. The scales represents the pixel positions in the images. The actual diameter is approximately 4 mm.
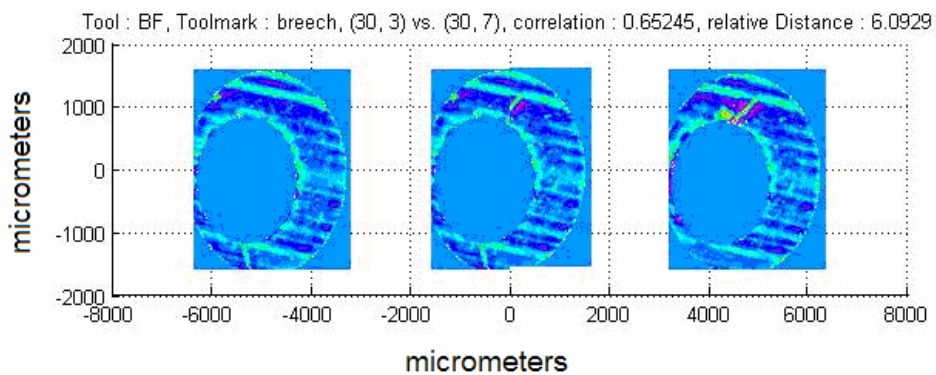

Figure 8-7. Results of alignment and similarity computation for the pair of matching breech face impressions shown in Fig. 8-6. Left–casing 3, right–casing 7, middle–a split screen image comparing the two images. Note: these images have been rotated by about 90° from those of Fig. 8-6.
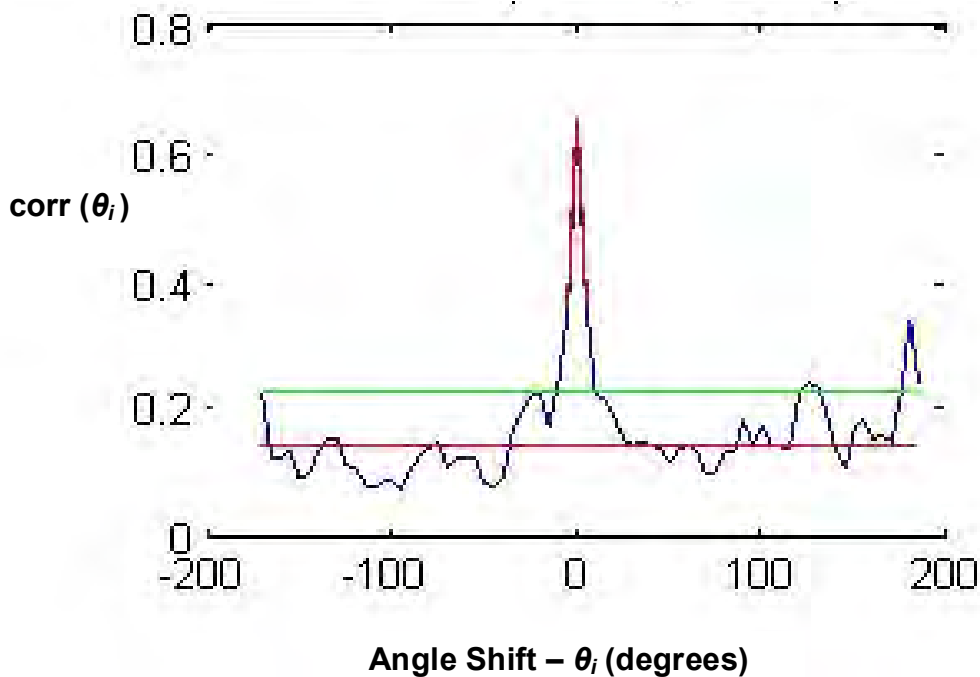
57

Figure 8-8. corr (θ_i) vs. alignment angle shift θ_i for a pair of topography images, showing a clear peak near 0°.

8.5 Consideration of Additional Metrics

In order to improve the ability of the code to produce accurate matches of casings fired by the same gun, we consider other metrics besides $ACCF_{max}$ for assessing similarity. These are described briefly below.

8.5.1 Relative Distance

The correlation value of 0.65, shown in Fig. 8-8, seems to be rather significant because it is "much higher" than the correlation values for all other rotation angles. In order to quantify the degree of significance of this peak, we introduce an additional metric, developed at IAI, which we refer to as the "relative distance." We define the relative distance as follows:

$$\text{relDist}(ACCF_{max}) = \frac{ACCF_{max} - \text{median}(\text{corr}(\theta_i))}{\text{std dev}(\text{corr}(\theta_i))}, \forall \, \theta_i = I_\theta. \tag{10}$$

Having computed the correlation between a pair of cartridge case breech face signatures for all relative orientations of interest, we define the relative distance as the difference (relative to one standard deviation) between the maximum correlation value and the median correlation value for all relative orientation angles under considerations. Shown in Fig. 8-9 is a histogram of the correlation values plotted in Fig. 8-8. From these data, the relative distance is computed to be 6.09, a very high value indicating that the peak correlation value is significantly greater than the median of correlation values obtained for all other rotational angles under consideration.
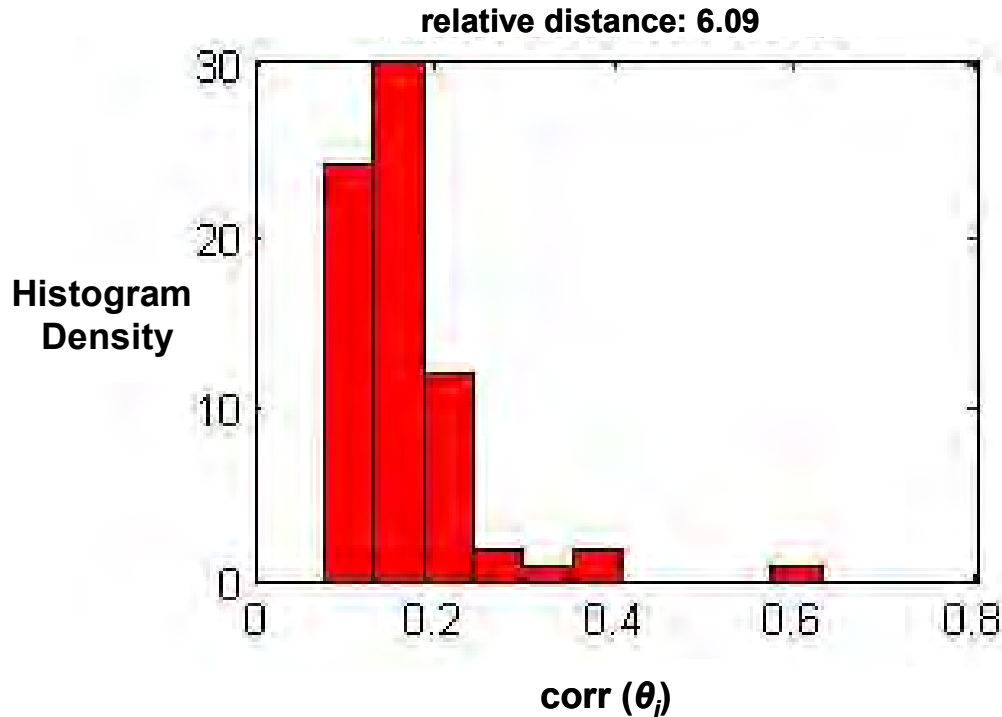
58

Figure 8-9. Histogram of cross-correlation values for the data shown in Fig. 8-8.

The relative distance can be interpreted in a number of ways. One interpretation is that the relative distance provides a metric of how "out of the ordinary," or significant, a particular relative orientation is with respect to a representative sample of random orientations. Another interpretation is that the median of the correlation values obtained for all rotational angles provides us a "baseline" of the correlation values, which would be achieved by pairs of non-matching breech face signatures. In other words, if we assume that the median of the correlation value for a representative sample of relative orientations is approximately the same for pairs of matching and non-matching cartridge cases (which will be true if the optimal peak for matching pairs is sufficiently narrow), then the relative distance provides a metric of the probability of obtaining the particular maximum correlation value given that the pair under consideration is a non-matching pair. Viewed from this perspective, a relative distance of 6.09 is a very convincing indication of a matching pair.

Figure 8-10 shows a graphical representation of results obtained from a comparison of matching and non-matching signatures of breech face impressions. The peak correlation and relative distance results from the comparison of each pair of such signatures is indicated by either a blue square (for non-matching pairs) or a red rhomboid (for matching pairs). The horizontal axis of Fig. 8-10 corresponds to $ACCF_{max}$ while the vertical axis corresponds to relDist. As expected, the majority of matching breech impression pairs achieve both high maximum correlation and high relative distance values, while the non-matching pairs only achieve relatively low values. Nevertheless, there are a few matching pairs that fail to achieve high values and a few non-matching pairs that achieve relatively high values. These pairs might become false negative identifications and false positive identifications respectively.
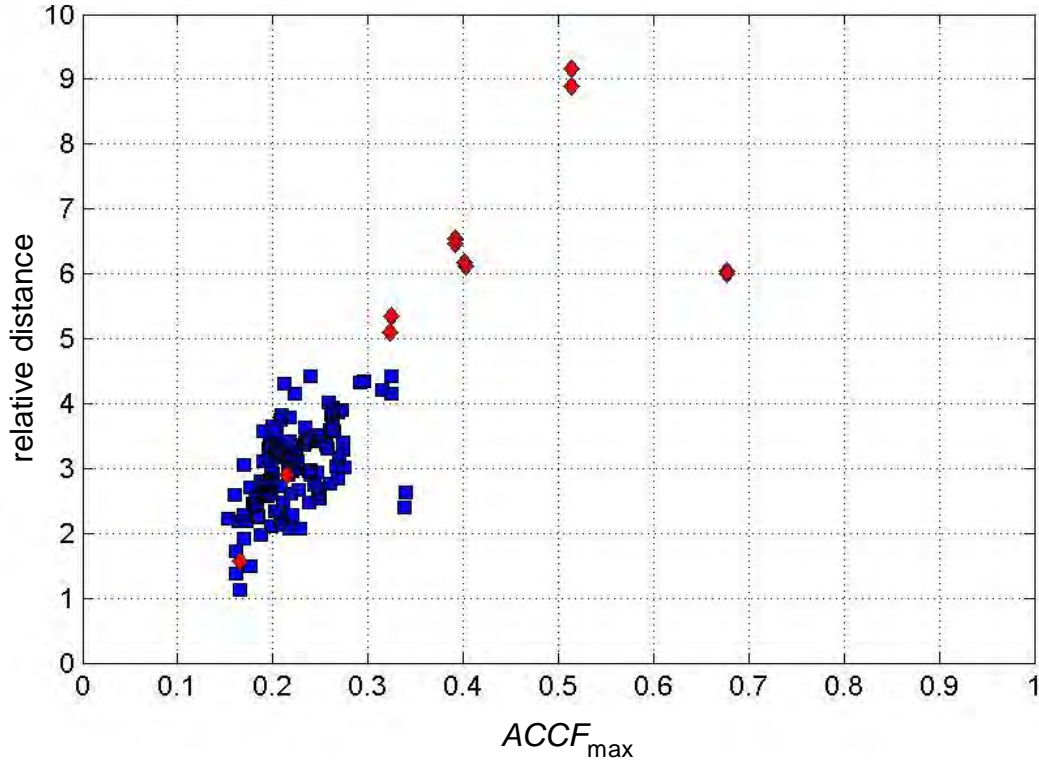
59

Figure 8-10. Graphical representation of correlation ($ACCF_{max}$) and relative distance results for the comparison of matching (red) and non-matching (blue) pairs of breech face signatures of Remington casings fired from guns 007, 009, 213, 375, 430, and 535 in the De Kinder collection.

The graphical representation shown in Fig. 8-10 lends itself to the definition of non-overlapping regions of the 2D space corresponding to matching, non-matching, and possibly "undetermined" pairs of signatures. One can construct an overall "score" metric based on a linear combination of the maximum correlation and the relative distance, which effectively projects the two dimensional results into a single dimension. The linear transformations used for this study (based on empirical observations) were:

$$s = Gd * \text{relDist} + Gc * (ACCF_{max} - \Delta\text{corr}),$$

(11)

with the empirical constants $Gd = 0.77$, $Gc = 6.4$ and $\Delta\text{corr} = 0.1$ for firing pin impressions, and $Gd = 0.89$, $Gc = 4.5$ and $\Delta\text{corr} = 0.1$ for breech face impressions. The constant Δcorr is the x-intercept in Fig. 8-10. We have used the relative distance and the combined parameter s shown in Eqn. 11 above for one of the correlation calculations shown later in Sec. 8.

8.5.2 Parameters Related to rms Roughness

Although the $ACCF_{max}$ can be used for signature comparison, we observed during previous 2D bullet signature measurements [17] that $ACCF_{max}$ does not characterize the uniqueness of a topography image. Based on the definition of the cross-correlation function, if two compared signatures have the same shape but different vertical scales, their $ACCF_{max}$ is still 100 %. We have, therefore, developed a parameter called the signature difference, D_s, which is highly

correlated with $ACCF_{max}$ but which directly quantifies bullet signature differences [26]. For 3D topography comparisons, the 3D version of the parameter D_s is calculated in the following way:

- At the registration position where the $ACCF_{max}$ value between topography B and A occurs, after shifting along the x- and y-directions and rotating around the z-axis, construct a new topography $Z_{B-A}(x)$, which is equal to the difference of the compared topography Z_B and the reference topography Z_A.
- Calculate the areal rms roughness for the new topography $Z_{B-A}(X)$, $Sq(B - A)$.
- Calculate the topography difference D_s between topography B and A defined as the ratio

$$D_s(3D) = Sq^2(B-A) / Sq^2(A). \qquad (12)$$

A similar parameter to D_s is the difference in mean square roughness between the two surfaces being compared. This metric is given by $Sq^2(B) - Sq^2(A)$. Both metrics are directly related to scale differences in roughness between two surfaces that might have otherwise similar shapes. We have calculated these parameters in some of the studies undertaken here but have not yet performed a systematic appraisal of their usefulness as metrics for correlating two surfaces, which would effectively supplement information provided by the cross-correlation maximum.

8.6 Uncertainty Arising from Topography Measurements

As suggested by the NA Committee, uncertainty in the $ACCF_{max}$ results was estimated by repeating topography measurements on two of the casings over four days. The correlations between topography images of the same surfaces and the variations in those correlations enable us to estimate uncertainty in the correlation results due to variation in the topography measurements. The two casings were the Remington casings #3 and #7, both fired by the Sig Sauer 535 gun in the De Kinder collection.

The results for $ACCF_{max}$ correlation values for pairs of these casings are shown in Fig. 8-11 for breech face impressions and Fig. 8-12 for firing pin impressions. The $ACCF_{max}$ values are shown plotted along the x-axis, and the relative distance values are plotted along the y-axis. The results of Fig. 8-11 clearly separate themselves into three groupings: correlations among the breech-face-impression images of Remington casing #3, correlations among the breech-face-impression images of Remington casing #7, and correlations between the images of #3 and #7. The results also include the images from these two casings taken when they were first measured as two of the casings in the set of 70 casings. The grouping of diamonds in Fig. 8-11, for example, includes the pair-wise correlations between the four topography images of the Remington #3 breech face impression measured over the four days and the original image measured about half a year earlier. Altogether there were 20 correlation values calculated over all of the pair-wise comparisons for the five images. Likewise the grouping of squares comprises the 20 pair-wise correlations for images of the Remington #7 casing fired by the Sig Sauer 535. The triangles comprise the intercomparisons of the #3 casing images discussed above with the #7 casing images. There are 50 of these. For all groupings, each pair-wise comparison is included twice, with the reference and compared casings switching places.
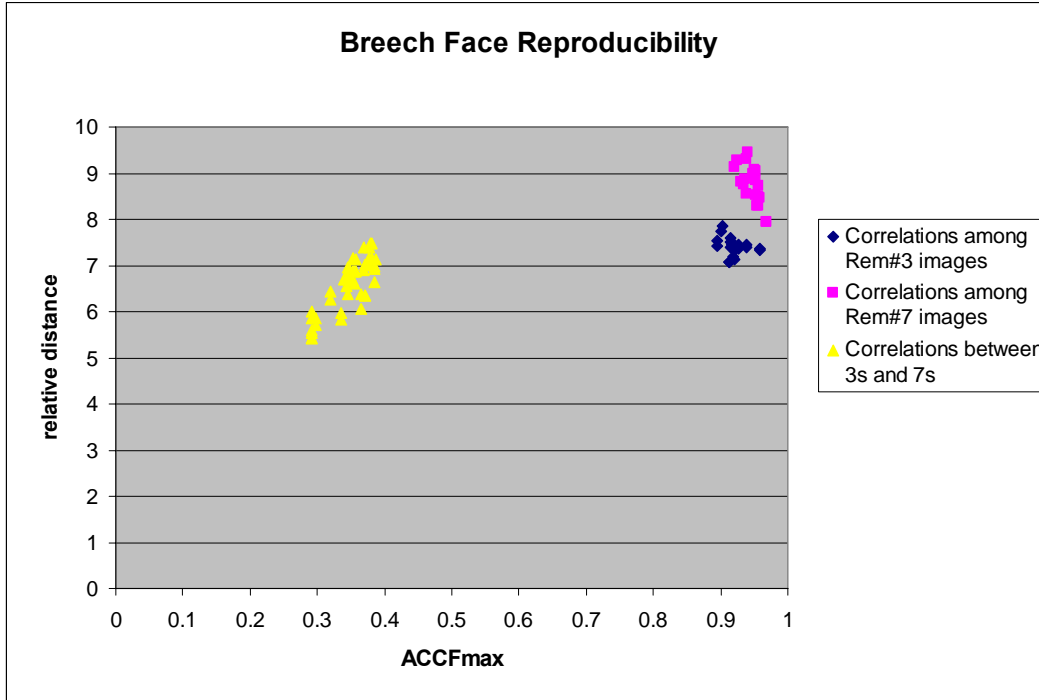
Figure 8-11. Reproducibility of topography measurements of Remington #3 and #7 breech face impressions as given by the $ACCF_{max}$ and the relative distance parameters.
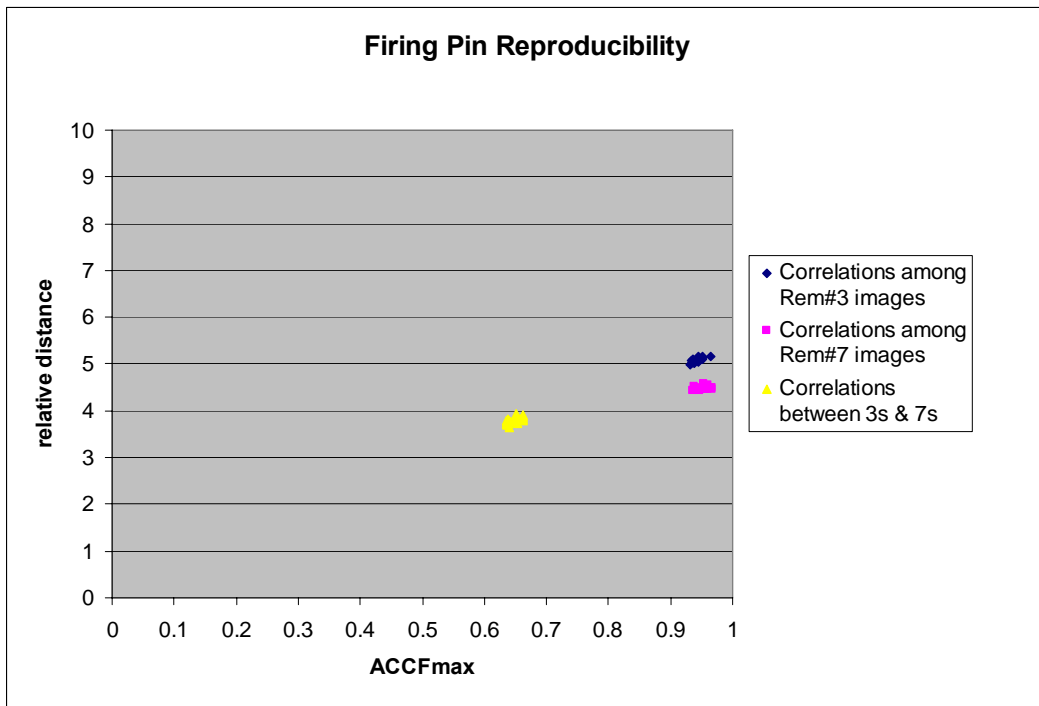

Figure 8-12. Reproducibility of topography measurements of Remington #3 and #7 firing pin impressions as given by the $ACCF_{max}$ and the relative distance parameters.

Reproducibility of the firing pin impressions is shown in Fig. 8-12. There are a smaller number of data points here because the original data for the firing pin impressions were measured on a different model of the confocal microscope with the camera having a different pixel spacing along the y-axis. Therefore we did not include those two images in this comparison. The clusters of diamonds and squares represent twelve correlations among four images each and the cluster of triangles represent 32 correlations between the two sets of four images.

We now discuss estimated uncertainties in the $ACCF_{max}$ values arising from sources of error in the topographic measurements. We expect that uncertainties in the relative distance parameter are based on similar considerations but do not discuss those here. The mean $ACCF_{max}$ values for these groupings and their standard deviations are shown in Table 8-1:

Table 8-1. Average correlation values and standard deviations for topography measurements of De Kinder Remington casings fired from Sig Sauer 535.

Components Being Compared	Breech Face Impressions $ACCF_{max} \pm 1$ std. dev.	Firing Pin Impressions $ACCF_{max} \pm 1$ std. dev.
Rem #3, Sig Sauer 535	0.920 ± 0.018	0.944 ± 0.010
Rem #7, Sig Sauer 535	0.947 ± 0.013	0.954 ± 0.011
Rem #7 vs.Rem#3	0.349 ± 0.030	0.647 ± 0.010

The results clearly indicate a high similarity between different topography measurements of the same casing but significant differences between measurements of the two different casings, even though those casings were fired by the same gun. The correlation values of 0.920, 0.944, 0.947, and 0.954 are both high and consistent with one another and have small standard deviations. However, they are slightly less than unity. This suggests that there are variations between topography images of the same object giving rise to an attenuation of the $ACCF_{max}$ value by 5 % to 8%. This variation arises from variation in the measured surface topography along all three coordinate axes, x or y or z. The fact that the standard deviations are so low suggests that the variation between images is consistent and random, probably with high spatial frequency components. By contrast the low scores, averaging 0.349 when comparing the two breech faces and 0.647 for the two firing pins, indicate large differences between the #3 and #7 surfaces, even though the casings were fired by the same gun.

We, therefore, conclude that the correlation value is biased below unity by about 5 % to 8 % by measurement-related differences and noise in the topography images, but the statistical uncertainty of this bias is only about 1.8 % (standard uncertainty). This source of uncertainty is much smaller than the changes in $ACCF_{max}$ due to the topography of the surface, which are the principal sources of the differences we aim to observe.

8.6.1 Uncertainty Budget for Breech Face Impressions

For the breech face impressions, the trimming process leads to a second source of variation and bias. Figure 8-13 shows the topography of the breech face impressions of Remington casings #3 and #7 from Sig Sauer #007. It reveals a prominent ridge around the firing pin impression. If such inner ridges are trimmed out of the breech face impressions to be matched, then the average correlation value is 0.349 for gun #535 as shown in Table 8-1. If, however, the inner ridges are included in the breech face impressions, the correlation score is more influenced by the prominent ridge, and the average score for the topography comparisons of the #3 casing with the

#7 casing increases by about 0.06 to 0.402 with a variation that is several times larger than the 0.01 levels we record in Table 8-1 above.
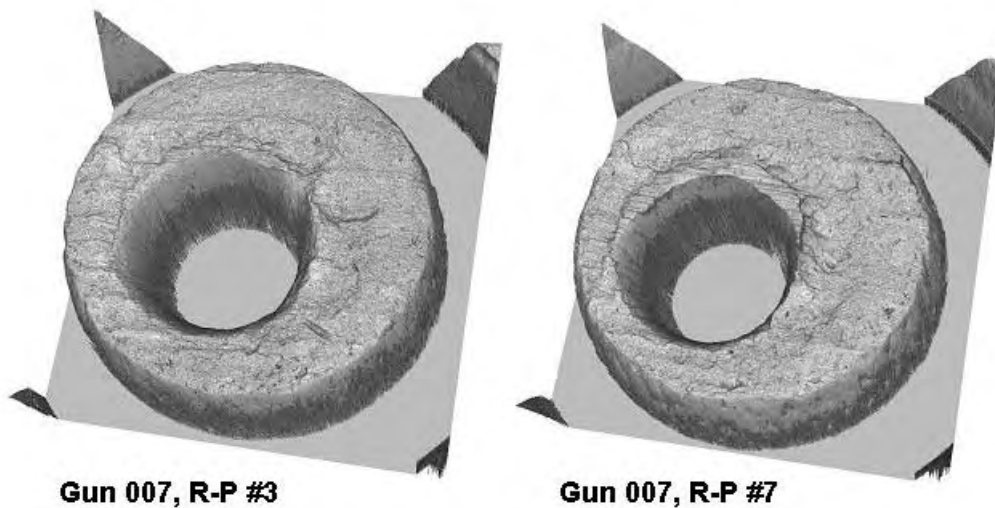

Figure 8-13. Topography images of the breech face impressions of Remington casings #3 and #7 fired in the Sig Sauer gun 007 showing the raised section around the inner radius.

As a result, one might include another factor of 0.06 as an added component of uncertainty for $ACCF_{max}$. However, including the inner ridge or other form deviations in the breech face topography is not regarded as good practice because this form component obscures the real individual characteristics of the surfaces present in the fine roughness structure. A good area-selection procedure should not include the ridge. The trimming procedure leading to Fig. 8-11 and the entries for breech face impressions in Table 8-1 correctly excluded the ridge from the topography images that were correlated. The average difference of 0.06 is therefore regarded as an upper limit of the variation in $ACCF_{max}$ that may arise from variation in the selection of areas to correlate for the breech face impression. This is a limitation of current algorithms. Further development could lead to automated algorithms for area selection that are more accurate.

Because the $ACCF_{max}$ cross-correlation parameter depends on relative differences between images, we believe that there are very few other significant sources of uncertainty arising from the measurement process, as long as a systematic protocol for measurement and the areas to be measured have been determined. If instrument-related errors occur in one of the images being compared, those errors produce a change in the $ACCF_{max}$ value. These types of changes should be captured by the Type A [46,47] statistical uncertainty discussed above.

The uncertainty budget for $ACCF_{max}$, for the breech face impressions therefore contains two components. The first (u_1) is the statistical type A uncertainty arising from instrument variations. We estimate this to be 0.018 ($k = 1$) by using the largest of the standard deviations shown in Table 8-1. The second source of uncertainty (u_2) arises from selection and variation of the measured areas. Using the value of 0.06 discussed above as an outer limit of error and assigning a uniform probability distribution to the error, we derive a Type B uncertainty [46,47] for u_2 equal to 0.034 ($k = 1$). Combining these two components quadratically, we arrive at a combined

standard uncertainty u_c ($k = 1$) for $ACCF_{max}$ equal to 0.038. In addition, we expect the $ACCF_{max}$ value to be consistently biased low by 0.065. The biases of 0.05 to 0.08 recorded in connection with Table 8-1 are consistent with that estimate and with the Type A uncertainty component discussed above.

8.6.2 Uncertainty Sources for Firing Pin Impressions

The data in Fig. 8-12 suggest that the uncertainties for firing pin impressions are smaller than the uncertainties for breech face impressions. However, correlation of the firing pin impressions is more susceptible to two sources of uncertainty, relocation error of the measured area and dropouts and outliers, than correlation of the breech face impressions. The measured area of the firing pin impressions is 800 μm × 800 μm, several times smaller than the 4 mm × 4 mm area of the breech face impressions. Modest changes in the measurement location between two matching images can produce relatively large offsets between the compared areas, thus reducing the overlap of matching features and leading to a decrease of the $ACCF_{max}$ value. Relocation error is especially likely when the firing pin impression is wide and flat, as seems to be the case for the Rem image in Fig. 6-1. In addition, because a large part of the field of view of the firing pin impression is on steeply sloped walls of the impression, outliers are more difficult to distinguish and eliminate while preserving the good data. Due to these effects, the uncertainty for calculating the $ACCF_{max}$ metric for firing pin impressions may be larger than the uncertainty for breech face impressions. In future work, we intend to estimate these effects by comparing matched pairs of images independently measured on each of the 70 De Kinder firing pin impressions.

8.7   Some Cross-Correlation Results

We now examine several cases for pair-wise correlation of the De Kinder firing pin impressions.

First, the same gun with different ammos can produce similar firing pin impression topographies producing a high $ACCF_{max}$ value. Figure 8-14 shows an example of a pair of casings fired from the same gun, Sig Sauer 007. The reference topography is obtained from a Remington casing and the compared topography from a Speer casing. The $ACCF_{max}$ value is about 95 %. The Sq values of the two topographies are quite similar, 1.43 μm and 1.47 μm. The topography difference D_s is 6.9 %.

The same gun with different ammos can also produce firing pin impressions with different forms yielding a low $ACCF_{max}$ value. Figure 8-15 shows an example from the Sig Sauer 139 gun. The reference topography was obtained from a CCI casing and the compared topography from a Federal casing. The $ACCF_{max}$ is about 46 %. Their Sq values are quite different, 0.46 μm and 0.86 μm, respectively. The topography difference D_s is large, about 222 %.

By contrast, different guns with different ammos can produce firing pin impressions with similar topographies leading to a high $ACCF_{max}$ value. Figure 8-16 shows the topographies of the firing pin impressions from gun 314 with CCI ammo and gun 215 with Winchester ammo. The $ACCF_{max}$ value is about 82 %. However, their Sq values are 1.31 μm and 2.55 μm, respectively, and the topography difference D_s is about 164 %. From this observation, it is clear that a high

$ACCF_{max}$ value does not necessarily confirm a match. The D_s parameter provides supplementary information that can improve the accuracy of proposed matches.

Figure 8-14. Comparison of the topography images of the firing pin impressions from a Remington casing (reference surface) and a Speer casing (compared surface) fired by the same Sig Sauer 007 gun. The value of $ACCF_{max}$ is approximately 95 %. All axes have units of μm.

Figure 8-15. Comparison of the topography images of the firing pin impressions from a CCI casing (reference) and a Federal casing (compared) fired by the same Sig Sauer 139 gun. The value of $ACCF_{max}$ is approximately 46 %, indicating that the same gun with different ammos can also produce firing pin impressions with different forms. All axes have units of μm.

Figure 8-16. Topography images of firing pin impressions from Sig Sauer 314 gun with CCI ammo and Sig Sauer 215 gun with Winchester ammo.

9.   Statistical Analysis: General

9.1   Introduction

This section analyzes the available data using several methods in order to answer the following questions:

1. Are the firing pin and breech face impressions left by individual guns on casings distinct and repeatable enough to distinguish them from those of other guns?
2. How does a 3D topography technology, such as N-3D, perform relative to an I-2D system?
3. What factors (e.g. gun manufacturer, ammunition type, etc.) affect the performance of the systems?
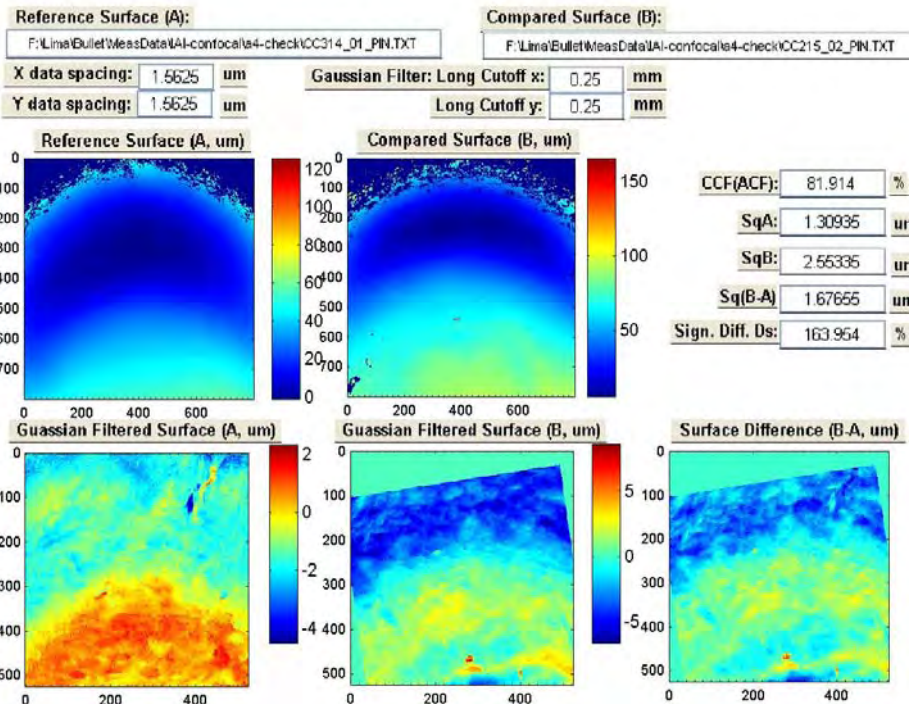4. Which are more helpful–firing pin or breech face data? Does the answer depend on the imaging system or the set of casings used?

The section begins with a descriptive recap of the experimental data obtained in this study. Then, using these data sets as test databases, the N-3D and I-2D imaging systems are tested using Top Ten searches of the type routinely output by the I-2D system. These tests give a clear picture of the performances of the systems over the various data sets.

Later sub-sections analyze the full round-robin correlation data obtained by the N-3D system. A probabilistic overlap metric p is introduced as a useful heuristic for comparing the empirical distributions of correlation scores. This metric is later used as an input for probability models that project how a system performs for very large databases.

9.2   Recap of Casings Databases

This study investigated two sets of casing data: "De Kinder" and "NBIDE". The De Kinder data are "historical," with the casings produced by De Kinder et al. [7] several years ago. NIST created the NBIDE database in 2005 as a part of this ballistics study.

9.2.1 De Kinder

What is referred to as the De Kinder casing set is actually a subset of a much larger experiment discussed in the paper by De Kinder et al. [7]. The casings described here involved 10 guns all of the same model (Sig Sauer 9 mm Model P226) and 7 ammunitions (cartridge types) of which two are repeats of the same brand (Remington):

1.   CCI
2.   Winchester
3.   Remington (Rem)
4.   Speer
5.   Wolf
6.   Federal
7.   Rem (a repeat)

A total of $10 \times 7 = 70$ test firings were used. The corresponding 70 casings were originally studied via the I-2D imaging method, and a report on the findings appears in Ref. 7. As part of the NIST study, these 70 De Kinder casings were also imaged for topography (3D) and analyzed via the N-3D method.

9.2.2 NBIDE

A central component in the NIST study was to determine the effect of gun type (manufacturer) on gun identifiability. As noted above, the De Kinder study was limited to a single gun type (Sig Sauer 9 mm). To go beyond this, a statistically designed experiment was developed as part of the NBIDE. The specifications for this experiment were as follows:

1. Number of gun types: 3 – Smith&Wesson, Ruger, Sig Sauer P226
2. Number of guns (total): 12
3. Number of ammo types: 4 – Remington, Winchester, PMC, Speer (extra)
4. Number of days: 3 – 48 firings per day
5. Total number of test firings: 144 = 12 guns × 4 ammos × 3 days
6. Total number of casings subsequently analyzed: 108 – only 3 ammos

The experiment was designed and conducted in accordance with rigorous statistical design principles and techniques. Within a given day, there were four sets of 12 gun firings. Across the three days, there were a total of 12 sets of 12 gun firings. In a latin square [48] fashion, for each time position (1 to 12) within a set, each gun was fired once and only once across the 12 sets. Each set of 12 firings used the same ammo. All four ammo types were used every day. The ordering of the ammos for the four sets within a day was balanced in an incomplete latin square fashion. Finally, after the 144 fired casings were collected and annotated, the casings were re-randomized for the first part of the analysis so as to assure that the analysis was done in a double blind fashion [21].

The advantage of the resulting NBIDE database is that it allows one to ascertain the existence and magnitude of gun type and ammo type on gun distinguishability. Because a given gun-ammo combination occurs three times across the experiment, this NBIDE database also allows for the assessment of how distinguishable or indistinguishable a given gun can be across the three firings. The limitation of the NBIDE database is that the conclusions are, strictly speaking, limited in scope to the three gun types utilized and the three ammo types. Also, the total number of firings (144) and analyzed casings (108) is still relatively small compared to the large sizes envisioned for a national ballistics database.

9.3   System Performance Analysis

When we compare a single reference casing to an existing database of casings, how do we determine if the comparison was a "success"? What is the criterion for "success"? In practice, an imaging/analysis system will yield a short (e.g., ten item) list of best-matching casings, which will then be subject to further examination by a human forensics expert. Thus for this single trial, the performance of an imaging/analysis system might be declared a "success" if the correct database casing, if existent, appears in the short list produced by the imaging/analysis system,

since frequently that list is what the human expert will limit his/her search to. The system will have "failed" if a matching database casing exists and the "TopTen" list does not include that casing.

The analysis of the performance of a system will thus be centered on the questions: Was it a "success" (for a given reference casing trial)? What percentage of the time was it a success (for a set of reference casings)?

For the two data sets at hand (De Kinder with 70 test firings and NBIDE with 108 test firings), we have many opportunities to achieve success. For the I-2D system, a Top Ten list is produced automatically via proprietary software. In the present study, we used the topography data, described in earlier sections, estimated the similarity of pairs of casings based on the cross-correlation function maximum ($ACCF_{max}$) described in Section 8, and formed Top Ten lists from the relative rankings.

We can use the 70 De Kinder test firings as an opportunity for 70 reference-casing trials. For each such reference casing, we use the remaining 69 test casings as our "database". Thus the casing from the first test firing was compared against 70 - 1 = 69 other casings from the remaining test firings. Out of those 69 casings, six came from the same gun as the reference firing, and 63 (= 9 guns × 7 ammos) came from other guns. Thus in an ideal world we would expect the system's Top Ten list to contain all six of those remaining casings. A less stringent criterion would expect five out of the six casings to appear in the Top Ten list, and so forth. A very weak criterion would expect at least one out of the six casings in the Top Ten list. A complete failure would yield none out of the six casings in the Top Ten list. Thus for a fixed inclusion criterion (6, 5, ..., 1), the imaging/analysis system will be judged as a "success" or a "failure".

In a similar fashion, we choose the casing from the second test firing and compare it to the 69 casings from the remaining firings. Again, for a fixed inclusion criterion, the imaging/analysis systems may be declared a success or a failure.

Repeating the process for all 70 casings yields a sequence of 70 successes or failures. The proportion of those 70 cases that were a success defines the performance of the imaging/analysis system for this particular casing database and a fixed inclusion (6,5,...,1) criterion.

For the NBIDE database, we have 12 guns, three ammos, and three days (repeats), and so the same sort of process would yield a comparison of the casing from the first test firing against the 108 - 1 = 107 casings from the remaining firings. Of those 107 casings, 8 (= 3 ammos × 3 days - 1) come from the same gun as the first test firing, and 99 (= 11 guns × 3 ammos × 3 days) come from different guns. Again in an ideal world, an excellent-performing imaging/analysis system would have all eight of those same-gun casings in the Top Ten list, a good performing system might have seven out of the eight in the Top Ten, and so forth, down to a weak performing system having only one out of the eight, and a poor-performing system having none out of the eight. Thus for a given reference casing and a fixed inclusion criterion (8, 7, ..., 1), the system may be judged as "successful" or "failing". Repeating the process for all 108 test-fire reference

casings will yield the percentage of the time the imaging/analysis system performed well; this will be our performance metric.

9.3.1 De Kinder Top Ten System Performance Analysis

In light of the above, the following listing shows the performance metric for the I-2D and N-3D imaging/analysis systems for the

Database = De Kinder (70 casings: 10 guns × 7 ammos)
Region = Firing pin impressions only
Total number of items in the system output list = 10 (i.e., Top Ten)
"Success" criterion: at least 1 of the 7-1 = 6 casings appears in the Top Ten
(a very weak criterion)

	$i \geq 1$	
	I2	N3
FP (%):	94	74

where I-2D is shortened to I2 here and N-3D is shortened to N3 and the casing region is designated by "FP" for firing pin.

This listing says that for the firing pin region of the De Kinder data, 94 % (= 66) of the 70 I-2D Top Ten lists were "successful"–containing one or more of the remaining six correct casings. It further says that 74 % (= 52) of the 70 N-3D "Top Ten" lists were "successful"—containing one or more of the remaining six correct casings. Thus for this particular case and (very weak) criterion, the I-2D system performed better.

The expanded listing, which contains stronger criteria, is as follows:

	$i \geq 1$		$i \geq 2$		$i \geq 3$		$i \geq 4$		$i \geq 5$		$i \geq 6$	
	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3
FP (%):	94	74	83	63	66	57	39	51	17	47	7	33

This listing shows that for the firing pin region of the De Kinder data, and for both imaging/analysis methods, as the inclusion criterion becomes more stringent, the success proportion decreases. For example, only 7 % (= 5) of the 70 I-2D Top Ten lists were fully "successful"—containing all six of the remaining six correct casings, and only 33 % (= 23) of the 70 N-3D Top Ten lists were "successful"—containing all six of the remaining six correct casings. Thus for this particular case and very strong criterion, the two imaging/analysis systems performed poorly, but with N-3D performing better than I-2D. Note that for the less stringent criteria, I-2D performs better than N-3D, but as the stringency increases, N-3D performs increasingly better relative to I-2D. The reason may be because N-3D performed very well on some guns and very poorly on others, while the I-2D performance tended to be between those two extremes.

The above analysis was for the firing pin region only. A similar analysis could be done for the breech face (BF) region only. Doing so yields the following augmented listing:

De Kinder/ Top Ten

	$i \geq 1$		$i \geq 2$		$i \geq 3$		$i \geq 4$		$i \geq 5$		$i \geq 6$	
	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3
FP (%):	94	74	83	63	66	57	39	51	17	47	7	33
BF (%):	67	90	30	74	4	57	0	37	0	19	0	6

This listing shows that for the breech face region of the De Kinder data, as the inclusion criterion becomes more stringent, the success proportion decreases for I-2D from 67 % to 0 % and for N-3D from 90 % to 6 %. For the most stringent criterion requiring all six of the six matching casings to be in the Top Ten list, none of the 70 I-2D Top Ten lists were "successful", and only 6 % (= 4) of the 70 N-3D Top Ten lists were "successful." Thus for this particular case and very strong criterion, the I-2D system failed and the N-3D system performed only slightly better.

General De Kinder/Top Ten findings (valid across both Firing Pin and Breech Face) that may be drawn from the above listing are as follows:

1.   Success proportions for both I-2D and N-3D decrease as stringency increases, a mathematical necessity.
2.   The most drastic decrease is I-2D/Breech Face.
3.   The least drastic decrease is N-3D/Firing Pin.
4.   I-2D does not do well for Breech Face.
5.   For Firing Pin, I-2D performs better than N-3D in the three least stringent cases, but N-3D performs better than I-2D in the three most stringent cases.
6.   For Breech Face, N-3D performs better than I-2D in all 6 out of 6 cases.
7.   For I-2D, Firing Pin is a better discriminator than Breech Face. For N-3D, Firing Pin is a better discriminator for the most stringent cases, but not for the least stringent cases.

9.3.2 De Kinder: Table of Top Ten Results

As a reference, the following table of Top Ten testing results gives some more information on how the systems performed with respect to the area imaged and the technology used.

Table 9-1. Comparison of the numbers of *correct* matches (out of the ten highest *scoring* matches) for the I-2D correlation metric and the N-3D $ACCF_{max}$ metric applied to firing pin and breech face impressions for 10 De Kinder guns . Rem sep denotes the second Remington casing filed separately.

Number of Correct Matches (Maximum 6)

Ref. Casing	Ammo	Firing Pin		Breech Face	
		I-2D	N-3D	I-2D	N-3D
Means		3.06	3.26	1.01	2.83
007-01	CCI	4	4	0	2
007-02	WIN	5	6	0	1
007-03	Rem	6	6	2	2
007-04	SPEER	6	6	1	0
007-05	WOLF	6	5	1	1
007-06	FC	6	5	2	1
007-07	Rem sep	6	6	1	4
009-01	CCI	3	5	0	2
009-02	WIN	0	1	0	0
009-03	Rem	2	5	0	4
009-04	SPEER	4	5	1	0
009-05	WOLF	4	5	0	1
009-06	FC	4	5	1	3
009-07	Rem sep	4	5	2	3
117-01	CCI	3	1	1	5
117-02	WIN	0	3	0	1
117-03	Rem	4	2	2	5
117-04	SPEER	5	4	1	3
117-05	WOLF	5	4	1	4
117-06	FC	5	3	0	4
117-07	Rem sep	3	0	1	6
139-01	CCI	3	1	0	4
139-02	WIN	3	0	1	1
139-03	Rem	4	0	2	3
139-04	SPEER	3	0	2	3
139-05	WOLF	2	0	0	2
139-06	FC	1	0	1	5
139-07	Rem sep	3	0	3	5
213-01	CCI	1	0	2	4
213-02	WIN	4	0	0	0
213-03	Rem	4	1	2	5
213-04	SPEER	3	0	0	3
213-05	WOLF	3	0	1	3

213-06	FC	1	0	0	1
213-07	Rem sep	0	0	2	5
215-01	CCI	1	3	1	6
215-02	WIN	2	0	0	4
215-03	Rem	3	0	1	6
215-04	SPEER	1	3	0	4
215-05	WOLF	1	1	2	3
215-06	FC	2	2	0	3
215-07	Rem sep	2	1	1	5
314-01	CCI	4	1	2	4
314-02	WIN	2	0	0	0
314-03	Rem	5	0	1	4
314-04	SPEER	2	2	2	2
314-05	WOLF	3	1	1	1
314-06	FC	3	2	1	0
314-07	Rem sep	3	0	1	4
375-01	CCI	3	6	3	1
375-02	WIN	3	6	0	2
375-03	Rem	4	6	2	2
375-04	SPEER	5	6	2	0
375-05	WOLF	4	6	1	2
375-06	FC	4	6	2	1
375-07	Rem sep	4	6	2	4
430-01	CCI	2	5	1	5
430-02	WIN	0	6	0	2
430-03	Rem	3	6	2	6
430-04	SPEER	3	6	1	1
430-05	WOLF	3	6	1	4
430-06	FC	2	6	1	3
430-07	Rem sep	5	6	1	5
535-01	CCI	1	5	0	3
535-02	WIN	4	6	1	2
535-03	Rem	1	6	3	3
535-04	SPEER	3	6	0	3
535-05	WOLF	2	6	0	2
535-06	FC	2	6	0	2
535-07	Rem sep	2	6	2	3

The table shows large performance differences between individual guns.

System Gun Differences: Firing Pin
For further analysis of the performance differences, we define for the firing pins:
Y = (# of N-3D correct matches) – (# of I-2D correct matches).

We find that: Median (Y) = 0, Mean(Y) = 0.2, Standard Deviation (Y) = 2.5.

This implies a slight advantage for N-3D. However, one wonders if that holds across the guns or whether it depends on the particular gun. Figure 9-1 charts the difference variable Y by reference gun.
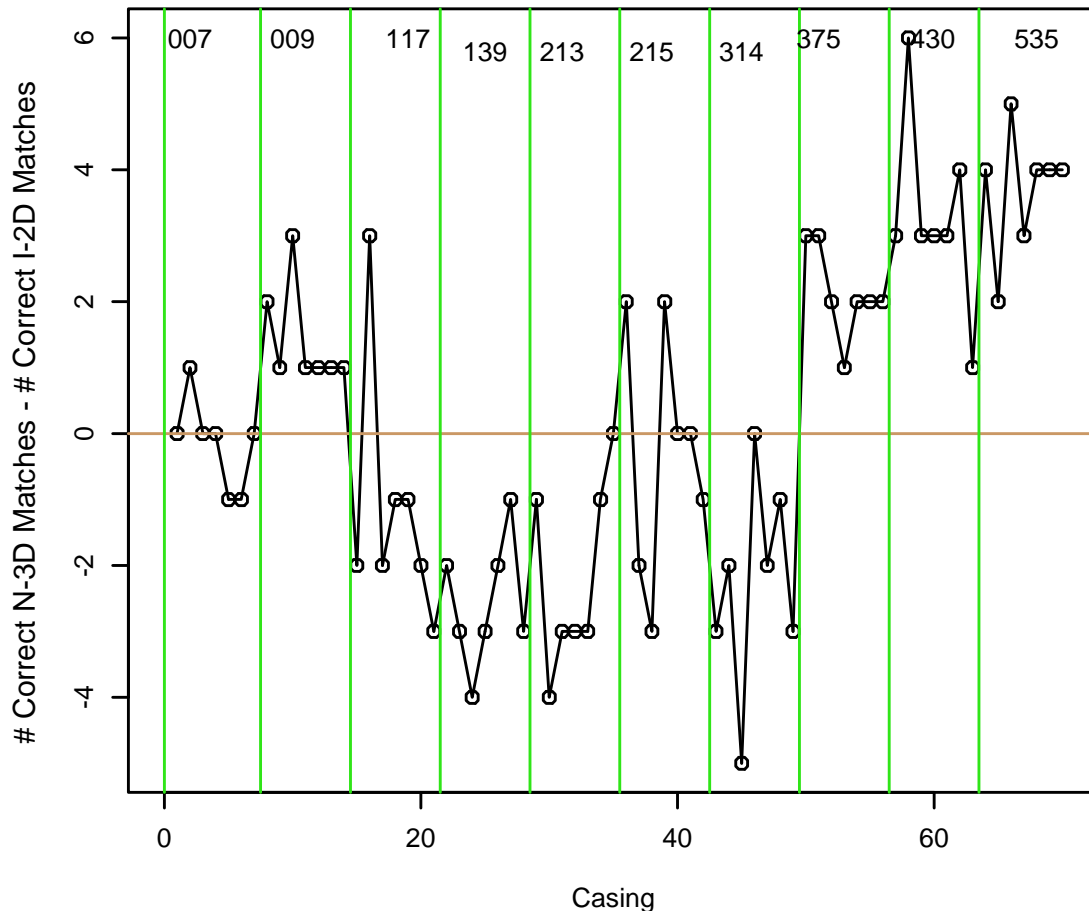

Figure 9-1. Graph of Y = (# of correct N-3D matches) – (# of correct I-2D matches) of De Kinder firing pin impressions arranged by reference gun. The points above the zero-line depict those casings where N-3D performs better than I-2D, while the points below the zero-line depict those where I-2D does better.

The performance differences are very gun-dependent. It is clear that the 3D and 2D methods are not making the same mistakes; the 3D method does better for guns 009, 375, 430, and 535, while the 2D method does better on 139, 213, and 314. Further inspection of the actual images for these casings may give clues as to what characteristics drive the differential performance of the two methodologies. A detailed analysis of differences between the guns is given in Sec. 10.

System Gun Differences: Breech Face

For further analysis of the performance differences, we define for the breech faces:
Y = (# of N-3D correct matches) – (# of I-2D correct matches). The summary statistics for Y are:
Median (Y) = 2.0, Mean (Y) = 1.9, Standard Deviation (Y) = 1.8.

This implies a clear advantage for the 3D method. Again, one wonders if that holds across the guns or whether it depends on the particular gun. Figure 9-2 charts the difference variable Y by reference gun.
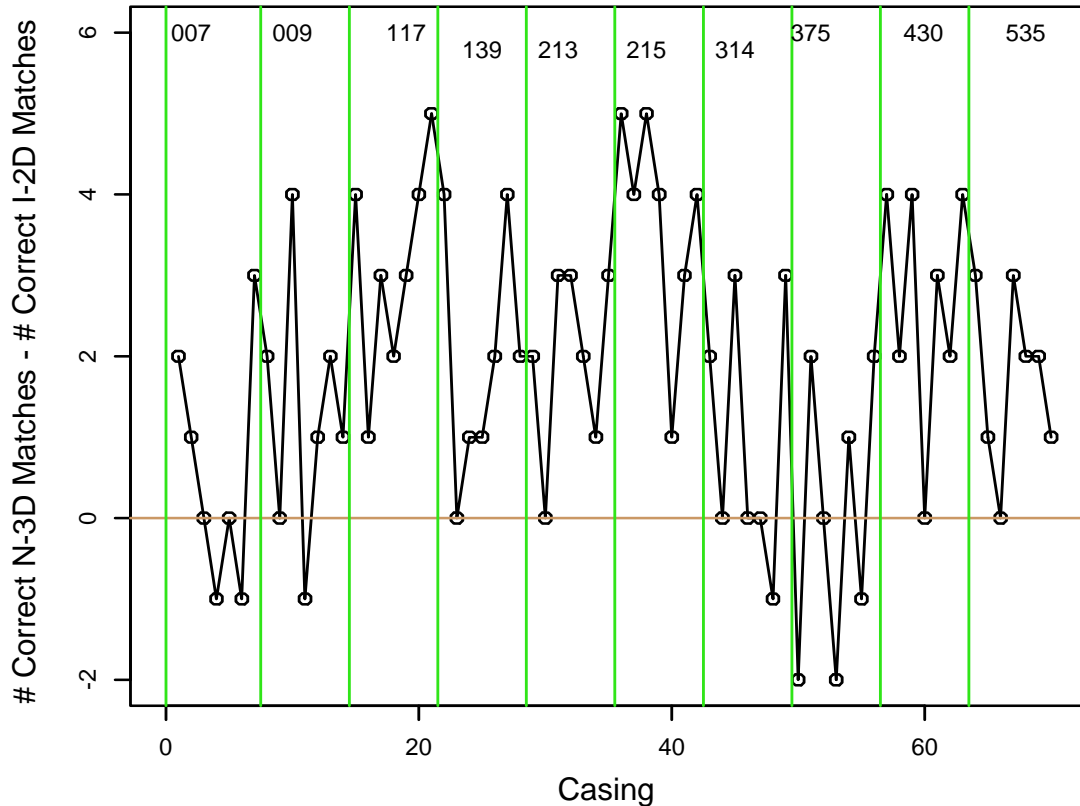

Figure 9-2. Graph of Y = (# of correct N-3D matches) – (# of correct I-2D matches) of De Kinder breech face impressions arranged by reference gun. The points above the zero-line depict those casings where N-3D performs better than I-2D, while the points below the zero-line depict those where I-2D does better.

The 3D method is doing somewhat better than the 2D method across the board except for gun 375. Again, a detailed analysis of the differences between the guns is given in Section 10.

9.3.3 NBIDE: Top Ten System Performance Analysis

In addition to the De Kinder data, the NBIDE study also created and analyzed its own casings database (NBIDE). Whereas the De Kinder data set had one gun type (Sig Sauer), ten guns, and seven ammo types (including a replicated Remington ammo) for a total of 70 (= 10 × 7) test firings, the NBIDE experiment had three gun types (Smith&Wesson, Ruger, and Sig Sauer), four physical guns of each type, for a total of 12 guns, three ammo types (Remington, Winchester, and PMC), and three repeat days, for a total of 108 (= 3 × 4 × 3 × 3) test firings. Carrying out a

76

similar analysis as that described above for the De Kinder data, we compare a given test fired casing with the remaining 107 test fired casings and compute an appropriate similarity metric (e.g., the $ACCF_{max}$) to yield a total correlation data base consisting of $108 \times 107 = 11\,556$ comparisons. For a given test fired casing, there will be eight (3 ammos $\times$ 3 days – 1) other compared casings that "match" (come from the same physical gun), and a total of 99 (= 11 guns $\times$ 3 ammos $\times$ 3 days) compared casings that do not match.

In carrying out a Top Ten analysis as above, we may thus define our comparison to be a "success" if (as before) one or more of the matching casings appear in the Top Ten list, or if two or more do, or if three or more do, all the way up to all 8 matching cases showing up in the Top Ten list. Summing up these "successes" over the 108 reference casings (and converting them to percentages out of the 108), we thus get the following listing:

NBIDE/ Top Ten

	$i\geq1$		$i\geq2$		$i\geq3$		$i\geq4$		$i\geq5$		$i\geq6$		$i\geq7$		$i\geq8$	
	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3	I2	N3
FP (%):	98	100	92	97	71	93	53	83	38	69	18	56	2	41	0	23
BF (%):	100	100	99	100	94	100	86	100	73	100	56	100	35	100	12	94

From this listing we conclude:

1. In stark contrast to the De Kinder data, Breech Face was the better discriminator over Firing Pin for both I-2D and N-3D.
2. For Breech Face, both I-2D and N-3D performed much better on NBIDE than on De Kinder.
3. Success proportions for N-3D and I-2D decrease as stringency increases—more dramatically for Firing Pin, less so for Breech Face.
4. For Breech Face, N-3D was perfect (= 100%) for 7 out of the 8 criteria.
5. Across Firing Pin and Breech face, N-3D performed better than I-2D.

9.3.4 NBIDE: Table of Results

The following table of Top Ten testing results gives detailed information on how the systems performed with respect to the area imaged and the technology system used for the NBIDE casings.

77

Table 9-2. Comparison of the numbers of *correct* matches out of the ten highest *scoring* matches for N-3D and I-2D applied to both firing pin and breech face. Note that the ammunition codes for the first column are [1=Winchester, 2=Remington, 3=PMC].

Number of Correct Matches in Top Ten (Maximum 8)

Gun – Ammo – RR#	Firing Pin		Breech Face	
	I-2D	N-3D	I-2D	N-3D
Means	3.72	5.63	5.57	7.94
Ruger 41 - 1 - RR 78	4	7	7	8
Ruger 41 - 1 – RR 102	4	6	6	8
Ruger 41 - 1 - RR 111	5	7	7	8
Ruger 41 - 2 - RR 45	1	5	3	8
Ruger 41 - 2 - RR 94	0	8	6	8
Ruger 41 - 2 - RR 134	2	7	1	8
Ruger 41 - 3 - RR 118	5	6	2	8
Ruger 41 - 3 - RR 129	5	7	2	8
Ruger 41 - 3 - RR 142	5	7	2	8
Ruger 42 - 1 - RR 28	6	8	7	8
Ruger 42 - 1 - RR 43	4	8	3	8
Ruger 42 - 1 - RR 75	4	8	5	8
Ruger 42 - 2 - RR 2	3	8	6	8
Ruger 42 - 2 - RR 35	6	8	5	8
Ruger 42 - 2 - RR 50	4	8	6	8
Ruger 42 - 3 - RR 16	5	8	5	8
Ruger 42 - 3 - RR 54	5	8	7	8
Ruger 42 - 3 - RR 72	5	8	6	8
Ruger 46 - 1 - RR 95	3	6	5	8
Ruger 46 - 1 - RR 120	3	5	5	8
Ruger 46 - 1 - RR 125	1	3	5	8
Ruger 46 - 2 - RR 1	2	4	6	8
Ruger 46 - 2 - RR 67	2	5	3	8
Ruger 46 - 2 - RR 82	4	5	5	8
Ruger 46 - 3 - RR 19	3	6	5	8
Ruger 46 - 3 - RR 53	3	6	5	8
Ruger 46 - 3 - RR 136	3	4	4	8
Ruger 48 - 1 - RR 31	7	8	7	8
Ruger 48 - 1 - RR 80	6	6	7	8
Ruger 48 - 1 - RR 96	7	8	6	8
Ruger 48 - 2 - RR 22	0	8	5	8
Ruger 48 - 2 - RR 130	2	3	6	7
Ruger 48 - 2 - RR 138	5	4	8	8
Ruger 48 - 3 - RR 49	5	8	6	8
Ruger 48 - 3 - RR 55	6	8	6	8
Ruger 48 - 3 - RR 139	6	7	8	8

Sig Sauer 30 - 1 - RR 40	1	4	4	8
Sig Sauer 30 - 1 - RR 60	1	4	5	8
Sig Sauer 30 - 1 - RR 89	2	4	6	8
Sig Sauer 30 - 2 - RR 8	2	4	3	8
Sig Sauer 30 - 2 - RR 10	3	5	3	8
Sig Sauer 30 - 2 - RR 17	2	1	5	8
Sig Sauer 30 - 3 - RR 21	2	4	6	8
Sig Sauer 30 - 3 - RR 30	3	4	4	8
Sig Sauer 30 - 3 - RR 135	2	6	5	8
Sig Sauer 31 - 1 - RR 27	4	2	4	8
Sig Sauer 31 - 1 - RR 48	4	5	7	8
Sig Sauer 31 - 1 - RR 114	1	1	4	8
Sig Sauer 31 - 2 - RR 15	3	3	7	8
Sig Sauer 31 - 2 - RR 65	1	2	2	8
Sig Sauer 31 - 2 - RR 92	2	2	7	8
Sig Sauer 31 - 3 - RR 20	3	3	6	8
Sig Sauer 31 - 3 - RR 62	1	1	5	8
Sig Sauer 31 - 3 - RR 119	5	5	6	8
Sig Sauer 32 - 1 - RR 87	5	7	7	7
Sig Sauer 32 - 1 - RR 90	3	5	7	8
Sig Sauer 32 - 1 - RR 91	6	6	8	8
Sig Sauer 32 - 2 - RR 12	5	7	7	8
Sig Sauer 32 - 2 - RR 25	5	7	7	8
Sig Sauer 32 - 2 - RR 115	3	6	7	8
Sig Sauer 32 - 3 - RR 42	3	8	3	8
Sig Sauer 32 - 3 - RR 56	5	8	7	8
Sig Sauer 32 - 3 - RR 100	5	8	7	8
Sig Sauer 33 - 1 - RR 23	6	6	4	8
Sig Sauer 33 - 1 - RR 66	5	6	5	8
Sig Sauer 33 - 1 - RR 99	4	5	4	7
Sig Sauer 33 - 2 - RR 32	4	2	3	8
Sig Sauer 33 - 2 - RR 34	4	2	3	8
Sig Sauer 33 - 2 - RR 141	2	6	3	8
Sig Sauer 33 - 3 - RR 61	4	7	4	8
Sig Sauer 33 - 3 - RR 79	5	6	4	8
Sig Sauer 33 - 3 - RR 128	4	6	4	8
S&W 305 - 1 - RR 57	6	8	8	8
S&W 305 - 1 - RR 64	6	7	6	7
S&W 305 - 1 - RR 97	6	7	8	8
S&W 305 - 2 - RR 24	6	8	7	7
S&W 305 - 2 - RR 103	2	7	5	8
S&W 305 - 2 - RR 137	2	8	5	8
S&W 305 - 3 - RR 4	6	8	7	8
S&W 305 - 3 - RR 5	6	8	8	8
S&W 305 - 3 - RR 59	6	8	8	8

S&W 306 - 1 - RR 7	4	6	8	8
S&W 306 - 1 - RR 26	2	4	7	8
S&W 306 - 1 - RR 71	3	3	8	8
S&W 306 - 2 - RR 106	2	4	8	8
S&W 306 - 2 - RR 121	2	3	7	8
S&W 306 - 2 - RR 131	2	3	7	8
S&W 306 - 3 - RR 13	3	6	8	8
S&W 306 - 3 - RR 41	3	7	8	8
S&W 306 - 3 - RR 143	2	3	8	8
S&W 314 - 1 - RR 85	3	7	4	7
S&W 314 - 1 - RR 112	6	8	5	7
S&W 314 - 1 - RR 127	5	7	6	8
S&W 314 - 2 - RR 36	6	6	5	8
S&W 314 - 2 - RR 39	6	7	6	8
S&W 314 - 2 - RR 116	2	4	6	8
S&W 314 - 3 - RR 6	5	7	6	8
S&W 314 - 3 - RR 29	5	7	7	8
S&W 314 - 3 - RR 117	5	5	6	8
S&W 401 - 1 - RR 3	2	5	6	8
S&W 401 - 1 - RR 46	3	4	4	8
S&W 401 - 1 - RR 63	2	5	7	8
S&W 401 - 2 - RR 9	5	5	6	8
S&W 401 - 2 - RR 84	3	3	2	8
S&W 401 - 2 - RR 110	4	3	7	8
S&W 401 - 3 - RR 44	4	5	6	8
S&W 401 - 3 - RR 51	3	4	7	8
S&W 401 - 3 - RR 76	3	4	4	8

The N-3D Breech Face results are much better than the other results. As an example of gun differences, all methods seemed to work relatively well with Ruger 42.

9.3.5 Findings: System Performance Analysis

A summary of some of the results of the previous sections is given in the following listing.

Proportion of Correct Matches in Top Ten Listings

I-2D De Kinder FP	3.06/6=0.51
I-2D De Kinder BF	1.01/6=0.17
N-3D De Kinder FP	3.26/6=0.54
N-3D De Kinder BF	2.83/6=0.47
I-2D NBIDE FP	3.72/8=0.46
I-2D NBIDE BF	5.57/8=0.70
N-3D NBIDE FP	5.63/8=0.70
N-3D NBIDE BF	7.94/8=0.99

Based on these results and others in the previous six sections, we thus extract the following observations:

1. For both the De Kinder and the NBIDE data, N-3D generally performed better on the average than I-2D.
2. For the De Kinder data, Firing Pin was usually a better discriminator than Breech Face.
3. For the NBIDE data, Breech Face was a better discriminator than Firing Pin.
4. The worst discriminator case was De Kinder Breech Face.
5. The best discriminator (near perfect) was N-3D on NBIDE Breech Face.

9.4   Matching and Non-matching Distributions of Correlation Scores

The previous sections on system performance analysis have demonstrated performance variations between systems and between imaged regions. For instance, N-3D performed better on the NBIDE breech face impressions than anything else, meaning that the topographic breech face images of casings fired from the same gun almost always correlated more highly with each other than with breech face images of casings fired from different guns. In contrast, for the De Kinder breech face impressions, a casing would often correlate more highly with certain casings fired from other guns than with casings fired from the same guns. In this section, we describe a probabilistic model for such findings.

A pair of casings fired from the same gun is called a "match," and a pair of casings fired from different guns is called a "non-match." Therefore, for each particular casing in the De Kinder set, there are six other casings that produce a match with that casing, and 63 that produce a non-match with that casing. The correlations from the six matching pairs should ideally be considerably higher than the 63 non-matching correlations. Similarly, for each casing in the NBIDE set, there are eight matching correlations, which should be higher than the 99 non-

matching correlations. How well the different technologies succeed in differentiating matches and non-matches will be the topic of this section.

Supposing that a ballistics database contains a casing from the same gun that fired the reference casing, this database search will yield consistently good results only if a pair of objects that should match each other (i.e., fired from same gun) will correlate much more highly than a pair that should not match each other (i.e., fired from different guns). However, as has been seen from the Top Ten results from the previous sections, there can be considerable variability in behavior between guns and even between casings fired from the same gun. For instance, it is possible that two casings fired from different guns may correlate highly with each other, but two other casings fired from the same gun may not correlate as well. Thus, the marks left by guns and the correlations found between images are not deterministic but have random components.

A probabilistic interpretation of these variations is to envision correlation scores of pairs of casings fired from the same gun as being generated by one distribution (which we call the matching distribution), and correlations of casings fired from different guns as following another distribution (the non-matching distribution). These distributions can be purely empirical in nature, rather than having specific parametric forms. A large degree of overlap between these two distributions will result in a significant number of false matches occurring during a database search for a match. If there is almost perfect separation, as shown in Fig. 9-3, then there will be minimal mistakes. Note that for most of this section, the $ACCF_{max}$ values will be scaled so that the maximum (perfect correlation) is 100 %.
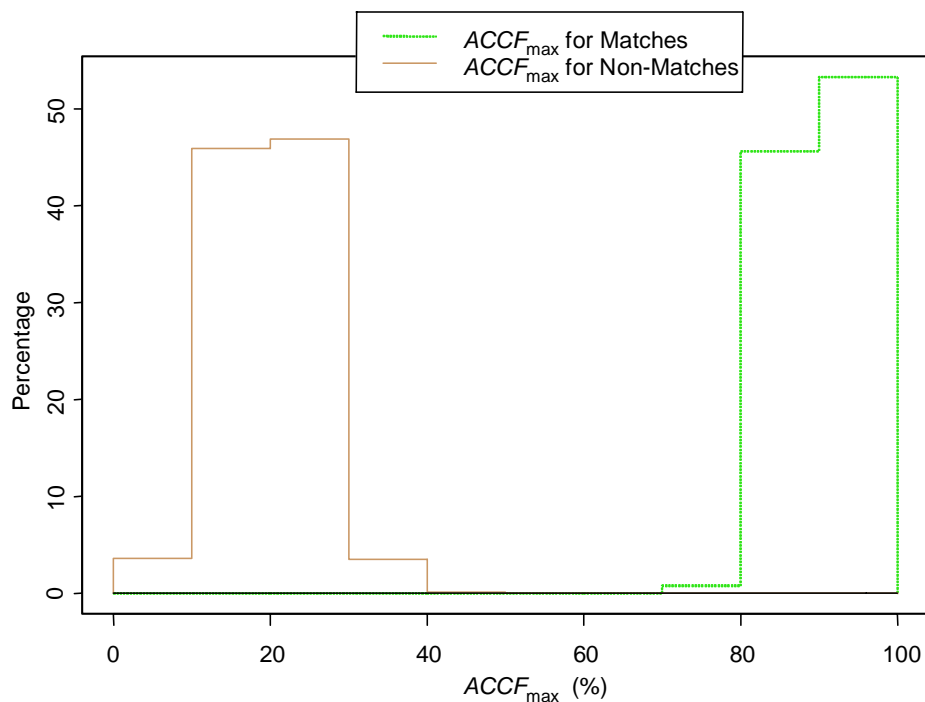

Figure 9-3. Idealized histograms of a matching distribution and a non-matching distribution of cross-correlation scores of ballistics surface topography.

The N-3D data contain the full round-robin of correlations between each pair of casings in the De Kinder 70-casing set, and also between each pair of casings in the NBIDE set. A complete round-robin correlation data set is usually not available, and indeed would not be feasible for much larger data sets because of the computational burden. These data provide an opportunity to explore the respective variabilities of matching and non-matching correlation scores and their dependence on underlying factors such as gun model.

For example, for each of the 70 De Kinder casings, the topography image of its firing pin impression was correlated with that of each of the 69 other casings. It is as if each casing were a reference casing found at a crime scene and were compared with 69 casings in a database (i.e. 69 *compared* casings). For arbitrary casings A and B, we denote $ACCF_{max}(A,B)$ as the maximum *ACCF* value between the reference casing A and the compared casing B. Each such casing pair, A and B, has two associated cross-correlations, since $ACCF_{max}(A,B)$ and $ACCF_{max}(B,A)$ are not necessarily identical (although they are usually very close). Thus, for the 70 casings, there are a total of 70×69 correlations = 4830 correlations that will serve as our basic data set for the De Kinder firing pin analysis. Note that $ACCF_{max}(A,A)$ is presumed to be 100 %.

It is natural to display such round-robin correlation matrix data as a color matrix. For instance, if the 70 De Kinder casings are labeled from 1 to 70, the correlation matrix is depicted as a 70×70 pixel chart, where the color of the pixel in, say, row 31 and column 45, is indicative of the size of the $ACCF_{max}$ between casing 31 as the reference and casing 45 as the compared casing.

Figure 9-4 displays what a fictional hypothetical $ACCF_{max}$ matrix for the De Kinder guns would look like with ideal discrimination properties. Here the casings in the matrix are ordered by reference gun, with the white lines partitioning different guns but grouping together the casings fired by the same gun. Those correlations depicted by the pixels inside the boxes with white numbers are $ACCF_{max}$ values of those casing pairs where both the reference casing and compared casing are fired from the same gun. Hence, those pixels depict the matching scores. The pixels outside the numbered boxes depict the non-matching scores.

For the hypothetical idealized case represented by Fig. 9-4, the matching scores are depicted by the orange and reddish pixels inside the numbered boxes; hence the matching scores are almost all above 70 %. A casing is presumed to correlate perfectly with itself, leading to the dark red line of pixels along the diagonal having $ACCF_{max} = 100$ %. The non-matching scores are represented by the bluish pixels outside the numbered boxes; thus, the matching scores are almost all below 25 %. There appears to be no overlap between matching and non-matching scores, which precludes mistakes by Top Ten selection procedures as discussed in previous sections.
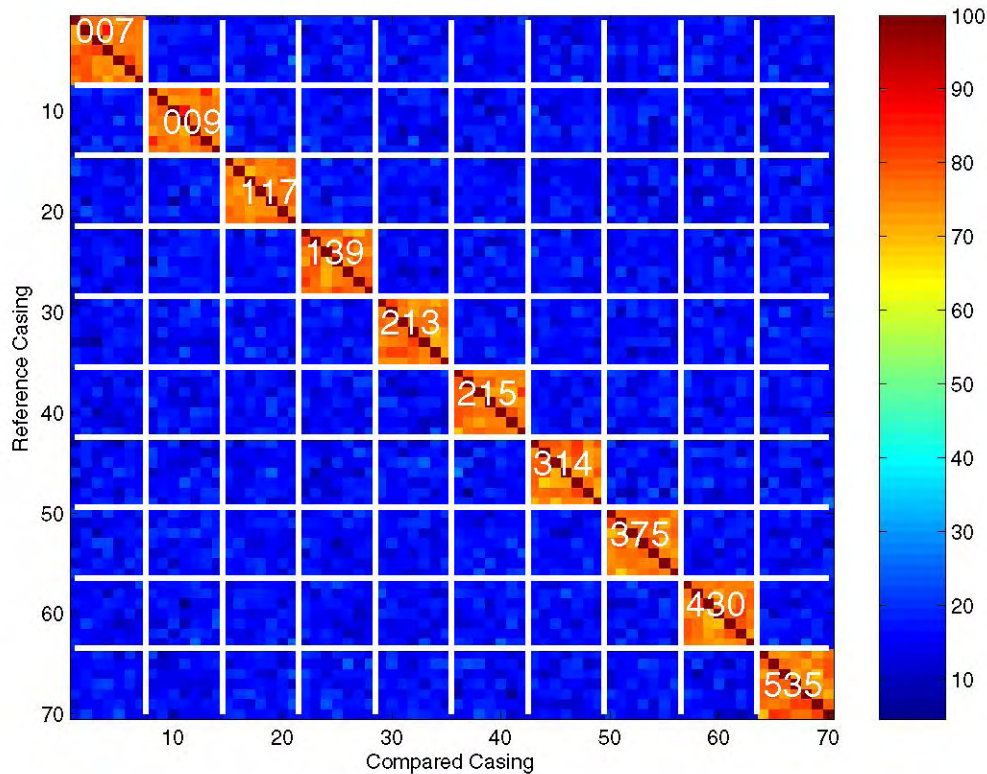
Figure 9-4. $ACCF_{max}$ matrix for the 70 De Kinder casings with ideal discrimination between guns. The color of the pixel in row I and column J indicates the value of the $ACCF_{max}$ (in %) between casing I as the reference casing and casing J as the compared casing. A casing correlates perfectly with itself. The casings in the matrix are ordered by reference gun. The white lines partition the casings into groups that are fired by the same reference gun or compared gun. Those correlations depicted by the pixels inside the boxes with white numbers are $ACCF_{max}$ values of those casing pairs where both the reference casing and compared casing are fired from the same gun. Hence, those pixels depict the matching scores. The pixels outside the numbered boxes depict the non-matching scores.

9.5 Overlap Metric p

There are many possible ways to quantify the degree of separation, or distance, between the matching and non-matching $ACCF_{max}$ distributions. One way we use here is to compute an *overlap metric p,* which is the probability that the $ACCF_{max}$ value of a randomly chosen member of the non-matching distribution is larger than a randomly chosen member from the matching distribution.

Since matching scores should ideally be near 100 % and non-matching correlations should ideally be near 0, the probability that a non-matching score exceeds a matching score should be at or near zero. If the two distributions were the same, then p would be 0.5. Such an overlap metric is commonly used in other fields; and it is referred to in the psychometric literature as the Probability of Superiority [49,50]. It is also related to the area under a Receiving Operating Characteristic (ROC) Curve [51] in the statistics literature.

A simple way to estimate p is to look at all pair-wise comparisons between single observations from each of the matching and non-matching samples and calculate the proportion in which the non-matching observation score is higher than the matching observation score. Such calculations are similar to those used in the Mann-Whitney Test [52]. Unfortunately, it is very difficult to accurately estimate very small p using sample sizes as small as those used in this study, leading to too many estimates of zero for p. Another way to estimate p is to fit a continuous distribution, such as the normal distribution, to the empirical $ACCF_{max}$ distributions. This ameliorates the problem by using tail values of the continuous distribution to provide estimates of very small probabilities. We will do this especially in those cases where there is little overlap, leading to very small estimates of p.

In the following sub-sections, we use the p and average p of $ACCF_{max}$ values grouped by gun as a descriptive method to compare the guns and the imaging methods. Since the behavior of casings fired from even the same gun can differ (as will be seen in some of the $ACCF_{max}$ matrices below), the p's estimated from individual casings will also be explored. Since the sample sizes for individual casings are smaller, fitting the matching and non-matching distributions using parametric distributions makes sense here. Again, the average and median of the casing p's are displayed for descriptive and comparison purposes. How they will be used and combined to predict the performance of the $ACCF_{max}$ measures in Top Ten lists (similar to those produced by I-2D) is a more complicated procedure that will be described in the later sub-section on binomial models.

The calculation of p for different levels of grouping can be visualized in terms of comparisons made within the $ACCF_{max}$ matrix. The hypothetical $ACCF_{max}$ matrix in Fig. 9-4 will be used for an intuitive view of the overlap calculations. We discuss three types of groupings below.

Single p

Suppose the same matching and non-matching distributions can be used for all casings and guns. In terms of the $ACCF_{max}$ matrix in Fig. 9-4, the sample of scores inside all of the numbered boxes constitutes the matching scores, and all those outside the numbered boxes are the non-matching scores. It turns out for this idealized case that all the matching scores are large in magnitude (orange-red colors), while all the non-matching scores are small (bluish color). Unfortunately, there usually will not be this much separation between matching and non-matching scores.

Gun specific p

There can be different matching and non-matching distributions for each gun, resulting in a different p for each gun. Referring back to Fig. 9-4, the matching scores for gun 007 are those represented by the pixels inside the white-bordered box numbered "007"; the corresponding non-matching scores are those depicted by the pixels contained in the nine other white-bordered boxes in the first row (the same level as the box labeled "007"). For gun 007, there are $7 \times 6 = 42$ matching scores (not counting the self-correlations on the main diagonal) and $7 \times 7 \times 9 = 441$ non-matching scores.

Casing specific p

There can be different matching and non-matching distributions for each casing, resulting in a different p for each casing. Referring back to Fig. 9-4, suppose that the pixels on the first row of the matrix depict the $ACCF_{max}$ scores when casing 1 is the reference casing. Thus casing 1 is compared with six other casings fired from the same gun (gun 007), and with $7 \times 9 = 63$ casings fired from different guns. Having only six matching scores can make the estimation of p problematic, especially if p is very small. For this reason, estimating the matching and non-matching distributions here can be useful, although estimating the parameters of a distribution using only 6 observations is problematic as well.

9.6   Data Analysis of Correlation Distributions

For each of the N-3D data sets (De Kinder Firing Pin, De Kinder Breech Face, NBIDE Firing Pin, NBIDE Breech Face), we thus present the data on the round-robin correlations in several stages: first, with a color depiction of the entire data set using a color $ACCF_{max}$ matrix, then by examination of the matching and non-matching empirical distributions at three levels of grouping: 1) overall, 2) by gun, and 3) by casing..

9.6.1 De Kinder Firing Pin Correlation Analysis

This subsection contains analysis of the firing pin image data of the De Kinder casings. Recall that the De Kinder set consists of 70 casings fired from ten guns, with seven casings fired from each gun. Thus, given any particular casing, the other 69 (= 70 − 1) casings include six (= 7 − 1) casings fired from the same gun, and 63 (= 70 − 7) casings that were fired from different guns. Therefore, for each particular casing in the De Kinder set, there are six other casings that produce a match with that casing, and 63 that produce a non-match with that casing. The correlations from the six matching pairs should ideally be considerably higher than the 63 non-matching correlations.

Recall the hypothetical $ACCF_{max}$ matrix of Fig. 9-4 in which there was no overlap between the matching and non-matching scores. Is such clear separation between matches and non-matches present with actual data? Figure 9-5 contains the color depiction of the actual $ACCF_{max}$ matrix for the De Kinder firing pin topography images.
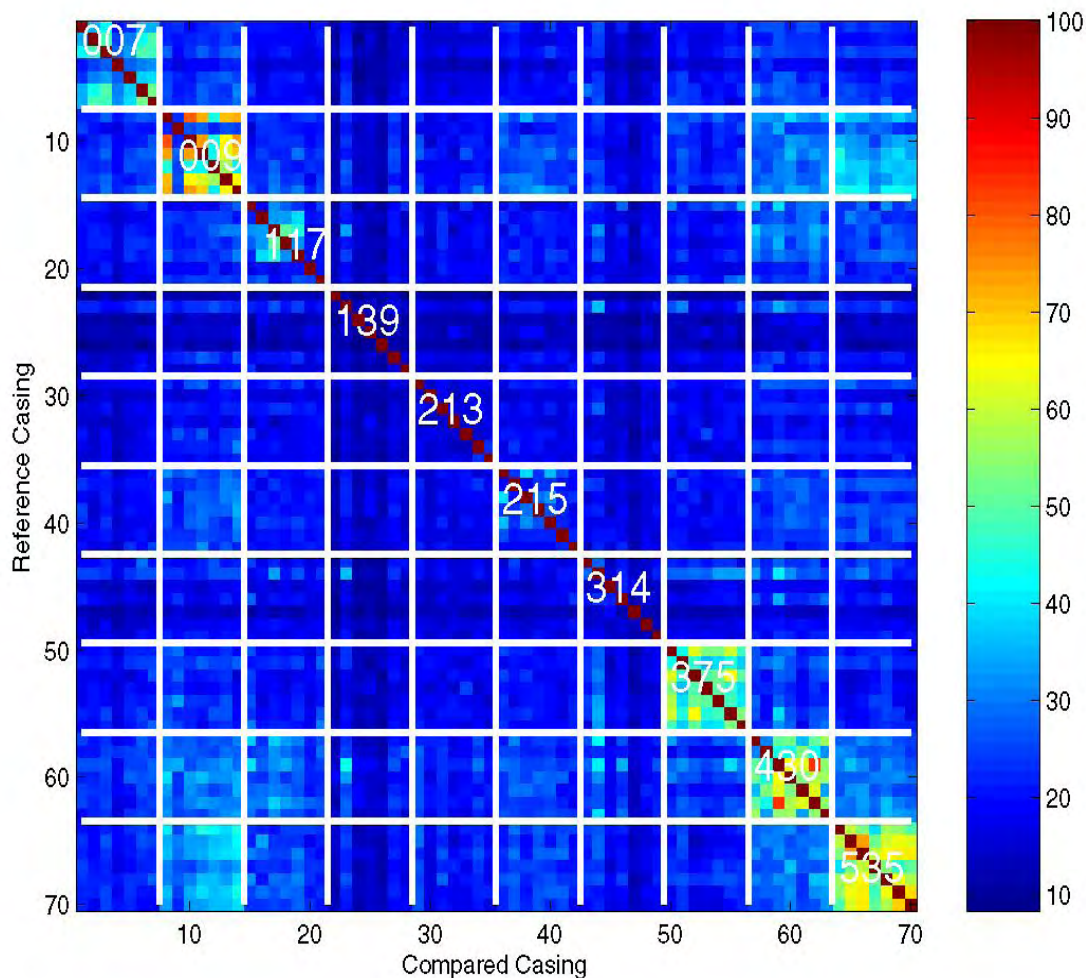
Figure 9-5. $ACCF_{max}$ matrix of the De Kinder firing pin data. The color of the pixel in row *I* and column *J* indicates the value of the $ACCF_{max}$ (in %) between casing *I* as the reference casing and casing *J* as the compared casing. The pixels are ordered by gun ID, and within gun by ammunition in the following order: CCI, Winchester, Speer, Wolf, Federal, Remington, Remington.

Looking at the color $ACCF_{max}$ matrix for the actual data shows that separation between matching and non-matching scores is not close to the ideal situation depicted in the hypothetical $ACCF_{max}$ matrix of Fig. 9-4. There are many casings that should correlate highly with each other that do not, especially for guns 139, 213, 215, and 314. Guns 007, 009, 375, 430 and 535 have much better (but not perfect) separation. One can see many other patterns from the $ACCF_{max}$ matrix. For instance, the Winchester casing from gun 009 does not correlate with the other casings from that gun. The highest non-match scores are between the casings from guns 009 and 535. The scores generated by gun 139 seem particularly low for both matches and non-matches.

Let's try to summarize the data depicted in Fig. 9-5 in useful groupings. How clearly separated are the distributions of the matching and non-matching correlation scores? Is it close to the ideal

situation of Fig. 9-3? Figure 9-6 contains a histogram of the matching scores and a histogram of the non-matching scores in a plot analogous to Fig. 9-3.
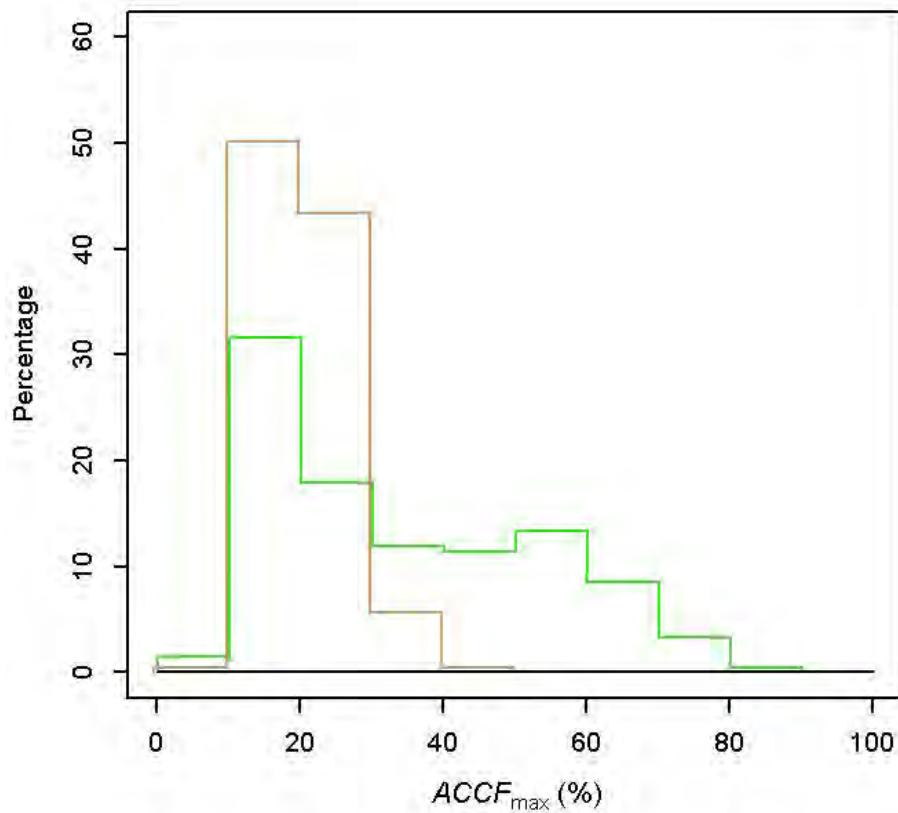

Figure 9-6. De Kinder firing pin correlations: The green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

Instead of the clear separation between distributions of Fig. 9-3, there is a large degree of overlap, suggesting that there would be many misidentifications using these correlation data.

The patterns in Fig. 9-5 as well as the multimodal nature of the matching scores in Fig. 9-6 strongly suggest that the matching and non-matching distributions may be gun-dependent. Figure 9-7 breaks down the results by reference gun, with the overlap metric p given for each grouping.

Figure 9-7. De Kinder firing pin correlations: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric p estimated for that group. In each plot, the green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores. The horizontal scale is the same as Fig. 9-6 (%). The symbol CCF stands for $ACCF_{max}$.

Figure 9-7 shows that while the guns have similar non-matching score distributions, they have differing matching distributions. Guns 139, 213 and 314 have match scores that are even more concentrated in the lower end of the scale than their non-match scores. Other guns show much better, though not perfect, separation.

Table 9-3 lists the overlap metric information for the firing pin data of individual gun groupings ordered by size of overlap metric. Small p values would indicate better discrimination between matching and non-matching distributions.

Table 9-3. Overlap metric p for the topographic signatures of the firing pin impressions of 10 De Kinder guns.

Gun ID	p
375	0.0003
535	0.0012
430	0.0161
007	0.0388
009	0.1250
117	0.3079
215	0.3378
314	0.6032
213	0.6621
139	0.7495
Mean	0.2842

Despite all guns being of the same Sig Sauer model, there are large differences in the overlaps produced by the different guns, with guns 139, 213 and 314 producing so much overlap that they are even worse than random chance. While the other guns have much better separation, the level of wrong matches for them are likely still too high for satisfactory performance in a very large database.

It makes sense to further refine overlap metric results by individual casings rather than just by guns. We will continue to use the overlap metric applied to gun scores as a convenient descriptor of gun properties, but we will also calculate p for individual casings to examine possible performance in very large databases. As an example, a look back at the $ACCF_{max}$ matrix in Fig. 9-5 reveals some casing-specific patterns. For instance, note that inside the boxes labeled "009" and "007", the matching scores are larger than the non-match scores except for the scores associated with particular pixels. Figure 9-8 breaks down the data by casing, by dividing the data into groups by reference casing. Each of the smaller plots depicts for each casing, its correlations with the 6 casings fired from the same gun (matches, lower triangles), and the correlations with the 63 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID and ammunition manufacturer of the reference casing. The main value of these plots (called "strip plots") is to see the degree of overlap between the six triangles depicting the matching scores with the mass of points representing the non-matching scores. Does the overlap vary greatly among casings fired from the same gun (which are on the same row of the diagram)? Again, ideally there should be clear separation between the matching and non-matching scores.

From Fig. 9-8, it can be seen that guns 375 and 535 produce the greatest separation. Note that because the triangles in Fig. 9-8 have non-zero width, some of the casing plots may give the impression of somewhat more overlap than actually exists.
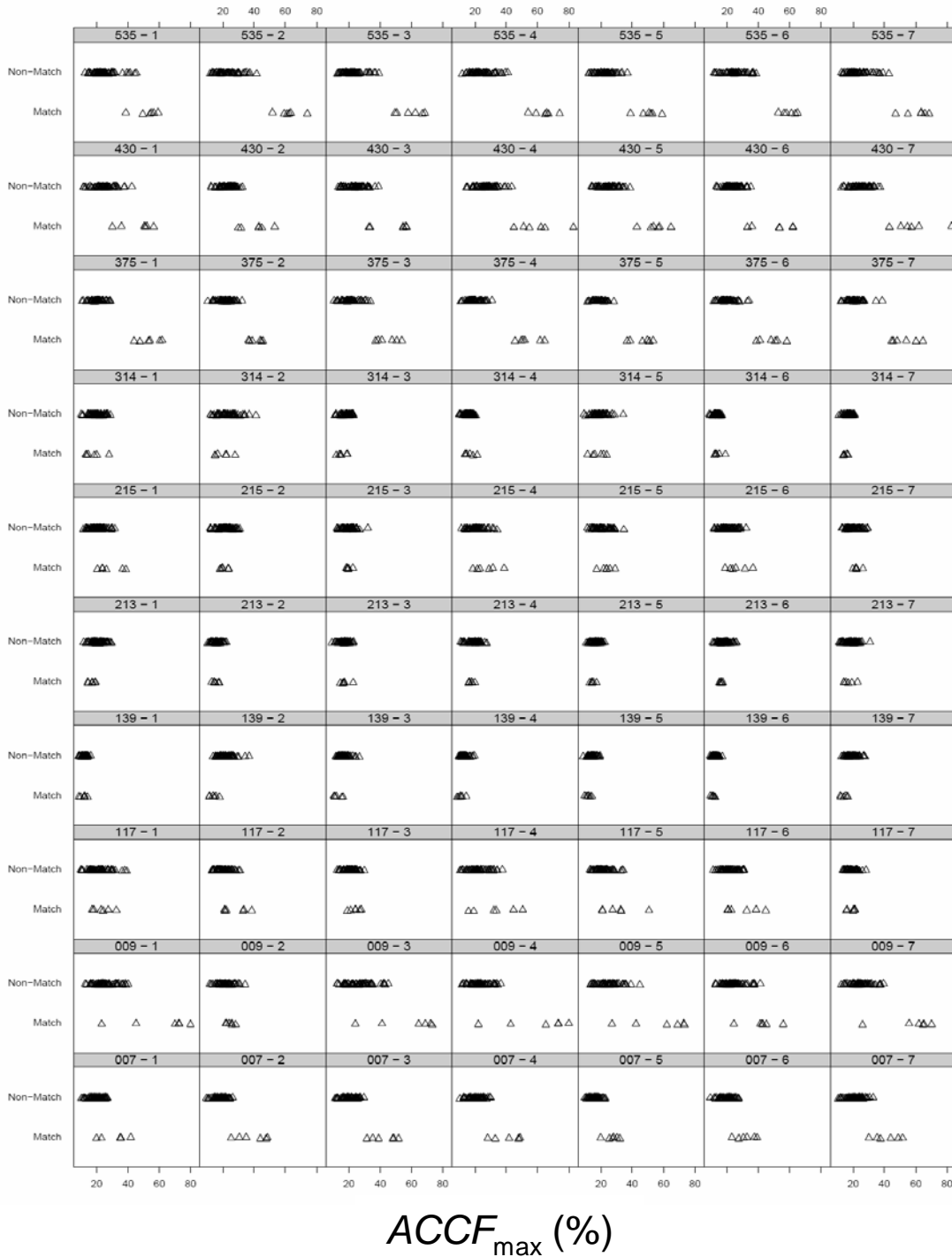
Figure 9-8. De Kinder Firing Pin: Correlations for individual casings. The above figure plots for each casing its correlations with the six casings fired from the same gun (matches, lower triangles), and the correlations with the 63 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID and ammunition manufacturer (1-CCI, 2-Win, 3-Rem, 4-Speer, 5-Wolf, 6-Fed, 7-Rem) of the reference casing.

For each of the 70 casings, an overlap metric for the matching and non-matching correlation scores produced by that casing can be estimated by looking at all pair-wise comparisons between

matching and non-matching correlations for it. The histogram in Figure 9-9 shows the empirical distribution of those overlap metrics. How do these estimates compare with estimates by gun?

The histogram in Fig. 9-9 shows that while some casings produce small overlaps, most produce substantial overlaps between matching and non-matching correlation scores that would lead to mistakes in a large database search.
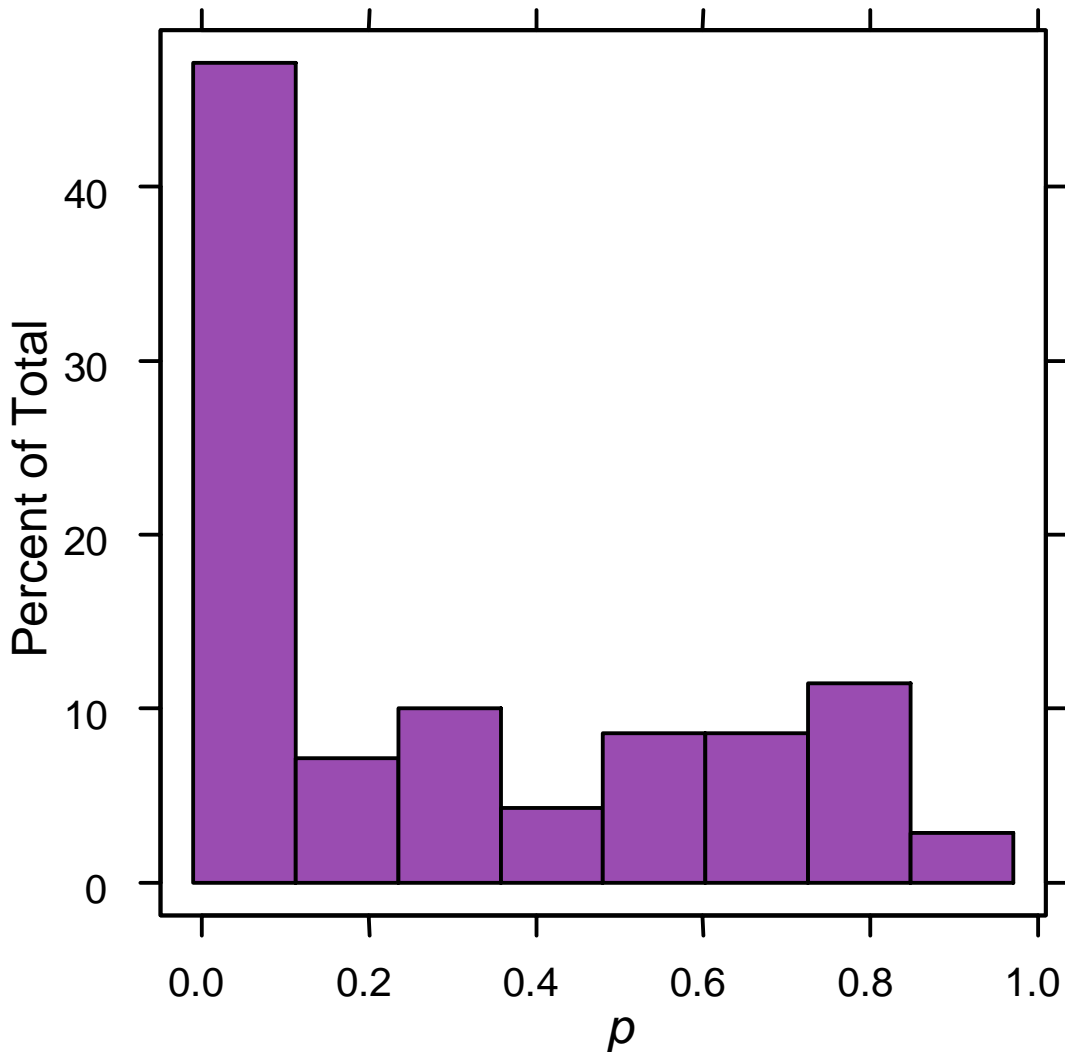

Figure 9-9. Histogram of the overlap metric p estimated for each of the 70 De Kinder casings using topographic signatures of the firing pin impressions and pair-wise comparison methods. The mean p is 0.29 and the median is 0.17. While 24 % of the estimates are zero, there are a considerable proportion of large estimates over 0.3.

9.6.2 De Kinder Breech Face Correlation Analysis

Figure 9-10 contains the $ACCF_{max}$ matrix for the topographic breech face analyses for the De Kinder casings. How does the pattern of separation and overlap between matching and non-matching scores compare with the firing pin data?
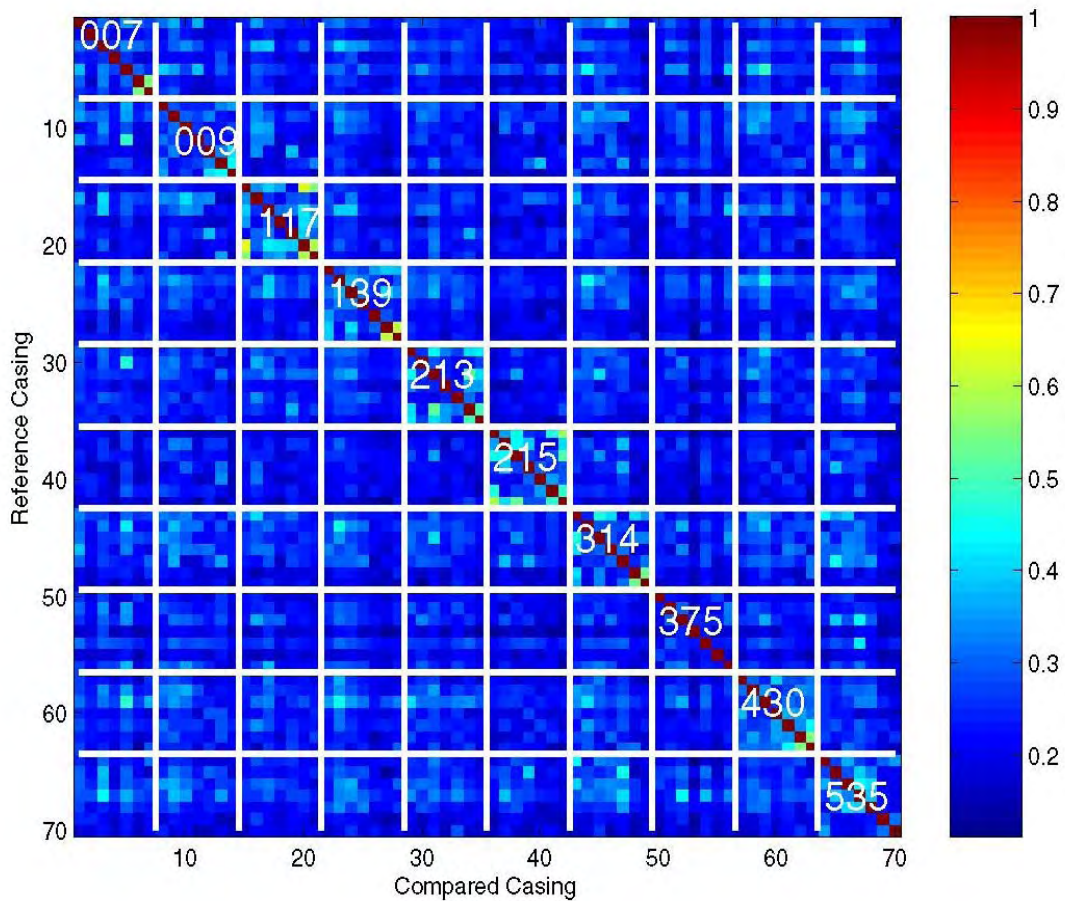
Figure 9-10. $ACCF_{max}$ matrix of the De Kinder breech face data. The color of the pixel in row *I* and column *J* indicates the value of the $ACCF_{max}$ between casing *I* as the reference casing and casing *J* as the compared casing. The pixels are ordered by gun ID, and within gun by ammunition in the following order: CCI, Winchester, Speer, Wolf, Federal, Remington, Remington.

It appears that most $ACCF_{max}$ values, including both matching scores and non-matching scores are in the 20 % to 40 % range, indicating more overlap and less separation than was seen with the firing pins. The green-colored pixels in the lower right corner of most boxes on the diagonal show that the breech faces from the two Remington casings are correlating more highly with each other than with the other casings fired from the same guns.

Figure 9-11 depicts histograms of the matching and non-matching scores for the De Kinder breech face impressions. What is the degree of overlap or separation between the two distributions?
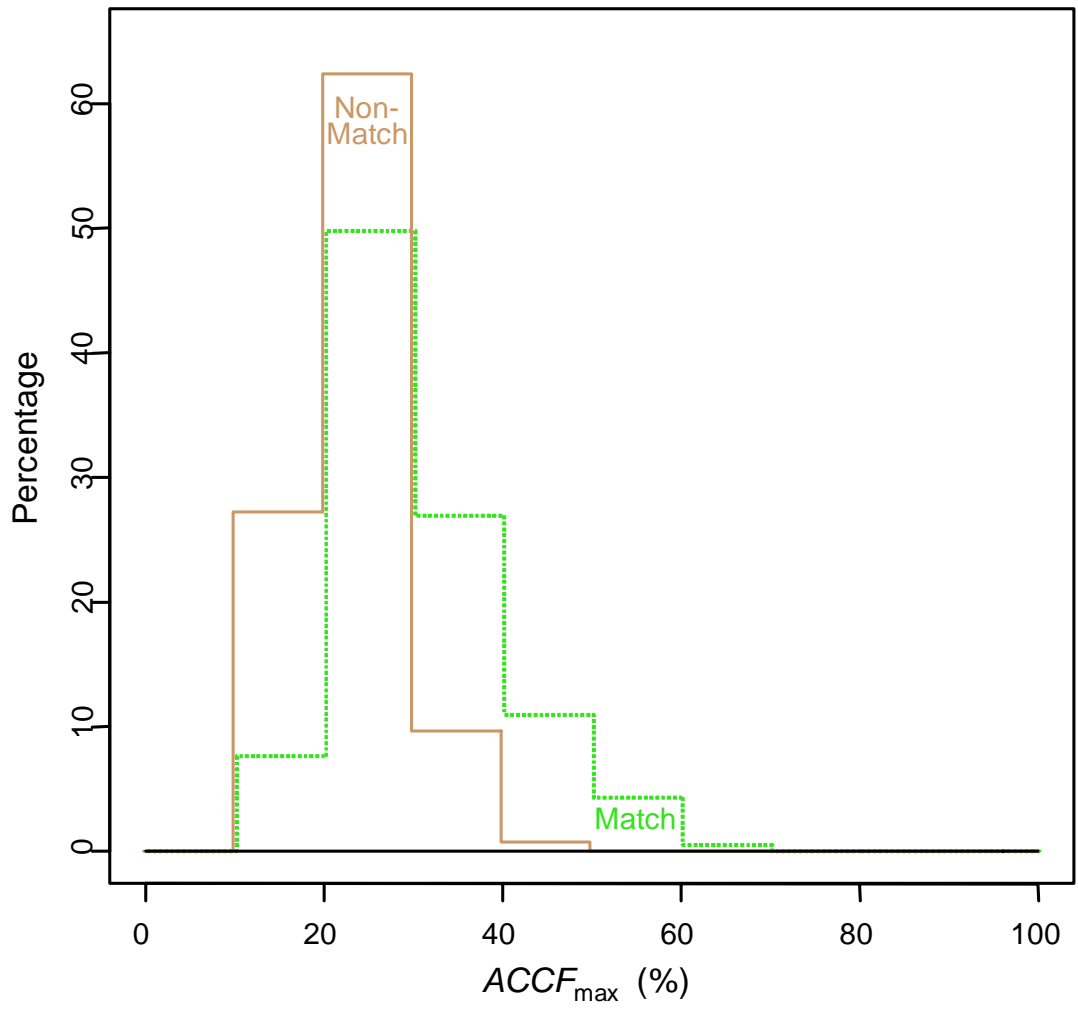
Figure 9-11. De Kinder breech face correlations: The green dashed lines depict a histogram of the matching scores, while the brown solid lines depict a histogram of the non-matching scores.

Figure 9-11 shows that there is a considerable degree of overlap between the matching and non-matching scores, which is not ideal. Figure 9-12 breaks down these results by reference gun, with the overlap metric p given for each grouping. For the firing pin impressions, there were considerable performance variations between guns. Does the same hold true for breech faces?
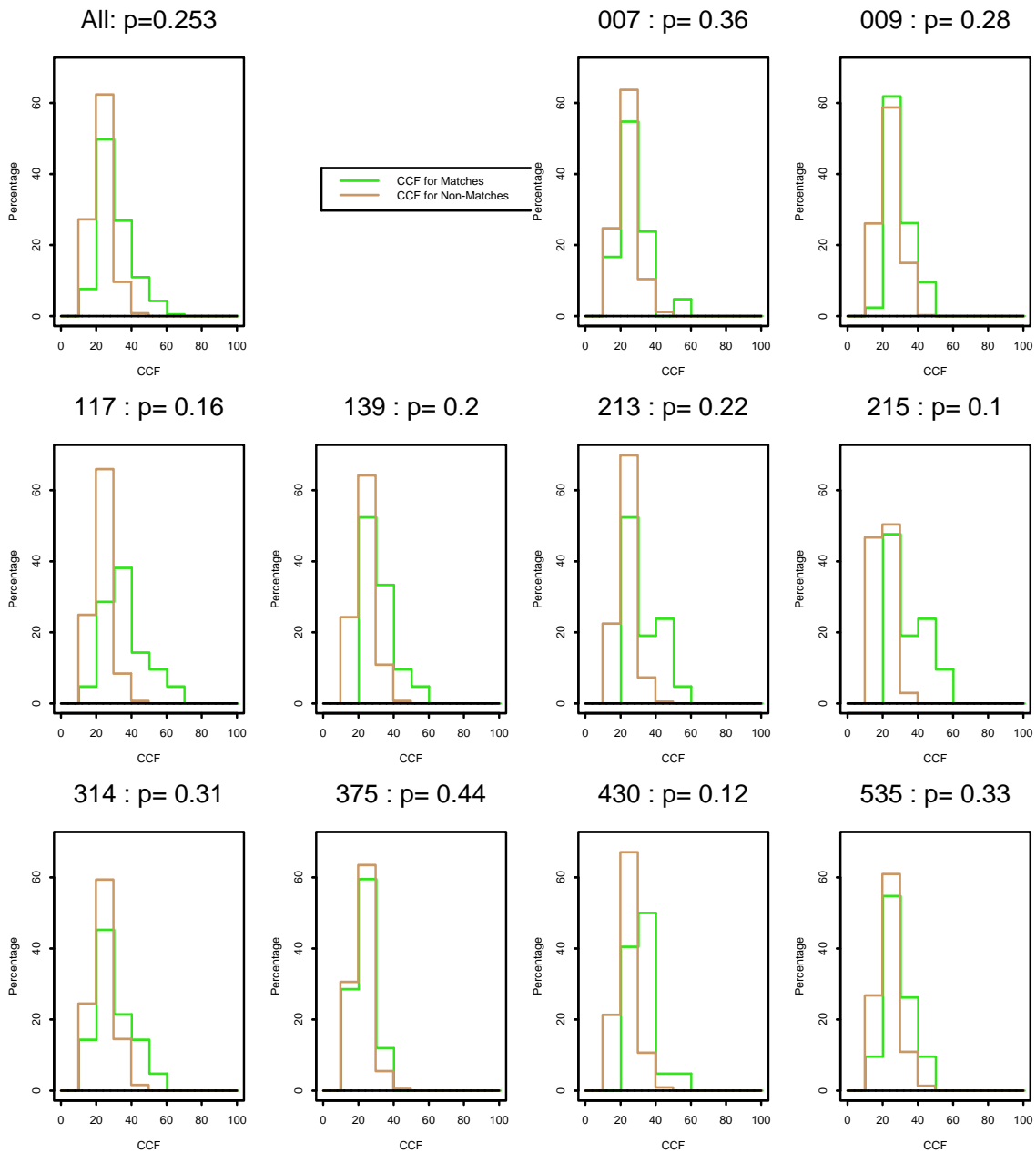
Figure 9-12. De Kinder breech face correlations: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric *p* estimated for that group. In each plot, the green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores. The symbol CCF stands for $ACCF_{max.}$ The horizontal scale for $ACCF_{max}$ is the same as Fig. 9-6 (%).

No gun has good separation between the matching and non-matching distributions. Also for the De Kinder breech face distributions in Fig. 9-12, we find less variation between the guns than we did for the De Kinder firing pin data. All the non-matching distributions look similar.
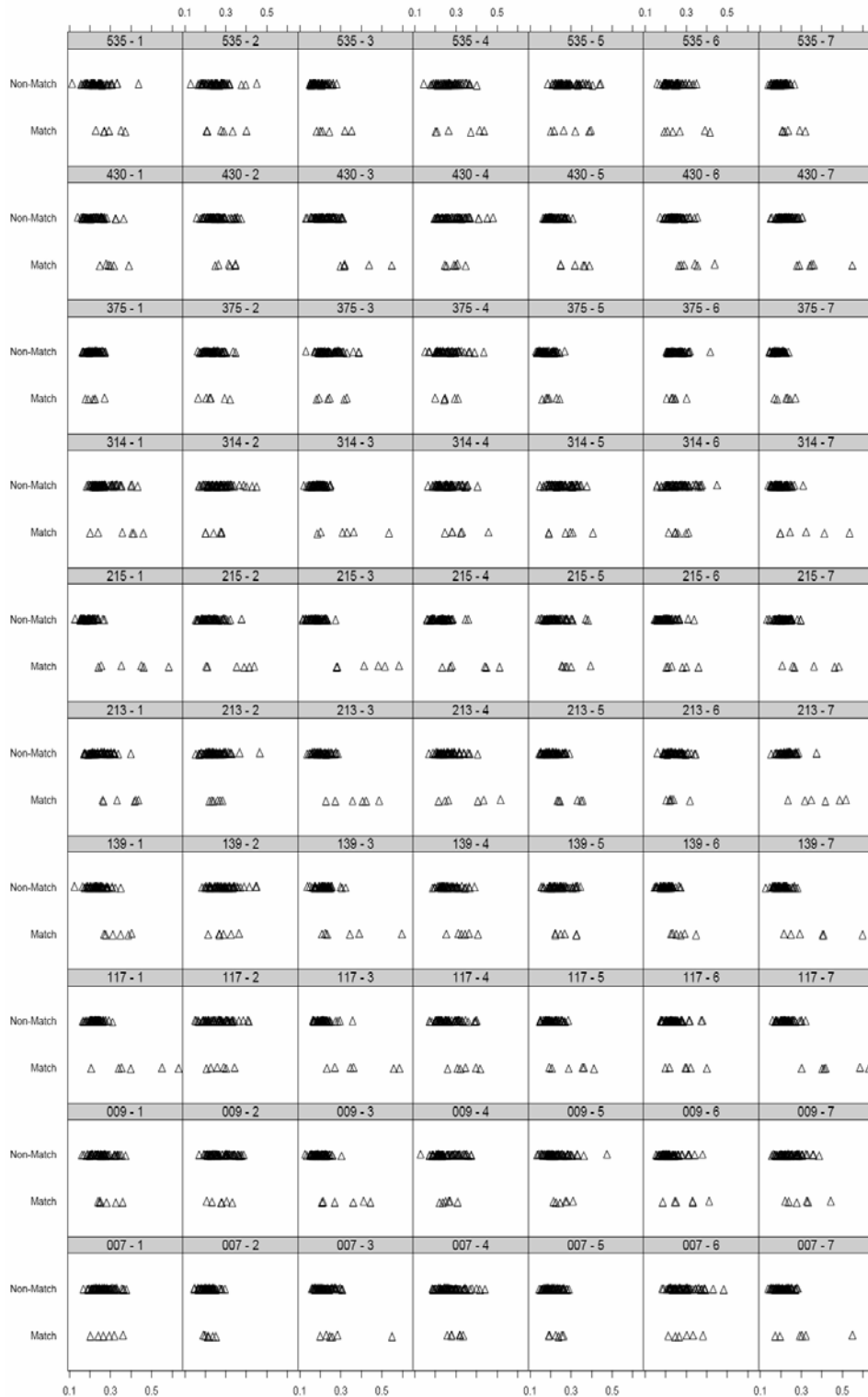
Table 9-4 contains the overlap metric statistics for the breech face data, ordered by performance of the gun. Small *p* values indicate better separation between matching and non-matching distributions. None of the guns have a large degree of separation. Even the gun with the most separation (215) still would produce many errors in a large database search scenario.

Table 9-4. Overlap metric *p* for the topographic signatures of the breech face impressions of 10 De Kinder guns.

Gun ID	*p*
215	0.098
430	0.124
117	0.158
139	0.201
213	0.219
009	0.283
314	0.314
535	0.332
007	0.358
375	0.442
Mean	0.253

Figure 9-13 breaks down the data more finely by dividing the data into groups by reference casing. Each of the smaller plots depicts, for each casing, its correlations with the six casings fired from the same gun (matches, lower triangles), and the correlations with the 63 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID and ammunition of the reference casing. Although there is some variation in the matching distributions, there is a consistently high degree of overlap or lack of separation between matching and non-matching casings for nearly all reference casings. From Fig. 9-13, it can be seen that only one casing, 215-3, has good separation between matching and non-matching scores, and even for that case, there is a very small margin between the smallest matching score and the largest non-matching score. Some of the other casings from gun 215 are among the next best performers in terms of separation between matching and non-matching scores.

For each of the 70 casings, an overlap metric for the matching and non-matching correlation scores produced by that casing can be estimated by looking at all pair-wise comparisons between matching and non-matching correlations for each casing. The histogram in Fig. 9-14 shows the empirical distribution of those overlap metrics. While some casings produce small overlaps, most produce substantial overlaps between matching and non-matching correlation scores that would lead to mistakes in a large database search.

Figure 9-13. De Kinder Breech Face: The above figure plots, for each casing, its correlations with the six casings fired from the same gun (matches, lower triangles) , and the correlations with the 63 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID and ammunition manufacturer (1-CCI, 2-Win, 3-Rem, 4-Speer, 5-Wolf, 6-Fed, 7-Rem) of the reference casing.
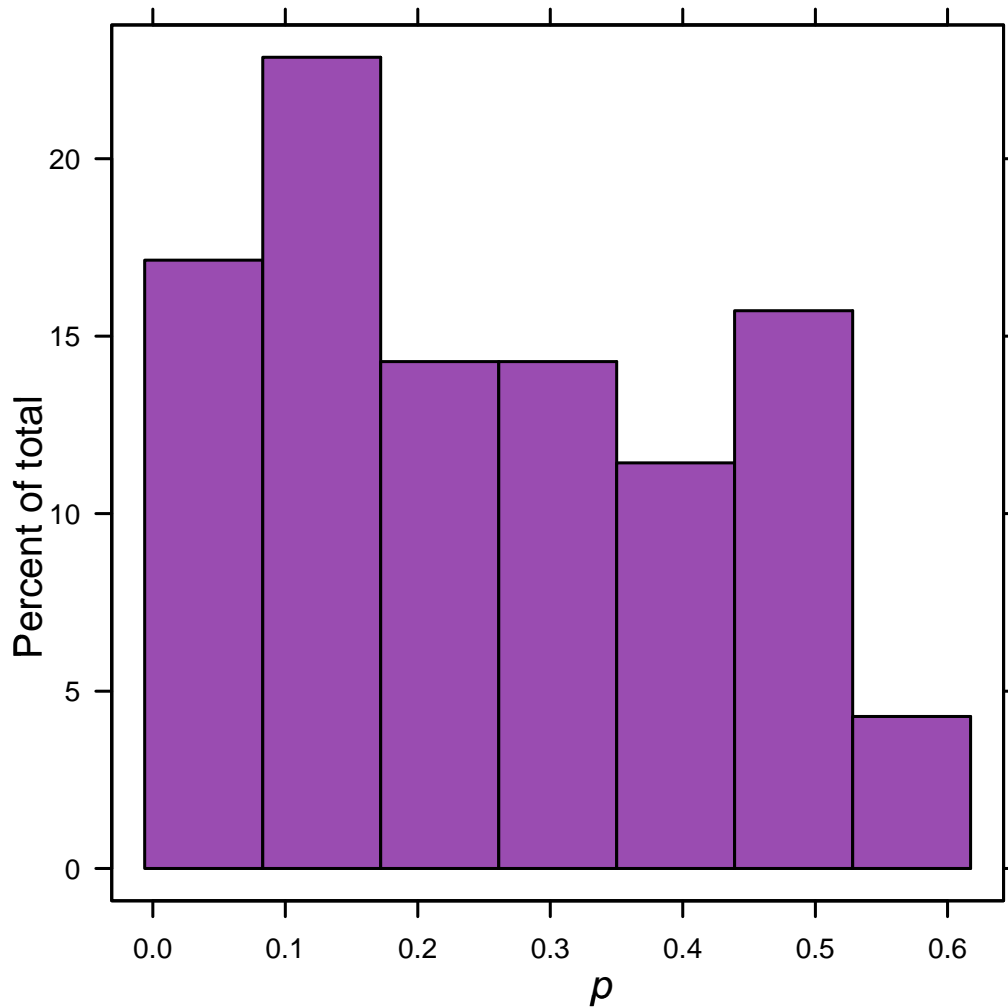
97

Figure 9-14. Histogram of the overlap metric *p* estimated for each of the 70 De Kinder casings using topographic signatures of the breech face impressions and pair-wise comparison methods. The mean is 0.26 and the median is 0.23.

9.6.3 De Kinder: Combining Firing Pin and Breech Face Results

A combination of the correlation metrics from both regions of the casings, e.g., firing pin and breech face, should perform better than the use of one region alone. There has been similar research on multi-modal biometrics, i.e. combining several fingerprints or combining face and fingerprint algorithms [53]. There are many possible methods of combining metrics, but we will just look at two of the simplest here. In the De Kinder article [7], casings that make the Top Ten of *either* I-2D breech face or I-2D firing pin correlations are included in the combined list. Such a combined list obviously must do at least as well as either metric alone, although it results in potentially twice as many candidate casings requiring manual examination. Most other combination schemes require more information than just Top Ten lists.

For example, if the $ACCF_{max}$ values are available, they can be combined in a multitude of ways. The simplest way is first to normalize the metrics, then add them together. Here, since the

$ACCF_{max}$ values are bounded by a perfect score of 100 %, we can produce a combined correlation metric by merely averaging the breech face and firing pin $ACCF_{max}$ values for each casing pair. Of course, the average and sum produce the same results, so here we call it the Sum Method, which is more prevalent in the literature. Unlike the 'make either list' method above, the Sum Method can perform worse for a particular casing than either or both of the constituent methods. Table 9-5 shows the results of the two combined metrics for the De Kinder casings.

Of course, using both metrics results in more, or at least no fewer, correct matches. The number of correct matches using both metrics cannot be more than the sum of correct matches for either, and it is equal to the sum only if there is no overlap. Thus, for I-2D, the strong measure (Firing Pin) cannot be helped much here by the weak method (Breech Face), especially since firing pin impressions caught all 6 correct matches for several casings. The 3D breech face impressions were more of a help to the 3D firing pin impressions.

Table 9-5. Results combining breech face impressions and firing pin impressions for 70 De Kinder casings. Columns 3 and 4 – number of correct matches appearing in the Top Ten List of either the breech face or the firing pin impressions. Column 5 – number of correct matches appearing in the Top Ten list for the sum of $ACCF_{max}$ values for the breech face and firing pin impressions.

Number of Correct matches (Max 6)

Ref.Casing	Ammo	I-2D Either List	N-3D Either List	N-3D Sum
Means		3.39	4.77	4.23
007-01	CCI	4	5	4
007-02	WIN	5	6	5
007-03	R-P	6	6	6
007-04	SPEER	6	6	5
007-05	WOLF	6	5	6
007-06	FC	6	6	4
007-07	R-P.sep	6	6	5
009-01	CCI	3	5	5
009-02	WIN	0	1	0
009-03	R-P	2	5	5
009-04	SPEER	4	5	5
009-05	WOLF	4	5	5
009-06	FC	5	6	5
009-07	R-P.sep	5	5	5
117-01	CCI	3	5	4
117-02	WIN	0	4	3
117-03	R-P	4	6	6
117-04	SPEER	5	6	3
117-05	WOLF	6	6	5

117-06	FC	5	5	3
117-07	R-P.sep	3	6	6
139-01	CCI	3	4	3
139-02	WIN	3	1	0
139-03	R-P	4	3	3
139-04	SPEER	4	3	2
139-05	WOLF	2	2	1
139-06	FC	2	4	3
139-07	R-P.sep	3	5	3
213-01	CCI	2	4	3
213-02	WIN	4	0	0
213-03	R-P	5	5	4
213-04	SPEER	3	3	3
213-05	WOLF	3	3	3
213-06	FC	1	1	1
213-07	R-P.sep	2	5	4
215-01	CCI	2	6	6
215-02	WIN	2	4	4
215-03	R-P	3	6	6
215-04	SPEER	1	6	5
215-05	WOLF	2	3	3
215-06	FC	2	5	4
215-07	R-P.sep	2	5	3
314-01	CCI	4	5	3
314-02	WIN	2	0	0
314-03	R-P	5	4	4
314-04	SPEER	3	4	2
314-05	WOLF	4	2	1
314-06	FC	4	2	0
314-07	R-P.sep	4	4	3
375-01	CCI	4	6	6
375-02	WIN	3	6	5
375-03	R-P	4	6	5
375-04	SPEER	5	6	6
375-05	WOLF	5	6	6
375-06	FC	4	6	6
375-07	R-P.sep	4	6	6
430-01	CCI	3	6	6
430-02	WIN	0	6	6
430-03	R-P	5	6	6
430-04	SPEER	3	6	6
430-05	WOLF	3	6	6

430-06	FC	3	6	6
430-07	R-P.sep	5	6	6
535-01	CCI	1	5	6
535-02	WIN	4	6	6
535-03	R-P	3	6	6
535-04	SPEER	3	6	6
535-05	WOLF	2	6	6
535-06	FC	2	6	6
535-07	R-P.sep	2	6	6

Table 9-6 shows the overlap metric statistics for the sum method of N-3D firing pin and breech face impressions.

Table 9-6. Overlap metric p for the sum of the $ACCF_{max}$ values for topographic signatures of the breech face impressions and firing pin impressions of 10 De Kinder guns, ordered by value of overlap metric.

Gun ID	p
535	0.0013
375	0.0056
430	0.0082
007	0.0706
117	0.0820
215	0.0843
009	0.0981
213	0.3303
139	0.3513
314	0.3660
Mean	0.140

The mean p was 0.284 for N-3D firing pins and 0.253 for N-3D breech faces, so the sum measure does somewhat better than either of the two measures. A comparison of the results of Table 9-1 and Table 9-5 shows that this is consistent with the marginal improvement of the mean number of correct matches in the N-3D top ten lists from 2.8 (breech face) and 3.3 (firing pin) to 4.2 (sum).

Although all the $ACCF_{max}$ values are on nominally the same scales, the firing pin $ACCF_{max}$ values are more spread out from 0-100 and thus tend to dominate the sum. This domination can be lessened by a different normalization scheme, e.g., using a z-score that divides by the standard deviation. However, that normalization is more difficult to apply because it requires knowledge of the population or the sample of scores to obtain the standard deviations.

Although it may be true that the use of more and more measures to be combined may always improve performance, there may also be diminishing returns, where each added measure adds only a marginal amount of improvement that may not be worth the extra effort. This will be especially true if one measure dominates or performs better than the others, or if there is dependence between the measures. If one measure performs extremely well by itself, combining it with much weaker measures may not help much.

9.6.4 NBIDE Firing Pin Correlation Analysis

Recall that the NBIDE set contains 108 casings fired from 12 guns, with nine casings fired from each gun. Thus, for any particular casing in the NBIDE set, there are eight (=9 – 1) casings fired from the same gun and 99 (=108 – 9) casings fired from different guns. Therefore, for each particular casing in the NBIDE set, there are eight other casings that produce a match with that casing, and 99 casings that produce a non-match with that casing. The correlations from the eight matching pairs should ideally be considerably higher than the correlations from the 99 non-matching pairs.

Figure 9-15 depicts the $ACCF_{max}$ matrix of the NBIDE firing pin data, where the casings are ordered by gun (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401) and within gun by ammunition (1-Win, 2-Rem, 3-PMC), and within ammunition by repetition number (RR#). Recall that there are four guns of each brand; each gun fires three shots of each of the three ammunition brands for a total of $4 \times 3 \times 3 \times 3 = 108$ firings.

Recall the $ACCF_{max}$ matrices for the De Kinder set (Figs. 9-5 and 9-10) and remember how greatly they differed from the "ideal" fictional case of Fig. 9-4. How do the results for the NBIDE firing pin images in Fig. 9-15 compare? It can be seen that most of the non-matching correlations are quite small in magnitude (bluish). Some of the guns have matching correlations that are higher in magnitude (greenish or orange). It looks more like the pattern of the ideal $ACCF_{max}$ matrix in Fig. 9-4, though still far from ideal.

Figure 9-16 depicts histograms of the matching and non-matching scores for the NBIDE firing pin N-3D data. It shows that there is still a considerable degree of overlap between the matching and non-matching scores, which is not ideal.
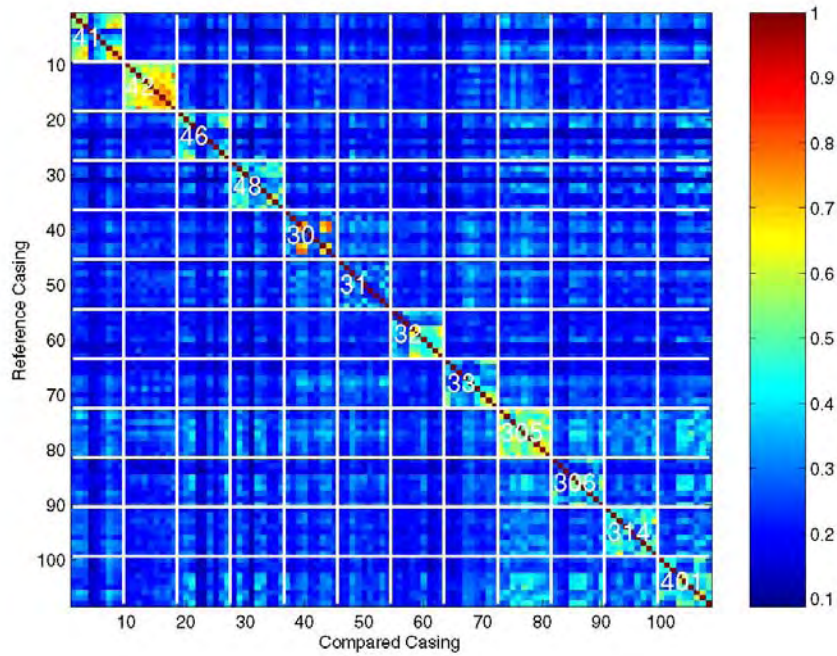
Figure 9-15. $ACCF_{max}$ matrix of the NBIDE firing pin N-3D data. The color of the pixel in row I and column J indicates the value of the $ACCF_{max}$ between casing I as the reference casing and casing J as the compared casing. The pixels are ordered by gun ID (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401) with Rugers in the upper left and S&W's in the lower right), and within gun by ammunition (1-Win, 2-Rem, 3-PMC), and within ammunition by RR#.

Figure 9-16. N-3D data for NBIDE firing pin impressions: The green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

Figure 9-17 breaks down these results by reference gun, with the overlap metric *p* given for each grouping. For the De Kinder firing pin impressions, there was considerable performance variation between guns. Does the same hold true for the NBIDE firing pin impressions?
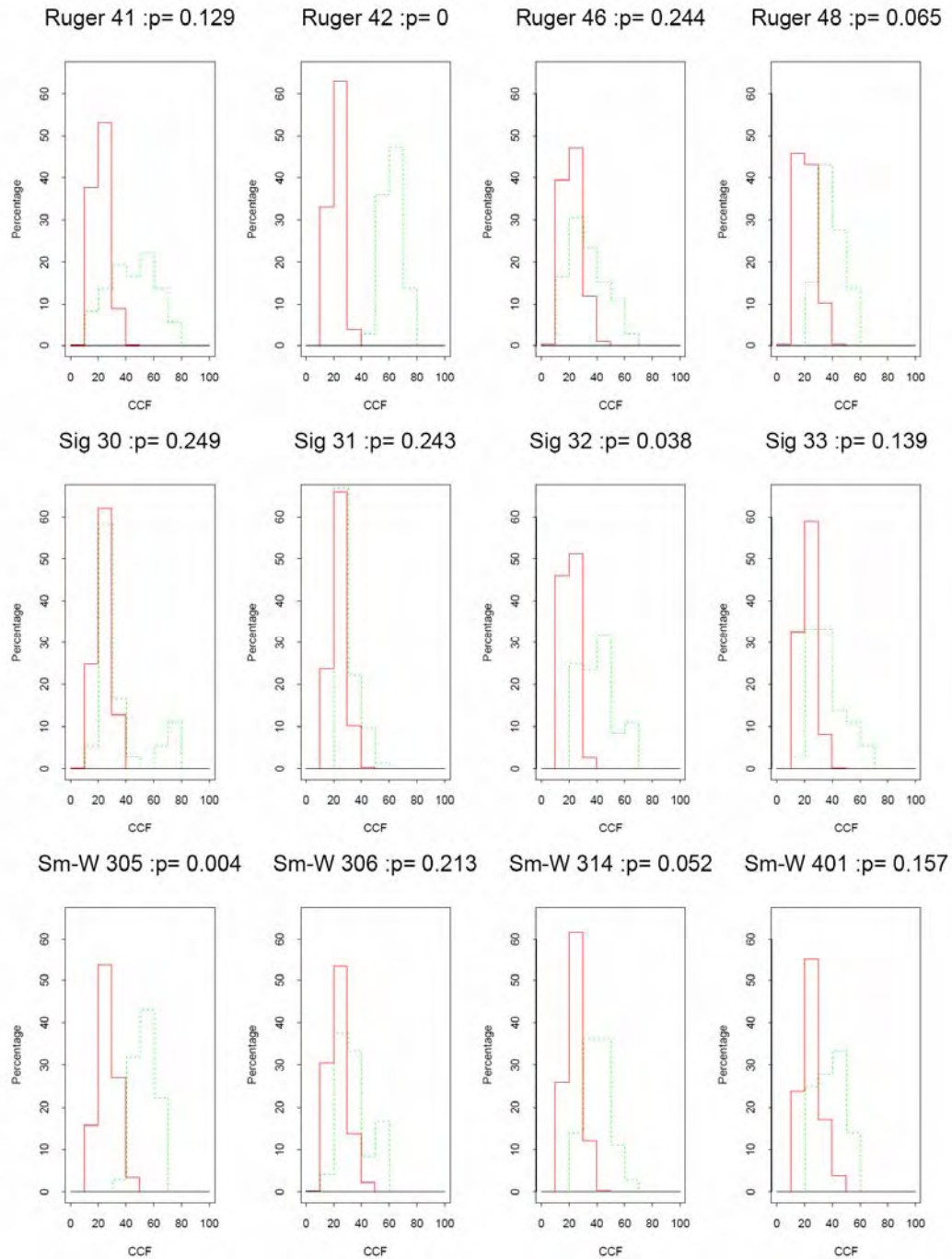

Figure 9-17. NBIDE firing pin N-3D data: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric *p* estimated for that group. In each plot, the green dotted lines depict a histogram of the matching scores, while the brown solid lines depict a histogram of the non-matching scores. The horizontal scale for $ACCF_{max}$ is the same as Fig. 9-6 (%).

Figure 9-17 shows considerable differences between guns. While the non-matching distributions appear similar for each gun, the matching distributions are quite different for each. There is variability within the gun brands: each brand (depicted by row) has one gun that has much more separation than the others of the same brand (Ruger 42, Sig Sauer 32, and S&W 305); and each brand has at least one gun with large overlap and p metric greater than 0.2.

Table 9-7 contains the overlap metric statistics for the firing pin data ordered by performance of the gun as estimated by the overlap metric.
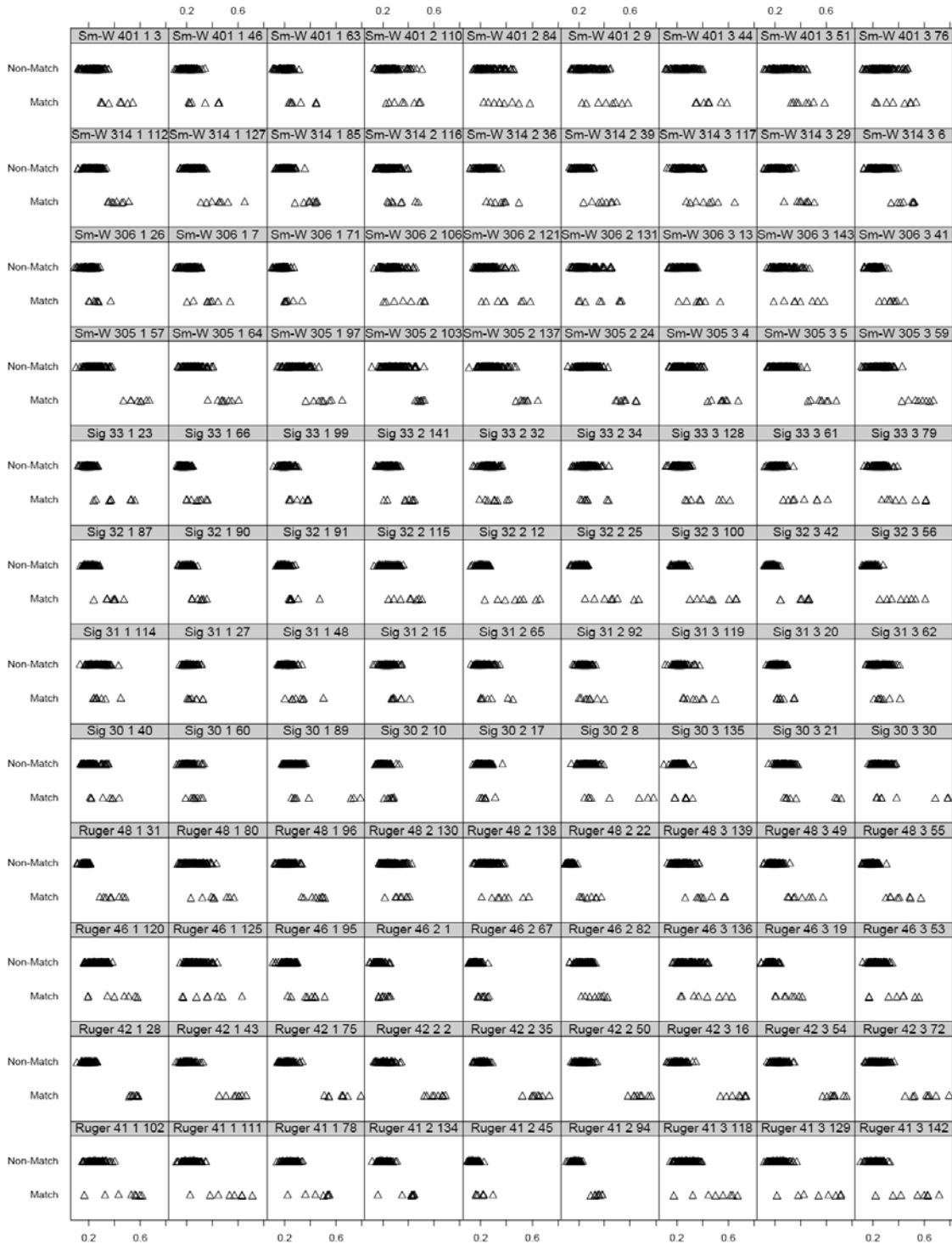
Table 9-7. Overlap metric p for the $ACCF_{max}$ values for topographic signatures of the firing pin impressions of 12 NBIDE guns.

Gun ID	p
Ruger 42	0
S&W 305	0.004
Sig Sauer 32	0.038
S&W 314	0.052
Ruger 48	0.065
Ruger 41	0.129
Sig Sauer 33	0.139
S&W 401	0.157
S&W 306	0.213
Sig Sauer 31	0.243
Ruger 46	0.244
Sig Sauer 30	0.249
Mean	0.128

There are considerable differences within guns of the same brand.

Figure 9-18 breaks down the data even more finely, by dividing the data into groups by reference casing. Each of the smaller plots depicts for each casing, its correlations with the eight casings fired from the same gun (matches, lower triangles), and the correlations with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID #, ammunition #, and RR # of the reference casing. How does the degree of overlap or separation between matching and non-matching vary among casings fired from the same gun, which are shown on the same row in the diagram?

From Fig. 9-18, it can be seen that Ruger 41 is not consistent because the row for this gun contains some casings that possess separation between matching and non-matching scores and other casings that have considerable overlaps. For the other guns, there is a lesser but still considerable variability within casings fired from the same gun.

Figure 9-18. NBIDE Firing Pin N-3D data: Correlations for each casing with the eight casings fired from the same gun (matches, lower triangles), and with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID and ammunition manufacturer (1-Win, 2-Rem, 3-PMC), and RR# of the reference casing

106

For each of the 108 casings, an overlap metric for the matching and non-matching correlation scores produced by that casing can be estimated by looking at all pair-wise comparisons between matching and non-matching correlations for each casing. The histogram in Fig. 9-19 shows the empirical distribution of those overlap metrics. How do these estimates compare with the estimates by gun?

From Fig. 9-19, it can be seen that the NBIDE firing pins produce greater separation between matching and non-matching distributions than do the De Kinder firing pins, but most still have a degree of overlap that will produce mistakes in a large database scenario.
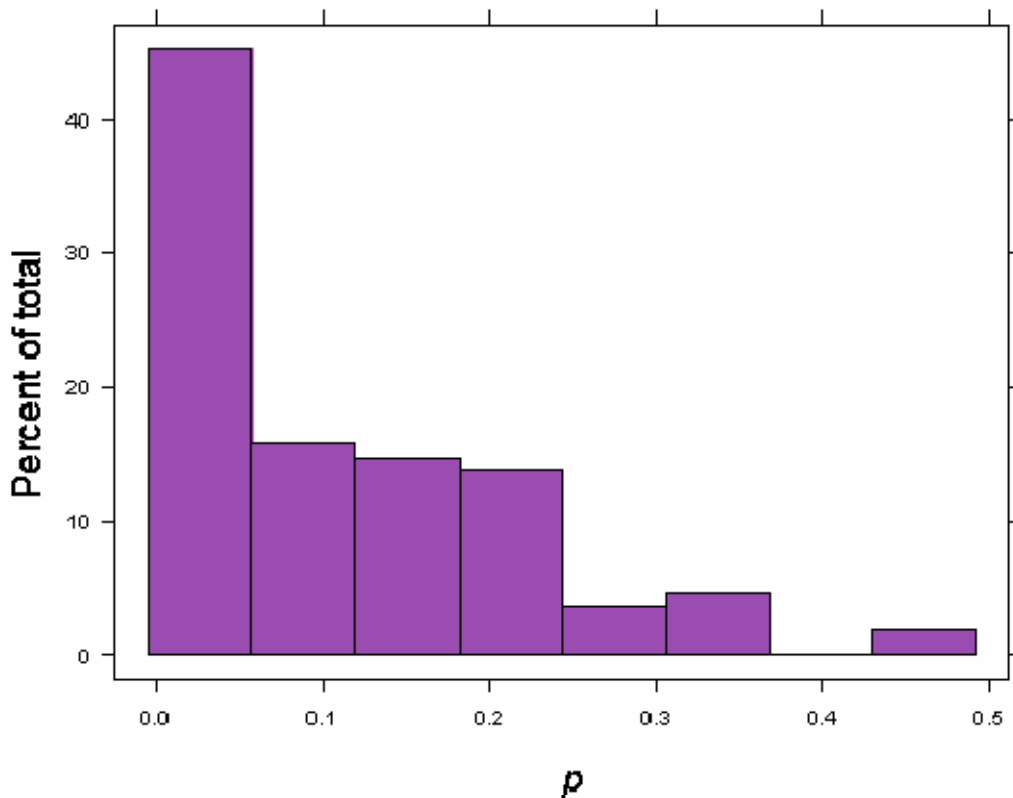

Figure 9-19. NBIDE Firing Pin N-3D data: Histogram of the overlap metric p estimated for each of the 108 NBIDE casings using pair-wise comparison methods. The mean value of p is 0.11 and the median value of p is 0.08. About 75% of the values are larger than 0.01.

9.6.5 NBIDE Breech Face Correlation Analysis

Figure 9-20 depicts the $ACCF_{max}$ matrix of the NBIDE breech face impression data, where the casings are ordered by gun (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401), and within gun by ammunition ((1-Win, 2-Rem, 3-PMC), and within ammunition by RR#. Recall that there are four guns of each brand; each gun fires three shots each of each of the three ammunition brands for a total of 4×3 ×3×3=108 firings.

Recall the $ACCF_{max}$ matrices for the De Kinder set (Fig. 9-5 and 9-10) and for the NBIDE firing pin impressions (Fig. 9-15) in the previous section. How do the results for the NBIDE breech face N-3D images in Fig. 9-20 compare?
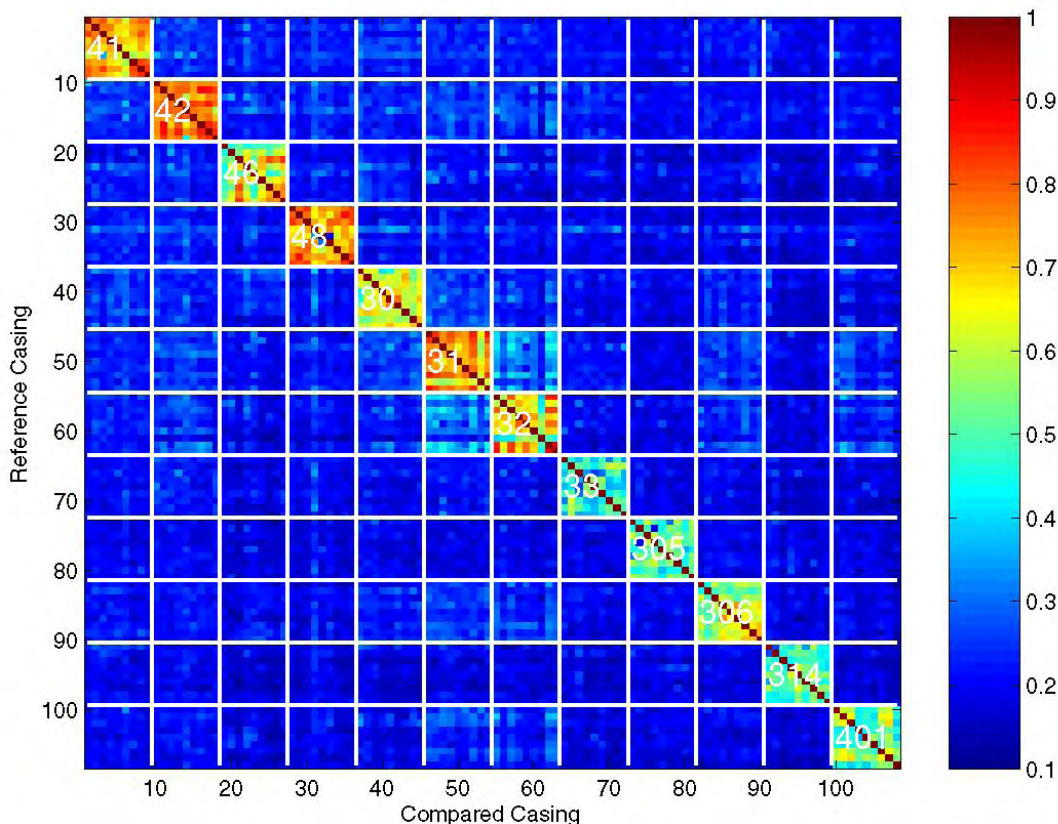

Figure 9-20. $ACCF_{max}$ matrix of the NBIDE Breech Face N-3D data. The color of the pixel in row I and column J indicates the value of the $ACCF_{max}$ between casing I as the reference casing and casing J as the compared casing. The pixels are ordered by gun ID (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401, with Rugers at upper left and S&W's in lower right), and within gun by ammunition ((1-Win, 2-Rem, 3-PMC), and within ammunition by RR#.

The $ACCF_{max}$ matrix in Fig. 9-20 is much closer to the "ideal" $ACCF_{max}$ matrix in Fig. 9-4 than anything else seen so far.

Figure 9-21 depicts histograms of the matching and non-matching scores for the NBIDE breech faces. Figure 9-21 reveals a much greater degree of separation between matching and non-matching scores than has been seen in the other data sets.
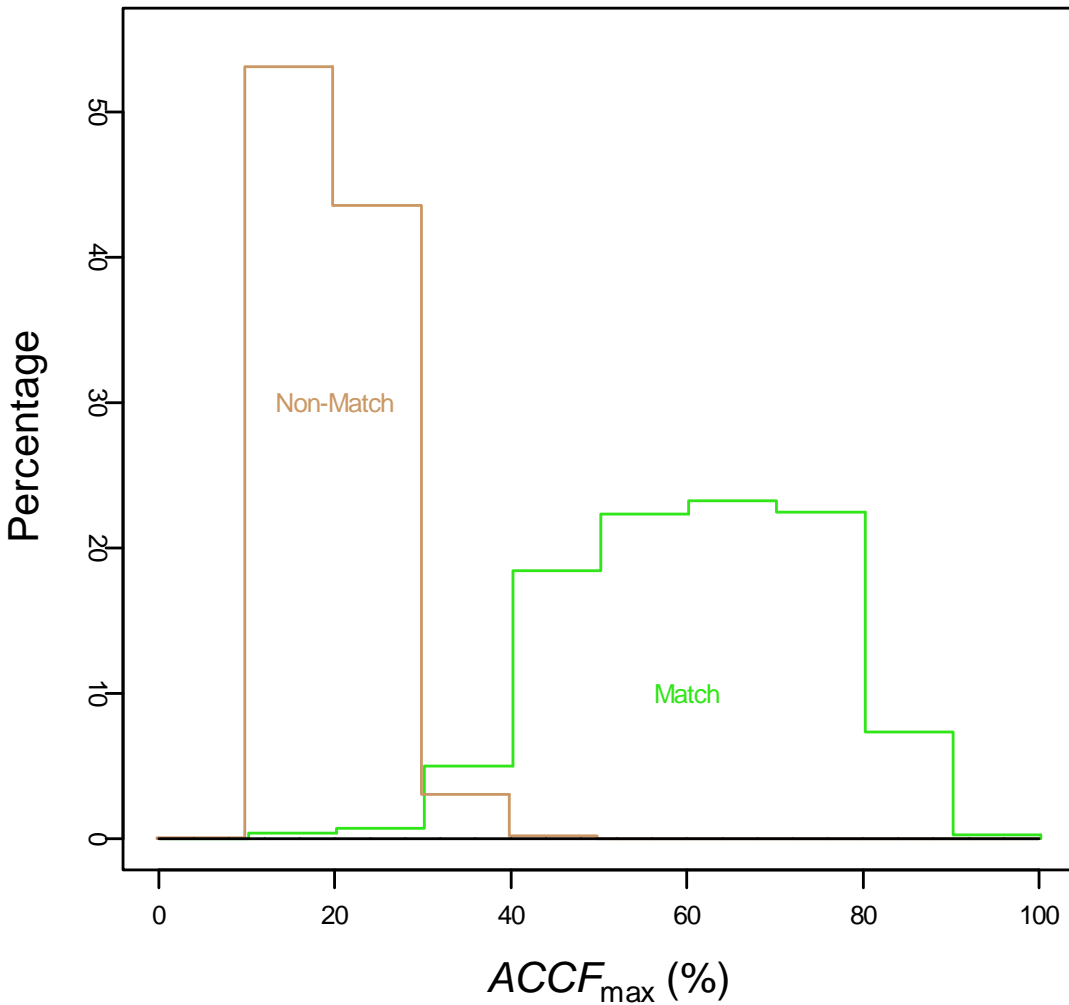
Figure 9-21. NBIDE Breech Face: The green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

Figure 9-22 breaks down these results by reference gun, with the overlap metric *p* given for each grouping. For the NBIDE firing pin impressions, there was considerable variation between guns. By comparison, Fig. 9-22 shows better separation between matching and non-matching distributions for all the guns. However, it should be noted than even for some cases where there is little or no overlap, there may not be the wide separation between distributions that we find ideal; examples would be Sig Sauer 33 and S&W 401.

Table 9-8 contains the overlap metric statistics for the breech face data, ordered by performance of the gun as estimated by the overlap metric, and calculated from all pair-wise comparisons. How do the guns perform?
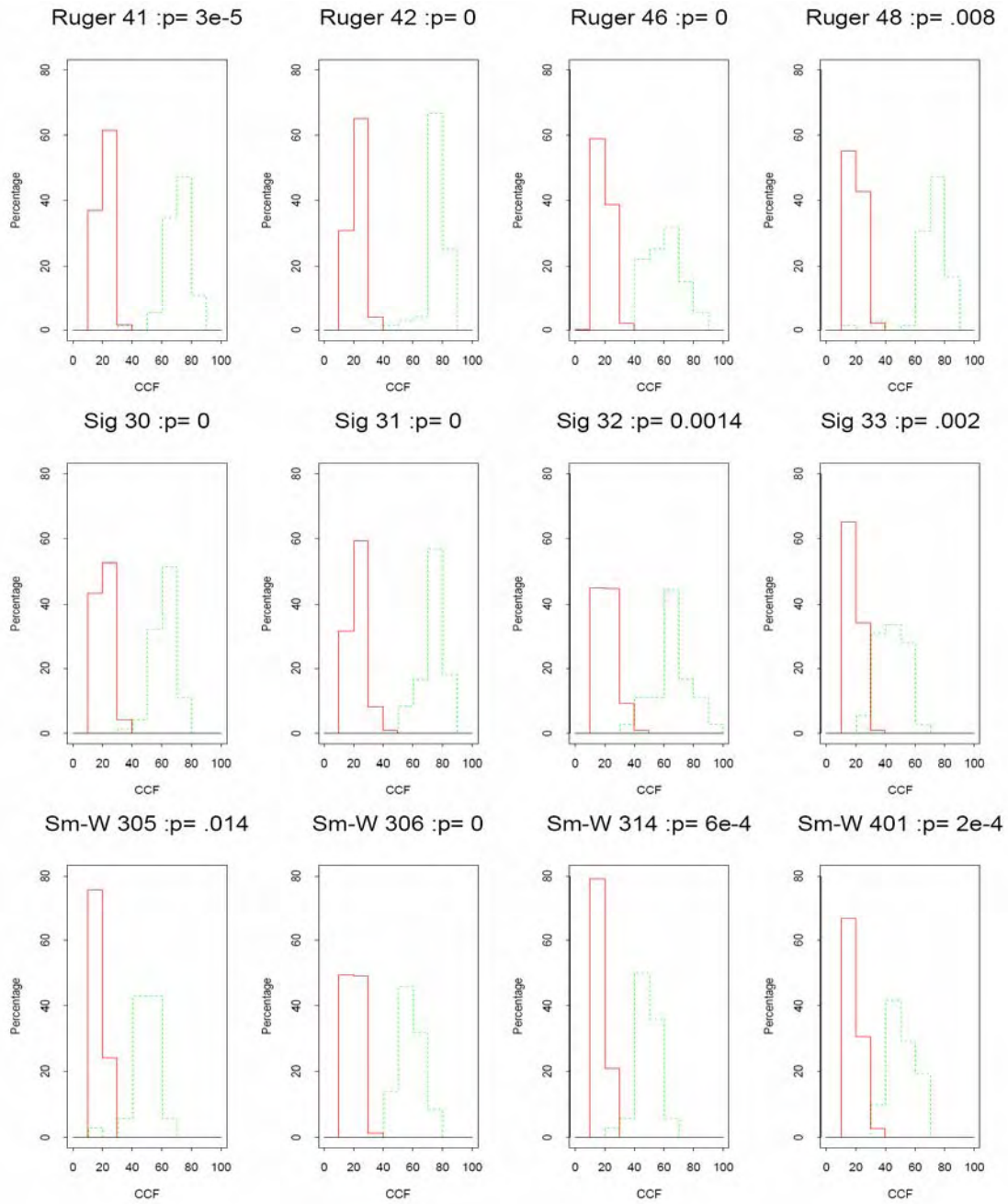
Figure 9-22. NBIDE Breech Face N-3D data: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric *p* derived for that group. In each plot, the green dotted line depicts a histogram of the matching scores, while the red solid line depicts a histogram of the non-matching scores. The symbol CCF stands for $ACCF_{max}$. The horizontal scale for $ACCF_{max}$ is the same as Fig. 9-6 (%).

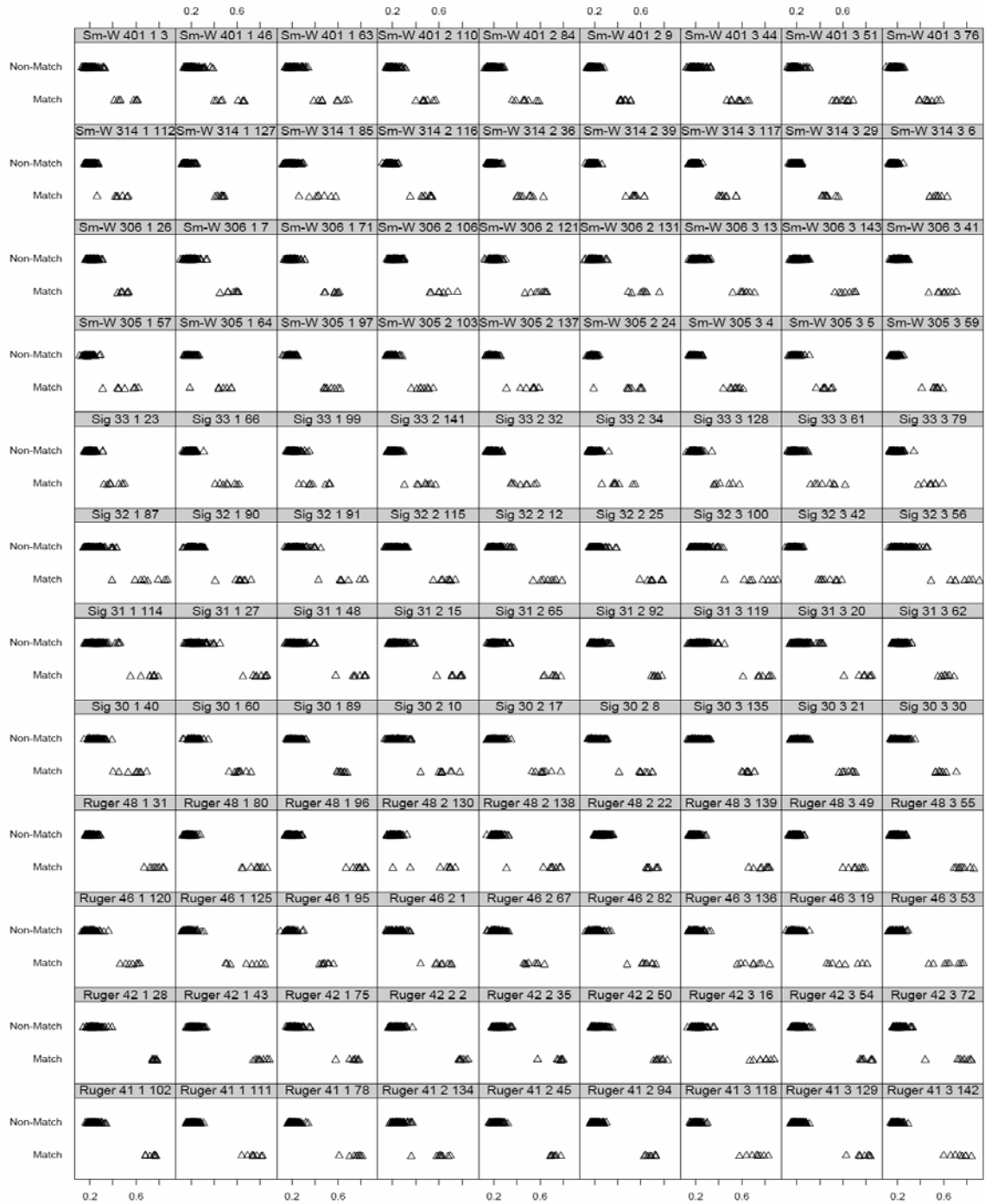Table 9-8. Overlap metric p for topographic signatures of the breech face impressions of 12 NBIDE guns

Gun ID	p
Ruger 42	0
Ruger 46	0
Sig Sauer 30	0
Sig Sauer 31	0
S&W 306	0
Ruger 41	0.00003
S&W 401	0.00017
S&W 314	0.00056
Sig Sauer 32	0.00142
Sig Sauer 33	0.00192
Ruger 48	0.00809
S&W 305	0.01353
Mean	0.00214

While some of the guns seem to have no overlap, note that as stated earlier, the estimation of very small probabilities by pair-wise comparisons can be problematic.

Figure 9-23 breaks down the data more finely by dividing the data into groups by reference casing. Each of the smaller plots depicts for each casing its correlations with the eight casings fired from the same gun (matches, lower triangles), and the correlations with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID, ammunition, and RR number of the reference casing.

From Fig. 9-23 it can be seen that for most of the guns the degree of overlap or separation between matching and non-matching casings varies significantly among casings fired from the same guns, which are shown on the same row in the diagram. For some casings (e.g. Ruger 42-1-28), there is considerable space between matching and non-matching scores. For others (e.g., those from Ruger 46), the matching and non-matching scores also do not overlap but may come close enough to cause overlapping observations for very large sample sizes.

The histogram in Fig. 9-24 charts the estimates of p of each of the 108 casings using the all pair-wise comparisons methods. How many are essentially zero and how high do the estimates go?

Figure 9-23. NBIDE breech face N-3D data: Correlations, for individual casings with the eight casings fired from the same gun (matches, lower triangles), and with the 99 casings fired from other guns (nonmatches, upper triangles). The label above each plot indicates the gun ID, ammunition (1-Win, 2-Rem, 3-PMC), and RR# of the reference casing.
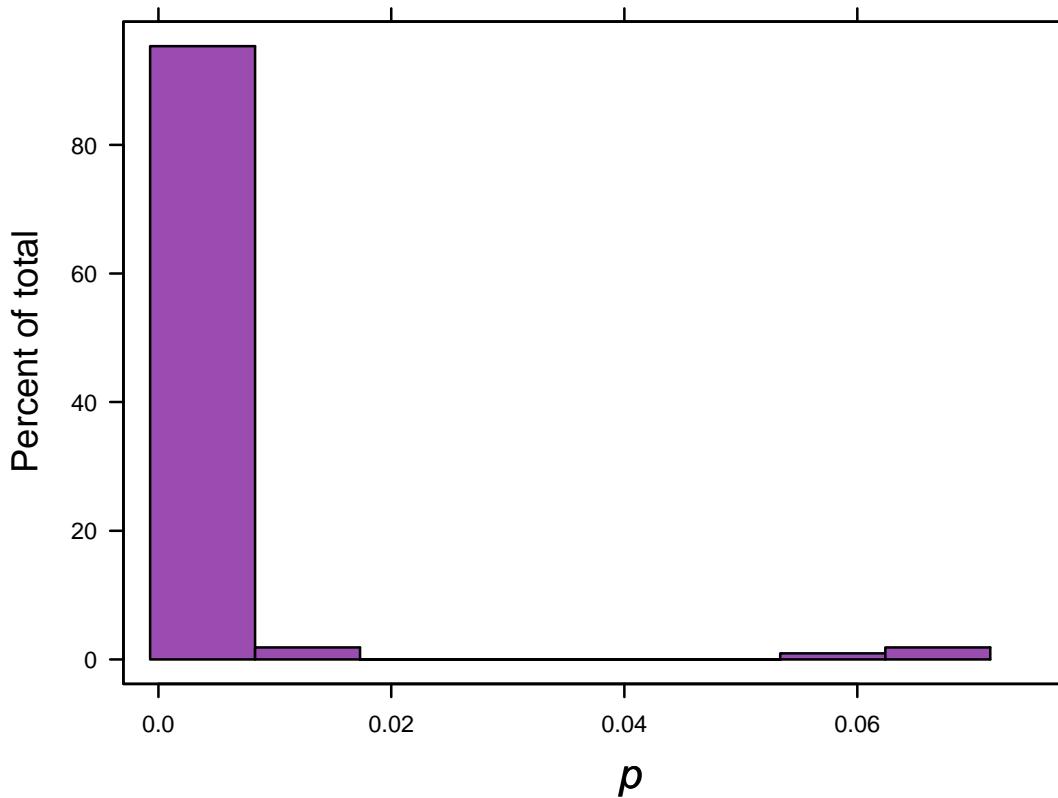
Figure 9-24. NBIDE breech face N-3D data: Histogram of the overlap metric p estimated for each of the 108 casings using pair-wise comparison methods. The mean value of p is 0.0022, and the median p is zero. In fact, 97 of the 108 estimates are exactly 0 (no overlap). The other eleven estimates range from 0.0013 to 0.0707.

The estimates of $p=0$ indicate no overlap between matching and non-matching distributions. It may be problematic to estimate very small probabilities when each casing has only eight observations in its matching sample and 99 observations in the non-matching samples. If a larger number of samples were available, there would likely be more overlaps and non-zero estimates.

Fitting continuous distributions to the samples involved may ameliorate the difficulty associated with low probability events and relatively small sample sizes. To assess this, we fit normal distributions to each of the matching and non-matching samples of correlation scores and derive the estimated means and variances of each sample of correlations. The effect of using different distributions is a topic for further investigation. Using this normal approximation, we obtain a new group of 108 estimates of p. Figure 9-25 contains a histogram of these estimates. Is it very different than that for the pair-wise comparison estimates shown in Fig. 9-24?
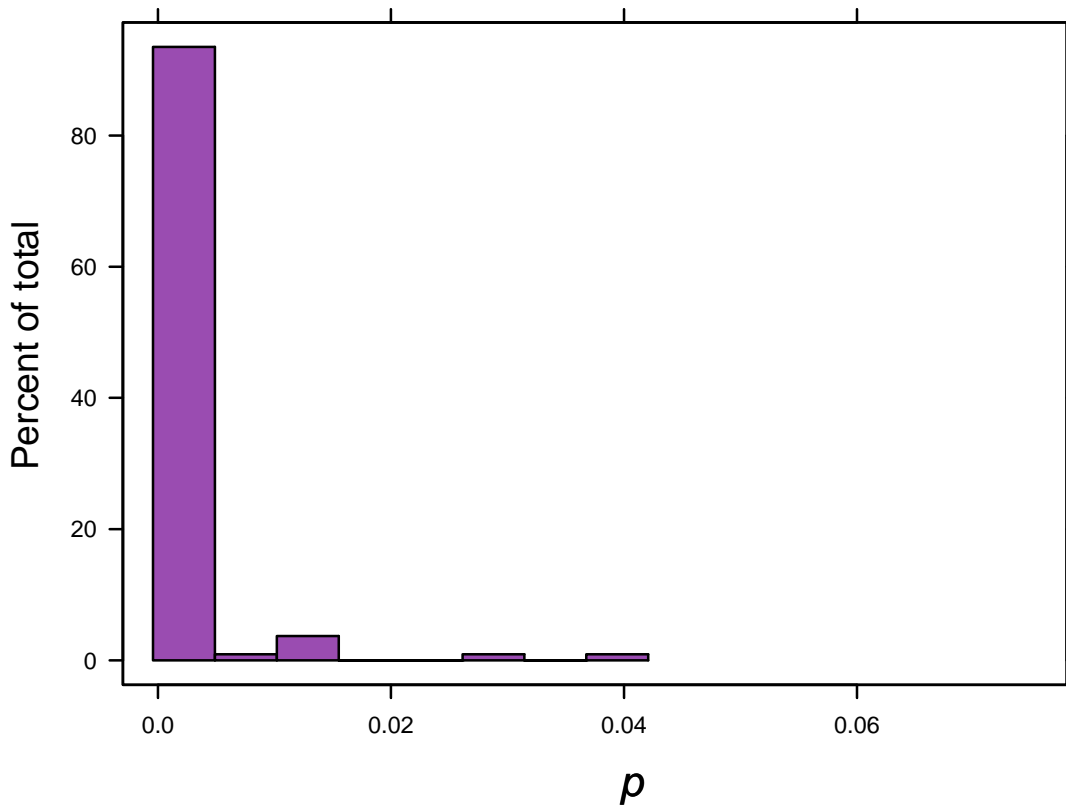
Figure 9-25. NBIDE breech face N-3D data: Histogram of the overlap metric p estimated for each of the 108 casings using normal distribution estimates. The mean value of p is 0.0016 and the median p is 2.8×10^{-7}.

A comparison of the histograms shown in Fig. 9-25 and Fig. 9-24 suggests that the two estimation methods produce very similar results; however, there are differences between the two that do not show up in the histograms. The primary difference is that the normal distribution method should enable a finer resolution in estimating very small probabilities. This difference is shown in the following listing, which breaks down the distribution of the estimates of p. It shows that 55 % of the estimates are less than or equal to 1×10^{-6}, while 94 % are less than or equal to 0.01.

Proportion of p estimates $\leq$	0	10^{-6}	10^{-5}	10^{-4}	10^{-3}	0.01	0.1
	0.11	0.55	0.63	0.72	0.81	0.94	1

The listing shows that most estimates of p for the individual casings are not exactly 0, but are close to 0. This is more useful since p needs to be extremely small for that casing not to call up many mismatches from an extremely large database. This is discussed further in Sec. 9.10 on probability models.

9.6.6   NBIDE: Combining N-3D Firing Pin and Breech Face Analyses

The NBIDE Top Ten results in Table 9-2 show that when using the N-3D breech face impression data, 101 of the 108 NBIDE casings had the maximum eight correct matches and the other seven casings each had seven correct matches, included in their corresponding lists of ten most highly correlated compared casings. When combining measures by including all casings that make the Top Ten correlation list of *either* breech face or firing pin, 107 of the 108 NBIDE casings had the maximum eight correct matches. Only one casing (RR #99, Winchester, fired from Sig Sauer 33) had only seven correct matches using either region. The compared casing that was omitted from both these Top Ten lists was RR #34 Remington. Conversely, when Casing RR #34 was the reference casing, its Top Ten list for breech face impressions did include RR #99, albeit in the 10th and final position on the list. One summary statistic of the benefit of using both regions is that the average number of correct matches per Top Ten list improved to 7.99 (out of a maximum 8 correct matches) from 7.94 for breech face alone and 5.63 for firing pin alone for N-3D.

9.7   I-2D Correlation Scores

9.7.1 I-2D Scores of NBIDE Firing Pin Impressions

Figure 9-26 contains the color score matrix of the I-2D correlation scores of the 108 NBIDE firing pin impressions. The I-2D correlations were performed using BrassCatcher Software Version 3.4.5.The I-2D results in this section are based on searches involving the extended NBIDE set of 144 casings, but any results involving the 36 Speer casings were omitted before the analysis. Are any gun and ammunition brand patterns evident?
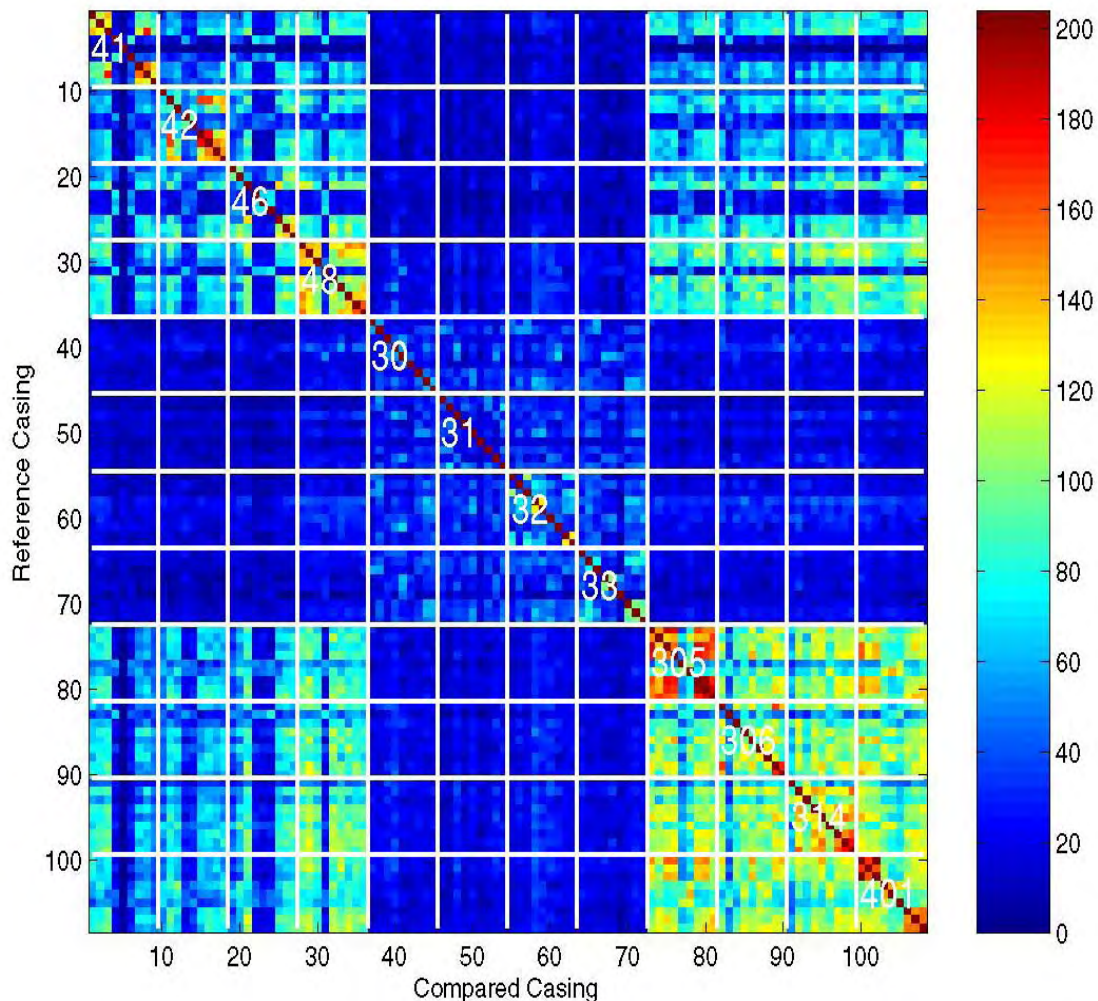
Figure 9-26. Color score matrix of the NBIDE Firing Pin I-2D scores. The color of the pixel in row *I* and column *J* indicates the value of the correlation score between casing *I* as the reference casing and casing *J* as the compared casing. The pixels are ordered by gun ID (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401, with Rugers at upper left and S&W's at lower right), within gun by ammunition (1-Win, 2-Rem, 3-PMC), and within ammunition by RR#. The self-correlation scores on the main diagonal were arbitrarily assigned the color of the maximum score present (204).

The Sig Sauers are responsible for the dark blue cross in the middle of the above figure. These guns do not correlate with guns of the other two brands. However, matching scores of the Sig Sauers are also lower than those from the other two gun brands. The opposite is true of the S&W's, in that they have higher matching scores, but also higher non-matching scores, especially with the other S&W's and with Ruger 48. Also, the matching scores for most of the guns appear to be higher if the ammunition brands as well as the guns are the same.

Figure 9-27 depicts histograms of the matching and non-matching scores for the firing pin impressions.
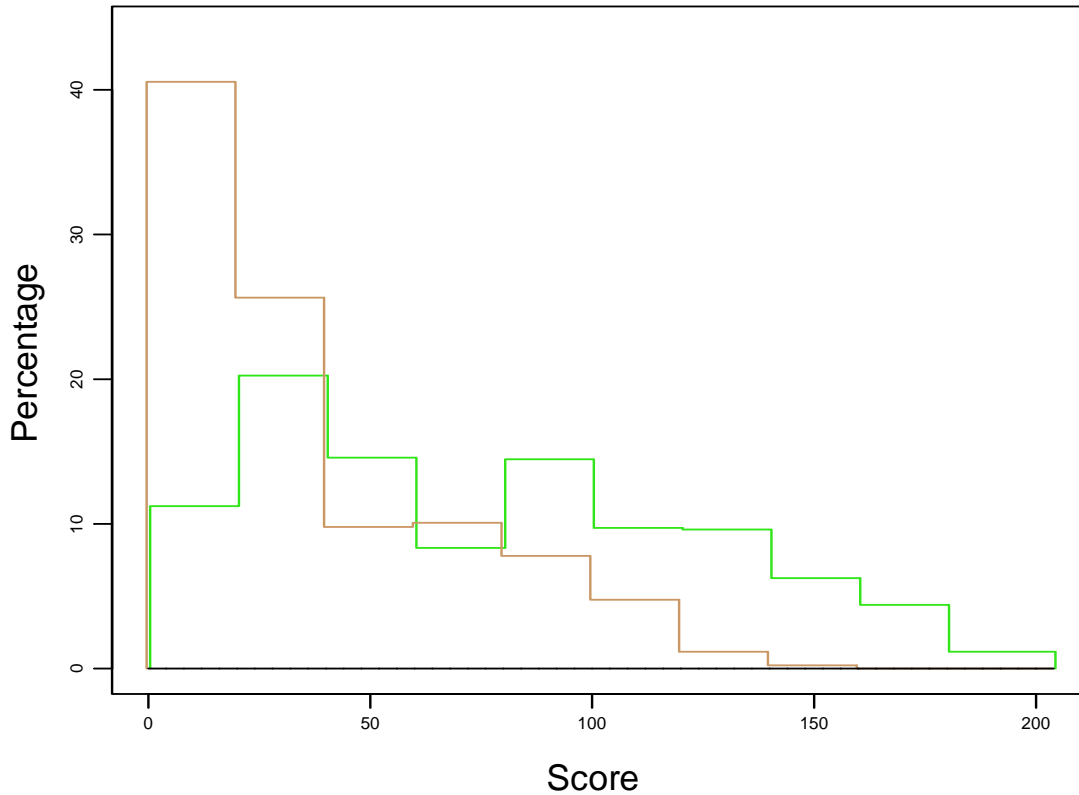
Figure 9-27. NBIDE Firing Pin I-2D data: The green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

There is a considerable degree of overlap between the two distributions of scores. Given the gun differences seen in Fig. 9-26, it makes more sense to examine Fig. 9-28 below, which groups these results by reference gun, with the overlap metric p given for each grouping. The figure shows some differences between gun brands. Clearly, the Sig Sauers have both matching and non-matching scores of low magnitude, while the S&Ws have matching and non-matching scores of higher magnitude, with the Rugers in between. All guns appear to have considerable overlap.

Table 9-9 contains the statistics for the firing pin impression data ordered by performance of the gun as estimated by the overlap metric. How do the gun brands differ?
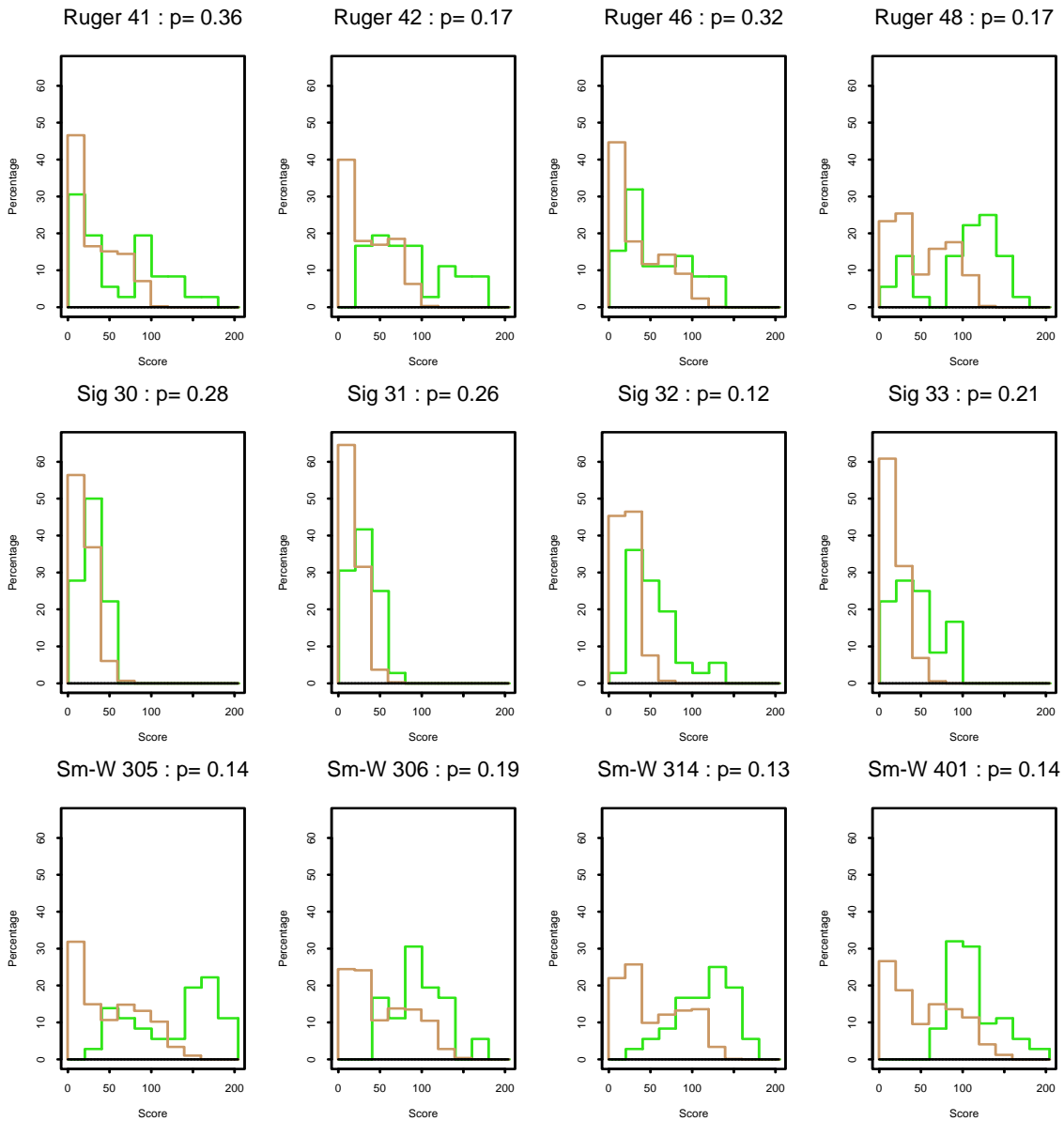
Figure 9-28. NBIDE Firing Pin I-2D data: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric *p* estimated for that group. In each plot, the green lines depict histograms of the matching scores, while the brown lines depict histograms of the non-matching scores.

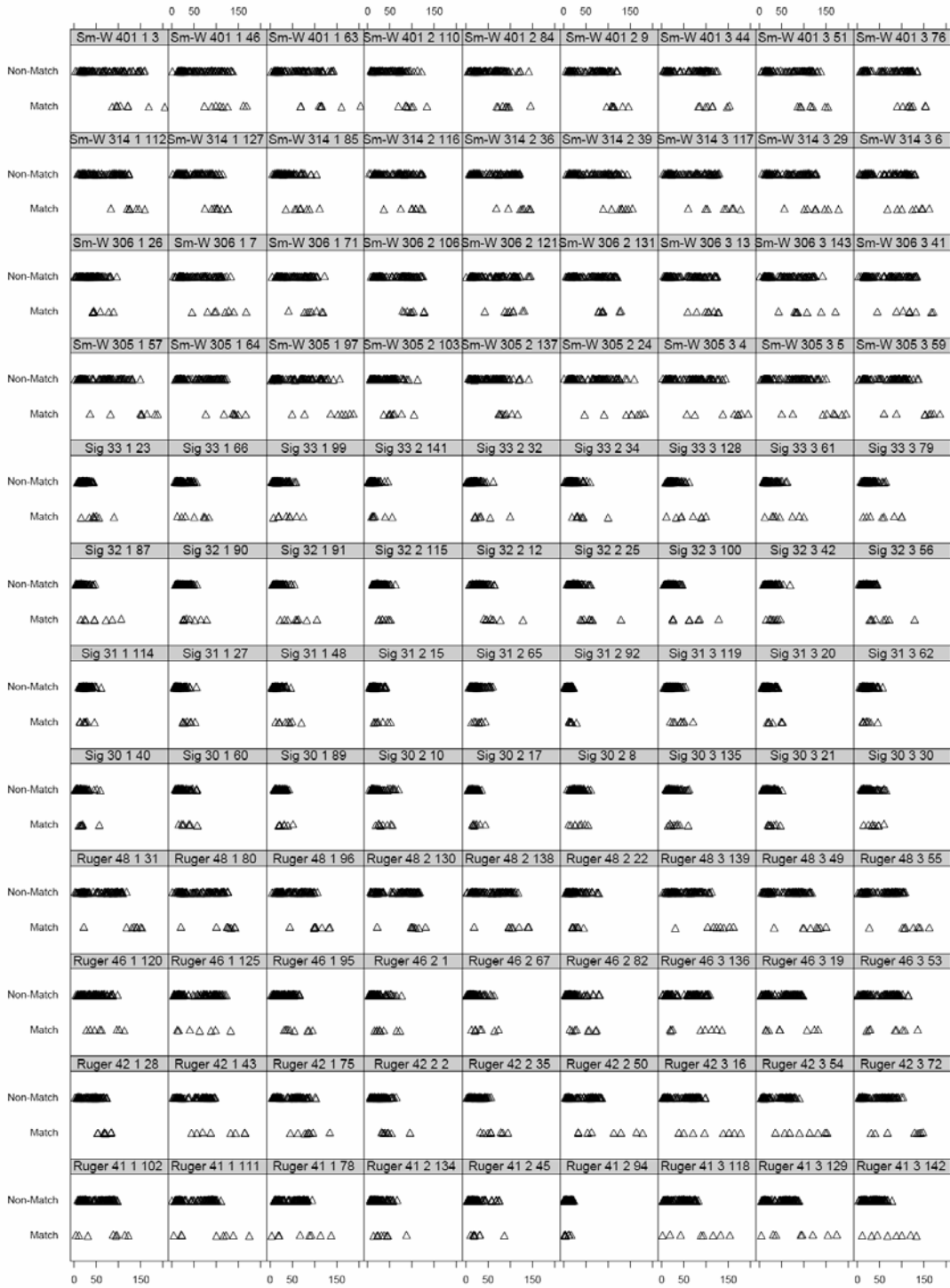Table 9-9. Estimated overlap metric p for the I-2D scores of the firing pin impressions of 12 NBIDE guns.

Gun ID	p
Sig Sauer 32	0.12
S&W 314	0.13
S&W 401	0.14
S&W 305	0.14
Ruger 48	0.17
Ruger 42	0.17
S&W 306	0.19
Sig Sauer 33	0.21
Sig Sauer 31	0.26
Sig Sauer 30	0.28
Ruger 46	0.32
Ruger 41	0.36
Mean	0.21

There are differences within guns of the same brand. However, the S&W's are on the average doing better than the Sig Sauers. A couple of Rugers are hurt by having some very low matching scores, as can be seen from Figures 9-26 and 9-28.

Given the differences even within the same gun, let's look at Fig. 9-29, which breaks down the data even more finely, by dividing the data into groups by reference casing. Each of the smaller plots depicts for each casing, its I-2D correlation scores with the eight casings fired from the same gun (matches, lower triangles), and the correlation scores with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun number and ammunition of the reference casing.

Most casings appear to have substantial overlaps between matching and non-matching scores. Ruger 48 produces better separation than most except for one casing, which does not correlate well with the other casings fired from that gun. Casing RR 94 fired from Ruger 41 resembles the Sig Sauers in having uniformly low matching and non-matching scores. For each of the 108 casings, an overlap metric for the matching and non-matching correlation scores produced by that casing can be estimated by looking at all pair-wise comparisons between matching and non-matching correlations for each casing. The histogram of Fig. 9-30 shows the empirical distribution of those overlap metrics.

The mean estimated p is 0.19, and the median estimated p is 0.16. The minimum estimated p is 0.04, and around 75 % of the estimates are larger than 0.11. All have a degree of overlap that would produce mistakes in a large database scenario.

Figure 9-29. NBIDE Firing Pin I-2D data: Correlations, for individual casings, with the eight casings fired from the same gun (matches, lower triangles), and with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID, ammunition (1-Win, 2-Rem, 3-PMC), and RR# of the reference casing.
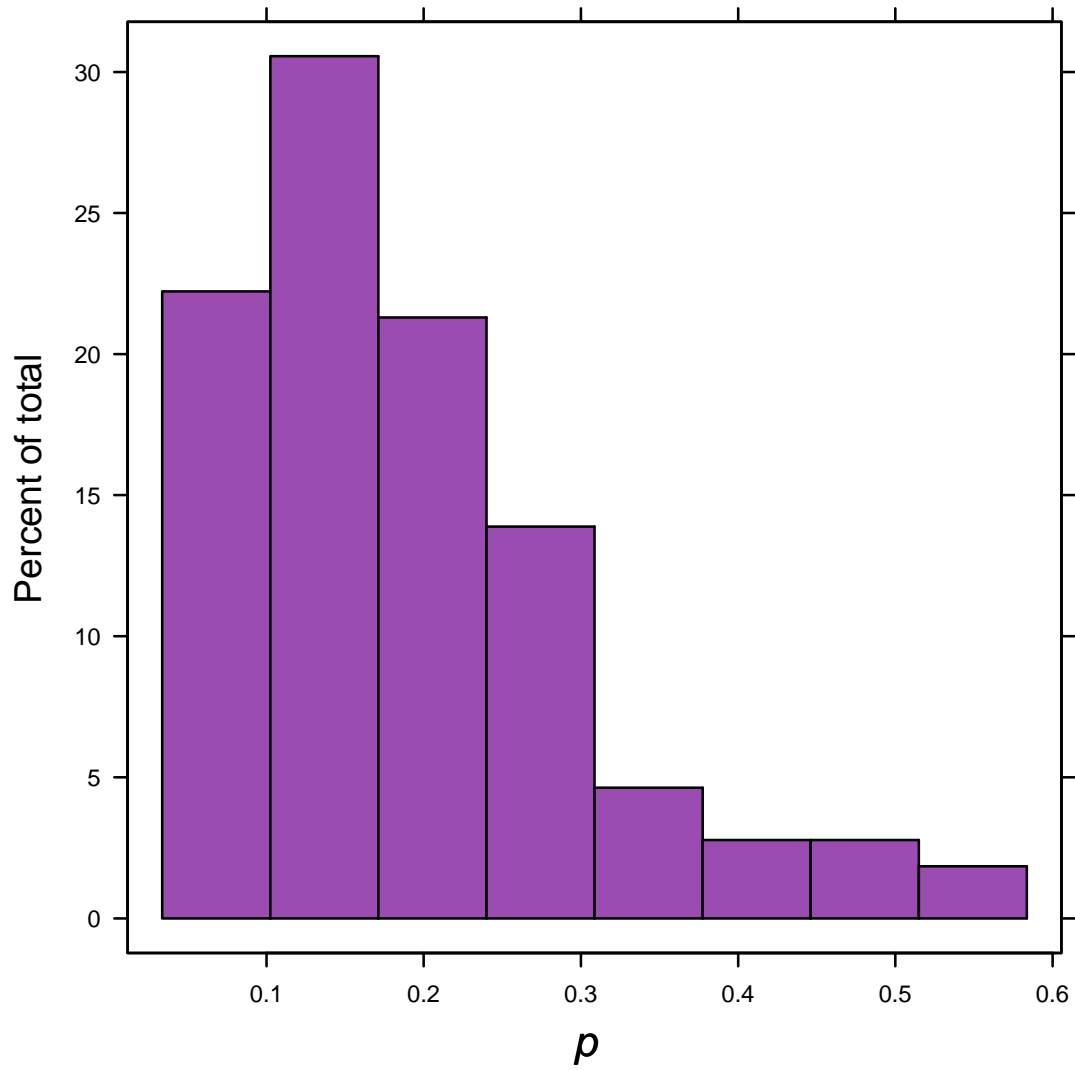
Figure 9-30. NBIDE Firing Pin I-2D data: a histogram of the overlap metric p estimated for each of the 108 NBIDE casings.

9.7.2 I-2D Scores of NBIDE Breech Face Impressions

Figure 9-31 depicts the I-2D scores of the 108 NBIDE breech face impressions in the form of a color matrix. Are the same gun and ammunition brand patterns evident as in the firing pin impressions?
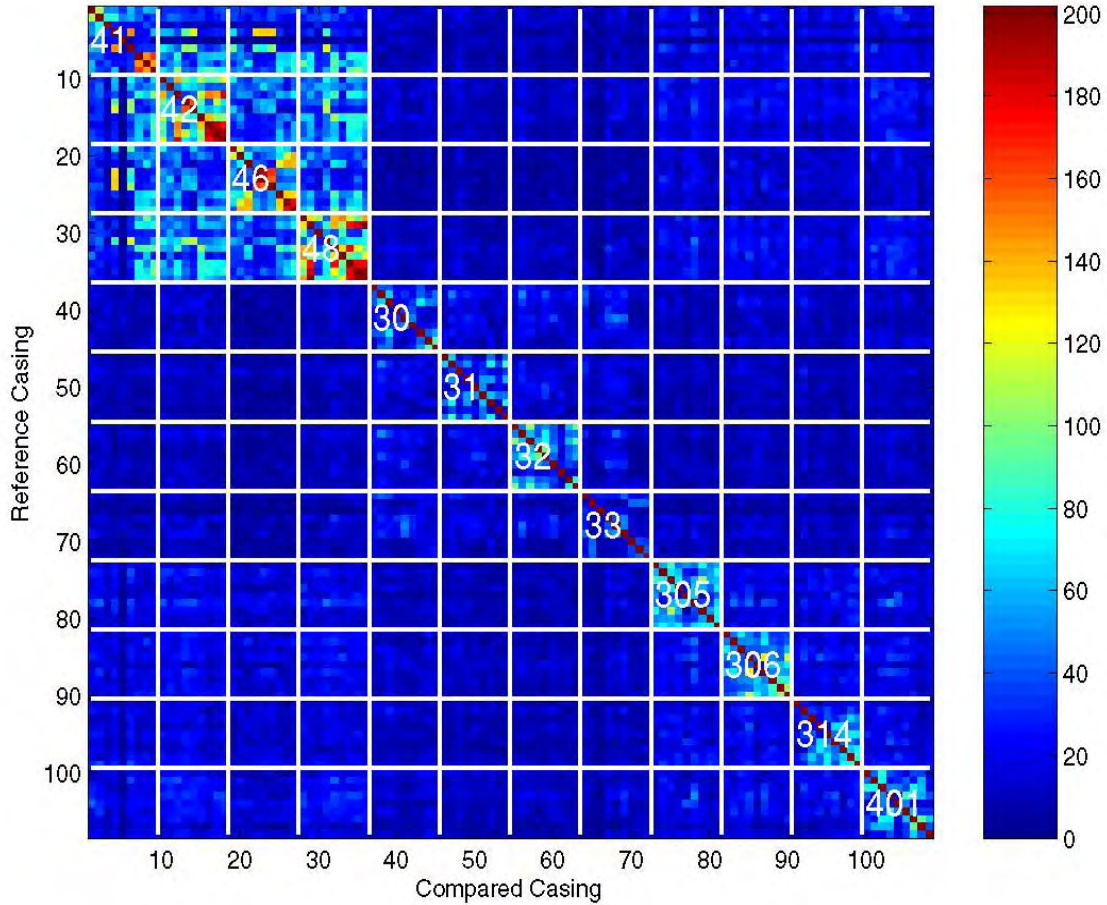

Figure 9-31. Color score matrix of the NBIDE breech face I-2D scores. The color of the pixel in row *I* and column *J* indicates the value of the correlation score between casing *I* as the reference casing and casing *J* as the compared casing. The pixels are ordered by gun ID (Ruger 41, Ruger 42, Ruger 46, Ruger 48, Sig Sauer 30, Sig Sauer 31, Sig Sauer 32, Sig Sauer 33, S&W 305, S&W 306, S&W 314, S&W 401, with Rugers at upper left and S&Ws at lower right), and within gun by ammunition (1-Win, 2-Rem, 3-PMC), and within ammunition by RR#. The self-correlation scores on the main diagonal were arbitrarily assigned the color of the maximum score present (202).

The Sig Sauers again produce lower matching and non-matching scores than do the other gun brands, although the pattern is not nearly as pronounced as for the firing pin impressions. The Rugers have the highest matching scores, especially when both the reference and compared casings are PMCs. However, the highest non-match scores occur when both guns are Rugers; Ruger 41 is an exception to that observation in having several very low matching and non-matching scores.

122

Figure 9-32 depicts histograms of the matching and non-matching scores for the breech face impressions. What is the degree of overlap or separation between the two distributions?

Figure 9-32. NBIDE Breech Face I-2D data: The green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

There is a considerable degree of overlap between the matching and non-matching scores, although less than for the firing pin impressions. Given the gun differences seen in Fig. 9-31, it makes more sense to examine Fig. 9-33, which groups these results by reference gun, with the overlap metric *p* given for each grouping. Are the gun brand differences evident?

Figure 9-33. NBIDE Breech Face I-2D data: Above each of the individual plots is a heading with the ID of the reference gun and the overlap metric *p* estimated for that gro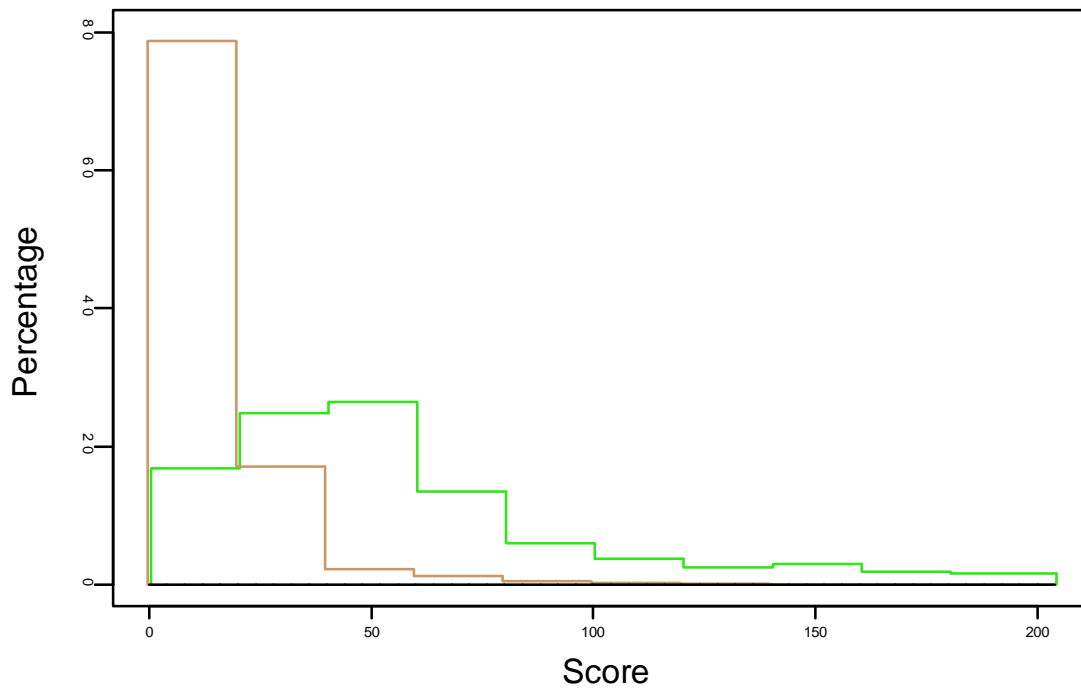up. In each plot, the green lines depict a histogram of the matching scores, while the brown lines depict a histogram of the non-matching scores.

As with the firing pin impressions, the Sig Sauers have both matching and non-matching scores of low magnitude. However, there is less overlap between both matching and non-matching scores than with the firing pin impressions.

Table 9-10 contains the statistics for the breech face data ordered by performance of the gun as estimated by the overlap metric. There is much greater separation between matching and non-matching scores than existed with the firing pin impressions. Also, there are again wide differences within gun brands.

Table 9-10. Estimated overlap metric *p* for the I-2D scores of the breech face impressions of 12 NBIDE guns.

Gun ID	p
S&W 306	0.02
Ruger 48	0.03
Ruger 42	0.04
S&W 305	0.05
Sig Sauer 32	0.05
Ruger 46	0.06
Sig Sauer 31	0.09
S&W 401	0.09
S&W 314	0.11
Sig Sauer 30	0.11
Ruger 41	0.16
Sig Sauer 33	0.22
Mean	0.08

Given the differences even within the same gun, let's look at Fig. 9-34, which breaks down the data even more finely, by dividing the data into groups 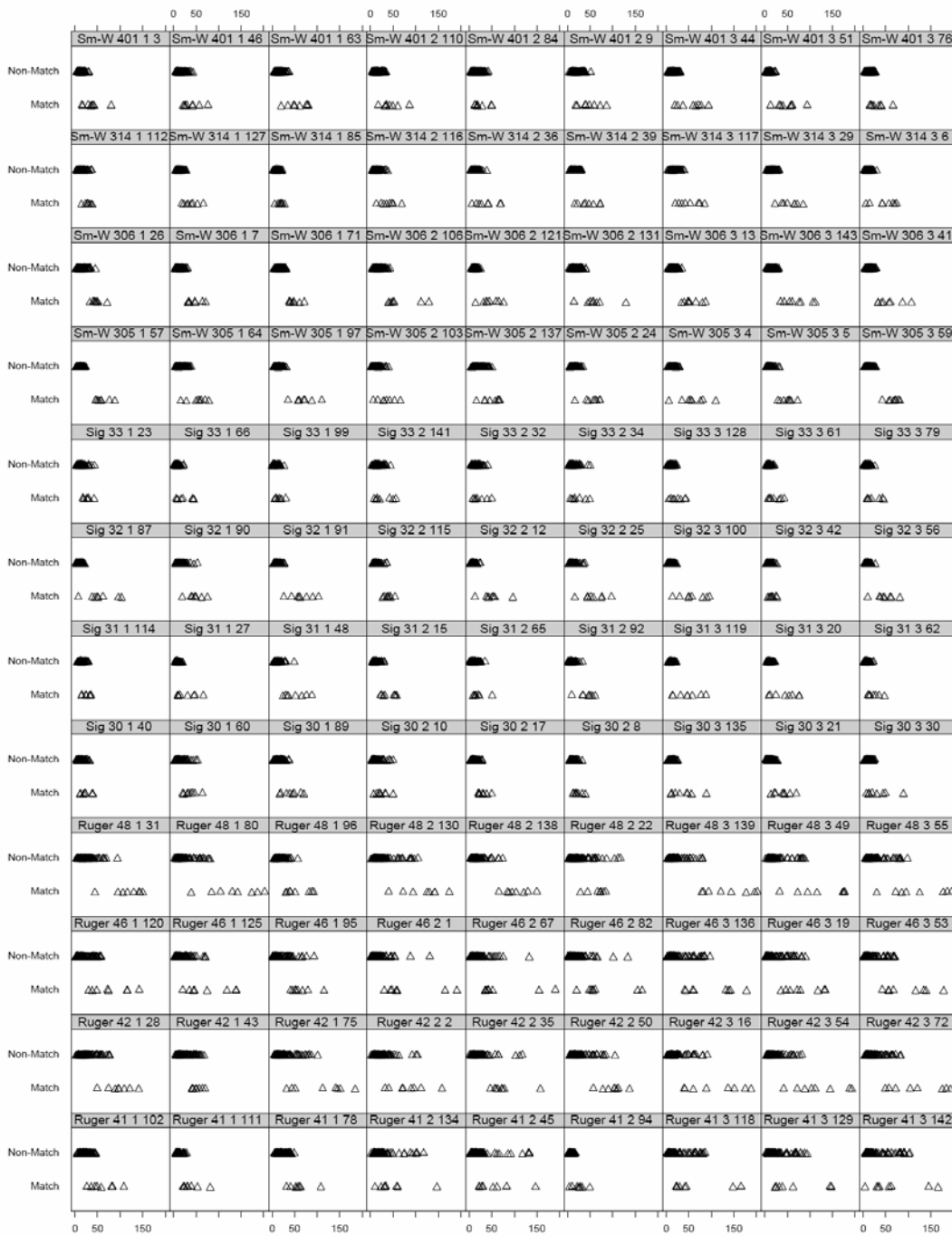by reference casing. Each of the smaller plots depicts for each casing, its correlations with the eight casings fired from the same gun (matches, lower triangles), and its correlations with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun # and ammunition of the reference casing.

From Fig. 9-34, most but not all casings have some overlap between matching and non-matching scores; however, there is less overlap than is present with the firing pin impression data. For each of the 108 casings, an overlap metric for the matching and non-matching correlation scores produced by that casing can be estimated by looking at all pair-wise comparisons between matching and non-matching correlations for that casing.

Figure 9-35 shows the empirical distribution of those estimated overlap metrics grouped by casing. How does it compare with that of the firing pin estimated scores shown in Fig. 9-30? The estimated overlap metrics are significantly better (closer to 0) than those for firing pin impressions. The mean of the estimated p is 0.09, and the median estimated p is 0.06. About 75 % of the estimated overlap metrics are larger than 0.027. Seven of the 108 casings have no overlap, leading to estimates of $p = 0$.

Note that because the triangles in Fig. 9-34 have non-zero width, some of the casing plots may give the impression of somewhat more overlap than actually exists. For example, casing RR# 41 fired from S&W 306 (ammunition 3) has no overlap between matching and non-matching scores. Several of those casings with no overlap have little separation between the largest non-matching score and the smallest matching score.

Figure 9-34. NBIDE Breech Face I-2D data: Correlations, for individual casings, with the eight casings fired from the same gun (matches, lower triangles) and with the 99 casings fired from other guns (non-matches, upper triangles). The label above each plot indicates the gun ID, ammunition (1-Win, 2-Rem, 3-PMC), and RR# of the reference casing.

126

Figure 9-35. NBIDE Breech Face I-2D data: Histogram of the overlap metric *p* estimated for each of the 108 NBIDE casings.

9.7.3 NBIDE: Combining I-2D Firing Pin and Breech Face Analyses

The I-2D Top Ten results in Table 9-2 show that the I-2D breech face impression data produce an average of 5.57 correct matches out of a maximum eight correct matches per Top Ten list. Using the I-2D firing pin impression data yields an average of 3.72 correct matches per Top Ten list. Combining measures by including all casings that make the Top Ten correlation list of *either* breech face or firing pin set yields an average of 6.20 correct matches out of eight, which is certainly better than either region alone, but still does not come close to eliminating the existing coverage mistakes.

9.7.4 Correlation Software and Algorithm Uncertainties

The I-2D correlations of the NBIDE casings were performed in two different ways. First, correlations were performed at the ATF shortly after the I-2D image acquisitions were made. Rankings of the top 30 to 40 breech face and firing pin correlations for each NBIDE casing were calculated and printed. The database for this correlation included all 144 NBIDE test fires and the 70 De Kinder casings. The top results for the NBIDE PMC, Remington, and Winchester ammunition were then entered manually into a spreadsheet and analyzed in a manner similar to that shown in Sec. 9.3 to produce Top Ten results among the 108 NBIDE casings. Afterwards, the data were moved to FTI, Montreal, where the entire collection of 144 NBIDE test fires were analyzed. This calculation produced an array of 144 by 143 I-2D correlation scores, which produced the results in Secs. 9.3 and 9.7.

Although the two correlations produced similar results, there were two types of differences. First, some pairs produced different scores in the ATF and FTI correlations. Although these differences were small in most cases, a few gave significantly different results. For example, the pair of breech face impressions from RR #16 as the reference casing and RR #28 as the compared casing produced an I-2D score of 89 in the ATF correlation but a score of 82 in the FTI correlation. Second, some pairs scoring in the Top Ten in the FTI correlation were completely missing in the ATF correlation.

The first type of difference suggests that there were slightly different versions of software used in the two correlation runs. The second type of difference is ascribed to a coarse filter procedure, which was likely used for the correlation run at the ATF but not for the correlation run at FTI. Because the correlations are generally used with large databases, the I-2D system routinely applies a coarse filter to correlation runs. That is, a preliminary correlation procedure is applied to all the entries in the database. Then for each reference casing only the highest scoring acquisitions, approximately 20 %, are used for a more extensive correlation calculation. The scores from the second correlation are recorded in the printed correlation results, but some of these high scoring pairs may be missing if they do not pass the coarse screening filter test. The frequency of missing entries in the correlation run at ATF was particularly significant when the reference casing was a Sig Sauer test fire because the database there included 36 Sig Sauer test fires from the NBIDE collection and 70 Sig Sauer test fires from the De Kinder collection. Therefore, the probability that a genuine match of casings would fall below the top 20 % (about 44 entries) in the first correlation pass was higher for a Sig Sauer test fire than for a Ruger or S&W, given that there were many similar Sig Sauer test fires in the ATF database.

The software differences discussed above produced differences in the Top Ten scores for the I-2D correlations. Referring to Table 9-2, the ATF average score for breech face impressions was about 5.1 correct matches versus the value of 5.57 given in the Table and derived from the FTI correlations. For the firing pin impressions the average scores were both about 3.7.

A third type of difference in the I-2D scores was also recorded. The matching score for a pair could change depending on which member was used as the reference. However, the differences were insignificant except for a few cases. This asymmetry was also present in the $ACCF_{max}$ scores for the topography correlations.

9.8 NBIDE Topographic Analysis for Gun Brand and Ammunition Effects

The NBIDE-designed experiment with its three gun types and four guns per type allows us to conduct a third kind of analysis that is not available from the De Kinder casings experiment with its single gun type. The experiment enables exploration of ammunition and gun brands on cross correlations. For instance, will casings fired from the same gun be more highly correlated if the ammunition brands are the same? Conversely, will the casings fired from different guns be more likely to be falsely matched if the ammunition are the same brand and/or the guns are of the same model?

For this analysis, we can divide the 108×107 casing permutations into six groups shown below with varying degrees of separation. Note that there are 108×107 pairs because the $ACCF_{max}(A,B)$ and $ACCF_{max}(B,A)$ are not necessarily the same, so both are included for each combination of casings A and B. This accounts for the final '$\times 2$' in each group enumeration below.

In that regard, we note the following categories of casing pairs:

1) **Same gun, same ammo:**
(12 guns $\times$ 3 ammos) $\times$ [3 casing combinations $\times$ 2]
$= 36 \times 6 = 216$

2) **Same gun, different ammo:**
(12 guns $\times$ 3 ammo combinations) $\times$ [9 casing permutations $\times$ 2]
$= 36 \times 18 = 648$

3) **Different guns of same brand, same ammo:**
(3 ammo brands $\times$ 3 gun brands) $\times$ [6 gun pairs $\times$ 9 casing permutations $\times$ 2]
$= 9 \times 108 = 972$

4) **Different guns of same brand, different ammo:**
(3 gun brands $\times$ 3 ammo combinations) $\times$ [12 gun permutation pairs $\times$ 9 casing permutations $\times$ 2] $= 9 \times 216 = 1944$

5) **Guns of different brands, same ammo:**
(3 ammo brands $\times$ 3 gun brand pairs) $\times$ [16 gun permutations] $\times$ 9 casing permutations $\times$ 2] $= 9 \times 288 = 2592$

6) **Guns of different brands, different ammos:**
(6 gun brand permutations $\times$ 3 ammo brand pairs) $\times$ [16 gun permutations $\times$ 9 casing permutations $\times$ 2] $= 18 \times 288 = 5184$

The six sub-totals above total to $108 \times 107 = 11556$.

In this section, we tabulate and explore the averages of the different groups in the first 5 categories above

9.8.1 NBIDE Firing Pin Images: Topographic Analysis of Ammunition Effects

Table 9-11 shows average $ACCF_{max}$ values for various matching casing pairs divided into groups by guns and ammunition combinations. The first four columns of numbers are for the groups where both casings are the same ammunition brand. The last four columns are when the casings are from different ammunition brands.

Table 9-11. Average $ACCF_{max}$ values from firing pin impressions for NBIDE casing pairs fired by the same firearm. All values are expressed in % where 100 % is the value 1.0, indicating perfect correlation.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Winchester	Remington	PMC	Average	Win-Rem	Win-PMC	Rem.-PMC	Average
Ruger 41	55	27	67	49.7	33	59	33	41.7
Ruger 42	54	66	69	63.0	60	60	66	62.0
Ruger 46	45	24	42	37.0	26	46	25	32.3
Ruger 48	42	27	45	38.0	36	46	35	39.0
Sig Sauer 30	33	24	41	32.7	32	40	37	36.3
Sig Sauer 31	27	27	29	27.7	29	30	30	29.7
Sig Sauer 32	32	53	51	45.3	31	32	51	38.0
Sig Sauer 33	34	42	58	44.7	24	41	33	32.7
S&W 305	48	53	63	54.7	54	54	53	53.7
S&W 306	24	52	37	37.7	27	34	41	34.0
S&W 314	43	33	48	41.3	37	47	37	40.3
S&W 401	45	50	46	47.0	26	38	43	35.7
Average	40.2	39.8	49.7	43.2	34.6	43.9	40.3	39.6

There is obviously a large degree of variability in the category-pair means between gun brands, between guns of the same brand, and even between categories from the same gun. There are ammo effects, but they are not consistent, e.g., lower $ACCF_{max}$ values for Remingtons with Rugers.

The standard deviations of $ACCF_{max}$ values of casing-pair groups, fired by the same gun and having the same ammunition brand, range from 1 % to 23 %, with the very largest standard deviations occurring with the Sig Sauer pairs. The matching casing-pair groups with different ammunition have group standard deviations ranging from 2 % to 20 %, with again the very largest standard deviations occurring with the Sig Sauer pairs.

Table 9-12 shows the average $ACCF_{max}$ values of casings from non-matching guns of the same brand. The columns are organized by ammunition brand combination. The standard deviations of the $ACCF_{max}$ groups in this table range from 3 % to 8 %, with the higher standard deviations occurring with the S&W's.

Table 9-12. Average $ACCF_{max}$ values from firing pin impressions for casing pairs fired by different firearms of the same brand. All values are expressed in %.

	Winchester	Remington	PMC	Win-Rem	Win-PMC	Rem.-PMC
Ruger	22	18	23	20	23	20
Sig Sauer	25	26	24	24	23	25
S&W	24	34	32	26	27	31

Table 9-13 shows the mean $ACCF_{max}$ values from firing pin impressions for casing pairs of the same ammunition brand, but fired from different brands of guns. The standard deviations of the $ACCF_{max}$ groups in this table range from 4 % to 7 %.

Table 9-13. Average $ACCF_{max}$ values from firing pin impressions for casing pairs fired by different brands of guns. All values are expressed in %.

	Winchester	Remington	PMC
Ruger-Sig Sauer	21	21	21
Ruger-S&W	23	22	26
Sig Sauer-S&W	20	26	24

For the large pool of pairs of casings that have both different gun brands and different ammunition brands, the mean of the $ACCF_{max}$ value is 23 % and the standard deviation is 6 %. These appear quite similar to the $ACCF_{max}$ values with the same brand ammunition but different brands of guns.

9.8.2 NBIDE Breech Face Images: Topographic Analysis of Ammunition Effects

Table 9-14 shows average $ACCF_{max}$ values for various matching casing pairs divided into groups by guns and ammunition combinations. The first four columns of numbers are for the groups where both casings are the same ammunition brand. The last four columns are when the casings are from different ammunition brands.

131

Table 9-14. Average $ACCF_{max}$ values from NBIDE breech face impressions for matching casing pairs fired by the same gun. All values are expressed in %.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Win.	Rem.	PMC	Ave.	Win-Rem	Win-PMC	Rem.-PMC	Ave.
Ruger 41	78	68	82	76.0	68	76	66	70.0
Ruger 42	76	80	83	79.7	76	78	74	76.0
Ruger 46	51	64	75	63.3	57	64	60	60.3
Ruger 48	81	50	77	69.3	70	80	69	73.0
Sig Sauer30	63	73	64	66.7	58	62	62	60.7
Sig Sauer 31	78	71	69	72.7	74	74	72	73.3
Sig Sauer32	67	74	62	67.7	66	65	67	66.0
Sig Sauer 33	45	55	51	50.3	41	48	40	43.0
S&W 305	51	54	50	51.7	44	52	49	48.3
S&W 306	51	67	70	62.7	54	56	63	57.7
S&W 314	40	57	53	50.0	47	46	50	47.7
S&W 401	62	48	58	56.0	44	56	50	50.0
Average	61.9	63.4	66.2	63.8	58.3	63.1	60.2	60.5

The standard deviations of $ACCF_{max}$ values of those casing pairs from the same gun and the same ammunition brand range from 1 % to 20 %, with the very largest standard deviations occurring with the Sig Sauer pairs. The matching casing pair categories with different ammunition have standard deviations ranging from 5 % to 11 %. The Rugers have higher $ACCF_{max}$ values, and the S&W's have lower $ACCF_{max}$ values. Sig Sauer 33 has particularly low $ACCF_{max}$ values.

Table 9-15 shows the average $ACCF_{max}$ values of non-matching casing pairs with common gun brands, organized by ammunition brand combination. The standard deviations of the $ACCF_{max}$ groups in this table range from 3 % to 7 %, with the higher standard deviations occurring with the Sig Sauers and the lower standard deviations occurring with the S&Ws. The one pattern that seems evident is that there are lower $ACCF_{max}$ values between non-matching S&W casings as opposed to non-matching casings of the other two brands.

Table 9-15. Mean $ACCF_{max}$ values from NBIDE breech face impressions for non-matching casing pairs fired by different guns of the same brand. All values are expressed in %.

Gun Brand	Ammunition			Ammunition Pair		
	Winchester	Remington	PMC	Win-Rem	Win-PMC	Rem-PMC
Ruger	22	26	21	24	21	23
Sig Sauer	26	23	24	24	24	23
S&W	19	19	18	19	18	18

Table 9-16 shows the mean $ACCF_{max}$ values of casing pairs of the same ammunition brand, but fired from different brands of gun. The standard deviations of the $ACCF_{max}$ values in this table range from 3 % to 5 %.

Table 9-16. Mean $ACCF_{max}$ values from NBIDE breech face impressions for casing pairs of the same brand fired by different brands of guns. All values are expressed in %.

	Winchester	Remington	PMC
Ruger-Sig Sauer	22	24	20
Ruger-S&W	19	21	19
Sig Sauer-S&W	22	19	20

For the large pool of casing pairs that have both different gun brands and different ammunition brands, the mean of the $ACCF_{max}$ values is 21 % and the standard deviation is 4 %. These values are similar to the $ACCF_{max}$ values with the same brand ammunition but different brands of guns.

Any variabilities due to ammunition are smaller than the variabilities due to guns. We have seen considerable variability even within guns of the same model. In any case, ammunition brand effects are relatively modest compared to gun effects, because the non-matching scores tend to be around the same magnitude regardless of whether the ammunition brands are the same or whether the gun brands are the same. An exception may be the higher firing pin non-matching $ACCF_{max}$ values for Smith&Wesson guns using non-Winchester ammunition. On the other hand, the matching scores vary most according to the gun brand and between individual guns within the same gun brand.

Some analyses of variance found that gun and ammunition effects, as well as interactions, were statistically significant for the matching scores. For the non-matching scores, gun and ammunition effects were also significant. In all cases, the largest effects were the gun effects. Section 10 contains an analysis of gun and ammunition effects using the Top Ten experiments.

9.9 NBIDE I-2D Analysis for Gun Brand and Ammunition Effects

9.9.1 NBIDE Firing Pin Impressions

Table 9-17 contains average I-2D score values for various matching casing pairs divided into groups by guns and by ammunition combinations. The first three columns of numbers are for the groups where both casings are the same ammunition brand. The columns on the right are when the casings are from different ammunition brands. What are the effects of the brand pairings for gun and ammunition?

Table 9-17. Average I-2D scores from firing pin impressions for casing pairs fired by the same weapon.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Win.	Rem.	PMC	Average	Win-Rem	Win-PMC	Rem.-PMC	Average
Ruger 41	125	39	136	100	15	102	25	47
Ruger 42	80	54	146	93	71	108	75	85
Ruger 46	66	70	122	86	32	89	22	48
Ruger 48	136	46	144	109	87	127	85	100
Sig Sauer 30	33	38	47	39	26	27	29	27
Sig Sauer 31	29	20	33	27	27	39	24	30
Sig Sauer 32	57	72	62	64	51	48	46	48
Sig Sauer 33	56	65	97	73	23	62	25	37
S&W 305	158	81	184	141	102	159	100	120
S&W 306	69	104	123	99	86	91	107	95
S&W 314	106	120	162	129	99	98	128	108
S&W 401	177	126	152	152	89	115	96	100
Average	91	70	117	93	59	89	64	70

For every gun except Sig Sauer 31, the average score is higher when the reference and compared ammunition brands are the same rather than different. When the ammunition brand is the same for both casings, the average scores are highest when both casings are PMC for every gun except S&W 401 and Sig Sauer 32. The Sig Sauers produce lower match scores, but as will be seen below, they also produce lower non-match scores. The standard deviations of the scores within each group vary widely (from 2 to 40).

Table 9-18 shows the average I-2D scores of casings from non-matching guns of the same brand. The columns are organized by ammunition brand combination. What are the effects of gun brand and ammunition brand?

Table 9-18. Average I-2D scores from firing pin impressions for casing pairs fired by different weapons of the same brand.

Gun Brand	Ammunition					
	Win.	Rem.	PMC	Win-Rem	Win-PMC	Rem.-PMC
Ruger	66	33	73	34	67	32
Sig Sauer	33	26	36	25	31	32
S&W	93	95	114	85	99	94

In a pattern that can seen from the score matrix in Fig. 9-26 and from the histograms in Fig. 9-28, the Sig Sauers produce lower non-match scores along with the lower match scores seen in Table 9-17. In contrast, different guns that are both S&W produce larger scores than for the other two gun brands. When the different guns are both Ruger, there are larger scores in some cases. The standard deviations of the scores within each group vary widely (from 10 to 50).

Table 9-19 shows the mean I-2D scores of casing pairs of the same ammunition brand, but fired from different brands of guns. The standard deviations of the I-2D groups in this table range from 5 to 27. Are there ammunition effects and gun brand interactions?

Table 9-19. Average I-2D scores from firing pin impressions for casing pairs fired by different brands of weapons with the same brand ammunition.

	Win.	Rem.	PMC
Ruger-Sig Sauer	16	16	17
Ruger-S&W	64	42	76
Sig Sauer-S&W	21	20	19

The Rugers and S&Ws correlate much more with each other than with the Sig Sauers, especially when PMC and Winchester ammunition is used for both casings. For the large pool of pairs of casings that have both different gun brands and different ammunition brands, the mean score is 31 and the standard deviation is 26. An analysis of variance (ANOVA) showed that ammunition and gun brand effects, as well as interactions, were significant.

9.9.2 NBIDE Breech Face Impressions

Table 9-20 shows average I-2D score values for various matching casing pairs divided into groups by guns and by ammunition combinations. The first three columns of numbers are for the groups where both casings are the same ammunition brand. The columns on the right are when the casings are from different ammunition brands. What are the effects of the gun and ammunition brands?

Table 9-20. Average I-2D scores from NBIDE breech face impressions for casing pairs fired by the same weapon.

Gun ID	Same Ammo Brand				Different Ammo Brands			
	Win.	Rem.	PMC	Average	Win-Rem	Win-PMC	Rem.-PMC	Average
Ruger 41	81	62	153	99	46	40	28	38
Ruger 42	80	108	185	124	73	101	84	86
Ruger 46	78	170	148	132	42	101	51	65
Ruger 48	89	89	183	120	93	118	102	104
Sig Sauer 30	47	40	63	50	23	38	18	26
Sig Sauer 31	38	42	35	38	25	42	38	35
Sig Sauer 32	71	65	35	57	52	51	35	46
Sig Sauer 33	22	48	37	36	19	24	13	19
S&W 305	72	49	69	63	46	66	43	52
S&W 306	56	61	91	69	48	41	64	51
S&W 314	26	44	75	48	24	33	54	37
S&W 401	66	51	67	61	34	39	38	37
Average	61	69	95	75	44	58	47	50

For every gun, the average score is higher when the reference and compared ammunition brands are the same rather than different. The scores are especially high when both casings are PMC and the gun is a Ruger. The Sig Sauers produce lower match scores, but as will be seen below, they also produce lower non-match scores, although not as much lower as for the firing pin impressions tabulated in the previous subsection. The standard deviations of the scores of those casing pairs from the same gun vary widely from 5 to 67.

Table 9-21 shows the average I-2D score values of non-matching casing pairs fired from different guns of the same brand, organized by ammunition brand combination. What are the effects of gun and ammunition brand?

Table 9-21. Mean I-2D scores from NBIDE breech face impressions for casing pairs fired by different weapons of the same brand.

Gun Brand	Ammunition					
	Win.	Rem.	PMC	Win-Rem	Win-PMC	Rem.-PMC
Ruger	36	51	66	34	41	35
Sig Sauer	19	23	13	17	14	11
S&W	18	24	19	18	15	18

The average non-match scores are highest when both the reference gun and the compared guns are Rugers. The standard deviations of the scores in the group run from 4 to 11 for the Sig Sauers and S&Ws, but from 14 to 40 for the Rugers.

Table 9-22 shows the mean I-2D scores of casing pairs of the same ammunition brand, but fired from different brands of gun. Are there ammunition and gun interaction effects?

Table 9-22. Mean I-2D scores from NBIDE breech face impressions for casing pairs of the same ammunition brand fired by different guns.

	Win.	Rem.	PMC
Ruger-Sig Sauer	11	10	9
Ruger-S&W	17	20	17
Sig Sauer-S&W	11	12	10

These average non-match scores are low, but again the lowest scores occur when one of the guns is a Sig Sauer. The standard deviations of the I-2D scores in this table range from 4 to 9. For the large pool of pairs of casings that have both different gun brands and different ammunition brands, the mean score is 12 and the standard deviation is 7. An analysis of variance showed that ammunition and gun brand effects, as well as interactions, were significant.

There is more evidence of gun and ammunition brand effects and interactions for the I-2D NBIDE data than for the N-3D data of the same casings. The most prominent effects in the I-2D data are:

1.   Matching scores are higher if the casings are the same ammunition brand, especially PMC.

2.   Sig Sauers give lower scores both for matches and non-matches. The Rugers and S&Ws are better correlated with each other than with the Sig Sauers.

For the N-3D NBIDE data, there are hints of similar trends, but ammunition and gun brand effects appear more pronounced for the I-2D data. Further study is needed for a fuller quantification of these effects.

9.10.   Probability Models

The first three subsections of this section discuss some theoretical models and their implications. In Sec. 9.10.4, these models are applied to the N-3D data of the NBIDE casings. It will be clear that all but one of the data sets examined, De Kinder or NBIDE, suggest that current technology is not good enough to support a very large ballistics database. The case of the N-3D analysis of the NBIDE breech faces is a special case and merits a separate discussion, including why it is so different from the De Kinder breech face results.

9.10.1 Simple Binomial Model

In this section we use a binomial model with the overlap metric parameter p to analyze the scenario of a casing from a crime scene being correlated with all the casings in a database. The casings in the database that are chosen for closer scrutiny by a ballistics examiner are those that correlate most highly with the crime scene casing, which will be called the reference casing.

Suppose that there is actually a casing from the same gun in the database, so that it should be a match for the test casing. Let there be N other casings in the database, where N is a suitably large number. For the real match to make a Top Ten list like those produced by the I-2D system, only nine or fewer of the N cross-correlations with non-matching casings may be greater than the cross-correlation with the real match.

For a first pass model, given several simplifying assumptions (Nair [54] has developed a formulation that goes beyond these assumptions), the number of casings in the database that yield a higher cross correlation with the reference casing than does the real match can be modeled as a binomial distribution, Binomial (N, p), where p is the relevant overlap metric. In layman's terms, this is akin to flipping N coins, each with p being the probability of tails, and hoping to get 9 or fewer tails.

In this model, the average number of non-matching correlations higher than the true matching correlation increases linearly with N. De Kinder found empirical evidence of such a linear relationship in his study when investigating the average rank of the true matching casing compared to the other casings.

9.10.2 Some Numbers

This crude probability model makes possible some approximate statements on how good the correlation methods have to be in order to be successful. For instance, suppose that the database has $N = 10\ 000$ members with the same class characteristics. How small does p have to be in

order to have the correct casing in the Top Ten at least 99 percent of the time? In probabilistic terms, how small does p have to be in order that, if X is a Binomial (N, p) [55] random variable, the probability,

$$\text{Prob}\,(X<10)=\sum_{i=0}^{9}\frac{N!}{i!\,(N-i)!}\times p^{i}(1-p)^{N-i}\geq\cdot 0.99\,?\qquad\qquad(13)$$

One concrete relationship to keep in mind is that by the properties of the Binomial distribution, if $N\times p=10$, then the probability of the matching casing being in the Top Ten is only around 0.46. Therefore, if N is very large, then p has to be accordingly small. In fact, p needs to be approximately 4/N to get the right match in the Top Ten 99 percent of the time. To get in the Top Ten 90 percent of the time, p can be around 6.2/N.

From this we can make statements of the sort, "If your database is *that* big, then your imaging and correlation techniques better be *that* good to have a reasonable probability of finding a match in it." For instance, if the database has 100 000 entries, then p needs to be on the order of 6.2×10^{-5} or smaller to have a 90 percent chance of getting the correct match in the Top Ten. 100 000 entries has been suggested as a reasonable size for a national database of 9 mm Luger type ammunition [56]. For sake of specificity, this is a typical, representative, and reasonable value for the population size."

9.10.3 Levels of Grouping for Casings and Guns

Note that all of the above applies to the chances for a single casing. Producing a model that describes the performance of a group of casings or a group of guns is more complicated. There are several levels to which the model can be refined. We consider three grouping levels here: a single grouping with a single value of p for all casings and guns, a different grouping for each gun, and a different grouping for each casing. Other types of grouping are also possible, such as grouping by casing brand or by gun brand or by a combination of those.

Single p
Suppose the same matching and non-matching distributions can be used for all casings and guns. Then the probability model in the previous sections can be used without modification to refer to all casings such that the same coin with the same p is being flipped for each reference casing. If X is a Binomial (N, p) random variable and we define

$P(N, p) = \text{Prob}\,(X<10),$

then $P(N, p)$ is essentially the probability that the real match successfully makes it to the Top Ten list, and thus we refer to it as a success rate. Suppose that the specified performance goal is that, given a single matching casing mixed with N non-matching casings in the database, the matching casing is included in a Top Ten list D % of the time. Then, performance is considered satisfactory if $P(N, p) \geq D/100$.

Here we usually set $D = 90$. The criterion of 90 % seems to be a conservative and reasonable criterion for estimating a desirable efficiency of a large database. This target success rate then

enables us to extrapolate to criteria that might be expected for an experimental database with a small number of entries such as the two collections we studied. The success rate for a small database needs to be extremely high in order to be consistent with the accuracy needs of a large database. More generally, the accuracy criterion depends on whether the gains in the number of matches and the aid to investigators will be sufficient to warrant the cost of developing and maintaining a large database.

Grouped by Guns
There is variability between guns. When there are multiple firings from each gun, we can form separate matching and non-matching distributions for each gun, resulting in a different p for each gun. Thus if a certain gun has an overlap metric p, its casings would tend to make the Top Ten list with probability $P(N, p)$. For a set of guns, each with its own p and $P(N, p)$, then the average success rate of the group of guns is the average of the gun success rates, i.e. the mean of the $P(N,p)$ values for each gun.

To give a simple example, suppose for a fixed database size N there are ten guns of which eight have perfect discrimination ($p = 0$ and $P(N, p) = 1$), and two guns have p so large that $P(N, p) = 0$. Then the average success rate $P(N, p) = 0.8$, so 80 % of the guns' casings would make the Top Ten list.

Another useful success criterion is to consider the proportion of guns that would satisfy a particular success rate for a given N. For the simple example above, suppose that the target success rate is 90 %; then eight of ten would meet the target success rate. For very large N, such as $N=100\ 000$, this success criterion will often be close to the average success rate because the individual gun success rates tend to be close to either zero or one for most target success rates.

Grouped by Casing
There can also be variability between casings fired from the same gun. When there are multiple firings from each gun, we can form separate matching and non-matching distributions for each casing, resulting in a different p for each casing. Thus if a certain casing has an overlap metric p, it would tend to make the Top Ten list with probability $P(N, p)$. For a set of casings, each with its own p and $P(N, p)$, then the success rate of the group of casings is the average of the casing success rates, i.e. the mean of the $P(N, p)$ values for each casing. The percentage of casings that satisfy the target success rate for a given N can also be used. Similar to the case for guns, this percentage tends to be close to the average success rate for very large N.

Note the requirement for multiple firings from the same gun. If each gun fired $m+1$ casings, then each casing has only m correct matches. It may be difficult to get good estimates of p using pair-wise comparisons because of the relatively small number of comparisons that can be made. This can be especially problematic because we are most interested in very small values of p, and the pair-wise comparisons can yield estimates of p only as multiples of $1/(mn)$, where n is the number of non-matching $ACCF_{max}$ values per casing. This may lead to too many estimates of zero for the value of p, as well as estimates of p that may be substantially high, as these estimates depend on how many values from the two samples overlap.

One solution to this problem is to fit continuous distributions to the matching and non-matching samples. These distributions can yield estimates of p that are non-zero but very small. Of course there would remain the problem of whether the fitted distribution is an appropriate fit, and how good the fit is, given the limited sample size. In this report, we fit normal distributions using the estimated means and variances of each sample. It is of course possible to use different distributions.

Discussion of Groups

Suppose that the different levels of groups produce substantially different overlap metrics. Then one can draw different conclusions depending on the level of grouping. In general, the more numerous and more refined groups will lead to more optimistic conclusions, while having fewer groups that are more pooled will lead to more pessimistic conclusions. That is because success in a very large database demands a very small value of p. Thus, individual casings that have bad distinguishability qualities will increase the estimated p of their member group to high levels. Having a smaller group limits the damage done by a single casing. To use a golfing analogy, playing extremely poorly on one hole is much more harmful in stroke play (where every stroke counts) than in match play (where only holes won or lost count). This phenomenon is seen with the NBIDE breech faces.

9.10.4 Experimental Results

Tables 9-23 and 9-24 recap some of the N-3D overlap metric results for the individual casings. Refer to Figs. 9-9, 9-14, 9-19, 9-24, and 9-25 for histograms of the overlap metric estimates.

Table 9-23. Distribution of pair-wise comparison estimates of the overlap metric (p) for the individual casing model.

Fraction of estimated p measures	=0	≤0.01	≤0.1	Data plotted in
De Kinder FP	0.24	0.31	0.46	Fig. 9-9
De Kinder BF	0.014	0.03	0.21	Fig. 9-14
NBIDE FP	0.18	0.25	0.56	Fig. 9-19
NBIDE BF	0.90	0.95	1.0	Fig. 9-24

Table 9-24. Distribution of the overlap metric (p) obtained from normal model estimates for NBIDE BF; data plotted in Fig. 9-25.

Fraction of estimated p measures	= 0	$\leq10^{-6}$	$\leq10^{-5}$	$\leq10^{-4}$	$\leq10^{-3}$	≤0.01	≤0.1
	0.11	0.55	0.63	0.72	0.81	0.94	1

The results are clear-cut for all pair-wise comparison estimates except the NBIDE breech face impressions, which will be discussed at the end of the section. For the other three, let's assume the case of the most optimistic grouping (by casing). For a very large database of size $N =$ 100 000, given the limited resolution of the estimates possible for limited sample sizes, the only estimated values of p small enough for a reliable ballistics identification system are those that are essentially zero. From Table 9-23, only 18 % to 24 % of the firing pin impressions satisfy this criterion; the percentage is much lower for the De Kinder breech face impressions. Essentially

the same results occur if N = 10 000. By making use of the relationship described in Sec. 9.10.2, we estimate that even if N is as small as 100, the proportion of casings meeting the target success rate of 90 % for Top Ten lists for the De Kinder sets is still less than 50 %. The analogous proportion of casings meeting the target success rate would be less than 56 % for the NBIDE firing pin impressions. For comparing the various technologies applied to different impression sites, the proportion of correct matches found in the Top Ten experiments described earlier can serve as a useful guide to system performance comparisons. These averages are summarized in Table 9-25.

Table 9-25. Proportion of correct matches in Top Ten experiments

I-2D De Kinder FP	3.06/6 = 0.51
I-2D De Kinder BF	1.01/6 = 0.17
N-3D De Kinder FP	3.26/6 = 0.54
N-3D De Kinder BF	2.83/6 = 0.47
I-2D NBIDE FP	3.72/8 = 0.46
I-2D NBIDE BF	5.57/8 = 0.70
N-3D NBIDE FP	5.63/8 = 0.70
N-3D NBIDE BF	7.94/8 = 0.99
I-2D De Kinder BF & FP	3.39/6 = 0.57
N-3D De Kinder BF & FP	4.77/6 = 0.80
N-3D De Kinder BF + FP	4.23/6 = 0.71 (Sum Method)
I-2D NBIDE BF & FP	6.20/8 = 0.78
N-3D NBIDE BF & FP	7.99/8 = 0.999

Again, both I-2D and N-3D did much better on NBIDE Breech Face than on De Kinder Breech Face, and N-3D did very well. The I-2D performance for Firing Pin was similar for the two data sets. For N-3D, the performance for NBIDE Firing Pin was somewhat better than for De Kinder Firing Pin. For the De Kinder set, I-2D was much better and N-3D somewhat better for Firing Pin than for Breech Face. In contrast, for the NBIDE data, both N-3D and I-2D were substantially better on Breech Face than on Firing Pin.

NBIDE Breech Face
The NBIDE Breech Face impressions are drastically different than anything else seen. Under the most optimistic scenario of grouping by individual casings, for a database of size N=100 000, 90 % of the casings meet the target success rate, using the pair-wise comparison estimates of p (See Table 9-23). For fitted normal model estimates of p, Table 9-24 suggests that between 63% and 72% of the casings have p metrics of 6.2×10^{-5} or less, a value of p small enough to be consistent with the accuracy requirement suggested in Sec. 9.10.2.

If instead, there is grouping by guns, then only about 50% of the guns have a p metric of 6.2×10^{-5} or less (see Table 9-8). Under the pessimistic scenario of a single group, the estimated mean value of $p = 0.002$ remains over 30 times too large, despite being orders of magnitude smaller than anything else seen.

Discussion

For a technology to be feasible for a very large database, its Top Ten lists should have obtained close to all possible correct matches in a relatively low sample size experiment like those described in this report. Nothing we have seen comes close to achieving such high performance standards except for the N-3D performance on the NBIDE breech face impressions, which suggests that topography methods are a significant advance for breech face analysis. However, any resulting claims for topography images of breech face has to be reconciled with the much less impressive performance of the same technology on the De Kinder casings. Gun differences may be a main cause of the differences, as the De Kinder casings all were fired from Sig Sauers. However, the NBIDE study also included casings fired from Sig Sauers, and the subset of topography results from those Sig Sauer casings are still much better than the topography results for De Kinder Breech Face. It also has been speculated that the higher quality ammunition used in the NBIDE study produced clearer breech face markings. How promising 3D technology is for very large databases depends on whether the NBIDE or De Kinder results are more representative of the challenges faced by a national database. Also, one must remember that only 108 casings were fired by only twelve guns from three different brands. Presumably a larger population of guns would make more likely the presence of unfortunate large correlations from non-matching guns. Also, gun brands and models not covered in the NBIDE study (e.g., low-cost guns that were no longer available as new purchases) may well be more difficult to distinguish and identify than those included.

In order to perform at levels necessary for very large databases, say around 100 000 guns of the same class, the error rates must be very low—so low in fact that for experiments with only 70 or 108 casings, as in this report, essentially the only way to achieve such low error rates is for there to be no overlap between the matching and non-matching samples. While there was considerable separation between matching and non-matching distributions for many of the reference casings, especially those fired from Rugers, others had much less margin for error in that the matching correlations were only slightly larger than the largest of the non-matching correlations. Those matching $ACCF_{max}$ values would be in danger of being overtaken by non-matching $ACCF_{max}$ values in a very large database with a much larger population of non-matching $ACCF_{max}$ values. For each individual casing in the NBIDE set, there were only eight $ACCF_{max}$ values of casings in the matching sample and 99 $ACCF_{max}$ values of casings in the non-matching sample. Thus, one can try to estimate the distributions by pooling the matching and non-matching samples for each gun; however, this likely makes the estimated distributions wider than they should be (and in fact would estimate that only half the guns would be successful using the NBIDE breech face impression data). Estimating very low probabilities with moderately low sample sizes continues to be a challenging problem. We used normal probability models for the correlation scores themselves in an attempt to ameliorate the problem. Use of the normal models lowered the success rate for the optimistic scenario of grouping by casing.

The topography methods are of relatively recent vintage, and as such are still being continually refined and improved. There are still questions on the best way of handling data processing and optimal cross correlations, including measures different from the $ACCF_{max}$. It is possible that refinements will result in great improvements in future. The topography methods look promising for breech face impressions, but improvements are still needed, as is investigation into the factors necessary to obtain the required accuracy.

9.10.5 Normal-Binomial Models

Nair's [54] probabilistic formulation goes beyond the simple binomial model described in Sec. 9.10.2 in dealing with dependencies. In addition, he produces a framing of the problem in terms of modeling the matching and non-matching score distributions by normal models, although he notes that the model does not rely on adherence to normality to produce useful results. Sections 9.8 and 9.9 of this report contain the means and standard deviations of groups of $ACCF_{max}$ values for various pairings of gun and ammunition types. The standard deviations of the combined groups tend to be larger if the groups are disparate. Nair notes that there will be better database performance if the standard deviations of matching scores are lower than those of non-matching scores (his 'optimistic scenario') and worse if the standard deviations are equal ('pessimistic scenario'). Unfortunately, we have found that in most cases, the matching scores contain more variability (and thus have higher standard deviations) than the non-matching score distributions. Thus, these results using his model would be even less promising than those for his pessimistic scenario. Note for some fingerprint algorithms, it is similarly the case that the non-matching scores are more tightly bunched, and the matching scores are more spread out [53].

10.   Statistical Analysis: Gun Distinguishability

10.1   Introduction

The main purpose of the present study is to determine if guns are uniquely distinguishable and identifiable, and if so—for the two data sets at hand (De Kinder and NBIDE)—what that implies about distinguishability/identifiability for a much larger (e.g., national) database of guns.

The starting point for this analysis is the Top Ten list that the N-3D analysis produced. For the De Kinder data, the N-3D cross-correlation analysis produced a Top Ten list for each of the 70 test fired casings. For each of these 70 reference casings, a correlation $ACCF_{max}$ was computed using all of the remaining 69 casings as a comparison group. Since the De Kinder experiment consisted of ten guns and seven ammos per gun, that means that of the 69 comparison casings, 63 (= 9 guns × 7 ammos) will not be a match, while six (= 1 gun × the 6 remaining ammos) will be a match. Hence ideally, a gun would be considered "distinguishable" if all 6 of the remaining ammo firings from that gun show up in the Top Ten list. On the other hand, if only a few (or none) of the six matching casings show up in the Top Ten list for a given reference gun, that particular gun has poor distinguishability. Thus for the De Kinder study, our distinguishability metric is the number of comparative matches (0 to 6) that a particular reference gun ID had in the N-3D Top Ten list. Six represents excellent distinguishability, while zero represents no distinguishability. Further, an individual gun would be considered distinguishable if all seven test firings that used that gun as a reference had all six of their correct matches in the seven Top Ten lists. If all ten guns behaved in this desirable fashion, we would have universal distinguishability for the ten guns (and seven ammos) used in this test.

Similarly, an analogous metric was used for the analysis of the NBIDE data. In this case, however, there were twelve distinct guns, 3 ammos, and 3 repeats (yielding a total of 108 firings) and so a gun would be considered "distinguishable" if for any of its fired casings, all eight (= 3 ammos × 3 repeats − 1) remaining casings fired from that gun appear in the Top Ten list when compared with it. If all twelve of the guns behaved in this fashion, then we would have universal distinguishability for the twelve guns (and three ammos) used in this test.

We now examine the individual gun distinguishability for each of the four combinations of two databases (De Kinder and NBIDE) and two imaging regions (Firing Pin and Breech Face).

10.1.1 Individual Guns (De Kinder / Firing Pin)

The De Kinder experiment had one gun type, ten distinct guns, and seven ammos, for a total of 70 firings. We address the following five questions:

Q1.   Are the ten guns distinguishable?
Q2.   Are some guns more distinguishable than others?
Q3.   What is the best (easiest) gun to distinguish?
Q4.   What is the worst (most difficult) gun to distinguish?
Q5.   What is the distinguishability ranking (best to worst) of the ten guns?

To address these questions, see the pair of plots in Fig. 10-1. The top plot is a scatter plot; the bottom plot is a mean plot. The horizontal axis of both plots is gun ID. There are ten De Kinder reference gun IDs: 7, 9, 117, 139, 213, 215, 314, 375, 430, and 535. These ten IDs reflect the fact that the De Kinder experiment had ten guns randomly drawn from a larger population of guns.
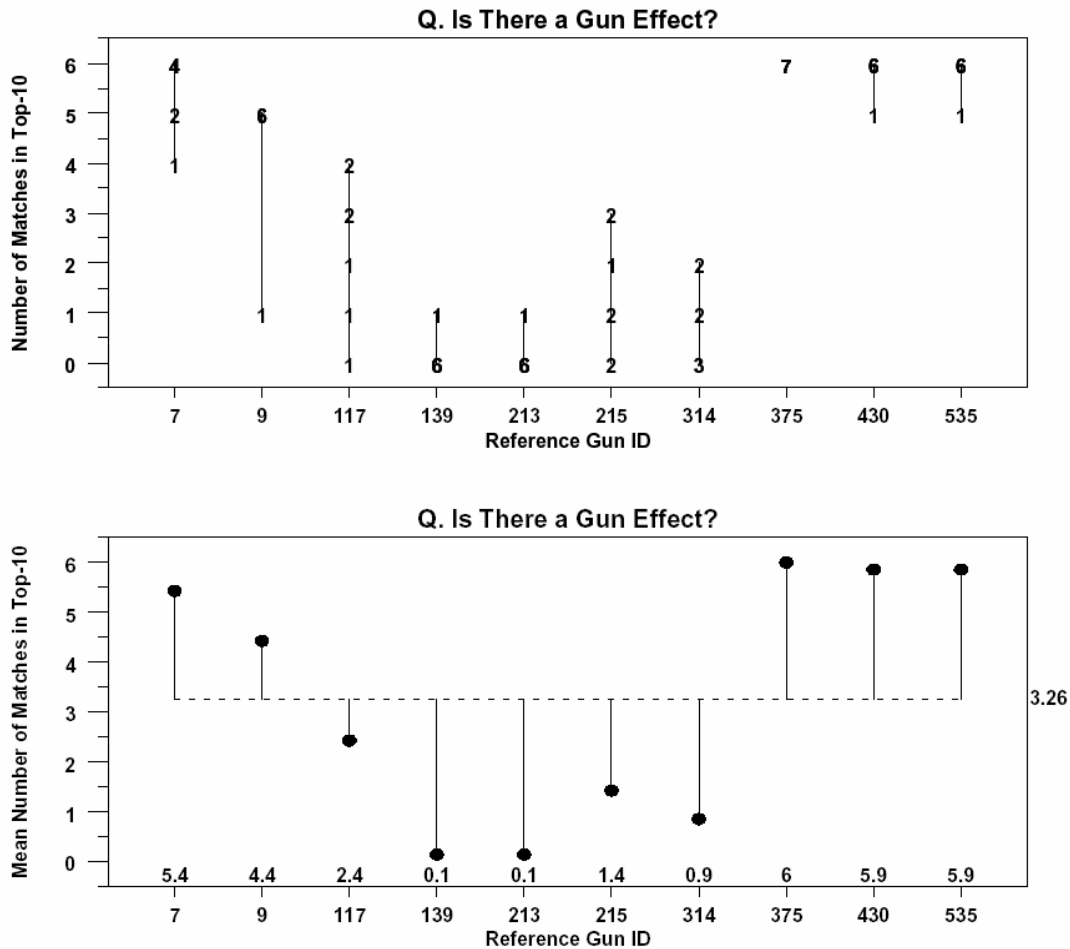

Figure 10-1. De Kinder / Firing Pin $ACCF_{max}$ Top Ten analysis for a gun effect .

The vertical axis of Fig. 10-1 is "score" for an individual gun as a result of the 70 De Kinder firings. Out of those 70 firings, an individual gun is involved in seven such firings. The casings for those seven firings —with each used as a reference with the remaining 69 casings used as a comparison set—result in seven Top Ten lists, and these seven Top Ten lists contain some number of its six matching casings. The vertical axis is 0 to 6. A score of 6 represents perfect distinguishability—for any reference casing, all six remaining casings fired by the same gun appear in the Top Ten list. A score of 0 says that none of the six matching casings appear in the Top Ten list, and hence that reference casing (and gun) is indistinguishable.

146

Above each reference gun ID on the horizontal axis, there should be 7 marks (one for each time the gun was used in a test firing). To accommodate overstriking, the plot character represents the number of times a score occurs. Hence for the first gun (gun 007), the "4" appearing at the vertical axis value of "6" means that out of the seven firings for this gun, exactly four of the seven Top Ten lists contain all six matches. Moreover for gun "007" there are two Top Ten lists that have five of six matches, and one Top Ten list which has four of six matches.

The ideal distinguishable gun would have "7" at $Y = 6$ —all seven test firings for this individual gun yielding Top Ten lists having all six of its matching casings. Universal distinguishability would have all ten guns with 7's at $Y = 6$.

The top plot is the scatter plot of raw scores. The bottom plot is the mean of the seven raw scores for each gun. A reference gun with high (= 6) mean score implies distinguishability; low mean score indicates non-distinguishability. The numbers above the lower horizontal axis (5.4, 4.4, 2.4,...) are the mean scores for each gun.

From Fig. 10-1 we conclude that the guns are not universally distinguishable. Only gun 375 has a perfect score of 6. Six of the ten guns have at least one score of 1 (= poor). The distinguishability ranking of the ten guns is as follows:

1. 375 (mean score = 6.0) The best gun in terms of distinguishability.
2. 430 (mean score = 5.9) This gun is near-distinguishable.
3. 535 (mean score = 5.9) This gun is near-distinguishable.
4. 007 (mean score = 5.4) Distinguishable for some ammos but not others.
5. 009 (mean score = 4.4) Distinguishable for some ammos but not others.
6. 117 (mean score = 2.4) Poor distinguishability for most ammos.
7. 215 (mean score = 1.4) Poor distinguishability for all ammos.
8. 314 (mean score = 0.9) Poor distinguishability for all ammos.
9. 139 (mean score = 0.1) The co-worst. This gun is not distinguishable.
10. 213 (mean score = 0.1) The co-worst. This gun is not distinguishable.

These De Kinder findings are poor. With ten nominally identical guns, one would expect more consistency across all ten guns. Such consistency is not in evidence. We note finally that the observed differences across the ten guns are statistically significant at the 5 % level (that is, the observed data could happen by chance at most 5 % of the time).

In the ideal, the ten guns should all have perfect scores of 7 at Y=6, and there should not be a statistically significant difference across the ten guns. Neither condition was observed for the De Kinder firing pin impression data, which implies insufficient distinguishability (and hence non-feasibility) for the issue of the much larger national database.

10.1.2 Individual guns (De Kinder / Breech Face)

To address the same five questions as in the previous section, but to focus on the De Kinder breech face data, we make reference to Fig. 10-2. From the figure we conclude that the guns are also not universally distinguishable. At the 5 % level, the differences between guns are statistically significant. Further, the best guns (375, 430, and 535) from the De Kinder firing pin analysis are not the best guns for the De Kinder breech face analysis. The best Breech Face gun (215) was the third worst Firing Pin gun. No Breech Face gun achieved a perfect score of 6, contrary to gun 375's perfect score for Firing Pin.

The average score for the breech face analysis is 2.83 (out of 6), which is smaller than the average score (3.26) for the firing pin analysis, which reaffirms that for the De Kinder data, Firing Pin was a better discriminator than Breech Face.

Figure 10-2. De Kinder / Breech Face $ACCF_{max}$ Top Ten analysis for a gun effect.

With 6 being a perfect score, the ranking of the guns for De Kinder / Breech Face is as follows:

1.   215   (mean score = 4.4)

2.	117	(mean score = 4.0)
3.	430	(mean score = 3.7)
4.	139	(mean score = 3.3)
5.	213	(mean score = 3.0)
6.	535	(mean score = 2.6)
7.	314	(mean score = 2.1)
8.	009	(mean score = 1.9)
9.	375	(mean score = 1.7)
10.	007	(mean score = 1.6)

Gun 215 is the most distinguishable gun, with high or moderate scores for all ammos. Guns 117 and 430 are next best, but their scores range from 6 to 1. The remaining guns have only moderate to poor distinguishability. The worst two guns are 007 and 375, which for some ammos had no matches in the Top Ten list.

10.1.3 Individual Guns (NBIDE / Firing Pin)

The NBIDE experiment had three gun types, four replicates of each gun type, a total of twelve distinct guns, three ammos, and three days(reps), for a total of 108 test firings. We here address five questions similar to those of Sec. 10.1.1:

Q1. Are the twelve guns distinguishable?
Q2. Are some guns more distinguishable than others?
Q3. What is the best (easiest) gun to distinguish?
Q4. What is the worst (most difficult) gun to distinguish?
Q5. What is the distinguishability ranking (best to worst) of the twelve guns?

To address these questions, we use Fig. 10-3. The horizontal axis of both plots in Fig. 10-3 shows the twelve NBIDE reference gun IDs: 1, 2, 3, ..., 11, 12. Below these horizontal axis tic labels is a second row of identifying labels: S1, SW5, R1, etc. The labels refer to the gun types as follows:

Gun 1	Sig Sauer 31	S1
Gun 2	S&W 305	SW5
Gun 3	Ruger 41	R1
Gun 4	S&W 306	SW6
Gun 5	Ruger 42	R2
Gun 6	Sig Sauer 32	S2
Gun 7	S&W 401	SW1
Gun 8	Sig Sauer 30	S0
Gun 9	Ruger 46	R6
Gun 10	Sig Sauer 33	S3
Gun 11	S&W 314	SW4
Gun 12	Ruger 48	R8

The last digit of each gun's serial number was used for the digit in the abbreviated ID name (e.g., gun 2 is shortened to SW5 where the 5 comes from the last digit of its serial number: PBV8305). With these identifiers and from Fig. 10-3, we can arrive at initial conclusions about the effect of the three gun types, but we postpone that gun type discussion to a later section (10.2.3), and continue now to focus on the core question at hand of distinguishing between the twelve individual guns.

The vertical axis of Fig. 10-3 shows the results of all 108 firings. There should be a total of 108 data points on the plot (and 9 data points above each reference gun ID), but again due to the over-striking, the plotted character represents the frequency of identical scores. The vertical axis is 0 to 8. A score of 8 represents perfect distinguishability—that is, the chosen reference casing has all eight matching comparison casings (3 ammos × 3 days minus the reference casing) appearing in the Top Ten list. A score of 0 indicates that none of the eight matching casings appeared in the Top Ten list and hence that reference casing/gun is indistinguishable.

As before for the ideal, if a gun were perfectly distinguishable, then there would appear on the plot above the gun ID a "9" at the $Y=8$ level. If all twelve guns were universally distinguishable, every one of the twelve reference gun IDs would have a "9" at the $Y=8$ level.

Also as before, the bottom half of Fig. 10-3 is not the nine individual scores, but rather the mean of the nine scores. A reference gun with high mean score implies distinguishability; low mean score indicates non-distinguishability.

From Fig. 10-3 we find that there is a (statistically significant) difference in the twelve guns and the average score is 5.63 (out of 8). Hence we find that the twelve guns are not universally distinguishable. Some guns are more distinguishable than others. Gun 5 has a perfect mean score of 8, followed by gun 2 with a mean score of 7.7. At the other extreme, gun 1 is poor, with none of its nine reference casings having more than five matches in the Top Ten lists. The ranked list of guns is as follows:

1.   5   (mean score = 8.0)
2.   2   (mean score = 7.7)
3.   6   (mean score = 6.9)
4.   3   (mean score = 6.7)
5.   12   (mean score = 6.7)
6.   11   (mean score = 6.4)
7.   10   (mean score = 5.1)
8.   9   (mean score = 4.9)
9.   4   (mean score = 4.3)
10. 7   (mean score = 4.2)
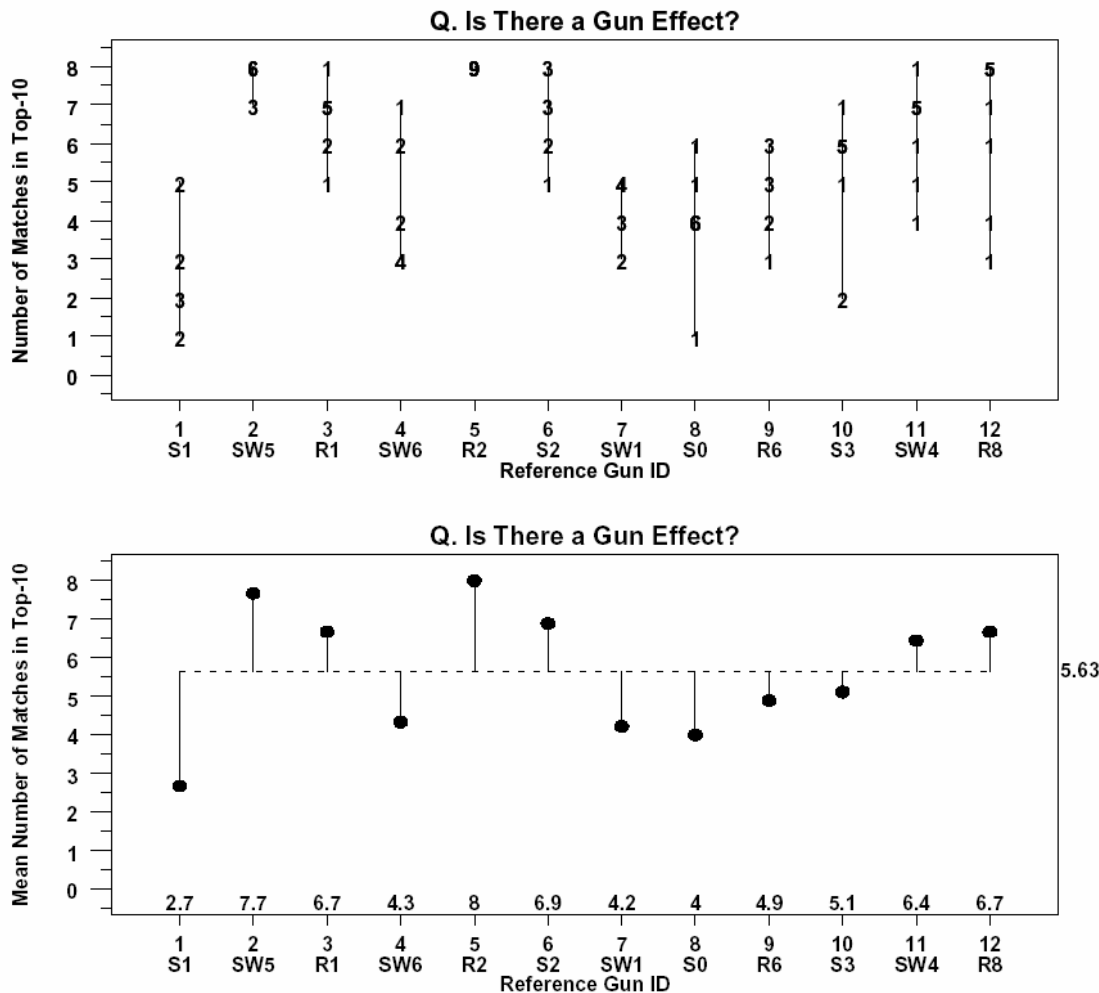11. 8   (mean score = 4.0)
12. 1   (mean score = 2.7)

Figure 10-3. NBIDE / Firing Pin $ACCF_{max}$ Top Ten analysis for a gun effect.

10.1.4 Individual Guns (NBIDE / Breech Face)

To address the same five questions as in the previous section, we make reference to Fig. 10-4. From this figure we conclude that the guns for NBIDE / Breech Face are very distinguishable. For seven out of the twelve guns, all nine of their reference test fires had all eight of the correct matching casings show up in the Top Ten lists. All twelve of the guns had at least seven test fires with eight correct matches. The ranked list of guns is as follows:

Rank 1 to 7.	Guns 1, 3, 4, 5, 7, 8, 9	(mean score = 8.0)
Rank 8 to 10 .	Guns 6, 10, 12	(mean score = 7.9)
Rank 11 to 12	Guns 2, 11	(mean score = 7.8)

The average score across all guns is high (7.94 out of 8), and the twelve guns are not statistically different. Hence, of the four data set / image region combinations considered, the NBIDE

Breech Face comes closest to satisfying the visual and statistical requirements for distinguishability.

Figure 10-4. NBIDE / Breech Face $ACCF_{max}$ Top Ten analysis for a gun effect.

10.1.5 Individual guns (Summary)

From the above four sections—and especially the mean plots—we find that distinguishability is highly dependent on the source of the data (De Kinder versus NBIDE) and the casing region being imaged (Firing Pin versus Breech Face). Based on the mean number of matches in the Top Ten list, the four cases are ranked as follows (best to worst):

1. NBIDE / Breech Face   (mean score = 7.94 out of 8 = 99.3 %)
2. NBIDE / Firing Pin   (mean score = 5.63 out of 8 = 70.4 %)
3. De Kinder / FiringPin   (mean score = 3.26 out of 6 = 54.3 %)
4. De Kinder / Breech Face (mean score = 2.83 out of 6 = 47.2 %)

As to the core question regarding the feasibility of a (large) national database and whether the imaging and analysis technology is accurate enough to make such a large scale database practical, it is clear that the only one of the four cases that might yield an affirmative is case 4 (NBIDE / Breech Face), with a mean score of 99.3 %. It is of research interest to determine and understand what made this particular combination perform so well. Even at that, major practical hurdles (processing/algorithm speed, gun wear/aging, etc.) would need to be addressed and overcome before the behavior for this small (= 108 firings) data set could be safely extrapolated to a large national system.

10.2   Other Factors Affecting Gun Distinguishability

The prior section deals with the central question of this study, namely whether it is feasible to identify (from a reference casing) the individual gun that fired the casing. We found that the NBIDE Breech Face data provided excellent identifiability, whereas the other cases (NBIDE / Firing Pin, De Kinder / Breech Face, and De Kinder / Firing Pin) yielded unacceptably poorer identifiability.

Given that, a related question arises as to what other factors affect the likelihood of matching a casing with an individual gun. We examine four such factors:

1.   Database
2.   Imaging Region
3.   Gun Type
4.   Ammo Type

10.2.1 Database

Database (De Kinder versus NBIDE) is a statistically significant factor. From the discussion in Secs. 10.1.1 through 10.1.4, we find that:

1.   The two databases (De Kinder versus NBIDE) are significantly different.

2.   The ranking of the two databases (better to worse) is:

NBIDE, mean of 7.94 and 5.63 (= 6.79 out of 8, or 84.9%),
De Kinder, mean of 3.26 and 2.83 (= 3.05 out of 6, or 50.8%).

3.   Overall, the best combination is NBIDE / Breech Face.

10.2.2 Imaging Region

Imaging Region is a statistically significant factor. Again, from the discussion in Secs. 10.1.1 through 10.1.4, we find that:

1.   The two imaging regions (Firing Pin versus Breech Face) are significantly different.

2.   The ranking of the two regions (better to worse) is:
      i.   Breech Face   (mean of 7.94 out of 8 (99.3%) and 2.83 out of 6 (47.2%)
                              = 73.2% overall),
      ii.   Firing Pin   (mean of 5.63 out of 8 (70.4%) and 3.26 out of 6 (54.3%)
                              = 62.4% overall).

3.   There is an interaction between region and database:
           Firing Pin is better than Breech Face for De Kinder, but
           Breech Face is better than Firing Pin for NBIDE.

4.   As before, the best combination is NBIDE / Breech Face.

10.2.3 Gun Type

Note that the question as to whether the matching scores are affected by gun type cannot be addressed from the De Kinder data, inasmuch as all 70 firings came from the same gun type, namely the 9 mm Sig Sauer P226. The NBIDE experiment does, however, shed light on this question with its 108 firings, twelve individual guns, and three gun types:

1.   Sig Sauer
2.   Smith&Wesson
3.   Ruger

Are these three gun types distinguishable? We address this question separately for each of the two imaging regions: Firing Pin and Breech Face.

Gun Type (NBIDE / Firing Pin)
For universal distinguishability, gun type should not have an effect—individual guns should be distinguishable irrespective of gun type. To assess the gun type effect for NBIDE / Firing Pin data, note Fig. 10-5.
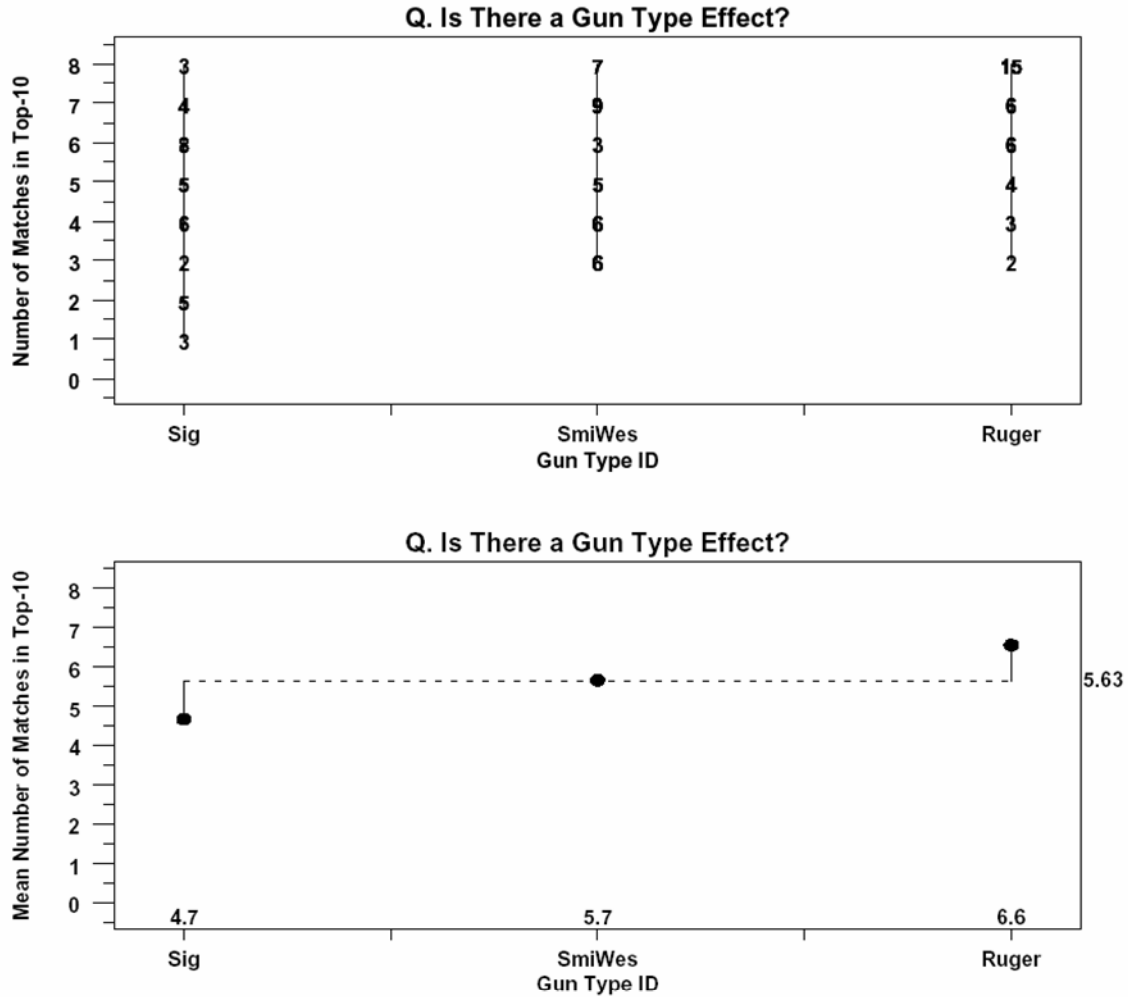
Figure 10-5. NBIDE / Firing Pin $ACCF_{max}$ Top Ten analysis for a gun type effect.

The three gun types are noted on the horizontal axis. As before, the numeric plot character indicates data frequency at that plot point. For example, for Ruger, out of its (108/3 =) 36 test firings, there were fully 15 instances in which all eight of the remaining Ruger casings showed up in the N-3D Top Ten list.

Figure 10-5 indicates that there is in fact a gun type effect—the three gun types are not equivalent in terms of their distinguishability. Some gun types are more amenable to being distinguishable than other gun types. The ranking (best to worst) of the three gun types is as follows:

1.   Ruger (mean score = 6.6),
2.   Smith&Wesson (mean score = 5.7),
3.   Sig Sauer (mean score = 4.7).

The difference across the three gun types is statistically significant at the 5 % level.

Gun Type (NBIDE / Breech Face)
The analysis of gun effects for the NBIDE / Breech face is given in Fig. 10-6.

Figure 10-6. NBIDE / Breech Face $ACCF_{max}$ Top Ten Analysis for a gun type effect.

For NBIDE / Breech Face, the three gun types are not statistically significant. This is not unexpected because for this case, almost all of the 108 firings yielded a full score of 8. NBIDE / Breech Face is the ideal: high distinguishability for the twelve guns, with other factors (in particular, gun type) not being statistically significant.

10.2.4 Ammunition Type

Conclusions about the distinguishability of gun type should ideally not be affected by ammunition type. However, it is useful common practice for examiners to use the same brand of ammunition for a testfire as that recovered from a crime scene. Because both the De Kinder and the NBIDE experiments were balanced with respect to ammo (seven ammo types for De Kinder and three ammo types for NBIDE), the analysis for ammo effects is straightforward. We assess the effect of ammo for the usual four cases: De Kinder versus NBIDE, Firing Pin versus Breech Face.

Ammunition Type (De Kinder / Firing Pin)
The De Kinder data set utilized ten guns, all of the same model (Sig Sauer 9 mm Model P226) and seven ammunition (cartridge types):

1.   CCI
2.   Winchester
3.   Remington
4.   Speer
5.   Wolf
6.   Federal
7.   Remington (a repeat)

Figure 10-7 examines whether an ammo effect exists. The horizontal axis shows the seven ammo types. The vertical axis is the usual matching score and mean matching score as used in previous figures. For the De Kinder data, each of the seven ammos should have ten points associated with it. If there were no ammo effect, Fig. 10-7 should be near flat, with about the same spread for all ammos. Visually, the ammos are near-equivalent. Quantitatively, at the 5 % level, the seven ammos are not statistically different.

Note that ammo types three and seven are both Remington and thus serve as an internal check on natural variability. As it turns out, the response for the two Remington ammos are near-identical and not statistically different.

Though not statistically different, the ranking of the 7 ammos is as follows:

        1.   Speer   (mean score = 3.8)
        3.   Federal   (mean score = 3.5)
        4.   Wolf   (mean score = 3.4)
        5.   Remington   (mean score = 3.2)
        6.   CCI   (mean score = 3.1)
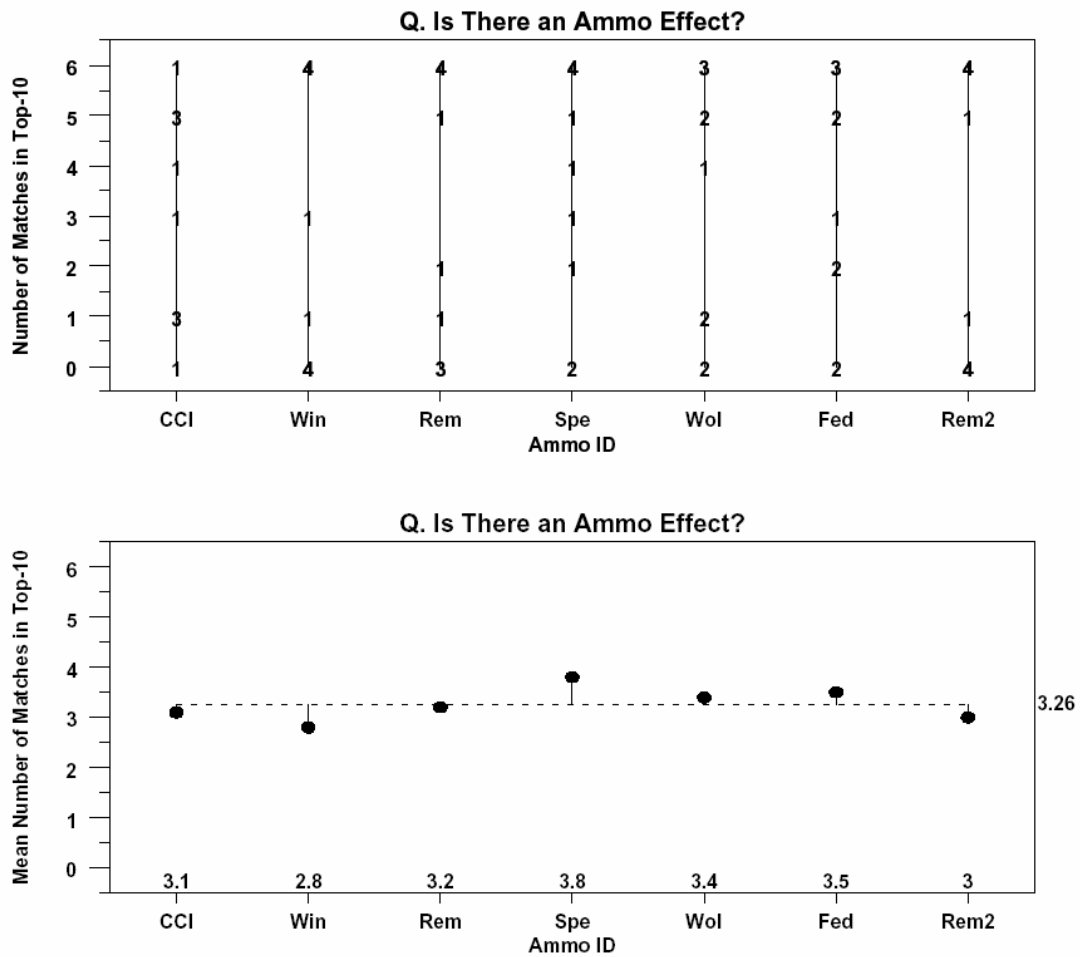   7.   Remington2   (mean score = 3.0)
        8.   Winchester   (mean score = 2.8)

Figure 10-7. De Kinder / Firing Pin $ACCF_{max}$ Top Ten analysis for ammo type.

Ammunition Type (De Kinder / Breech Face)

Figure 10-8 shows the De Kinder / Breech Face data for the question as to whether an ammo effect exists.
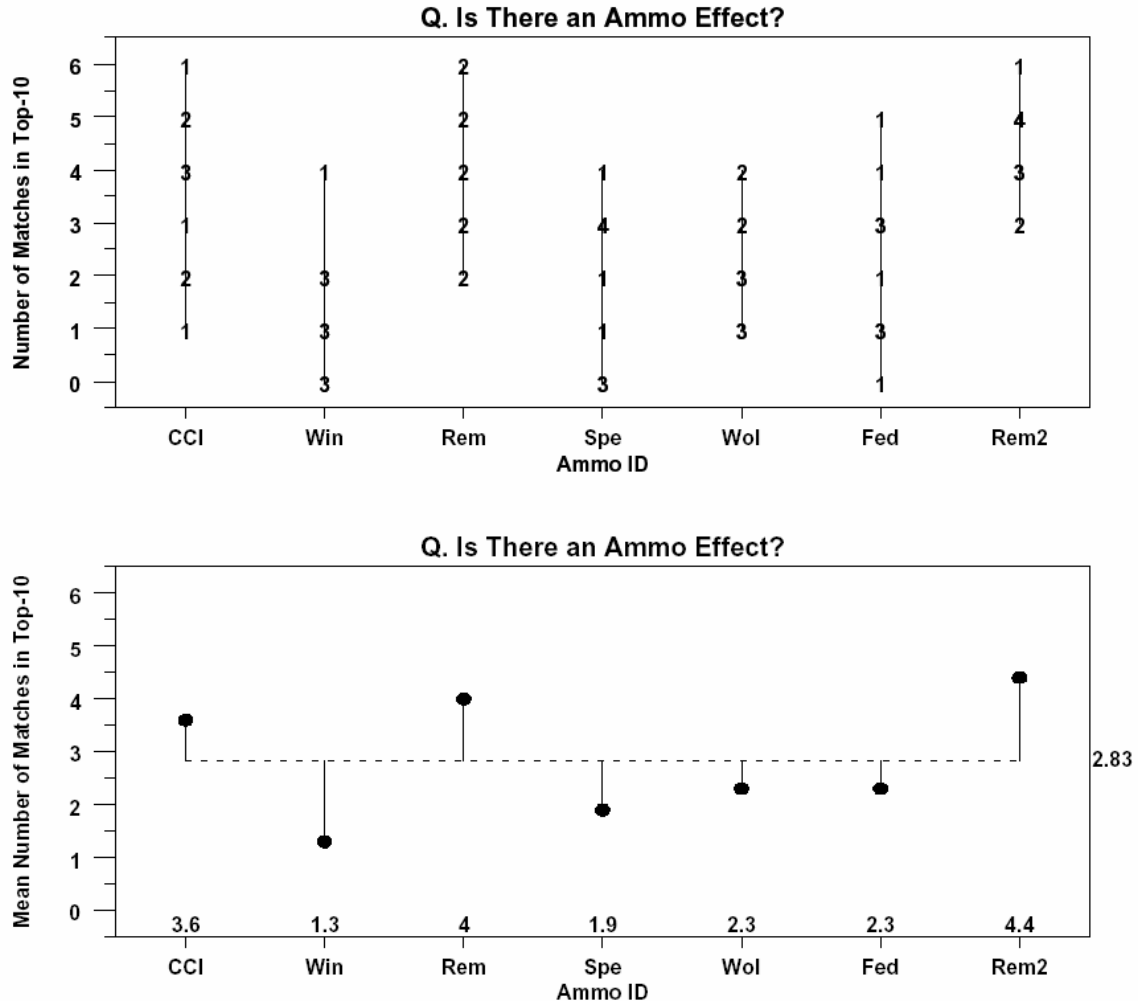

Figure 10-8. De Kinder / Breech Face $ACCF_{max}$ Top Ten analysis for ammo type.

The seven ammos are not equivalent. The seven ammos are statistically different at both the 5 % level and the 1 % level. The two Remington ammos are consistent and higher than the remaining five ammos. The ranking of the seven ammos is as follows:

1. Remington2 (mean score = 4.4)
2. Remington (mean score = 4.0)
3. CCI (mean score = 3.6)
4. Wolf (mean score = 2.3)
5. Federal (mean score = 2.3)
6. Speer (mean score = 1.9)
7. Winchester (mean score = 1.3).

159

Ammunition Type (NBIDE / Firing Pin)

The NBIDE experiment utilized twelve guns and three ammo types; three days of replication yielded a total of 108 firings. The three ammos were: 1-Remington, 2-Winchester, 3-PMC.

Figure 10-9. NBIDE / Firing Pin $ACCF_{max}$ Top Ten analysis for ammo type.

Figure 10-9 examines whether an ammo effect exists for NBIDE/Firing Pin. Though considerable overlap exists in all three ammos, the graph with PMC having an 11 at $Y = 8$ suggests that there may be a difference. Statistically, the ANOVA test statistic falls at the 94.5 % point, and so just misses significance at the 5 % level. In short, we reckon ammo type to be marginally significant, with PMC tending to yield more accurate matchings. The ranking of the three ammos is as follows:

1.   PMC   (mean score = 6.1)
2.   Remington   (mean score = 5.7)
3.   Winchester   (mean score = 5.0).

Ammunition Type (NBIDE / Breech Face)

Figure 10-10. NBIDE / Breech Face $ACCF_{max}$ Top Ten analysis for ammo type.

Figure 10-10 investigates for an ammo effect for NBIDE/Breech Face. Because for this case the twelve guns are very distinguishable, the ammo data will necessarily cluster near the $Y=8$ level. There appears to be little difference between the three ammos, although with PMC achieving a perfect score with all of its values at Y=8, there is again the hint that PMC may be doing slightly better than the other two ammos. Statistically, the ANOVA test statistic again falls at the 94.5 % point, and so just misses significance at the 5 % level, which would lead it to being marginally significant. With a maximum difference in the three averages being 0.1, this appears to be a case where the observed differences are statistically significant, but not practically different. The ranking of the three ammos is as follows:

1. PMC (mean score = 8.0)
2. Remington (mean score = 7.9)
3. Winchester (mean score = 7.9)

10.3   Relative Importance of Factors

10.3.1 Relative Importance of Factors (Graphical)

Subsection 10.2 examined individual factors and assessed whether they were significant or not. We finish this section on distinguishability by addressing what is the relative importance of the factors. In particular, we focus on the 3 factors:

1.   individual gun
2.   gun type
3.   ammo

Figures 10-11 through 10-14 examine the relative importance of factors for the usual four cases:

1.   De Kinder / Firing Pin`   (Fig. 10-11)
2.   De Kinder / Breech Face   (Fig. 10-12)
3.   NBIDE / Firing Pin   (Fig. 10-13)
4.   NBIDE / Breech Face   (Fig. 10-14)

Each individual plot has the multiple factors and the individual factor levels on the horizontal axis, and has the usual mean matching score on the vertical axis. Ideally, for universal distinguishability of gun type, the mean score for the gun should be high, there should be no statistical difference between the individual guns, and there should be no statistical difference among the secondary factors (gun type and ammo).

More to the point, previous analyses have indicated that the various factors are statistically significant in many cases, but here we would like to assess their relative significance. The two De Kinder plots will not have any gun type effect information, since there was only one gun type used (Sig Sauer) in that experiment. From the four plots we conclude:

1. De Kinder / FP: The individual gun effect is more important than the ammo effect.

2. De Kinder / BF: The individual gun effect and the ammo effect are about the same.

3. NBIDE / FP: The individual gun effect is more important than the gun-type effect and the ammo effect, both of which appear to be about the same.

4. NBIDE / BF: The individual gun effect, the gun type effect, and the ammo effect all appear to be negligible. Appearance-wise, this category is markedly different than the other three categories (mean = 7.94).
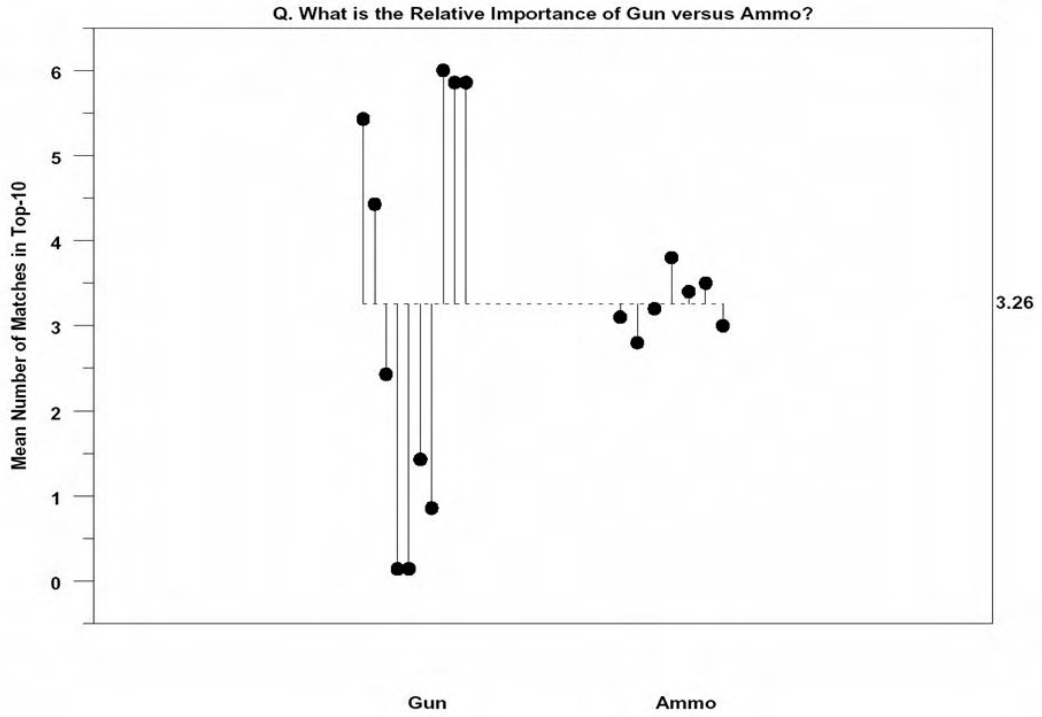
Figure 10-11. De Kinder / Firing Pin $ACCF_{max}$ Top Ten, relative importance of factors.
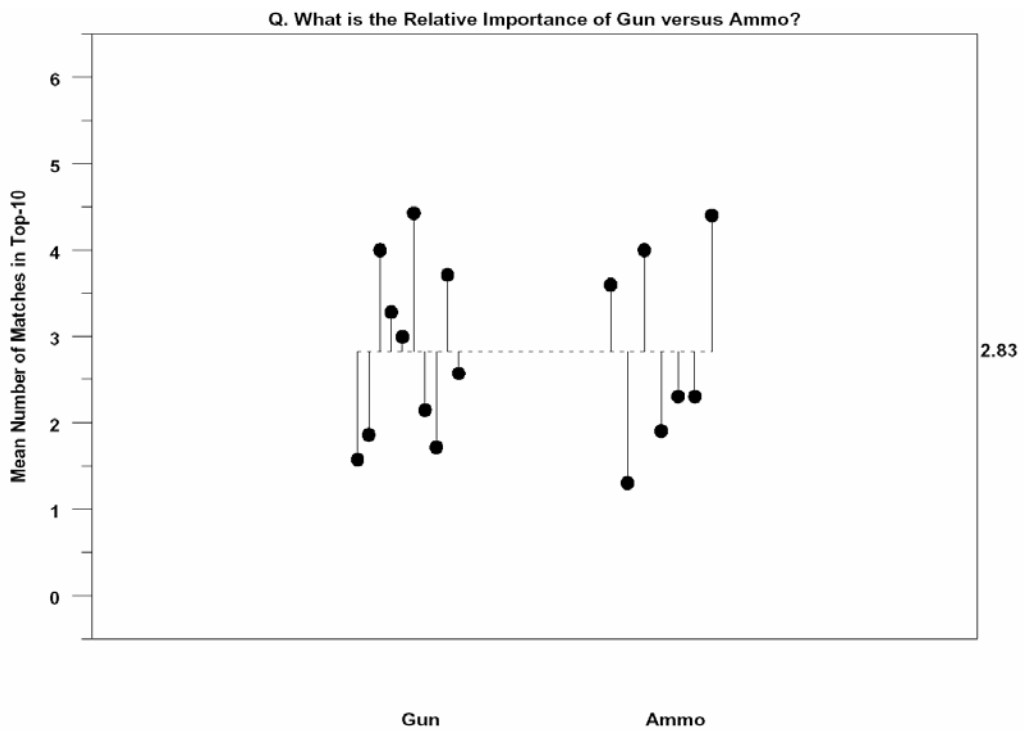

Figure 10-12. De Kinder / Breech Face $ACCF_{max}$ Top Ten, relative importance of factors.
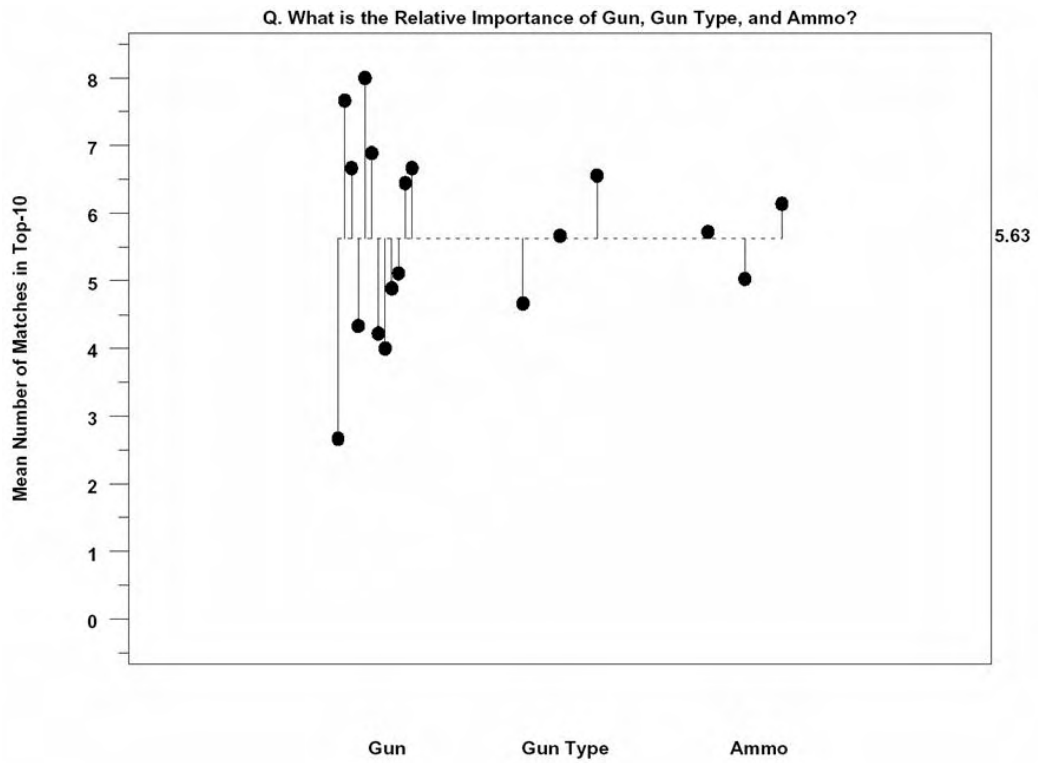
Figure 10-13. NBIDE / Firing Pin $ACCF_{max}$ Top Ten, relative importance of factors.
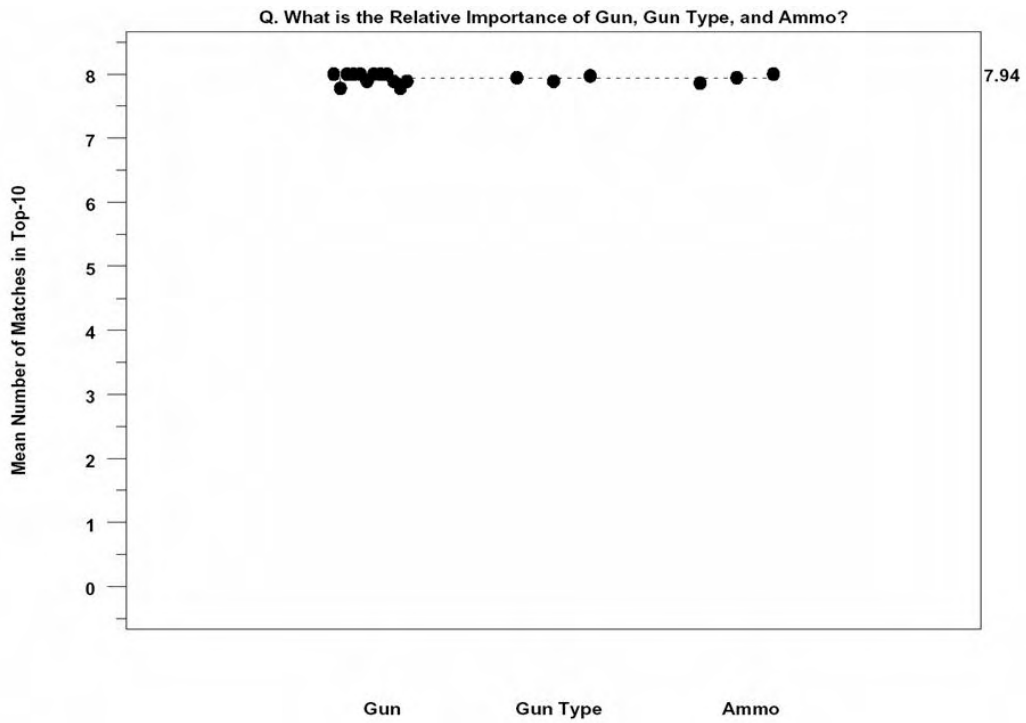

Figure 10-14. NBIDE / Breech Face $ACCF_{max}$ Top Ten, relative importance of factors.

10.3.2 Relative Importance of Factors (ANOVA)

The final subsection deals with examining the relative importance of factors by means of classical (fixed-effect) analysis of variance (ANOVA), which emphasizes the significance testing aspect, in contrast with the random-effects ANOVA, which emphasizes the variance components aspect. We present these ANOVA results for completeness because it allows the assessment of the relative importance of the various factors to be carried out in a formal statistically rigorous complementary fashion. The summary of the ANOVA results is given in Table 10-1 for the four cases graphically presented in Figs. 10-11 through 10-14. The ANOVA does not change the conclusions for the study as a whole. The table is organized as follows:

Column 1: Database and imaging region
Column 2: Summary statistics
Column 3: Results from 1-way ANOVA on gun
Column 4: Results from 1-way ANOVA on gun type
Column 5: Results from 1-way ANOVA on ammo
Column 6: Results from 2-way ANOVA on gun and ammo

From the table we conclude:

1. Gun: Significant at the 1% level for the first 3 cases.
             Not significant for NBIDE / Breech Face.

2. Gun Type: Significant for NBIDE / Firing Pin
             Not significant for NBIDE / Breech Face

3. Ammo: Not significant for De Kinder / Firing Pin
             Significant for De Kinder / Breech Face
             Marginally significant for NBIDE / Firing Pin & Breech Face

Finally, a reminder that "gun" being significant is not good *per se*. In the ideal situation of universal distinguishability, we do not want statistical significance. Rather, we want non-significance in combination with high mean match scores for each and every gun.

Table 10-1. ANOVA Summary Table.
y – the mean of Top Ten Scores.
s – the standard deviation of Top Ten scores.
FCDF – value of the cumulative distribution function for the F–statistics; values ≥ 95% imply significance.
ResSD – residual standard deviation.

Database & Imaging Region	Summary Stat	Gun (k=1)	Gun Type (k=1)	Ammo (k=1)	Gun & Ammo (k=2)
DeKinder Firing Pin	n = 70 Range: 0-6 y = 3.26 s = 2.53	FCDF = 100% ResSD =0.91 Highest Gun: 375 Lowest Guns: 139 & 213 **Significant**	N.A.	FCDF = 1.44% ResSD = 2.63 Highest Ammo: Speer Lowest Ammo: Win Not Significant	1. FCDF = 100% 2. FCDF = 78.27% ResSD =0.89 1. Gun: **Significant** 2. Ammo: Not Sig
DeKinder Breech Face	n = 70 Range: 0-6 y = 2.83 s = 1.63	FCDF = 99.64% ResSD = 1.51 Highest Gun: 215 Lowest Guns: 7 & 375 **Significant**	N.A.	FCDF = 99.98% ResSD = 1.85 Highest Ammo: Rem2 Lowest Ammo: Win **Significant**	1. FCDF = 100% 2. FCDF = 100% ResSD = 1.01 1. Gun: **Significant** 2. Ammo: **Significant**
NBIDE Firing Pin	n = 108 Range: 0-8 y = 5.63 s = 1.99	FCDF = 100% ResSD = 1.26 Highest Gun: Rug2 Lowest Gun: Sig1 **Significant**	FCDF = 99.98% ResSD = 1.85 Highest GT: Rug Lowest GT: Sig **Significant**	FCDF = 94.46% ResSD = 1.95 Highest Ammo: PMC Lowest Ammo: Win Marg. Significant	1. FCDF = 100% 2. FCDF = 99.95% ResSD = 1.17 1. Gun: **Significant** 2. Ammo: **Significant**
NBIDE Breech Face	n = 108 Range: 0-8 y = 7.94 s = 0.25	FCDF = 67.55% ResSD = 0.245 Highest Gun: Many Lowest Guns: SW4&5 Not Significant	FCDF = 65.01% ResSD = 0.247 Highest GT: Rug Lowest GT: S&W Not Significant	FCDF = 94.50% ResSD = 0.243 Highest Ammo: PMC Lowest Ammo: Rem Marg. Significant	1. FCDF = 70.89% 2. FCDF = 94.78% ResSD = 0.240 1. Gun: Not Sig. 2. Ammo: Marg. Sig.

11. Observations and Continuing Work

The topography data and analysis shown here indicate that surface topography measurements may significantly enhance the capability of matching casings fired from the same firearm, particularly for data gathered from breech face impressions. Although the error rate for matching casings was roughly 60 times smaller when topographic data of the NBIDE breech faces were analyzed than when the next most accurate metric was calculated, the error rate would have to decrease by roughly another factor of 35 to adequately support a large database with the assumptions of

- 100 000 guns having the same class characteristics,
- with a single probability distribution of correlation scores,
- and an accuracy goal of 90 %,
- for placing real matches in a Top Ten listing.

Two or three more independent metrics with equal or smaller error rate than that obtained here for matching NBIDE breech face impressions may need to be developed to bring the overall error rate to an acceptable level with the above assumptions.

Segmenting the databases using class characteristics, such as firing pin shape, has been proposed as a way to reduce the sizes of the datasets to be correlated within a large database in order to improve the efficiency of searching a national database for matches. Segmenting by demographic patterns, such as zip code, has also been proposed [57].

The results of the experimental N-3D approach were more accurate than the I-2D results for four experiments. This observation is consistent with results reported by Brinck indicating improved accuracy using IBIS BulletTrax-3D methods to find correct bullet matches as opposed to I-2D methods [58]. Another report by Roberge and Beauchamp reports the successful matching of ten pairs of bullets using IBIS BulletTrax-3D methods [59]. Topography (3D) methods have several advantages:

- Ballistics signatures are mainly geometrical topographies, so a method to measure topography directly should be preferable to reflection microscopy.
- The topography images are not as sensitive to the illumination conditions as reflection microscopy images indicating increased accuracy for 3D methods
- Topography measurements are traceable to dimensional metrology standards.

In addition, the N-3D analysis scheme of outlier removal, filtering, registration, matching, and statistics is non-proprietary, and this openness should facilitate development of improved algorithms by the technical community. For example, standard topography analysis methods [30] may be adapted to separate micro- from macro-topography and extract individual characteristics of the surfaces for correlation and identification.

A disadvantage of the current prototype topography approach is the time required to record the data, which is considerably longer for casings than that required for I-2D. The data gathering procedure and likely the analysis algorithms would need to become much more efficient for

practical application to a large database. The system described here is experimental and is not intended for commercialization.

A second metric might come from using the topography of the ejector marks. During this study we have not used the information from the ejector marks. In particular, we have not developed a technique for correlating different ejector marks because of their widely varying outer boundaries. It is difficult to develop automated software to correlate the shapes of such regions, particularly when some ejector marks are partially obliterated by labels imprinted on the casing by the manufacturer. Common practice for the I-2D is a manual operation whereby the users draw the ejector mark boundaries themselves when making entries. One of our tasks for future work is to develop a similar analysis program for the existing ejector mark data.

A second task not yet completed here is the correlation analysis for the 176 IAI bullets we are measuring. The correlation results could then be compared both with I-2D correlations from image acquisitions at ATF by Ols and Simmers, and with topography images previously measured by Bachrach et al. [15,35] using a single point confocal system. We will also be able to explore whether a sufficiently reliable metric can be developed for bullets to be consistent with a large database of bullet entries.

12. Acknowledgements

13. References

1.   P.L. Kirk, *Crime Investigation*, 2nd Ed. (Krieger Publ. Co., Malabar FL, 1985) p. 364.
2.   A.A. Braga and G.L. Pierce, Linking Crime Guns: The Impact of Ballistics Imaging Technology on the Productivity of the Boston Police Department's Ballistics Unit, Paper ID JFS2003205, *J. Forensic Sci.* **49**, 1 (2004).
3.   http://www.forensictechnologyinc.com/p7.html, IBIS Heritage Systems (2006)
4.   W.C. Boesman and W.J. Krouse, *CRS Report to the Congress, National Integrated Ballistics Information Network (NIBIN) for Law Enforcement,* Order Code RL31040 (Congressional Review Service, Library of Congress, Washington DC, 2001).
5.   http://www.nibin.gov/nibin.pdf, ATF's NIBIN Program, June 2005.
6.   F.A. Tulleners, Attachment A, Technical Evaluation: Feasibility of a Ballistics Imaging Database for all New Handgun Sales in B. Lockyer, *Feasibility of a California Ballistics*

Identification System, Assembly Bill 1717 (Hertzberg) (Stats. 2000, ch. 271) Report to the Legislature (California Department of Justice, Sacramento CA, 2003).

7.   J. De Kinder, F. Tulleners, and H. Thiebaut, Reference Ballistic Imaging Database Performance, *Forensic Sci. Int.* **140**, 207 (2004).

8.   W. George, A Validation of the Brasscatcher Portion of the NIBIN/IBIS System, *AFTE J.* **36**, 286 (2004) and A Validation of the Brasscatcher Portion of the NIBIN/IBIS System Part Two: "Fingerprinting Firearms" Reality or Fantasy, *AFTE J*. **36**, 289 (2004).

9.   A. Beauchamp and D. Roberge, Model of the Behavior of the IBIS Correlation Scores in a Large Database of Cartridge Cases, http://www.forensictechnology.com/d4.html, accessed 16 April 2007.

10.   R. Nennstiel and J. Rahm, A Parameter Study Regarding the IBIS Correlator, *J. Forensic Sci.* **51**, 18 (2006) and An Experience Report Regarding the Performance of the IBIS Correlator, *J. Forensic Sci.* **51**, 24 (2006).

11.   http://www.ncjrs.org/txtfiles/165476.txt, J. Travis, Guns in America: National Survey on Private Ownership and Use of Firearms, NIJ Research in Brief (National Institute of Justice, Washington DC, May 1997) p. 12.

12.   http://www.troopers.state.ny.us/Firearms/CoBIS/ , New York State Division of State Police, Combined Ballistic Identification System, Accessed, 06 Dec 2005.

13.   J.J. Tobin, Jr. *Maryland- IBIS Integrated Ballistics Identification System* (Maryland State Police Forensic Sciences Division, Pikesville MD, 2003).

14.   B. Lockyer, *Feasibility of a California Ballistics Identification System, Assembly Bill 1717 (Hertzberg) (Stats. 2000, ch. 271) Report to the Legislature* (California Department of Justice, Sacramento CA, 2003), including Attachments A (Ref. 6), B, C, and D.

15.   B. Bachrach, A Statistical Validation of the Individuality of Guns Using 3D Images of Bullets, http://www.ncjrs.gov/pdffiles1/nij/grants/213674.pdf, March 2006.

16.   J. Song, E. Whitenton, D. Kelley, R. Clary, L. Ma, S. Ballou, and M. Ols, SRM 2460/2461 Standard Bullets and Casings Project, *J. Res. Natl. Inst. Stand. Technol.* **109**, 533 (2004).

17.   L. Ma, J. Song, E. Whitenton, A. Zheng, T. Vorburger, and J. Zhou, NIST Bullet Signature Measurement System for RM (Reference Material) 8240 Standard Bullets, *J. Forensic Sci.* **49,** 649 (2004).

18.   A. Harasaki, J. Schmit, and J. C. Wyant, Improved Vertical Scanning Interferometry, *Appl. Opt.* **39**, 2107 (2000).

19.   M.A. Schmidt and R.D. Compton, Confocal Microscopy, in *ASM Handbook Volume 18 Friction, Lubrication, and Wear Technology,* P.J. Blau, ed., (ASM International, 1992), p. 357.

20.   T.R. Thomas, ed., *Rough Surfaces* (Longman, Harlow, UK, 1982), Chap. 2.

21.   J.M. Utts and R.F. Heckard, *Statistical Ideas and Methods* (Thomson Brooks/Cole, Belmont CA, 2006).

22.   J.M. Bennett and L. Mattsson, *Introduction to Surface Roughness and Scattering* (Optical Society of America, Washington DC, 1989).

23.   O.C. Wells, *Scanning Electron Microscopy* (McGraw-Hill, New York, 1974).

24.   J. Song, T. Vorburger, T. Renegar, H. Rhee, A. Zheng, L. Ma, J. Libert, S. Ballou, B. Bachrach and K. Bogart, Correlation of Topography Measurements of NIST SRM 2460 Standard Bullets by Four Techniques, *Meas. Sci. and Technol.* **17**, 500 (2006).

25. M. Bray, Stitching Interferometry and Absolute Surface Shape Metrology: Similarities, *Proc. SPIE.* **4451** (2001); http://www.mboptique.com/common/publications/ MBO_2001_SPIE_4451-40_(Stitching_Interferometry).pdf.

26. J. Song and T. Vorburger, Proposed Bullet Signature Comparisons Using Autocorrelation Functions, *Proc. 2000 Nat. Conf. Standards Laboratories* (Toronto, July 2000).

27. http://www.fti-ibis.com/en/s_4_1_5.asp, Bulletrax-3D, Accessed, 06 Dec. 2005.

28. H.-G. Rhee, T.V. Vorburger, J.W. Lee, and J. Fu, Discrepancies between Roughness Measurements Obtained with Phase-Shifting and White-Light Interferometry, *Appl. Opt.* **44**, 5919 (2005).

29. T.V. Vorburger, H.-G. Rhee, T.B. Renegar, J.-F. Song, and A. Zheng, Comparison of Optical and Stylus Methods for Measurement of Surface Texture, *Int. J. Adv. Manuf. Technol.* DOI 10.1007/s00170-007-0953-8, http://www.springerlink.com/content/0851313276m3t772/, online February 2007 (in press).

30. ASME B46.1-2002, *Surface Texture (Surface Roughness, Waviness, and Lay)* (Amer. Soc. Mech. Engrs., New York, 2003).

31. J.F. Song, T.V. Vorburger, R. Clary, E. Whitenton, L. Ma, and S. Ballou, Standards for Bullets and Casings, *Materials Today* **5** (11), 26 (2002).

32. P. Rubert, Properties of Electroformed Calibration Standards for Surface Topography Measurement Systems, in *Tenth International Colloquium on Surfaces*, edited by M. Dietzsch and H. Trumpold (Shaker-Verlag, Aachen, 2000) p. 245.

33. A. Moenssens et al., *Scientific Evidence in Civil and Criminal Cases*, 4th Edition (the Foundation Press Inc., 1995, New York) p. 375.

34. ISO 25178-6, Committee Draft, *Geometrical product specification (GPS)—Surface texture: Areal— Part 6: Classification of methods for measuring surface texture* (International Organization for Standardization, Geneva, 2005).

35. B. Bachrach, Development of a 3D-based Automated Firearms Evidence Comparison System, *Journal of Forensic Sciences* **47**, 1253 (2002).

36. J.H. Dillon, Three Dimensions: The Next Level for Firearms Examination, *Evidence Technology Magazine*, July-August, 34 (2005), www.EvidenceMagazine.com.

37. *Crime Gun Trace Reports (2000) National Report* (Department of the Treasury, Bureau of Alcohol, Tobacco, and Firearms, Washington DC, 2002).

38. µsurf, Non contact 3D-measurement of complex surfaces, http://www.nanofocus.info/ product_overview.php?productId=7, Dec 2005, accessed 06 Dec. 2005.

39. Infinite Focus,http://www.alicona.com/.

40. R. Simmers and M. Ols, private communication.

41. Y. B. Yuan, T.V. Vorburger, J. F. Song, T. B. Renegar, A Simplified Realization for the Gaussian Filter in Surface Metrology, in *X. International Colloquium on Surfaces*, M. Dietzsch, H. Trumpold, eds. (Shaker Verlag GmbH, Aachen, 2000), p. 133.

42. Bergen, J. R., Anandan, P., Hanna, K., and Hingorani, R., Hierarchical model-based Motion Estimation, in *Proceedings of Second European Conference on Computer Vision*, (Springer-Verlag, 1992), pp. 237-252.

43. Heeger, D., Notes on motion estimation, prepared by Prof. David J. Heeger for Courses Psych 267/CS 348D/EE 365, New York University, 20 Oct. 1996.

44. Heeger, D., MATLAB™ Software for Image Registration, Copyright 1997, 2000 by Stanford University, available at URL http://www.cns.nyu.edu/~david/registration.html.

45. G.M. Jenkins and D.G. Watts, *Spectral Analysis and Its Applications*, (Holden-Day, San Francisco, 1968) p. 171 ff.

46. *Guide to the expression of uncertainty in measurement* (GUM) (International Organization for Standardization, Geneva, 1995).

47. B. N. Taylor and C. E. Kuyatt, *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Tech. Note 1297, (1994).

48. Latin Square and Related Designs, in *Engineering Statistics Handbook*, Sec. 5.3.3.2.1, http://www.itl.nist.gov/ div898/handbook/, updated 18 July 2006.

49. R.J. Grissom, Probability of the superior outcome of one treatment over another, *Journal of Applied Psychology* **79,** 314 (1994).

50. D.M. Green and J. Swets, *Signal Detection Theory and Psychophysics* (John Wiley, New York, 1966).

51. J.A. Hanley and B.J. McNeil, The Meaning and Use of the Area under an ROC Curve, *Radiology*, **143**, 129 (1982).

52. H.B. Mann and D.R. Whitney, On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics* **18,** 50 (1947).

53. L. Hong and A.K. Jain, Integrating Faces and Fingerprints for Personal Identification, *IEEE Transactions PAMI* **20**, 1295 (1998).

54. V. Nair, private communication.

55. R.S. Burington, *Handbook of Mathematical Tables and Formulas* (McGraw-Hill, New York, 1965) p.357.

56. F. A. Tulleners, Ref. 6, Secs. 1.5 and 4.5.

57. Forensic Technology, Inc., Segmenting Tool Mark Image Reference Files (TMIRF) to Identify Crime Guns More Effectively, http://www.forensictechnology.com/d4.html, accessed 16 April 2007

58. T. Brinck, Comparing the Performance of IBIS and Bullet TRAX-3D Technology Using Bullets from Ten Consecutively Rifled Barrels, AAFS Meeting, San Antonio, February 2007, http://www.forensictechnologyinc.com/d4.html, accessed 16 April 2007.

59. D. Roberge and A. Beauchamp, The Use of BulletTrax-3D in a Study of Consecutively Manufactured Barrels, *AFTE Journal* **38**, 166 (2006).