Response to reviewers

Eric Hare, Heike Hofmann, Alicia Carriquiry

July 26, 2016

AE

1. The paper should better explain the role of the new methods for forensic science. The authors should consider referencing some of the previous work the reviewers mention.

   The AE makes two very good points. We have greatly extended our review of the existing literature and have included references to relevant papers. Where possible, we have also contrasted the proposed approach with existing methods. We have also added a discussion of the potential uses of a score such as ours in the forensic and legal applications.

2. Methodologically, the crux of the data analysis seems to be a generic application of decision trees. It would be good to clarify that aspect, and to more formally explain how the ensemble of the steps took by the authors to complete this work can be applied to other forensic science questions.

   The basic objective is to find the most discriminating combination of attributes to construct a score that provides a quantitative summary, or signature, for each bullet. In that sense, yes, random forests are a standard alternative. Random forests are extraordinarily flexible tools, and permit the use of both continuous and discrete variables in the construction of the score. This is one reason to think that they are likely to be useful in many other forensic applications. We have added a comment about this in the manuscript. One aspect of our work that is (sadly) novel, at least relative to the forensics community, is that we are making our algorithm and code completely public, so that the scientific community can evaluate the approach in detail.

Reviewer A

3. Random forests, as used and described on pages 18-21 of the paper comes out looking like a winner, but other than that success—worth reporting *somewhere*—, there is no new information on that method. No new insights. And certainly no comparisons with the tens of other things that could have been tried.

   We selected random forests because the method has desirable attributes and can be easily ported to other forensic applications. We were not intending to provide new insights on random forests. The focus of the paper was to show the potential of combining a large number ofimage attributes to discriminate among bullets. As the forest ultimately predicted every land-to-land comparison properly, we felt that a comparison to other learning methods was outside the scope of this paper. Had a wider variety of data been available, there would likely be more variability in the predictions and a comparison to other learning methods would be prudent.

Reviewer B

4. The authors should discuss their work in the context of Nicholas Petraco's work as well as other ongoing research endeavors in impression and pattern evidence. See https://www.ncjrs.gov/pdffiles1/nij/grants/239048.pdf The alignment method should have been compared with a standard method- See Patraco's NIJ report for a couple of examples. As presented the novel component of the paper is the "twicing"

for grooves combined with a loess fit to the remove curves from bullet images. I am not sure this is the authors intention. It would be interesting to compare this method with some of the classical methods from functional data analysis used for curve registration.

Thank you for the comments. Using the cross correlation function as a method of alignment is actually pretty standard - the appropriate references are now included. We have attempted to better clarify the novel component of this work, by explaining openly and transparently describing each individual step of the analysis process, and providing the source code used to reproduce the findings.

5. page 7- figure 4- What exactly are you bootstrapping? It is unclear in the paper whether you are bootstrapping residual or bullets or something else entirely. Please clarify.

   We have decided to remove the bootstrap confidence intervals for the loess fit residuals because they did not add enough information.

6. There should have been a comparison of different classification methods to show that the random forest is reasonable.

   A classification method should be judged by its success in predicting the outcome. Random forests are based on decision trees, which are an extremely flexible tool for classification, that can deal with quantitative and categorical types of variables and, in the implementation we used, also deal with some missing values (which was not needed in the data here).

7. On the random forest classifier- are the prior probabilities being estimated from the cross comparisons described on page 15? What exactly are the prior probabilities used in the random forest?

   No data-based adjustment was made to the prior probabilities, i.e. prior probabilities of 50:50 for matches/non-matches were used.

8. Equal Error Rates on the ROC curves would be a helpful metric as well as the AUC.

   Included

9. Would it be possible to stress the classification system by only using part of the profile? Such as would occur when a partial bullet is recovered? By stressing the random forest you would be able to compare it to other classifiers.

   Yes, it would be possible to stress the classification system by using only parts of the profile, but a systematic evaluation of this by far exceeds the scope of this paper.

10. What is the standard error of the error rates presented for the matching algorithms? The number of cross comparisons gives false impression the standard errors the when presenting the results.

    Under the assumption of land-to-land independence, the number of land-to-land matches is distributed according to a Binomial(N, p) distribution where N is the total number of bullet lands and p is the probability of a match between two bullet lands. However, the bullet lands from the Hamby study are all known to be either pairwise matches or non-matches. Thus, we present the Out-of-bag (OOB) error for the random forest, representing the error rate as determined from the observations not present in the particular bootstrap samples. This is discussed further on Page 19.

11. In the conclusions section the authors mention an assumption of "uniqueness". In general forensic science have moved beyond these questions- see the work of Saks and Kaye for details.

    The reviewer is absolutely correct, but we are talking about a different "uniqueness" concept here. The discipline of firearm and toolmark examining relies on two foundational assumptions: that every gun leaves a unique set of markings and that the marks are reproducible. This is still the credo among tool and firearms examiners. The reviewer is referring to the fact that in testimony, forensic scientists are beginning to move away

from declaring that, for example, an individual gun, to the exclusion of all others, was the one that fired the bullet. The work we present addresses the fundamental assumption of uniqueness, and will hopefully help move the practice away from statements of unique source.

12. A discussion on how the authors view the use of their matching algorithm would be useful to the forensic community. For example are they intending to use this to build a closed set identification system similar to IAFIS or are them intending this to work in a manner similar to a biometric verification system? etc.

    The problem with existing closed set systems is that it is unclear how the matching is exactly done and what statistical properties it has (number of parameters, distribution of them, sensitivities). Repeatability and reproducibility is a huge issue in this area. 'Matching' is also not just a single step - it consists, as we have laid out in the paper, of a series of steps, which all deserve attention. While it was pointed out to us to use 'standard methods' - there are no actual standards in place. The status quo is NIBIN, which is not accessible to anyone outside of law enforcement. There might be a set of alternatives, and some evidence as to which methods might work better in some situations than others, but there is a lack of an in-depth discussion of the alternatives at each step as well as an overall discussion. In this manuscript we are detailing one route to divvy up the process and provide the code for others to reproduce our findings.

Reviewer C

13. there is a lack of knowledge of the existing literature on statistical work performed on tool marks and firearms. For example, the work performed by Nicholas Petraco, Fabiano Riva, or Alan Zheng in the past 5 years is not mentioned. Instead, old research projects are mentioned, which mostly focus on manual identification. The authors need to discuss the novelty of their method (and how it potentially addresses the shortcomings of these recent methods)

    The point that the reviewer makes is valid. While we were familiar with the computer-assisted approaches that have been proposed, practice still relies on human examination of potential candidates presented by a system such as NIBIN. We have added what we hope is an acceptably complete review of the literature in the revised manuscript.

14. The authors would need to study (or at least discuss extensively) the ability of their method to capture the variability in bullet profiles over time (at interval of several dozen of bullets fired by the same weapon, when the weapon has been cleaned, or when the bullet was degraded by hitting a target). In addition, a dataset of 10 firearms (even consecutively manufactured) does not provide any information on the rarity of a particular bullet profile, and therefore does not address the second probability statement. Alternatively, the proposed metrics could be tested in a database searching situation, which has not been done.

    At the time of the submission of this manuscript, the 35 bullets from 10 barrels from Hamby set 252 were the only ones accessible to the public. They are part of Dr Zheng's efforts at NIST to provide a ballistics reference database. In the future more results will be available, but until then we are unable to provide a data-informed discussion of the very important issues raised by the reviewer.

15. On a related subject, I believe that a random forest provides the probability that two bullets were fired by the same gun given the set of observations (Fig 19), not the probability of interest: the probability of two sets of observations given that the two bullets were fired by the same gun. The authors may want to revise some of the statements in their discussion and conclusion.

    The reviewer is correct in stating that two probabilities of interest are the probability of observing the data under the "same source" (same gun) hypothesis or under the "different source" hypothesis. An empirical approach to get at those two probabilities is to derive

the distribution of scores when two bullets are known to have been fired by the same gun and the distribution of scores when two bullets are known to have been fired by different guns. We have actually derived those to distributions and include them in the revised manuscript.

16. I am concerned by the assumption of independence of between the different lands on a given bullet that the authors seems to use in their error rates calculations (and ROC curves). The assumption seems unreasonable since as soon as a land has been paired between two bullets, it seems that it becomes more likely to pair the other lands. In other words, I believe that the authors need to redo their error rate measurements, but using bullets as their experimental units and not lands. I am also concerned by the overwhelming majority of non-matches, this must certainly artificially increase the rate of correct exclusion and distort the data.

    Assuming independence between the lands of each barrel is certainly not realistic in practice, however, by assuming independence we make the task of matching harder for the algorithm. This also has the benefit of potentially being more useful for degraded data, where we may not have a full set of six groove-to-groove scans. Based on the actual numerical values of features derived from each of the lands, an independence assumption is not unrealistic, because the features do not exhibit any clustering associated to barrel (Bachrach 2006). The technique suggested by the reviewer could be solved using a hierarchical approach to matching, and we agree with the assessment that it would likely decrease the error rates.

17. the authors attempt to define a "stable" profile (see Figure 9) and use the stable profile for comparison. What happens if such profile does not exist. Would it not be best to compare surfaces instead of sections? a 2D version of the proposed algorithm should be easily implementable (and has been done at least by Riva).

    Figure 11 shows images of all bullet lands we encountered in the Hamby set where no stable profile could be identified. All four of these images show massive signs of tank rash - this motivates our suggestion to not consider bullet lands without a stable region for the further analysis. Concentrating on one-dimensional profiles is quite common in the literature, and we have expanded the discussion on this choice with appropriate citations.

18. L323: how do the authors know that there are "172 known matches" since the 15 "trace bullets" are supposed to be of unknown sources? which raise the question on how where the ROC and other density graphs generated? only based on the bullets of known origin?

    For the Hamby study there is a known ground truth of known matches and known non-matches, which we used for determining the accuracy of the algorithm. As state in the paper, the ROC curves are based on comparisons of 88 lands from unknown bullets (90 minus the two lands that showed too much tank rash) and 118 lands from known bullets (120 minus two bullet lands affected by tank rash), resulting in a total of 10,384 land to land comparisons.

19. If would be helpful if the legend from Fig15 could include a definition of the CMS, CNMS, CCF, ... acronyms such that the reader does not need to flip back and forth between several pages.

    Done for figure 15 and 16

Reviewer 1

20. First, the term "land" is used in the title but not defined until p. 3. I recommend changing the title, to ensure broad comprehenson; perhaps "Automatic Matching of Bullet Marks," or something of that kind.

To clarify the title, we have modified the abstract to provide a brief definition of bullet lands.

21. p. 3, l. -10: Be explicit that grooves and striations are synonyms.

    Grooves and striations are actually not synonyms in this case. The grooves are impressions induced by the rifling of the barrel, and properties associated with them are class characteristics not unique to individual gun barrels. We've taken care to ensure the distinction is clear between the striae of interest, which are extracted from the lands of the bullet (between grooves), and the grooves themselves.

22. Early on in the paper, it might be helpful to mention the typical width of a striation, the number of striations typically seen on a fired bullet, and so forth.

    We have added the typical width of both the groove-to-groove (land) width, as well as the typical width of a striation.

23. p. 2, l. -1: It would be good to reword so the sentence doesn't start with "x3p". Also p. 3, l. 2.

    DONE

24. p. 3, Fig. 1: The numbers on the y-axis are superimposed.

    We have removed the values from the y-axis, as they are not necessary in interpreting the image.

25. p. 7, l. 1: Make this "... can be described in the form ...".

    DONE

26. p. 7, l. 9: Remove the comma after "here" in "... function used here, ...".

    DONE

27. p. 7, l. -6: I'm confused. The text says "bullets 2 are shown" but the caption refers to bullet 2- which I think should be a specific land on a single bullet.

    DONE

28. p. 12, l. -11: The use of a colon to transition to a new paragraph seems odd. I recommend just rewording so is a single paragraph and does not use a colon at all. It is almost as if you are trying to quote rather than state.

    DONE

29. p. 13, l. -3: You write "Given that striation patterns are typically much larger, ..." and I think it would be more clear to write "Given that striation widths are typically much larger, ...".

    DONE

30. p. 16, l. -1/-2: Could you elaborate on why a "large number of maximal CMS by itself is not indicative of a match"? I think the reason is that there could be other marks that are completely incompatible and preclude a match, based on the next sentence, but I am not sure and a few words would be helpful.

    In the Hamby study, with 10,212 land-to-land comparisons which are known non-matches, two of these comparisons still yielded a CMS of 12. Figure 13 emphasizes this. In our particular data, all comparisons with a CMS of 13 or more were in fact matches. Still, it is likely due to pure chance alone that, particularly if more bullets were used in comparison, a small portion would have high pairwise CMS despite being non-matches. Adding features beyond just CMS will decrease this error rate - In other words, a high CMS value is not a sufficient condition for a match.

31. p. 17, l. -13: I'd prefer another term than "correlation" here. In fact, the entire sentence could be improved.

DONE

32. p. 17, l. -2: Remove the comma after "indicates".

    DONE

33. p. 18, l. -4: Change "this study" to "that study", so the reader is certain you mean Ma et al. (2001) and not your own.

    DONE

34. p. 20, l. -3: I really like the idea of including an inconclusive zone.

    Thank you for the comment. We have chosen to leave off a more detailed discussion on an inconclusive zone. Certainly, formulating such a zone would have important implications in terms of the criminal justice system, but it is one that we felt was outside the scope of this particular paper. Certainly, should an automated algorithm like we propose be used as an aid in the conviction or the exoneration of suspects, much attention would be needed on the extent of this zone.

35. p. 21, Fig. 20: Most of the features plotted have been defined, but I don't recall x1 and x2. You may want to remind the reader about all the definitions, or have a table of features appear early in the random forests discussion.

    We have removed references to x1 and x2. Thanks for pointing them out. These were features representing the height of the bullet at which the signature was extracted for each of the two bullets, but they were not intended to be part of the random forest.

36. p. 22, l. -9: Change "bottomline" to "bottom line".

    DONE

37. p. 23, l. 4: Remove the comma after "found".

    DONE

38. p. 23, l. -13/-5: I don't think this mention of the Aadhar project adds much, and it distracts from the points you want to highlight about the work in this paper.

    We have removed references to the Aadhar project.

39. Supplement, p. 8, l. 18: Change "de-facto" to "de facto".

    DONE

Reviewer 2

40. So, my suggestion here is to change the sentence in the abstract to something like "Firearm examination is a forensic tool used to help the court determine whether two bullets were fired from the same gun barrel." Or "...help the trier of fact determine...".

    DONE

41. the concept of "match": The primary goal in forensic science is NOT to attempt to match two bullets. It is to quantify the value of the observations made on the two bullets with regard to the two competing propositions of interest to the court (i.e., assign a likelihood ratio or Bayes factor).

    The reviewer is correct. Here, we talk about "match" only as a means to test whether our algorithm is actually pairing the samples that are true pairs. This is a term used only in this development stage and is not intended for the court. As such, we have clarified that matching two bullets is one of our goals, but not the sole or primary goal of this procedure in forensic science.

42. I strongly encourage the authors to not follow this trend and use the term "ballistics" only when you are actually talking about ballistics.

We have removed references to the term "ballistics" except in the context of citing prior work which uses the term.

43. 2nd to last line in the abstract: In point (a) "correctly identify lands with too much damage to be suitable for matching" I suggest replacing the word "matching" with "comparison" to use the same terminology as the other pattern evidence disciplines.

DONE

44. Line 55: Replace "in the forensic sciences" with "in forensic science". (See Margot P, Commentary on The Need for a Research Culture in the Forensic Sciences, UCLA Law Review (2011) 58: 795-801.)

DONE

45. Lines 156-158: The sentence "The blue area around the signature are 95% bootstrapped confidence intervals based on 1,000 bootstrap samples." is confusing. The term "signature" is defined two sentences further. Is the sentence referring to Figure 4b? If yes, it should probably come at the end of this paragraph.

DONE

46. Figure 5: In this figure, where is the limit between bullet 1-5 and bullet 2-1?

Light grey shading in the background was added to help with the distinction between the two flattened bullet lands.

47. Line 246: Replace "this representations" with "this representation" or "these representations"

DONE

48. Caption of Figure 8: "At a lag of -17 the correlation peaks, indicating the largest amount of agreement between the signatures." is not a complete sentence.

We have revised this caption to remove the ambiguity from the word "peaks", which we had intended to mean "maximizes."

49. Line 266: Capitalize the "b" in "by".

DONE