Response to the reviewers

Ganesh Krishnan and Heike Hofmann

8/6/2018

We would like to thank the reviewers for their in-depth reviews and the suggestions regarding the manuscript. Please find our detailed response addressing each of the points below.

## Editorial changes

We would like to also thank the editor for helping us identify the formatting changes. We have addressed all the following issues as pointed out

* "Table of contents - please delete : Done
* Introduction header - please delete: Done
* References by name - these need to be cited numerically: Done
* Tables in text - please include the tables and their headers in the - Main Document as separate pages after the References when uploading the revision. Please delete from the text: Done
* Figures in text - all figures must be uploaded as separate TIF, PNG or EPS files with at least 300 dpi. Please delete from the text: Done
* Headers, sub-headers and sub-sub-headers numbered - please remove the numbering: Done
* Sub-headers bolded and not italicized - these need to be italicized only: Done
* Sub-sub-headers bolded - please do not bold: Done
* References incorrect format - please follow the formats included in the Information for Authors: Done
* Figures loaded as separate PNGs but these are not 300 dpi - see above: Done
* Two files uploaded - one as bib file and one as tex - I am not sure what these are? However, we only except text in Word files. If these were intended to be included in the manuscript file, please include them and do not load the information as separate files.: Removed the Latex files uploaded earlier
* Title Page, Manuscript and Appendix uploaded as PDFs - please upload as Word files." : Done

It is imperative that the format guidelines outlined in the JFS Information for Authors be followed in preparing the revision. I have to say it was disappointing that you made no attempt to follow the correct format.

Please also note that when loading figures into Manuscript Central it is important to Insert FIG. X before each figure legend. In addition, you must load each figure as a separate file.

Please note that all figure files (TIFF, PNG or EPS) MUST be at least 300 dpi and that some of your figures do not satisfy this requirement. Please modify your figures accordingly.

Please include the tables and their headers in the Main Document as separate pages after the References section when uploading the revision.

There is also a need to strengthen the Abstract. Although we do not yet use a structured abstract, we do expect the same information - background, brief description of methods and results, and conclusions.

Upon acceptance for publication, manuscripts become the copyright property of AAFS. The corresponding author is responsible for transferring copyright to AAFS through the Wiley Author Licensing Service (or WALS). In doing so the corresponding author confirms that all authors have contributed to the manuscript per the JFS Information for Authors and all have agreed that the accepted manuscript will be published in JFS.

To revise your manuscript, log into <https://mc.manuscriptcentral.com/jofs> and enter your Author Center, where you will find your manuscript title listed under “Manuscripts with Decisions.” Under “Actions,” click on “Create a Revision.” Your manuscript number has been appended to denote a revision.

You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Once the revised manuscript is prepared as a MS Word document, you can upload it and submit it through your Author Center.

When submitting your revised manuscript through your Author Center, you must respond to the comments made by the reviewer(s) in the space provided. Please use this space to document all changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s). It is also imperative that you follow the required format provided in the JFS Information for Authors, particularly those for the References. As noted above, if your revised mansucript is not in the JFS required format, it may be returned for further revision.

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission. Your revision submission MUST include your complete manuscript, i.e., title page, main document, tables and figures (if applicable).

Because we are trying to facilitate timely publication of manuscripts submitted to the Journal of Forensic Sciences, your revised manuscript should be uploaded within 21 days of receipt of this email. If it is not possible for you to submit your revision in this time frame, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to the Journal of Forensic Sciences and I look forward to receiving your revision.

## Reviewer #1

The article “Adapting the Chumbley Score to match striae on Land Engraved Areass (LEAs) of bullets” extends an existing comparison algorithm for striated marks to the analysis of bullet LEAs. The paper makes two primary contributions. First, they modify the deterministic work of Hadler & Morris (2017) which itself is an extension of Chumbley et al. The authors’ contribution is a modification of Hadler & Morris to reduce the number of scenarios which ‘fail’ under the Hadler & Morris approach. Second, they evaluate the work on two sets of bullet LEA data and explore some variants of the algorithm parameters to tune it for bullets rather than screwdrivers. Overall the paper presents a reasonable project to extend a nice algorithm for striated toolmark comparison.

Adaptation of the Chumbley el al algorithm from screwdrivers to bullets makes sense as both analyses involve striated marks. The authors here nicely point out that compared to screwdriver marks (as in the original Chumbley paper), LEAs are typically smaller both in length and width. LEAs also have a curved cross-section baseline which must be corrected.

The authors do a nice job covering previous work. I appreciate the summary tables 1 and 2. Table 1 is good as-is. Table 2 can be a bit misleading because we don’t know what bullet datasets were used. Because these different papers analyzed different bullet data sets the underlying difficulty of matching impacts the performance. That is, bullets from very well marking firearms are ‘easier’ to match than those from firearms which mark poorly. So the FP and FN reported should be taken with a grain of salt. The authors should add a statement regarding different performance on different datasets.

*Authors: Thank you for pointing this out. You are right, we have added more information about the data, wherever available. We have also added a sentence to emphasize different difficulty levels and resulting differences in error rates of different brands of firearms and ammunition*

I’m also confused how the FP rate for Hare CNMS is 6.26% whereas all other FP are 6.25% This seems to likely be a typo. 6.25% imply close to 10,000 comparisons which I don’t think the datasets have. I could be wrong but I wanted to point out this potential typo that should be checked.

*Authors: Thank you for spotting the typo. With respect to the sample size, the data set being used by Hare et al. 2017 has about 10384 land-to-land comparisons and the error rates being reported here are computed from Receiver operating characteristic (ROC) curves. We decided to report False Negative error rates for a fixed value of False Positives (FP) = 6.25% to make them comparable to the observed values in Figure 10 (bottom right)*

The authors describe the two methods of ‘failure’ for the method (Sec 2.3). I agree with the authors statement that this failure ‘should’ only happen for non-matches. It could also happen for matches if the optimal profiles are incorrect. That is, if there are two possible optimal match pairs (with very similar correlation scores), where the top scoring pair is incorrect but the second best scoring pair is correct. I’m not sure how often that happens. Overall, I would think that failing for non-matches would be ok. It’s not clear that the described failures are the biggest cause of performance hits but improving the algorithm to handle these failures seems reasonable.

*Authors: You are right, a failure might occur for a match if the initial optimization fails. The new method is still susceptible to that. It would be interesting to further investigate this problem in the future. We have made a note of that in the conclusions.*

The following sections list specific points that should be addressed before acceptance. They are grouped into clarification/grammar/preciseness, minor, and major revisions.

Some examples of clarification / grammar / preciseness issues:

* p3 “confocal microscopy allow \* to measure 3D surfaces in a high-resolution digital form” \* should be “one”?
* p3 “Digitized images of 3D surfaces of form the data basis\* of statistical analysis” Awkward phrasing.
* p3 “Faden et al and Chumbley et al have been analyzing screwdriver marks” perhaps change to “have analyzed” because the work is completed and published.
* p9 Seems to be a broken latex figure reference around line 28/29. These are not all the issues of this type, we encourage the authors to review the manuscript carefully clean issues like the above.

*Authors: Thank you for pointing out these issues. We have fixed those accordingly.*

Minor edits should be made to accuracy and clarity:

* p3 “Methods for matching marks for a variety of tools” should be changed to “Methods for matching striated marks for a variety of tools”. This is an important point in that it helps make clear that the authors are studying and comparing striated toolmarks and not impressed toolmarks. I would also propose that the authors make this distinction earlier in the introduction.

*Authors: We have added emphasis on striae as opposed to impressed toolmarks.*

* p3 “Analysis of these digitized markings require the use of statistical methods”. Analysis does not *require* statistical methods. Useful information regarding common origin can be obtained from non-statistical methods. This being said, statistical methods are of course great and we should be working towards them. Authors could change to analysis “benefits from the use of statistical methods” or something similar.
* p14 how are the false negative and false positive rates computed?

*Authors: We have added two paragraphs on the Chumbley Score test for clarification: the first paragraph expicitly states the hypotheses involved, the second defines error rates.*

Major Changes Required before acceptance:

* p8 “This severe limitation in the amount of available data poses the main challenge in”. It’s unclear what ‘this’ refers to. Do the authors mean the number of bullets and LEAs in the dataset or do they mean the number of sample points per striation profile (due to the fact that LEAs are shorter than screwdriver profiles)? Both would seem to be limitations but just because the profile is shorter doesn’t inherently mean that there are fewer or less informative striaes. You could have a shorter profile with more informative features.

*Authors: We were indeed referring to the shorter length of the marks, which given the same amount of information decreases the power of the test, a purely technical consequence. Obviously you are right that this power calculation changes with a change in the amount of information*

* Eq 1 uses a variable i, which is defined above as having the values of 1 or 2. I therefore don’t understand how i<0 as written in Eq 1. This should be improved. Possibly this is a sloppy reuse of variables?

*Authors: thank you for pointing this out - we accidentally changed notations. this is fixed now. Index k=1,2 refers to the two different signatures, index i in Z refers to window locations.*

* How are the original profiles (eg. Figure 2 Left) identified from the 3D surface (eg. Figure 1 Top). That is, does the user simply draw a line from one end of the LEA to the other? How is the original 3D surface resampled to the 1d profile? This is very important because if you ‘cut’ the 3D profile the wrong way (at the wrong angle) then you will dilate or compress the extracted profile. This needs a lot more description.

*Authors: We followed the Hare et al. 2017 paper in extracting the profiles from scans. We have expanded on the discussion at the begining of section titled “Scans for land engraved areas” in the paper.*

* Section 2.4 is not clearly written. I’m not confident that I understand their new method. This section should be rewritten to explain what the authors are doing. My guess is that rather than taking windows in integer number of steps away from the optimal position matching a left shift on one profile with a right shift on the other, the authors simply walk from the right-end to the left-end of one profile and from the left-end to the right-end of the other profile being careful to not compare same shift windows. Note that my description is not clear either as this is a tricky thing to put into words, but is something the authors need to do.

*Authors: Thank you for poining this out, you are right in your understanding of the method. We have changed figure 5 for clarification and expanded our discussion, which hopefully makes it easier to understand the method.*

* The results presented on page 14 list the correct identification of “same-source and difference-source toolmarks of 93.5 to 94.1%”. Does this mean the true positives are 93.5% and the true negatives are 94.1%? I don’t think so because the next sentence lists the false negative rate of between 0.28 and 0.36. The authors need to explicitly list the true positive rate and the true negative rate. These must be separate because there are many more true negatives than true positives in any matching dataset. That means that an algorithm that always predicts “different source” would be correct most of the time. In fact, for the author’s dataset the “always predict different” algorithm would be correct approximately 94% of the time (35 bullets where each has appx 2 matches). The lack of a true positive rate and true negative rate is is a major point that must be corrected. To be clear, the true positive rate, true negative rate, false positive rate, and false negative rate should all be reported for the first dataset, perhaps added to Table 3 or another table like Table 3.

*Authors: Thank you for bringing this to our notice. We have included a paragraph with a definition of error rates, and switched to a table to show the results as suggested below.*

* The true positive rate, true negative rate, false positive rate, and false negative rate should also be presented for the second dataset (Hamby 44 scans, Sec 3.5). The authors do present AUC and EER curves but these are less informative and can often be used to hide weaker performing models.

*Authors: We have added a table with error rates for the Hamby 44 scans.*

* The results should all be listed as either fractions or all as percent (which is preferred) and not mix them.

*Authors: We have changed the reporting to error percentages.*

* p21 the authors state “this kind of assessment is only feasible in the setting of a large study, such as the one we presented”. I strongly disagree with the use of the word “large”. The presented work is not a large study. Ten firearms were used in each of two studies (it’s not clear if both datasets used the same 10 firearms). A large study would include at least 100 different firearms and even 100 is not likely to provide statistical confidence. This is a very important point. If the paper is allowed to state that the described work is a large study it will set a precedent that will be detrimental to the field.

*Authors: We agree the data available to us is by no means large. Our comment was referring to the availability of 35 bullets rater than just 2 0r 3 as might be seen in actual case work. We have changed the sentence accordingly.*

Overall, this paper presents a nice incremental improvement to a promising algorithmic approach. A number of major revisions are required prior to acceptance as described above. Note that these revisions all relate to the writing and require no additional experiments to be performed. They improve the clarity and preciseness of the manuscript and serve to address questions that readers will have.

## Reviewer #2

You may not fully appreciate and how critical their phase orientations used to identify fired bullets (LEAs). This important consideration should have been an important aspect of the experiment and not instead focus primarily only LEA comparisons. While each area is in practice independently compared microscopically or by surface profile algorithms, consideration as to their orientation in their position in the barrel rifling, and correlations of LEAs cannot ignore the phase of positions. Including the phase consideration of the fired bullet comparisons, the success of finding a match in a particular position between two bullets is significantly increased and demonstrated in their reference Chu, et al. If you are to base independent single LEA scoring to test the method, you should make this clearer in the beginning of your paper and hypothesis, and not discuss it finally in the last sentence of the document body. If you are going to compare method performance between chisel/tool stria to LEAs on bullets, you should have a detailed discussion as to the large differences in the toolmark widths that are in consideration. The Chumbley scoring parameters may have been optimized for chisel widths but may not be optimal for the LEAs.

*Authors: Thank you for your comments, we appreciate your insight. You are right, for the purpose of matching bullet to bullet, all lands of each of the bullets and their relative positions on the bullet should be considered. This is not the problem we tried to address in this manuscript. To get from land-to-land comparisons to bullet-to-bullet comparisons, an additional step is required. In Chu et al. 2010, the average correlation value is taken. In this manuscript, the land-to-land comparisons result in a p-value. There are different options to aggregate these values at this stage to quantify a bullet-bullet comparison.*

*We agree on the point that the toolmarks have large differences in width and the parameters of the chumbley algorithm are optimized for tools, which is why we first do a parameter study to identify the best set of parameters that would work for the LEAs.*

Another very important point; in none of the cited texts should there be an impression that if a ground truth match is not indicated by sufficient match score, that it falls into a miss-identification, or Type 2 error “a failed test”. A type 2 error (false positive) would occur if a comparison has a match score over a threshold for identification when the ground truth is two sources. Therefore, the hot-button term “error rates” in this case is not accurate. The manuscript is using this erroneous description throughout, is other words, “Missed identifications” and are improperly indicated as described as “miss-identifications”. This miss-characterization is a major weakness in the manuscript.

*Authors: You are right - the term ‘error rate’ is highly charged emotionally for firearms examiners. The Chumbley score test is peculiar in the sense, that the null hypothesis is stated as ‘two markings coming from a different source’. For clarification, we have added a paragraph at the end of the section on the the Chumbley score that details the test hypotheses and a paragraph at the end of the following section with the corresponding (statistical) errors. We have also added a sentence to the effect that all of the comparisons run on the Hamby 252 set only included lands that were deemed suitable for a comparison (e.g. not affected by tank rash), similar to Hare et al (2016). Lands leading to inconclusive results by FTEs are therefore not included in the comparisons and would not lead to a ‘missed identifaction’.*

Coarseness of the surface is not a direct indicator of sub-class influence. More literature review is necessary to understand the earmarks of subclass. For the surfaces compared in the Hamby sets, there was no indication of subclass influence between fired bullet sets from consecutively rifled barrels. If the term is instead a parameter of form (curvature) then that should be more fully described, and how it is used. (I suspect that this is how it was used for fired bullets in this research.)

*Authors: A coarse Gaussian filter was indeed used to remove some class characteristics in Bachrach et al (2010) for toolmark striae. In this manuscript, the filter is used to remove bullet curvature. However, subclass characteristics expressed in form of ‘waviness’ in the barrel (e.g. introduced by specifics in the firing process) can be addressed in the same way as for toolmarks.*

Additionally, consider including a firearm examiner who is acquainted with objective methods of comparisons for 3D toolmark profiles to review your updated manuscript draft.

*Authors: Our research center is working closely with forensic examiners, including firearms examiners. Thank you for the advice.*