Quantitative Evaluation of Footwear Evidence: Initial Workflow for an End-to-End System

Gautham Venkatasubramanian, MSc, (Former) Research Associate
Vighnesh Hegde, MSc, (Former) Research Associate,
Steven P. Lund, PhD
Hari Iyer, PhD
Martin Herman, PhD
Information Technology Laboratory
National Institute of Standards and Technology

Some of this material was presented at a talk at the 104th International Association for Identification's International Forensic Educational Conference, Reno, Nevada, August 2019.

Acknowledgements

We greatly appreciate discussions and initial implementations of RAC similarity metrics provided by Weiqing Chen from Everspry when she was a Guest Scientist at NIST. We thank Brian McVicker from the FBI for some of the impressions and Mike Gorn from the FBI for helping us collect mock crime scene impressions. We thank Brian McVicker and Jeremy John from USACIL for testing and providing feedback on an early implementation of the system. We also thank Günay Doğan, Adam Pintar, Yooyoung Lee, Sarala Padi, Clément Driard, Ali Daoudi, Sarah Hood, Nicholas Vollbrecht, and Akul Sareen, current and former members of the NIST forensic footwear research team, for discussions, ideas and feedback. We are also grateful for the valuable feedback received from the anonymous reviewers and from the editor. Funding for this research was provided by the National Institute of Justice (NIJ) under Award DJO-NIJ-17-RO-0202. The NIST Special Programs Office provided funding for fundamental research in image analysis and forensic science whose results were applied to the research described here.

Disclaimer

Certain commercial entities, equipment, or materials may be identified in this paper in

order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

The opinions, recommendations, findings, and conclusions presented in this paper do not necessarily reflect the views or policies of NIST or the United States Government.

ABSTRACT

In the U.S., footwear examiners make decisions about the sources of crime scene shoe impressions using subjective criteria. This has raised questions about the accuracy, repeatability, reproducibility, and scientific validity of footwear examinations. Currently most footwear examiners follow a workflow that compares a questioned and test impression with regard to outsole design, size, wear and randomly acquired characteristics (RACs). We augment this workflow with computer algorithms and statistical analysis so as to improve in the following areas: (1) Quantifying the degree of correspondence between the questioned and test impressions with respect to design, size, wear, and RACs, (2) Reducing the potential for cognitive bias, and (3) Providing an empirical basis for examiner conclusions by developing a reference database of case-relevant pairs of impressions containing known mated and known non-mated impressions. Our end-to-end workflow facilitates all three of these points and is directly relatable to current practice. We demonstrate the workflow, which includes obtaining and interpreting outsole pattern scores, RAC comparison scores and final scores, on two scenarios - a pristine example (involving very high quality Everspry EverOS scanner impressions) and a mock crime scene example that more closely resembles real casework. These examples not only demonstrate the workflow but also help identify the algorithmic, computational and statistical challenges involved in improving the system for eventual deployment

in casework.

KEYWORDS

forensic footwear, footwear evidence, shoeprints, pattern evidence, quantitative evidence evaluation, quantitative analysis, weight of evidence

HIGHLIGHTS

- Presented an end-to-end workflow for footwear impression comparisons.
- Using simulated casework example, demonstrated that workflow is implementable in practice.
- Uses examiner annotation of impressions with algorithmic, quantitative evaluation.
- Uses a multi-stage automated comparison that includes size, design, wear and RACs.
- Uses case-relevant ground-truth-known reference score distributions to interpret comparison scores.

In the U.S., footwear examiners make decisions about the sources of crime scene shoe impressions using their training and experience but based on subjective criteria. This has raised questions about the accuracy, repeatability, reproducibility, and scientific validity of footwear examinations. The 2009 National Academy of Sciences report (1) noted that forensic footwear identifications rely on aspects of an examiner's training and experience that have not been empirically tested for accuracy, repeatability, and reproducibility. The 2016 report by the President's Council of Advisors on Science and Technology (PCAST) (2) states, in addition, that subjective methods used in forensic footwear comparisons need to be transformed to objective methods. PCAST defines objective methods as those "consisting of procedures that are each defined with enough standardized and quantifiable detail that they can be performed by either an automated system or human examiners exercising little or no judgment." Subjective methods are defined as "methods including key procedures that involve significant human judgment - for example, about which features to select within a pattern or how to determine whether the features are sufficiently similar to be called a probable match." (See p. 5 in (2).) PCAST found that the scientific validity of footwear analysis had not been established and encouraged conducting "black-box" studies to assess "error rates" similar to what had been done by the forensic latent fingerprint community. Several black box studies in forensic footwear have been performed (e.g., (3, 4)) and more are still needed.

Developing high-performance computer-based systems that can discriminate impressions made by different footwear, with very small error rates, has been of great interest for some time. A survey of the literature reveals that nearly all of the research on the use of algorithms for footwear impression comparisons have "database retrieval" as their focus and are primarily concerned with discriminating between different outsole design patterns (e.g.,

(5-9)).

During the investigative phase, given a crime scene impression, often referred to as the *questioned impression,* law enforcement agencies would like to determine, for starters, the make and model of the shoe that produced it. This information could help narrow down the list of potential perpetrators of the crime in question. The collection of images against which the crime scene impression is to be compared may therefore consist of a library of outsole design images from various manufacturers. In some situations, the library may consist of test impressions obtained from shoes belonging to persons previously apprehended in connection with different crimes; this could lead to a (non-exhaustive) list of possible perpetrators of the current crime. It is generally impractical for a human examiner to compare the questioned impression against every image in a gallery or library of impression images. Recognizing this fact, researchers have focused on automating the "database search and retrieval" task and developed various algorithms for this purpose.

While algorithms designed to assess which test impressions among a fixed collection are most similar to a crime scene impression or whether two impressions come from shoes sharing a common outsole design are useful, they represent an incomplete mapping of the process examiners use for evaluating footwear impression evidence. When given a questioned impression and test impression from a shoe of interest, examiners routinely conduct additional analyses to assess the extent to which wear or randomly acquired characteristic (RAC) patterns seen in a questioned impression may single out the particular shoe of interest from an envisioned crowd of other shoes of the same make, model, and size. Several publications have discussed methods for assessing the rarity of RACs (10-15). Quantitative assessments of similarity between two RACs using various features such as shape, perimeter, and area were discussed in (16). Models for describing the spatial distribution of RACs were proposed and investigated in (17, 18). Bayesian approaches for using wear information in interpretation of footwear evidence have been discussed in (19, 20). Though our focus in this paper is on

describing a workflow and demonstrating an end-to-end process for footwear comparisons that could be used in casework, it is expected that previous research by others will play a role in the development of a system that is casework ready.

In actual cases involving footwear impressions, triers of fact will assess whether the specific shoe that made the test impression also made the crime scene impression. It is important they receive as much potentially helpful objective information as possible to aid their decisions. Our proposed system replaces some of the subjective components in the footwear comparison process with objective steps that use computer algorithms. When using a computer-based system to assign an ordinal similarity score between two impressions, information that answers the following questions are important:

- How similar are the test and crime scene impressions?
- How similar to the crime scene impression are test impressions from other shoes?
- What similarity levels have we seen in the past when comparing test and crime scene impressions known to have come from the same shoe?
- What similarity levels have we seen when comparing test and crime scene impressions from two different shoes of arbitrary make, model and size?
- What similarity levels have we seen in the past when comparing test and crime scene impressions from "close non-matches" (e.g., two different shoes of the *same make, model, size, side, and outsole manufacturing mold)?*

Though the answers to these questions, by themselves, do not tell us whether or not the questioned impression was made by the shoe of interest, their answers form an empirical basis for subsequent subjective interpretation and can help investigators, lawyers, judges, and jurors alike with the decisions they make throughout the investigative and judicial processes.

Workflow

Currently a footwear examiner in the United States will often follow the following workflow. Compare the questioned impression Q and the test impression K with respect to outsole design (same design or different designs), the sizes of shoes responsible for the impressions (same size or different sizes), the degree and location of wear observed, and any apparent RACs. Based on these comparisons the examiner will arrive at a conclusion, often following the guidelines outlined by the Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD) (21). This conclusion is a key component of the examiner report and, when called for, is offered to the court during testimony. The concerns expressed by the NAS and PCAST reports pertain to the 'subjective' nature of the comparison process used by forensic footwear examiners. As mentioned above, both reports have expressed the need for quantitative assessments of footwear evidence using scientifically valid methods, including the use of algorithmic, automated methods.

We believe the current workflow performed by examiners can be augmented with computer algorithms so as to improve in the following three specific areas and thus increase the overall level of objectivity:

1. Quantification of the degree of correspondence (or lack thereof) between Q and K for each attribute (design, size, wear, RACs).

2. Reducing the potential for cognitive bias without affecting overall performance in separating mated impression pairs from close non-mates.

3. Providing an empirical basis for examiner conclusions by developing a reference database of ground-truth-known pairs of impressions containing known mated impressions and known nonmated impressions (preferably using close non-matches as well as arbitrary pairs of impressions) and demonstrating how the quantitative summaries of the degree of correspondence (or lack thereof) perform in casework-like scenarios in their ability to discriminate be-

tween impressions made by the same shoe versus impressions made by different, but closely matching, shoes. Currently, examiner interpretations are primarily supported by unspecific phrases like "training and experience" or "expert judgment." This step would increase the support available to an examiner to include a demonstrable collection of data showing how successful the current analytic step has been in past, similar instances. Champod et al. discuss how to establish the most appropriate database for source level propositions (22).

Therefore, the focus of this paper is to propose a workflow that may be viewed as a fine-tuning of the current practice and that points (1), (2), and (3) above.

Point (1) is addressed by discussing approaches for quantification of the degree of agreement/disagreement between features in Q and K for each of the four areas of focus (design, size, wear, RACs). We do not identify the *best* metrics to be used for each area of comparison as this requires, within each area of comparison, in-depth studies of candidate metrics, those already available as well as new proposals, and this is deferred to future work.

Point (2) is addressed by incorporating, in our workflow, the capability for computing the comparison metrics without a side-by-side comparison of Q and K. When more than one examiner is available, the proposed workflow allows for Q and K to be annotated by different examiners. When only one examiner is available, the examiner is to first fully annotate Q before seeing K or the corresponding shoe. This is somewhat akin to the practice by fingerprint examiners where they mark minutiae on Q and on K independently, without looking at them side by side. This is a major contribution of our proposed workflow to reduce potential cognitive bias. It is well-known that cognitive bias can create significant problems in forensic science (23).

Point (3) is a major project in itself. In this paper we only show what is possible if an adequate reference database is available. This is demonstrated by using an example database of ground-truth-known pairs of impressions created in our laboratory primarily for purposes

of demonstration.

In addition to the three specifically targeted points of improvement, we believe our proposed workflow can integrate current practice with algorithms to improve repeatability and reproducibility.

Forensic evaluation of evidence (including algorithmic evaluation) generally consists of steps that may be grouped into two categories: (1) measurement and (2) interpretation.

In the context of footwear impression comparison, the measurement step (when performed algorithmically) consists of analyzing the footwear impression images and computing numerical summaries of the degree of correspondence/non-correspondence between Q and K. Although there are many different approaches for arriving at such numerical summaries one can generally identify methods that perform well in terms of their ability to discriminate between mated and nonmated pairs of impressions. The numerical summaries can be evaluated empirically using ground-truth-known data.

The interpretation step involves judging the weight of the presented evidence which includes examiner findings communicated via verbal descriptions and numerical summaries of degree of correspondence. Reference distributions associated with the numerical summaries of interest based on casework-similar, ground-truth-known, comparisons provide the context to help the decision maker assess the strength of evidence. This is a personal assessment by the decision maker and there is no single, unique, weight of evidence value that can be computed by others for universal use. Furthermore, computing and presenting a strength of evidence number, such as a forensic likelihood ratio (LR), rather than providing the empirical information available to help triers of fact and other decision makers assess the strength of the presented evidence, is a controversial topic that has yet to be settled, at least in the U.S.

Therefore, this paper does not focus explicitly on characterizing weight of evidence, using

either a number or a verbal scale. By focusing on developing repeatable and reproducible methods for quantification of the agreement/disagreements between Q and K, we hope to provide additional tools to examiners who can then better support their findings since they can provide empirical bases for their judgements.

While we prefer to focus on empirically grounded summaries in favor of subjective probabilistic interpretation, nothing specific to the footwear comparison discipline precludes one from developing LR systems for footwear comparisons; in fact, the metrics used as part of this workflow may themselves serve as inputs to any LR system that one may develop in the future.

Fingerprint and firearms evaluations have made greater progress toward the quantitative and automation goals set forth by the NAS and PCAST reports than have footwear evaluations. Unlike fingerprints, footwear outsoles are mass produced, with many replicates produced by the same physical cast. Much like questioned bullets and cartridge cases, there are a large number of close non-matches for nearly every questioned shoe impression. This requires focus on very subtle details, such as wear patterns and RACs, to help separate true matches and close non-matches. Whereas firearms consistently impress discriminating information into the metal of bullets and cartridge cases they discharge thereby resulting in questioned impressions that are frequently of good quality, shoe impressions deal with imperfect reproductions subject to sub-optimal recovery methods. Additionally, the critical information from shoe impressions at crime scenes occurs in a wide variety of substrates (e.g., ceramic, vinyl, paper, wood) and matrices (e.g., blood, dust, water). This variability makes it all the more difficult to develop algorithms that consistently extract the small but critical details that can help distinguish whether a given shoe actually produced the questioned impression or is simply another shoe of the same make, model, and size.

The workflow described in this paper relies on some level of human pattern annotation as inputs to the algorithmic comparisons. Other researchers have also presented footwear

evidence analysis methods that use human annotation. Examples include (10, 11, 14-16), where the focus is on analysis of RACs. Examples of research where footwear evidence analysis does not use human annotation include (6, 7, 24, 30, 31), where the focus is on make/model identification as well as evidence evaluation.

In our approach, users may highlight apparent RAC regions, corner points, regions that exhibit contact with the outsole surface, or other features. Much like marking minutiae in latent prints, the markups provide a more stable input for comparison algorithms from one case to another, hopefully allowing for better algorithm development, training, and ultimately discrimination performance, than what has been previously published. As mentioned above, potential biases in these manual annotations may be alleviated either by having two different examiners annotate K and Q, respectively, such that neither examiner sees both the shoe of interest and the questioned impression or by having a single examiner first annotate Q before seeing K or the shoe of interest.

The eventual goals of the work described in this paper are to develop quantitative, "more objective" methods than currently used, to support examiners in each phase of footwear impression comparisons. By "more objective," we mean that some comparison process components that are typically performed using human judgement have been implemented as algorithms that do not require human oversight. Our workflow cannot be considered fully objective as it relies on human annotation of features in the impressions, which is a subjective process. However, the subsequent comparison algorithms that quantify the correspondence between two impressions based on these features is objective in the sense that given the same input set of annotated features, the algorithm will always produce the same value.

To the best of our knowledge, there is no discussion in the literature regarding a detailed, systematic, end-to-end process that can provide quantitative, empirical support for each step of the footwear impression comparison process currently used by examiners. This is the topic addressed in this paper. With further, ongoing refinements, it is envisioned that the system

will evolve into one that can be deployed in routine casework in the near future.

End-to-End Systems

Our view of an end-to-end system for footwear evidence is a system whose input consists of two images, a questioned impression (Q) from a crime scene and a test impression (K) from a known shoe of interest. The output of the system is quantitative information provided to the examiner, judge, jury, investigator or other interested party that can provide empirical support for their opinions and conclusions.

In order to gain acceptability by the practitioner community and to conform to existing SWGTREAD guidelines, it would be desirable for the quantitative evaluation process of the end-to-end system to closely follow the conventional examiner evaluation process (e.g., as reflected in the SWGTREAD (21) conclusion scale). The examiner should be able to understand how the quantitative information produced by the system is related to the main elements considered in the conclusion scale – design, size, wear, and RACs. A new footwear conclusion scale is currently being developed by the Organization of Scientific Area Committees (OSAC) for Forensic Science. The workflow presented in this paper can provide quantitative support to footwear impression examiners who conduct their evaluation by sequentially considering design, size, wear, and RACs, regardless of which conclusion scale they use to summarize their findings.

Initial Prototype

The initial version described in this paper is intended to demonstrate the general steps in the end-to-end workflow. The actual algorithmic components or modules in this version will undoubtedly be replaced as better performing algorithms continue to be developed.

To conduct quantitative comparisons between questioned and test impressions, footwear-related structure in these impressions, that is, design and wear features that derive from the contact of the shoe with the substrate, need to be identified. Reliably identifying such features in questioned impressions can be extremely difficult to do automatically because such features are often partial, occluded, smeared, noisy, distorted, low contrast, derived from multiple impressions, or occur on a variety of cluttered or highly structured backgrounds. Automated annotation of such impressions is generally beyond current technology.

Our approach is therefore to have the human examiner annotate the features in the questioned impression that are necessary to (a) facilitate an alignment or registration between the questioned impression Q and the test impression K so that corresponding features overlap well in impression pairs that are mates or close non-matches, and (b) provide as much discriminating power as possible between true matches and close non-matches. Algorithmically discriminating between different outsole designs is of interest for database retrieval but of much less interest for evidence assessment since the human examiner will typically be able to readily discriminate between different outsole designs in casework (assuming there is enough footwear-related structure in the crime scene impression).

We aim to optimally utilize the respective strengths of human visual perception and algorithmic computation, resulting in a hybrid process combining a human/automated feature extraction step with a fully automated evaluation of correspondences and discrepancies among compared impressions resulting in a "comparison score." Automatically identifying features in test impressions is often not as difficult as in crime scene impressions because the quality and clarity of test impressions is usually much higher and the image signatures associated with contact (often black ink or dark powder) and non-contact (generally white) are under the control of the examiner. However, reliable identification of RACs in test impressions generally requires examination of both the test impression and the physical shoe outsole itself. When assessing whether a potential RAC seen in a test impression seems

more likely to be a reproduction of a true RAC visible on the outsole or more likely due to a manufacturing defect or artifacts generated during the creation of the test impression, trained examiners carefully study the outsole and multiple test impressions from it. Therefore, rather than attempting to automate the RAC finding process using algorithmic approaches, our workflow relies on identification of RACs in the test impression by a trained human examiner.

Once features have been identified in each impression of a given pair, the act of comparing impressions will require no human involvement. The output of the algorithmic evaluation is a comparison score. Examiners may use these scores in conjunction with a collection of scores from known match and known non-match (including close non-match) comparisons performed under casework-similar scenarios to provide empirically supported statements of their findings during report writing and courtroom testimony.

The steps in the end-to-end workflow are the following:

1. The human annotates the questioned impression. All human annotation of Q is performed without being guided either by the shoe of interest or by a test impression from it. However, to help identify outsole design features, the human may be guided by a test impression from a different shoe of the same make and model, or an image of the shoe outsole of the same make and model, e.g., obtained from the internet. In general, the human may annotate contact and non-contact regions in Q to aid in subsequent comparison with test impressions. In the workflow proposed here the human is also asked to oversee placement of "corner points" in Q that are used to achieve alignment between Q and K. These are points that represent corners of 2D features in the impressions. If the user provides a contact/non-contact annotation, or if Q is exceptionally clear, corner points can be automatically extracted in the same manner as with K (see item 3 below). Otherwise, corner points must be manually placed by

the user. We plan to enhance our workflow by including examiner annotation of the quality or clarity of regions in the impression (allowing an analysis of alignment and comparison scores to be weighted by the confidence that the user has in the extracted features) and the annotation of potential RACs visible in Q.

2. The human marks up RACs in K (with the aid of the shoe). Each RAC in K is marked by placing a vertical bounding box around the RAC. RACs are human-annotated in K only and not in Q. However, in future versions, potential RACs visible in Q may be annotated as well (as mentioned above). Also, in future versions, RACs in K may be marked by tracing their boundaries; wear regions in K may be annotated.

3. An algorithm aligns the two impressions. Corner points are automatically chosen in K, though a user can opt to manually select corner points if desired. The method of corner detection used here is discussed in (24), which is a minor modification we made to the FAST algorithm (25). We then use an image registration technique based on an algorithm for finding "maximum cliques" in graphs (24, 26). This approach has been previously suggested for matching a questioned fingerprint to a reference fingerprint (27). The maximum clique approach provides a rigid transformation (rotation and translation) to align the two impressions. Details of the maximum clique image alignment can be found in (24, 28). Note that, once the two impressions are aligned, the RAC boxes marked in K may be projected onto Q, thus resulting in a pair of corresponding regions for each RAC.

4. Algorithms conduct a multi-stage comparison. As with manual examinations, the automated comparison workflow considers design, size, wear, and RACs. In our current version, the comparison process is conducted in two stages: (a) outsole pattern comparison and (b) RAC comparison. Correspondences and discrepancies in outsole design, size, and wear are summarized by an "outsole pattern comparison score."

   Many similarity metrics have been developed and/or suggested for comparison of out-

sole design impressions – Normalized cross correlation, Phase-only correlation, Fourier-Mellin Transform, Multi-channel normalized cross correlation, Random forest scores based on selected features, to name a few. A more complete list of similarity metrics that have been explored in the literature may be found in (29). See also (24, 30). Our system uses a similarity metric which is based on the Resnet-50 pre-trained convolutional neural network model (layer ResNet-2bx, as done by Kong et al. (31), for which the network weights are publicly available. This results in 256 "feature maps" for each image. We have implemented the Average Phase-Only Correlation ($AvPOC$) metric on these feature maps (see (24) for details). It must be emphasized that this paper is not focused on which similarity metric should be used but on what the role of a similarity metric or a comparison score is in the overall process. A detailed comparison of the performances of available metrics on various casework-similar footwear impression comparisons will eventually determine the choice of the similarity metric. The point to note is that the workflow will not be affected regardless of which metric is selected during deployment.

The RAC comparison is performed separately on each pair of corresponding RAC regions. The comparison metric used is the Normalized Cross Correlation (NCC) (24), although there are many other metrics we could have used as well. Equation (1) provides the formula for NCC, where x_{ij} represents the grayscale value at pixel position (i, j) in the RAC region of K and y_{ij} represents the grayscale value at *matching* pixel position (i, j) in the corresponding region in Q. It is worth noting that NCC is essentially what is commonly referred to as the Pearson Correlation Coefficient (32) in statistical literature.

$$NCC = \frac{\sum_{i,j}(x_{ij} - \overline{x})(y_{ij} - \overline{y})}{\left(\sqrt{(\sum_{i,j}(x_{ij} - \overline{x})^2}\right)\ \left(\sqrt{\sum_{i,j}(y_{ij} - \overline{y})^2}\right)} \tag{1}$$

If human annotation of RAC boundary regions is available (in both Q and K) in the form of curves that describes the boundary pixels of a RAC, one may use, as an example, *shape distance* described in (33) in place of (NCC).

5. Scores from ground-truth-known case-relevant reference comparisons are used to provide context for all the scores obtained using the casework pair of impressions. A score by itself cannot help assess strength of evidence. In almost all situations, a score (even if it is labelled as a likelihood ratio) has to be interpreted in the context of other scores obtained by applying the comparison process (e.g., algorithm) to similar kinds of (Q, K) pairs where the ground truth regarding the relationship between the shoe(s) that left the impressions is known, i.e., known match or known close non-match or known non-match. When deciding what scores to use as context, one must consider the relevance of the task completed in obtaining the scores for reference pairs to the task completed in obtaining the casework score. For instance, reference scores obtained when comparing RAC regions are irrelevant to interpreting casework scores obtained when comparing entire outsoles. Similarly, we should not compare scores from a case with a blurry, partial questioned impression to scores from reference comparisons where Q is clear and complete.

   We consider three ground-truth classes of reference comparisons: *known match* (KM), where both impressions are from the same shoe; *known non-match* (KNM), where the impressions are from shoes that differ in outsole design and/or size; and *known close non-match* (KCNM), where the impressions being compared are at least from shoes of the same make, model, size, and side. Note that there are several potential classes of impression pairings that could constitute "close non-matches." Ideally one would be able to show that the comparison metrics strongly discriminate between mated pairs and even the strictest definition of close non-matches, impression pairs from different shoes of the same make, model, size, and side, produced from the same mold and worn by different individuals. If an algorithm cannot be shown to discriminate effectively

between mated pairings and this "closest non-match" group (e.g., due to poor algorithm performance or limited availability of image pairs meeting this strictest definition of close non-match pairings), it is also informative to assess how effectively mated pairs can be discriminated from looser definitions of close non-matches, such as shoes of the same make, model, size, and side, potentially produced from different molds, as this still helps inform the general sense of how many shoes are capable of having produced the questioned impression. When the available number of close non-match comparisons is small, one might consider including "flipped" impressions from shoes for the opposite foot to create additional close non-match comparisons, though care must be taken to assess whether the manufacturing process results in discrepancies in the outsoles of left and right shoe pairs before treating pairings with flipped impressions as close non-matches. For the illustrations presented in this paper, we consider impression pairings from different shoes of the same make, model, and size to be close non-matches. Due to limited available data, we primarily use comparisons between impressions from both shoes in a right and left shoe pair, with one of the impressions flipped on the vertical axis, so that the impressions appear to be from shoes of the same make, model, size, and side.

Each of the provided reference score collections can be specified by task (i.e., outsole comparison or RAC comparison), questioned impression clarity (i.e., pristine or more realistic), and ground truth (i.e., KM, KCNM, or KNM). We provide reference collections for each combination of these factors, with the exception that we do not provide KNM RAC collections because examiners only conduct RAC analysis for impressions that could plausibly be of the same make, model, size and side.

Score collections are labeled *pristine* when they contain results from comparisons where both the impressions were made from the Everspry EverOS scanner (34). While these comparisons do not generally reflect the quality and complexity of comparisons involving actual crime scene impressions, they do allow us to highlight the role of RAC

comparisons in the workflow. Score collections are labeled *more realistic* when they result from comparisons involving questioned impressions collected as mock crime scenes that more closely reflect casework. Note that each of the aforementioned reference score collections are taken from our collection of impressions and are used to illustrate our workflow. It is anticipated that, by collecting a large number of KM, KCNM, and KNM reference comparisons, reflecting a variety of casework-like scenarios, one will be able to construct a database that could be used to extract more precisely targeted case-relevant subsets for actual applications.

For outsole pattern comparison scores in the workflow, the KM, KCNM, and KNM reference sets all play a role in mimicking the comparison process used by examiners. Initially, examiners will assess whether the shoe of interest can be eliminated on the basis of design or size. This corresponds to seeing whether the outsole design comparison score from the current case clearly falls outside the range of scores from the KM reference collection and within the range of scores from the KNM reference collection (see, for example, Figure 5(a)). If it appears plausible that the two impressions share a common design and size, examiners may then evaluate whether observable wear patterns increase or decrease the overall strength of correspondence. Using the reference score collection (e.g., Figure 5(a)), wear patterns that increase the overall strength of correspondence would correspond to the observed outsole design comparison score appearing more likely among the KM reference scores than among the KCNM reference scores. Wear patterns that decrease the overall strength of correspondence would correspond to the observed outsole design comparison score appearing more likely among the KCNM reference scores than among the KM reference scores.

After evaluating design, size, and wear information, if an examiner finds it plausible that two impressions are made by the same shoe, then RAC information will be considered. Because RACs often fail to reproduce in conditions of crime scene impressions, failing to find a RAC in Q that was apparent in K is generally not considered exclusion-

ary. However, as the SWGTREAD criteria for an *identification* conclusion indicate, the scenario in which one or more RACs are visible in Q and appear in the same location and similar to RACs observed in K are viewed as substantially increasing the strength of overall correspondence. To reflect this process in the algorithmic workflow, we consider a KM RAC comparison score collection, obtained from comparing corresponding RACs in known matches, and a KCNM RAC comparison score collection, obtained from comparing an observed RAC with the same outsole region in a KCNM shoe (see Figure 6 (Bottom)). As with the outsole design, the overall strength of correspondence would increase when the current case RAC comparison score is high enough to appear far more likely among the reference KM scores than among the reference KCNM scores. This would occur when the similarity score for a given RAC exceeds what has been observed when comparing KCNM impressions. Among reference comparisons involving Qs that reflect casework, there are expected to be many KM instances in which a RAC observed in K does not reproduce in Q. This will cause the KM distribution to produce many scores falling within the range of KCNM scores. This mimics current subjective RAC evaluations in that a low similarity score for a RAC comparison would not provide strong exclusionary evidence because low scores can plausibly occur for KM or KCNM comparisons.

6. Comparison scores from each stage are combined into a final score. Our method to combine the various scores is to compute a score-based likelihood ratio (SLR) for each type of score and multiply them together to obtain a final score. This simple approach is inspired by the Naive Bayes classification strategy (i.e., combining different facets of evaluation by pretending their results are mutually independent). While there are limitations to considering SLRs as strength of evidence (35), the present application only uses probabilistic arguments as the inspiration for developing scores, not to imply any direct probabilistic interpretation of that score. To obtain SLRs for individual scores, we use kernel density estimation (36, 37) to convert the background distributions

into smoothed probability density curves, and then compute, at the point representing the score obtained for the current case pair, the ratio of the height of the curve for KM scores to the height of the curve for KCNM scores (see Figure 5 (b)). The logarithm (base 10) of the SLR, $\log_{10}(\text{SLR})$, is then taken, allowing the various scores to be combined by adding the $\log_{10}(\text{SLR})$s (rather than multiplying the SLRs).

Combining evidence is an age-old topic that has been discussed by a number of authors (see, for instance, (38, 39) and references contained therein). We have chosen to use a naive-Bayes approach to illustrate the step of combining evidential information contained in comparison scores from the different stages of the process. Additional research comparing different methods for combining evidence will point to the approach that will be used in the deployment stage.

7. All results are displayed visually. This will make it easier for the examiner to understand, report, and present the quantitative results obtained using this approach. The visual reports will draw attention to the existing body of relevant empirical information that supports more subjective casework interpretations. The visual displays will be demonstrated throughout the examples shown in the following section.

End-to-End Examples

The workflow described in the previous section is now demonstrated through two examples. The first example falls under the pristine clarity level and involves comparing two test impressions as K and Q, respectively. This example is intended to illustrate the workflow in an ideal application where RACs are visible and our current comparison metrics are effective. The second example involves comparing a test impression with a mock crime scene impression, representing the more realistic clarity level. This example is intended to be more recognizable as a potential casework comparison. Collectively, these two examples will

highlight the importance of choosing reference comparisons relevant to the case at hand.

Pristine Clarity Example

Figure 1 shows the two images considered in the first example - a test impression on the left and a "questioned" impression on the right. We use these two Everspry EverOS impressions to illustrate the various steps in the workflow proposed in this paper, including RAC comparisons. The long-term goal of our workflow is to help examiners assess the strength of the evidence by algorithmically comparing outsole size, design, wear, and RACs. This information should help them support conclusions using, for example, the SWGTREAD conclusion scale (21).

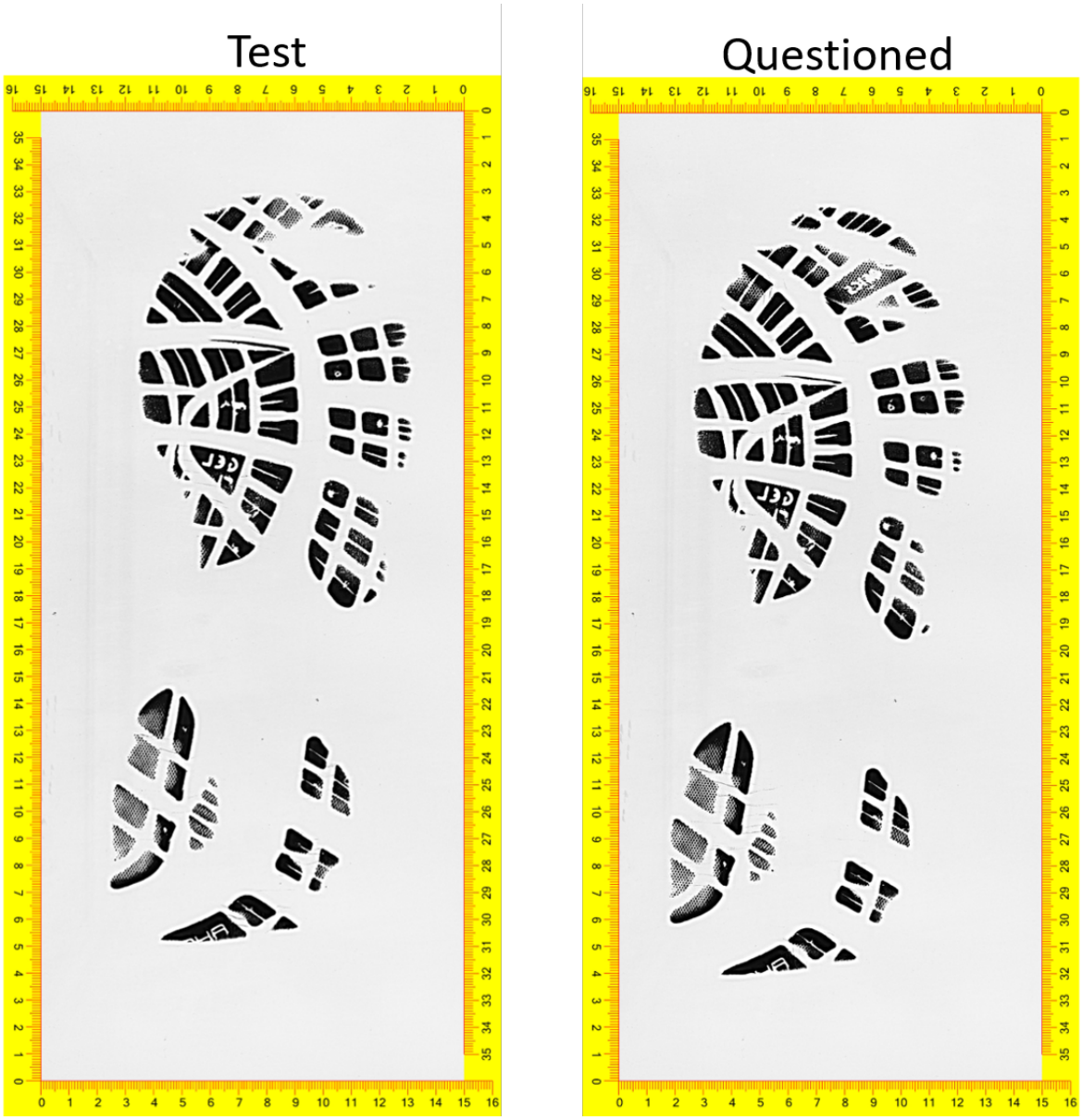

Figure 1: *Example of two impressions that are input to the end-to-end system: a test impression (on the left) and a questioned impression (on the right). This example is intended to demonstrate the entire workflow, including RAC comparisons, in an ideal albeit unrealistic scenario. In this case, the two impressions are made by the same shoe.*

The first step is to annotate Q. The user must decide whether or not to provide a manual markup of the contact surface seen in the questioned impression to aid the comparison

algorithms. If contact regions are manually marked, the result will be a binary contact image which will be used in downstream analysis. For instance, corner points and the region of interest (i.e., shoeprint boundary) may be automatically extracted from the binary contact image or the pattern score may be based on comparing the binary contact image with K. If a binary contact markup is not provided, then pattern comparisons will be performed between the raw Q and K and the user will need to manually mark the shoeprint boundary and corner points (see Figure 2(a)). It is not necessary that all corner points be marked but it would be beneficial to spread points throughout much of the contact region. The reason for marking corner points is to enable the alignment step. The more spread out the corner points are, the better the alignment is likely to be. The manual annotation is done in Adobe Photoshop.

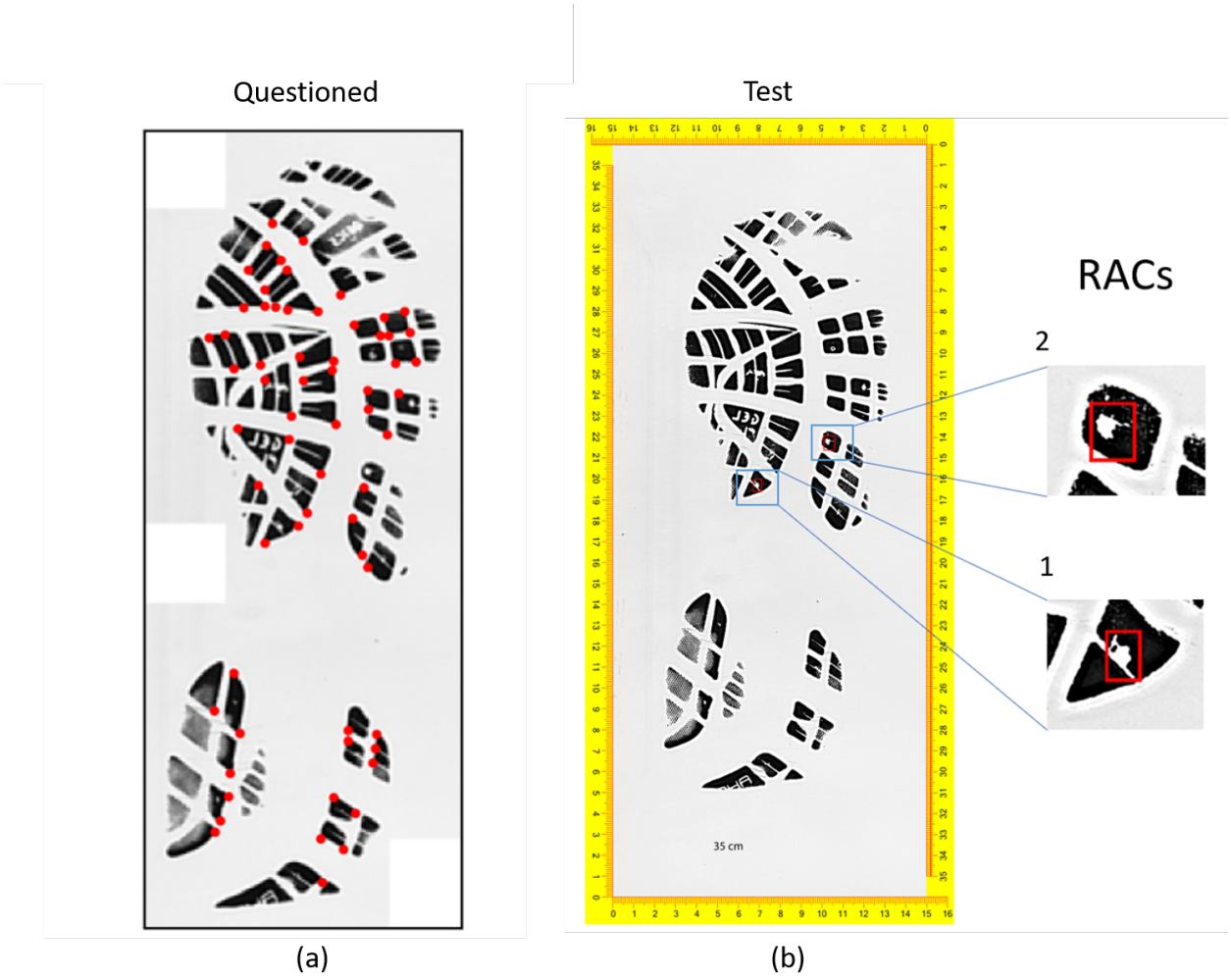

Figure 2: *(a) Manual annotation of the questioned impression with corner points. (b) Manual annotation of the test impression with bounding boxes around apparent RAC areas.*

The next step is to manually annotate K (see Figure 2 (b)). This consists simply of placing vertical bounding boxes around any apparent RAC areas. The examiner will likely have the physical shoe available to help do this. Again, this annotation is done in Adobe Photoshop.

After each impression is annotated, the scale visible in each image (in both Q and K) is used to determine the resolution of the image. A user marks a straight line along the visible scale (in Adobe Photoshop) and reports the length in centimeters. This length is used to calculate the number of pixels per centimeter so that the two images can be scaled to the

pixel density of the image with lower resolution, facilitating pixel by pixel comparisons. In addition, the region of interest in each impression is manually annotated by marking a closed polygon around the shoeprint area (again using Adobe Photoshop). By marking the region of interest, regions of the image known to be outside the shoeprint can be excluded during comparison.

The next step is to automatically align Q and K. The corner points automatically found in K for this example are shown in Figure 3.

Figure 4 shows the result of automated alignment using the maximum clique algorithm. The algorithm obtained its best alignment using 11 corresponding points. These points are shown in red in the left image and in blue in the middle and right images. Corresponding areas in the two images for RAC regions are also shown.

Figure 3: *Automatic extraction of corner points in K. (Note that RAC boxes marked manually are also shown as green rectangles.*

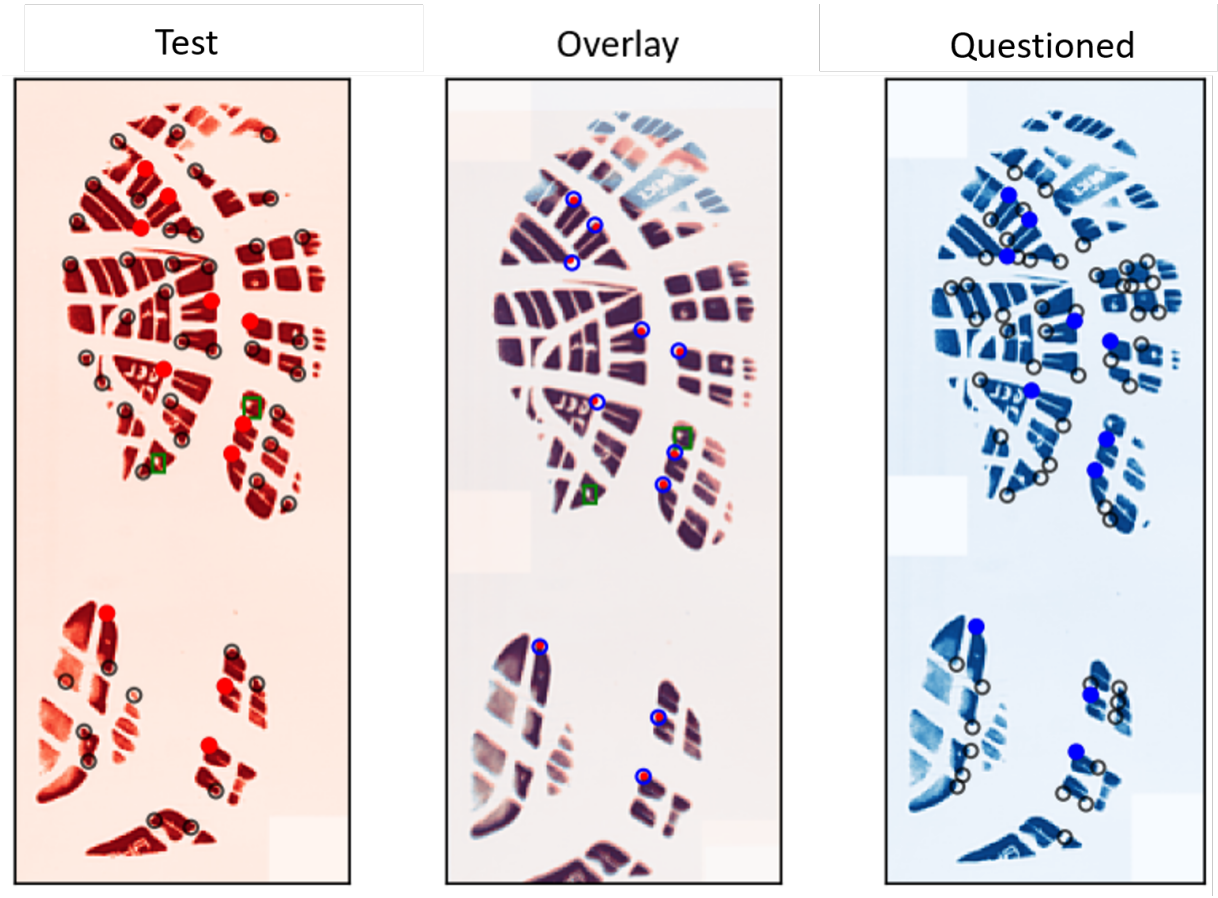

Figure 4: *Automated alignment. The middle image shows the overlay obtained, after alignment, of K (on the left) and Q (on the right). The overlay of the two RAC boxes in the middle image shows how RAC boxes marked in the left image can be projected to potential RAC regions in the right image.*

Outsole Pattern Score Evaluation

Figure 5(a) shows the results for automated computation of the $AvPOC$ pattern comparison scores using the Resnet-50 features. The scores are computed only within the overlapping regions of interest in Q and K. Image pixels outside these regions are ignored.

Histograms shown in Figure 5(a) display the reference distribution of scores from 50 KM pairs (red), 125 KCNM pairs (blue), and 210 KNM pairs involving comparisons of impres-

sions from shoes of different design and/or size (gray). All 125 KCNM comparisons in this illustration are between impressions from opposite shoes in the pair of shoes with one of the impressions flipped to resemble an impression from the same side as the other.

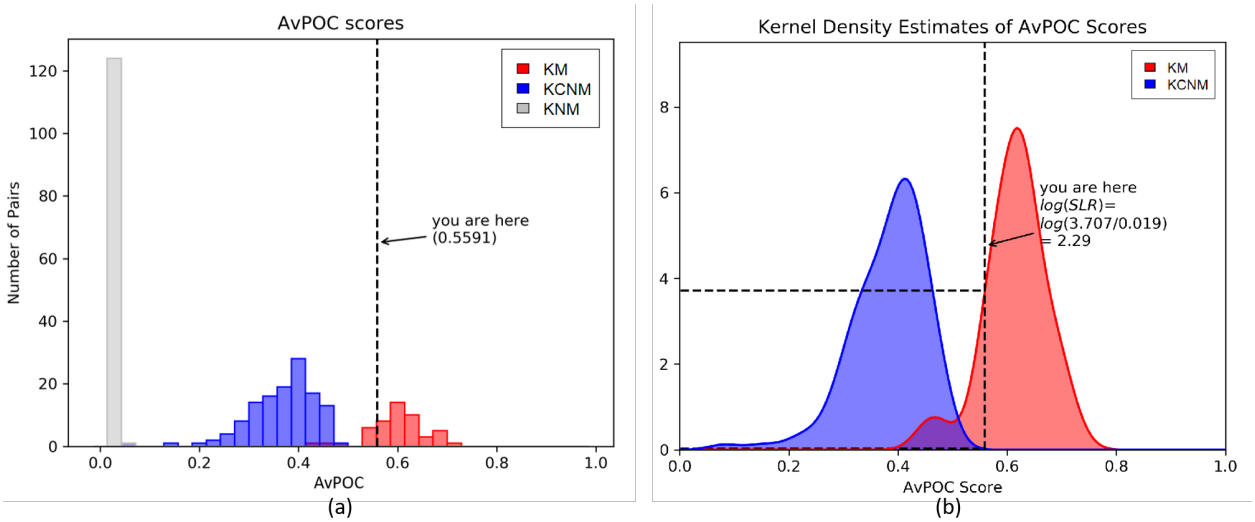

Figure 5: *(a) Outsole pattern scores for pristine comparisons. The score obtained using Average Phase-Only Correlation (AvPOC) for the casework comparison is 0.5591, and is mapped onto the three score reference distributions in the chart. The red histogram corresponds to pristine known match (KM) scores, the blue histogram to pristine known close non-match (KCNM) scores and the gray histogram to pristine known non-match (KNM) scores. The AvPOC metric easily discriminates between KNM (gray) and KM (red) comparisons. (b) The AvPOC score for the casework comparison (0.5591) is mapped on the two fitted score background distributions obtained by smoothing the KM and KCNM histograms in the left panel using kernel density estimation.*

All impressions for these comparisons were obtained using the Everspry EverOS scanner (34) and were processed as outlined in the above paragraphs. As described in step 4 under Section 2.1, we used Resnet-50 features to compute *AvPOC* as the outsole pattern score for each comparison. The role of the background reference distributions is to provide context for evaluating the *AvPOC* score obtained in a pristine comparison.

It is clear from Figure 5(a) that the scores for KM comparisons (red) are well separated from those for comparisons involving pristine impressions from two shoes of different make and/or model and/or size (gray). The general separation exhibited between the KM (red) and KCNM (blue) scores indicates that the discriminating information used by the $AvPOC$ metric includes not only size and design, but also wear and, likely to a lesser extent, RACs. The slight overlap observed between the KM (red) and KCNM (blue) confirms the challenge of discriminating between the two.

The outsole pattern comparison score for our example comparison is 0.5591, which is seen to be within the range of scores obtained from the pristine KM reference comparisons and is higher than any of the scores obtained from the pristine KNM comparisons. This result would provide empirical support for an examiner's opinion that design and size evaluations for the current case do not provide a basis for exclusion; otherwise the current case score would fall within the KNM range of scores. The obtained score is also higher than any of those from pristine KCNM reference comparisons. This result would provide empirical support for an examiner's opinion that observed wear correspondence for the current case exceeds what is generally observed among pristine KCNMs. Overall, Figure 5(a) would support an examiner's opinion that the current case comparison has strong correspondences in design, size, and wear.

We next quantify the contribution of the observed outsole pattern comparison score to the final comparison score in light of the displayed reference comparisons using an SLR. Figure 5(b) shows smoothed versions of the mated and close non-match histograms in Figure 5(a), obtained using kernel density estimation (36, 37). These smoothed histograms are in fact probability densities (i.e., the total area beneath either curve is 1). For the probability density function representing mated comparisons, the height of the curve for an $AvPOC$ score of 0.5591 is 3.707. For the close non-match curve, the height is 0.019. The SLR is given by the ratio of these two heights, computed as 3.707/0.019=195.11 or, equivalently,

a $\log_{10}$(SLR) value equal to 2.29. Figure 5(b) is annotated to show these values. As with Figure 5(a), Figure 5(b) provides empirical support for an examiner's opinion regarding the outsole pattern comparison score. A score that lies mainly within the KM scores (red region) indicates that this score value falls in a region that has been more frequently seen among results obtained when comparing KMs than when comparing KCNMs; the $\log_{10}$(SLR) value in such cases will tend to be positive. A score that lies mainly within the KCNM scores (blue region) indicates that this score value falls in a region that has been more frequently seen among results obtained when comparing KCNMs than when comparing KMs; the $\log_{10}$(SLR) value in such cases will tend to be negative. A score in a region that has occurred nearly equally often among KM and KCNM comparisons will not have a large effect on the final comparison score, as the $\log_{10}$(SLR) value in such cases will be nearly zero.

It is important to note that the probability densities used in this workflow represent just one possible way of translating the observed reference comparison scores into smoothed distributions. For example, applying kernel density estimation with different levels of smoothing (i.e., different bandwidths) will produce different distribution curves and different score-based likelihood ratio (SLR) values. Additional uncertainties come from the fact that adding (or removing) one or two data points from the underlying data set can strongly affect the estimated densities, especially when the underlying data set is not very large. We illustrate the sensitivities related to bandwidth choice and small alterations to the underlying dataset for this SLR in Appendix A. If one intends the SLR value to be accepted as an interpretive end point literally providing the ratio between the respective probabilities of obtaining the observed score under the mated and close non-match comparisons, one should carefully investigate and characterize the range of plausible values this ratio may have under different modeling approaches (including alternatives to kernel density estimation) and different criteria for whether a given model is plausible. This concept is described as an "uncertainty pyramid" and discussed in detail in Lund and Iyer (40). We do not consider or use the SLR value computed here as an interpretive end point. Rather it is a numerical input required

to evaluate the overall comparison score, whose discrimination ability we would like to optimize. For this reason, we do not repeat sensitivity evaluations for other SLR computations presented in this paper. Additionally, to reduce the influence of rounding errors on our final overall score, we retain many significant figures for each scoring component instead of rounding each result based on its uncertainty.

RAC Score Evaluation

Recall that an examiner annotates the test impression for RACs by drawing bounding boxes around features judged to be RACs (see Figure 2(b)). This section discusses the RAC comparison workflow for Everspry EverOS images as in Figure 1.

Once Q is aligned with K, we can locate the regions or patches in Q that correspond to the RAC bounding box regions marked in K. For a given RAC region, the normalized cross correlation (NCC) between the corresponding patches in Q and K is computed according to Equation (1). The value of NCC is always between -1 and +1 (inclusive). If Q and K were made by the same shoe we expect the value of NCC to be closer to +1 when computed for these corresponding patches of pristine impressions. (Note that in the second example, discussed in a later section, that involves more realistic questioned impressions, we use the absolute value of NCC as the RAC metric in case the color scale of Q is inverted relative to K.)

The general steps in the RAC comparison workflow for Everspry comparisons is as follows. For each RAC marked in K, determine if the corresponding patch in Q is within the shoeprint boundary (i.e., region of interest). If Q is a partial impression, some portions of the physical outsole may not have generated pattern information in the impression and some RAC regions may be outside the portion of the physical outsole that generated the impression. If a given RAC patch is inside the region of interest for Q, a search is performed in a local neighborhood around this patch to find the area that has the highest similarity score, in this case normalized

cross correlation (NCC). The local neighborhood in Q is determined as follows. After Q has been rotated and translated to fit K, the RAC box in K is projected onto Q, where it will be aligned with the $x - y$ axes. The resulting box is then enlarged by 1% of the whole image on each side of the box. This means the total length and total width of the box are increased by 2% of the image length and width, respectively. This helps account for any local distortions that may be present in Q relative to K. The maximized NCC value is the RAC comparison metric used.

Reference KM and KCNM RAC score sets for pristine comparisons were obtained by applying the above process to the Everspry impression pairs and used to generate the KM and KCNM distributions in Figure 5(a). This resulted in 640 pristine KM RAC scores and 1600 pristine KCNM RAC scores, which are represented in Figure 6 (Top). To compute the contribution of a RAC similarity score to the final overall score, we apply kernel density estimation to these reference score sets and compute the $\log_{10}$(SLR) for each RAC identified in K. Figure 6 (Bottom) shows the results of RAC comparison for the two RACs annotated in K (see Figure 2(b)) by the examiner. RAC #1 produced a normalized cross correlation (NCC) score of 0.980. The corresponding score-based likelihood ratio (SLR) value is $3.130/0.047 = 66.596$ and the $\log_{10}$(SLR) is 1.82. RAC #2 in the example comparison produced an NCC score of 0.760, which corresponds to an SLR of 1.267/0.354=3.579 and a $\log_{10}$(SLR) of 0.55. The charts in these figures provide the examiner with information that helps facilitate an explanation of contributions from the RAC comparison scores.

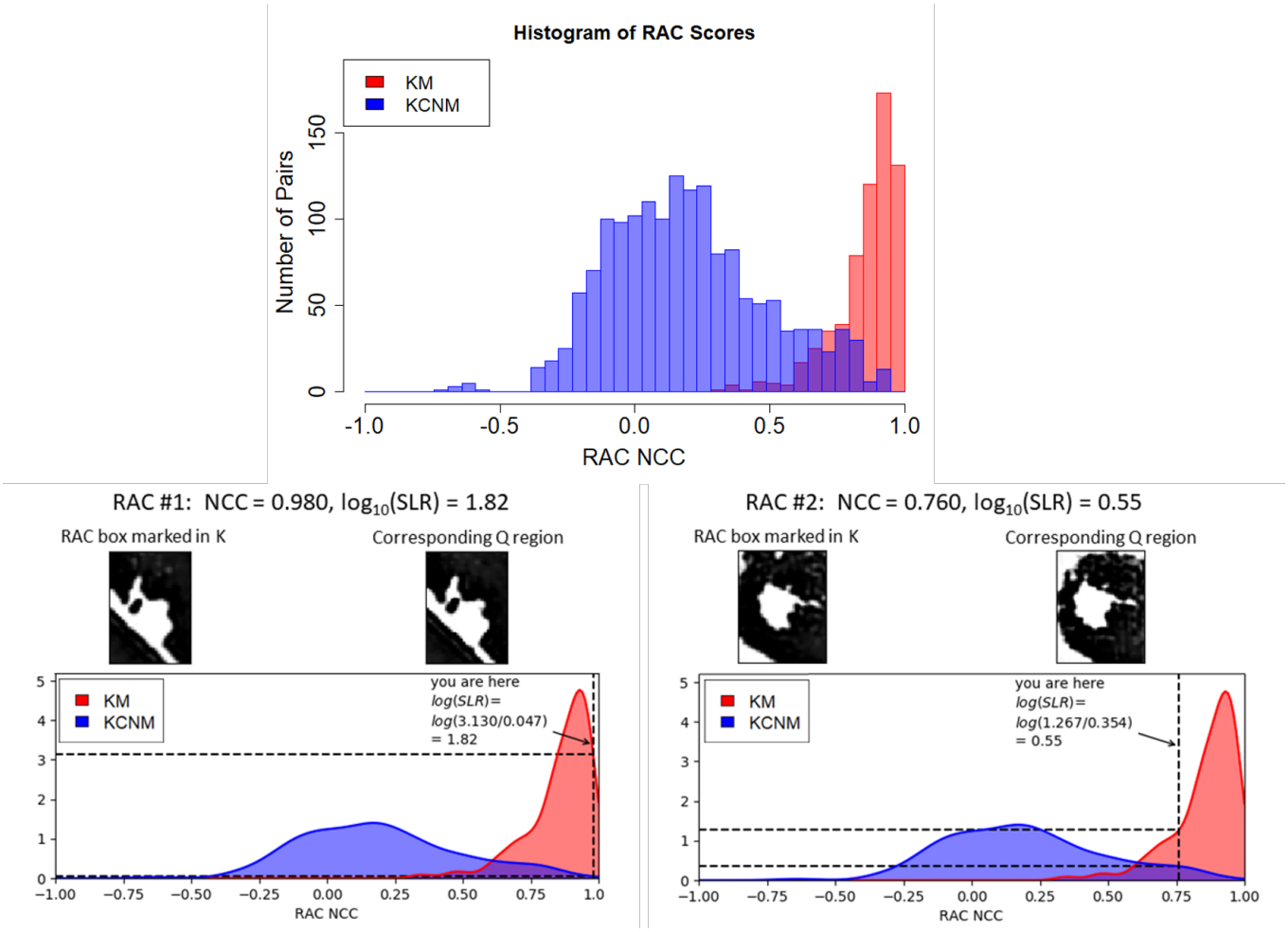

Figure 6: *(Top) Normalized cross correlation (NCC)-Score histograms for pristine comparisons of known match (KM) (red) and known close non-match (KCNM) (blue) RAC regions. (Bottom Left, Bottom Right) Algorithm computes comparison scores for RAC #1 and RAC #2. The RAC regions from Q and K that are being compared are shown at the top of each panel. The obtained score is placed in the context of KM and KCNM score distributions obtained from the top panel.*

Different Q versus K comparisons will involve different numbers of RACs. Each of these RAC comparisons has an NCC value and a corresponding $\log_{10}$(SLR) value as explained above. Results from individual RACs are combined into a composite RAC score, formed by summing the $\log_{10}$(SLR) values across all RACs in the currently considered Q versus K

comparison. In our current example, two RACs have been annotated on the test impression, leading to $\log_{10}$(SLR) values of 1.82 and 0.55, respectively. Thus, for this example the composite RAC score is 1.82+0.55=2.37.

Final Comparison Score

The final comparison score is computed by summing the outsole pattern $\log_{10}$(SLR) score and the composite RAC $\log_{10}$(SLR) score. In our Everspry example (Figure 1) this value is 4.66. Figure 7 shows this score in the context of reference KM and KCNM distributions of final scores for similar comparisons.

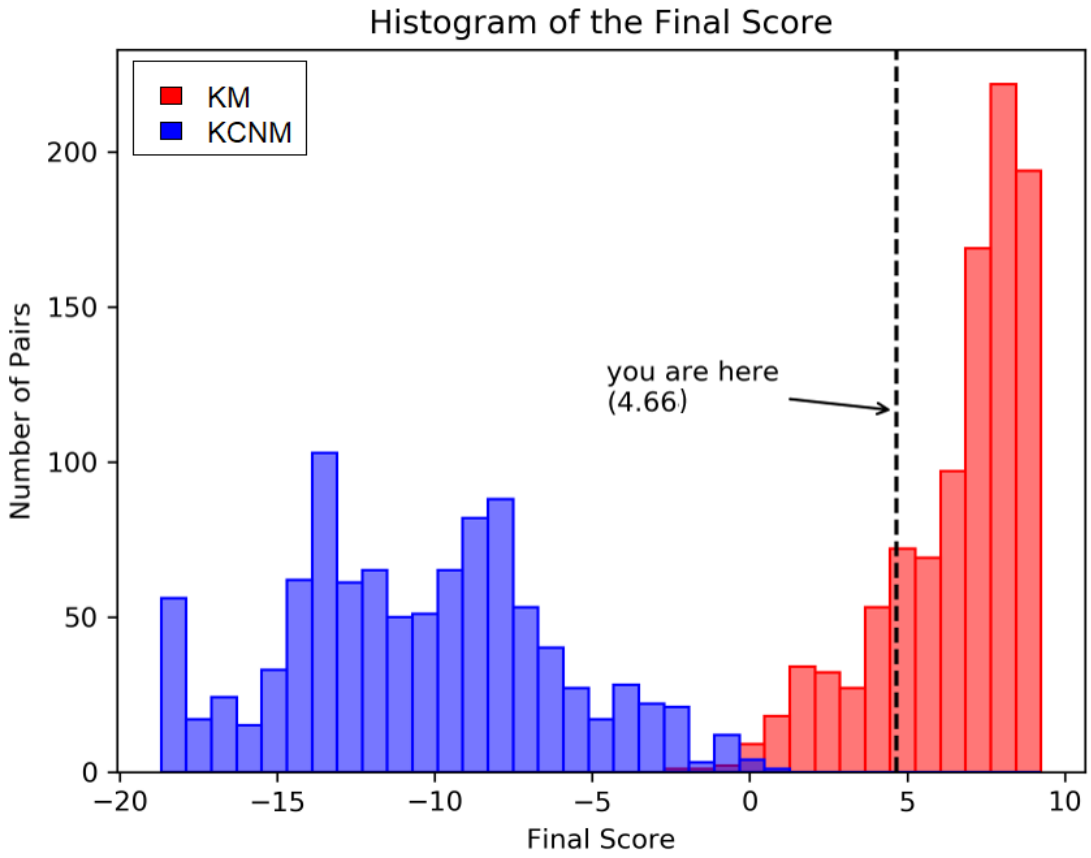

Figure 7: *The final score is the sum of the outsole pattern $\log_{10}(SLR)$ score and the composite RAC $\log_{10}(SLR)$ (score-based likelihood ratio) score; its value is 4.66. This chart shows that value mapped on the two reference final-score histogram distributions corresponding to known match (KM) and known close non-match (KCNM) comparisons in which two RAC regions were identified on the test impression and for which Q is pristine.*

These reference distributions are based on the number of RACs being compared, as follows. The final score linearly depends on the number of RACs. Because the final score is the sum of the outsole pattern and RAC scores, a score which is high if only two RACs are compared may be low if four RACs are compared. So we generate reference distributions for the final score which depend on the number of RACs compared. Ideally, the database of comparisons would

have a collection of pairs sufficiently large to build reference distributions for a comparison with k RACs. For now, we simulate such distributions using the reference distributions of outsole pattern and RAC scores. This simulation is described in Appendix B.

The reference distribution in Figure 7 is for 2 RACs. As before, Figure 7 provides the examiner or any other stakeholder with information for assessing the strength of the evidence provided by the footwear impression comparison analysis. A score that lies mainly within KM scores indicates support for the proposition that the casework pair of impressions come from the same shoe; a score that lies mainly within KCNM scores indicates support for the proposition that the casework pair of impressions were made by different shoes. A score that occurs nearly equally often among KM and KCNM scores does not indicate support for either proposition.

Mock Crime Scene Example

We now consider a more realistic mock casework comparison between the impressions shown in Figure 8. Figure 9 shows the manual annotation of Q, consisting of the binary contact region and corner points marked in the contact region.

Figure 8: *Example of two images that are input to the end-to-end system: K (on the left) and a (mock) Q (on the right). In this case, the two impressions are made by the same shoe. (Impressions courtesy of Brian McVicker, FBI.)*

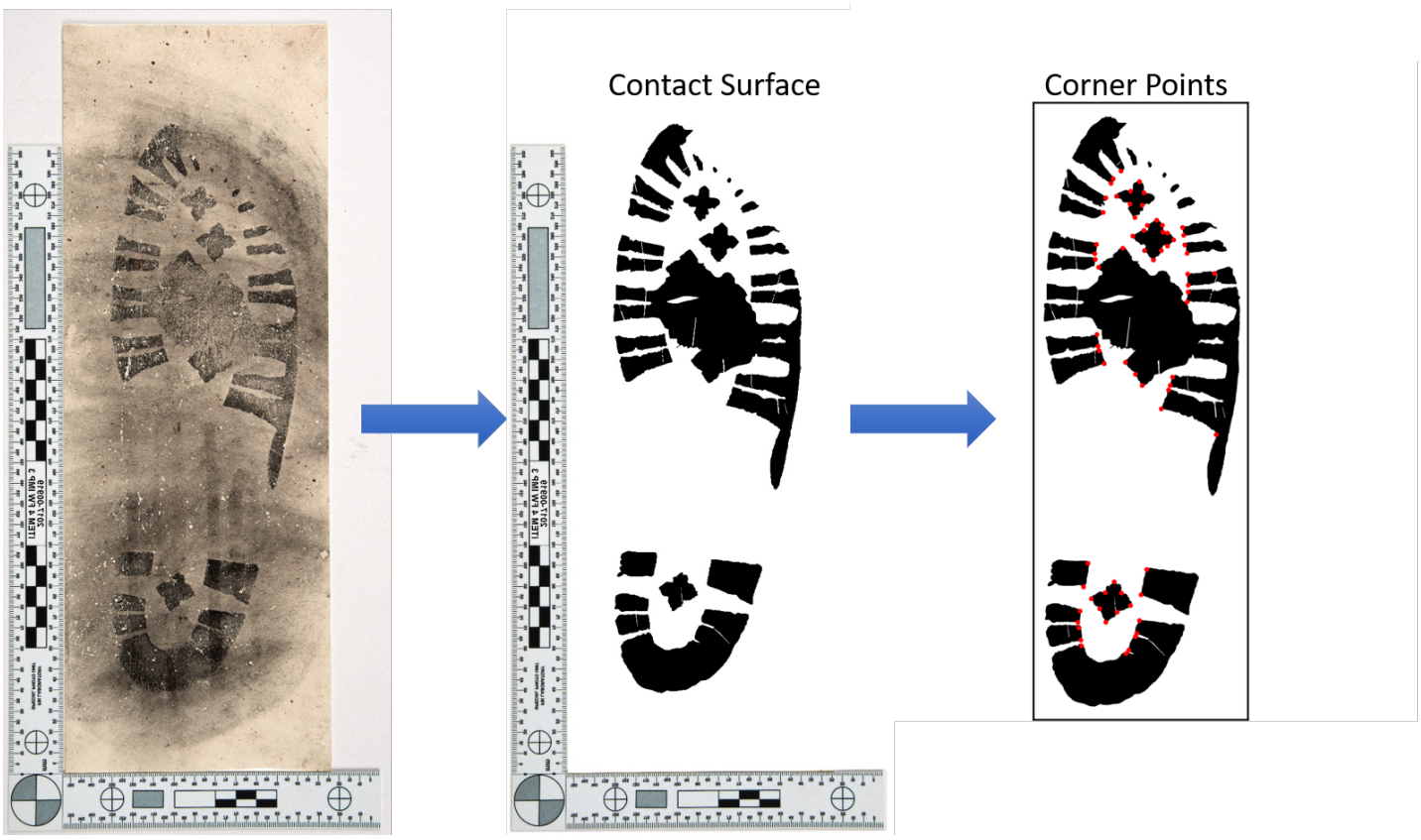

Figure 9: *Manual annotation of the (mock) Q (left) with contact surface (middle) and corner points in red (right). (A set of corner points could also have been automatically extracted in this case due to the high-quality contact surface markup.)*

The next step is to manually annotate K (see Figure 10) by placing vertical bounding boxes around any apparent RAC areas. Corner points are then automatically extracted in K, and then K and the binary contact image from Q are automatically aligned using the maximum clique algorithm.

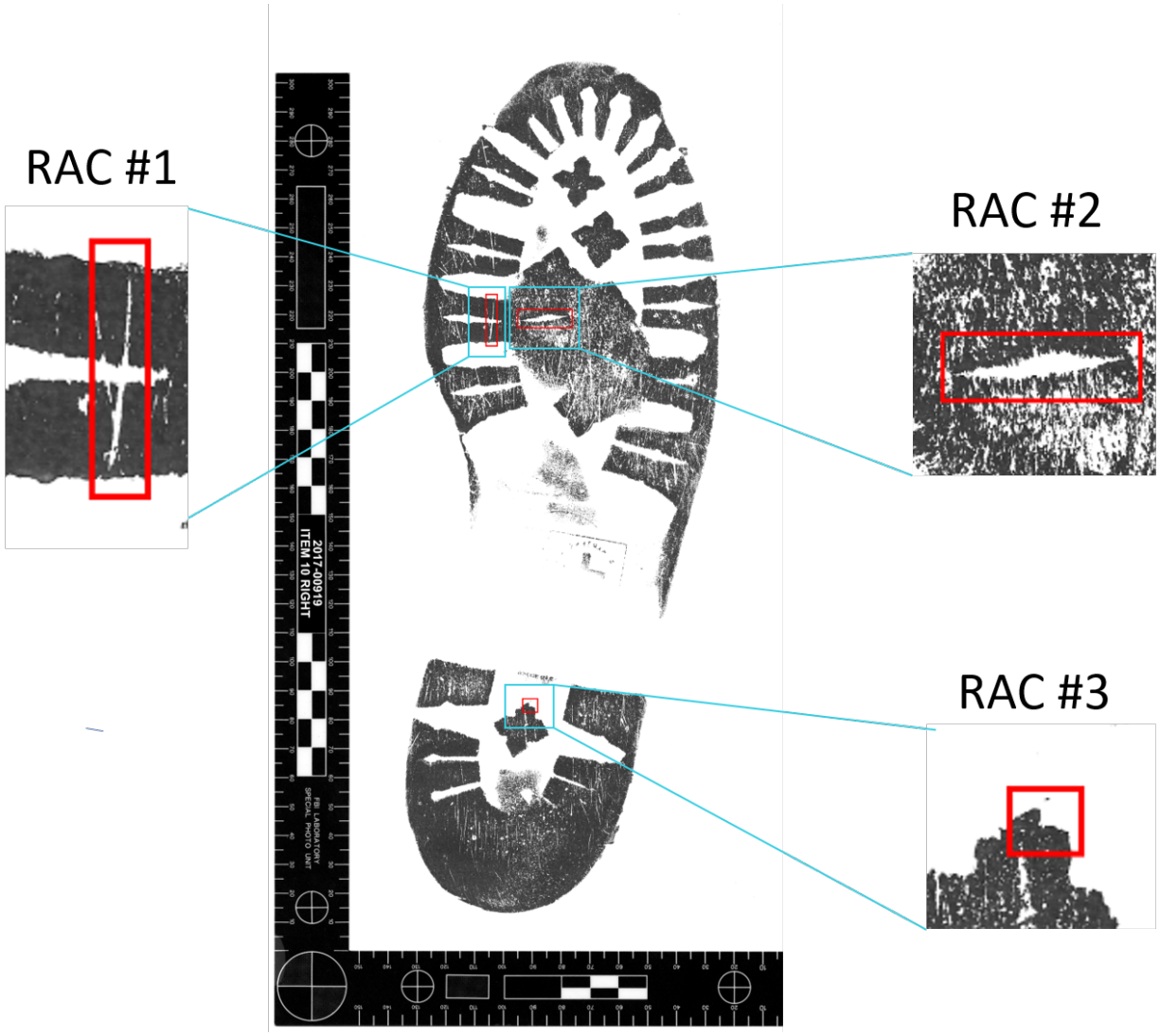

Figure 10: *Manual annotation of K with bounding boxes around three apparent RAC areas.*

Figure 11 shows the *AvPOC* outsole pattern scores obtained from comparisons involving more realistic Qs. These scores consider the Resnet-50 features only within overlapping regions of interest in Q and K. Image pixels outside these regions are ignored. Histograms shown in the left panel of Figure 11 display the distribution of scores from KM pairs (red), KCNM pairs (blue), and KNM pairs (gray). These reference distributions are generated from comparisons involving 8 realistic mock Qs. An example is shown in Figure 12. The KM scores result from comparing the binary contact markup from each mock Q with 5 Ks

from the shoe that created Q, producing a total of 40 KM outsole pattern scores. The KCNM scores result from comparing the binary contact markup from each Q with 5 Ks from each of 5 different shoes of the same make, model, and size, as the shoe that created Q. Two of these shoes are from the same side as the shoe that created Q. Impressions from the other three shoes were flipped to appear as originating from the same side as the shoe that created Q. In total, we produced 200 KCNM outsole pattern scores (120 of which involved flipped impressions). The KNM scores result from comparing the binary contact markup from each Q with a K from each of 10 shoes with different outsole designs from the shoe that created Q, producing 80 KNM outsole pattern scores.

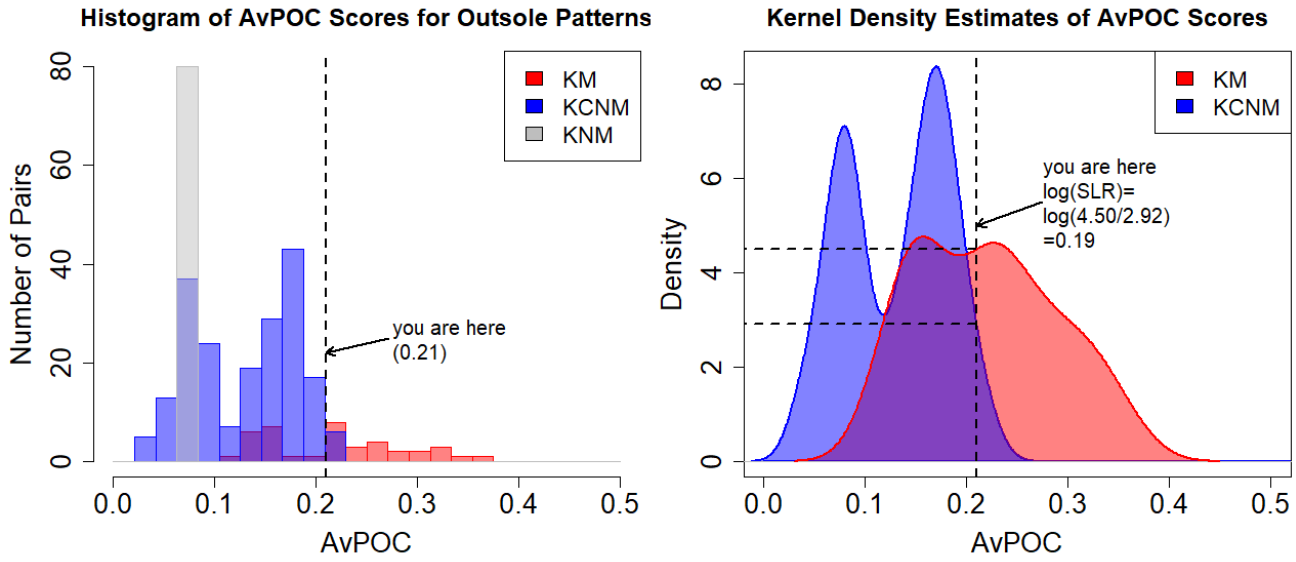

Figure 11: *Outsole pattern comparison scores from mock crime scene examples. (Left) Histogram of known match (KM), known close non-match (KCNM) and known non-match (KNM) outsole pattern scores. The score obtained using Average Phase-Only Correlation (AvPOC) for the casework comparison is 0.2100, and is mapped onto the score reference distributions in the chart. (Right) Kernel density estimates for the KM and KCM outsole pattern scores. The observed casework score of 0.21 has a corresponding* $\log_{10}(SLR)$ *of 0.19.*

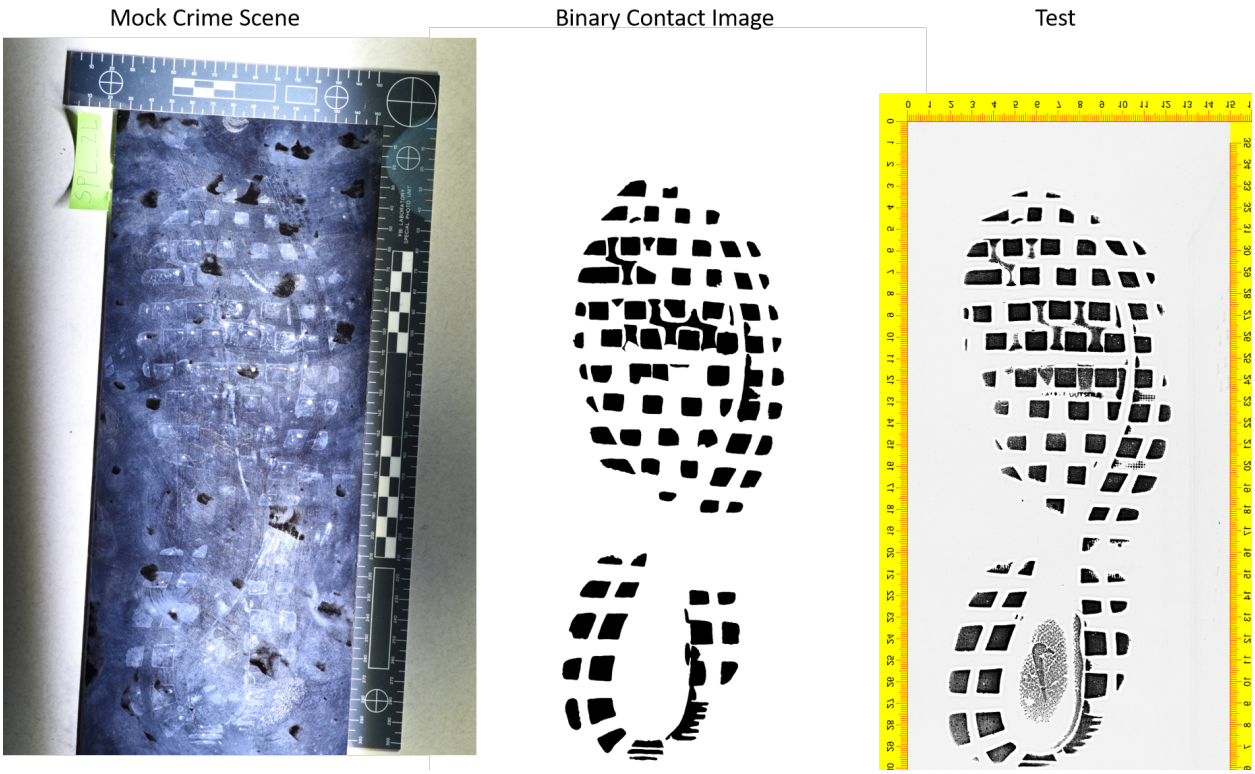

Figure 12: *Example of Nike shoe Qs used to generate the reference distributions in Figure 11. Shown on the left is a mock Q obtained by stepping in water to wet the shoe outsole, then stepping on tile, sprinkling aluminum powder on the tile, and then using a gel lift. In the middle is a binary contact image manually generated from the left impression. On the right is K, an Everspry test impression obtained from the same shoe.*

It is clear from Figure 11 (left panel) that the scores for the KM, KCNM, and KNM comparisons involving more realistic Qs are not as well separated as in Figure 5(a), where the Qs are pristine. This result is consistent with expectations that comparisons are easier when Qs appear like test impressions than when they reflect typical crime scene conditions.

The outsole pattern score for the case comparison in Figure 8 is 0.2100. As annotated in the right panel of Figure 11, this translates to an SLR value of 4.50/2.92=1.54 or, equivalently, a $\log_{10}$(SLR) value equal to 0.19. Had the scores obtained from comparisons with pristine Qs been used for context, as in Figure 5(b), the score of 0.21 would have had a severely negative

$\log_{10}$(SLR) value, since 0.21 is lower than most KCNM scores and much lower than all KM scores among the pristine comparisons. This highlights the importance of selecting reference comparisons that reflect the general conditions (e.g., completeness, complexity, and clarity) of the currently considered case.

As with the pristine comparisons, each RAC region annotated on K by the user is compared to the corresponding region of Q after alignment. Note that the corresponding RAC region used for comparison is obtained from the original Q, not from the binary contact markup of Q. The top left of Figure 13 provides the KM and KCNM RAC scores obtained from comparisons involving more realistic Qs. The two distributions are seen to be nearly perfectly overlapping, although our collection includes far more KCNM RAC comparisons than KM RAC comparisons. As such, most RAC scores appear nearly equally likely under either set. This means that these RACs will have little influence since $\log_{10}$(SLR) values will generally be close to 0.

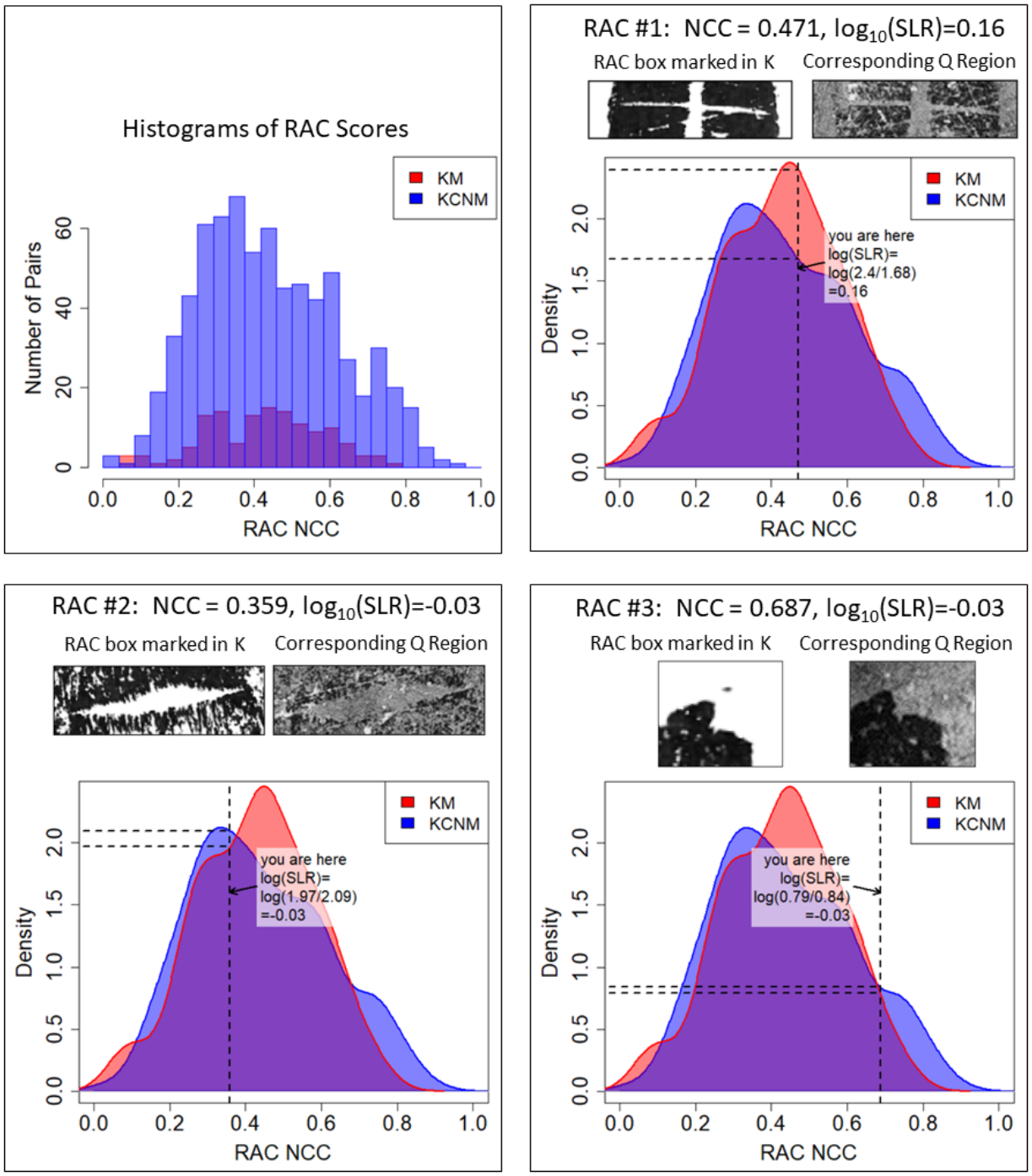

Figure 13: *RAC comparison scores from mock crime scene examples. (Top Left) Histograms of known match (KM) and known close non-match (KCNM) RAC comparison scores. (Top Right, Bottom Left, Bottom Right) Results from RAC regions 1, 2, and 3 in the current example, placed in the context of kernel density estimates from the KM and KNM collections of RAC comparison scores shown in the top left panel.*

For instance, as illustrated in Figure 13, the three considered RAC regions in the current example produced normalized cross correlation (NCC) values of 0.471, 0.359, and 0.687, respectively. These NCC values have corresponding $\log_{10}$(SLR) values of 0.16, -0.03, and -0.03 for a composite RAC score of 0.1.

Upon visual inspection, none of the eight mock crime scene impressions used to form these reference collections show clear signatures of RACs. Adding scores from instances where the mock crime scene impressions have clear RAC signatures would likely increase the KM density associated with high correlation values.

Additionally, although effective in pristine comparisons, the normalized cross correlation (NCC) metric is less effective for RAC comparisons involving realistic Qs. Although each of the three RAC regions shown in Figure 13 provide clear visual correspondence between K and Q, yet these regions produce lower NCC values than in pristine comparisons. We will eventually replace NCC with a more powerful RAC comparison metric to provide better separation between the KM and KCNM score collections, at least for comparisons involving high-quality, but realistic, Qs like the example considered in this section. RAC comparisons involving low-quality Qs are expected to remain very difficult, as these comparisons are generally challenging even for trained examiners.

The final comparison score for this example is 0.29, formed by summing the $\log_{10}$(SLR) value from the outsole pattern comparison (0.19) and the composite $\log_{10}$(SLR) value for the RAC comparisons (0.10). Figure 14 displays this score in the context of final scores corresponding to KM and KNM comparisons in which three RAC regions are identified on K and for which Q is more realistic. As seen in Figure 15, and because the KM and KNM RAC scores are not well separated for the more realistic comparisons, the separation between KM and KCNM final score distributions does not noticeably improve as the number of considered RACs increases.

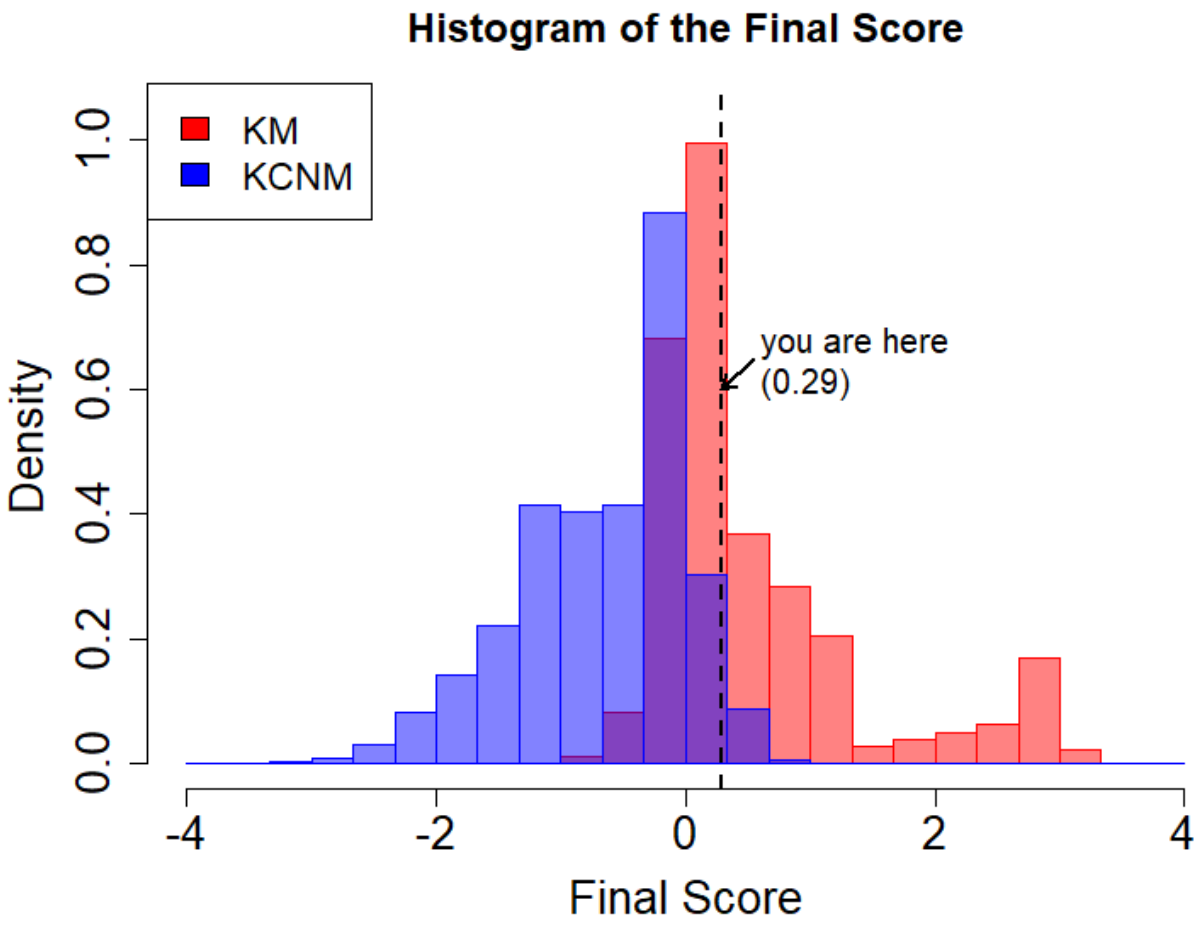

Figure 14: *The final score is the sum of the outsole pattern $\log_{10}(SLR)$ (score-based likelihood ratio) score and the composite RAC $\log_{10}(SLR)$ score; its value is 0.29. This chart shows that value mapped on the two final-score histogram reference distributions corresponding to known match (KM) and known close non-match (KCNM) comparisons in which three RAC regions are identified on K and for which Q is more realistic.*

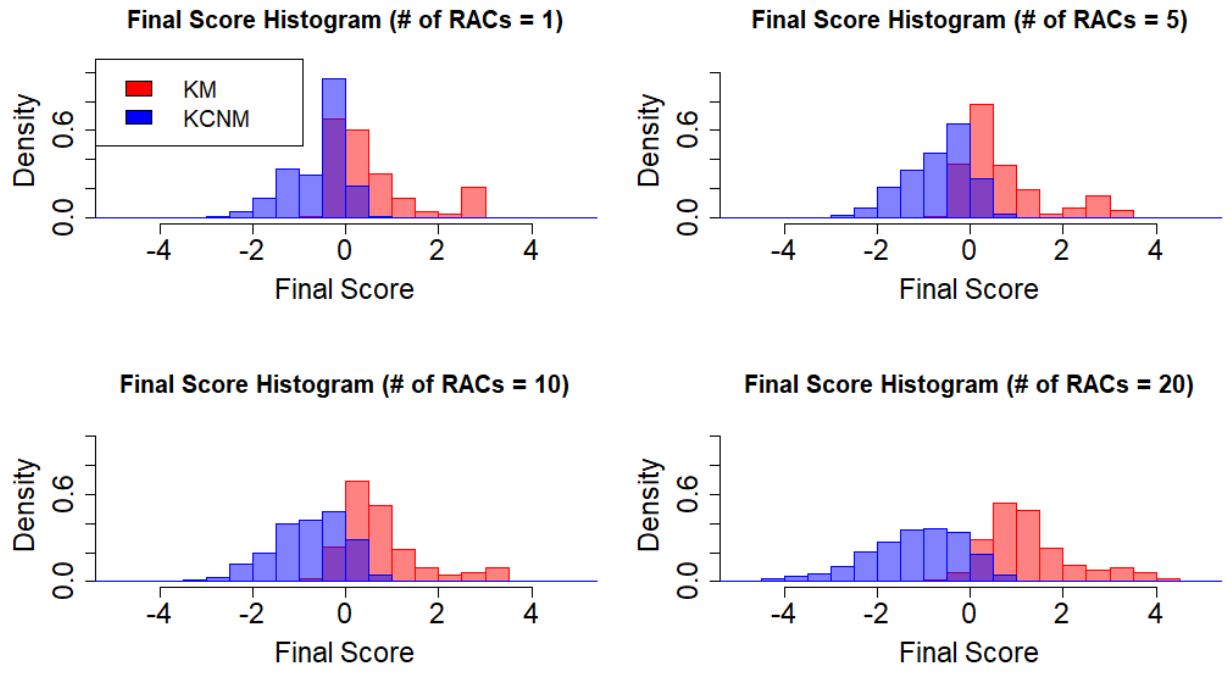

Figure 15: *Simulated final score distributions for known match (KM) and known close non-match (KCNM) comparisons involving a more realistic Q and 1, 5, 10 or 20 RAC regions annotated in K, respectively.*

The reference distribution in Figure 14 is for 3 RACs. As before, Figure 14 provides the examiner or any other stakeholder with information for assessing the strength of the evidence provided by the footwear impression comparison analysis. A score that lies mainly within KM scores indicates support for the proposition that the casework pair of impressions come from the same shoe; a score that lies mainly within KCNM scores indicates support for the proposition that the casework pair of impressions are made by different shoes. A score that occurs nearly equally often among KM and KCNM scores does not indicate support for either proposition.

Conclusions

This paper has presented an end-to-end workflow for quantitative evaluation of footwear evidence. The workflow includes human annotation of the crime-scene impression, human markup of RACs and other features in the test impression, automated alignment of the two impressions, a multi-stage automated comparison that includes size, design, wear and RACs, the use of relevant reference ground-truth-known score distributions to provide context to interpret the comparison scores, automatically combining all the comparison scores into a single composite final score, and visual displays to help the examiner understand, document, and present the quantitative results. The paper has demonstrated the workflow, which has included obtaining and interpreting outsole pattern scores, RAC comparison scores and final scores both for comparisons involving pairs of Everspry EverOS scanner impressions and for comparisons involving more realistic questioned impressions.

Future versions of the system will demonstrate additional and improved components of the workflow, with the goal of eventually achieving versions that are robust enough to be deployable in casework. The following are some of the additional components and improvements that require research.

- The comparison metrics we have used, Average Phase-Only Correlation ($AvPOC$) and normalized cross correlation (NCC), have been found to be effective for good clarity impressions, both when comparing Q and K impressions from arbitrary shoes as well as when comparing impressions from close non-matching shoes (24). However, research is still required to obtain improved metrics for comparing lower clarity impressions and for comparing RACs.
- As we see with the mock crime scene impressions, the clarity of questioned impressions varies greatly. In this paper, crime scene clarity is only considered at the level of "pristine" versus "more realistic." However, a more nuanced approach for taking clarity

into account is necessary for evaluating how similar, or relevant, reference comparisons are to casework comparisons and to indicate how challenging it is to tell contact regions apart from non-contact across different regions of a crime scene impression. We plan to investigate methods for examiner annotation of the quality/clarity of the crime scene impression.

- Future implementations will include quantitative metrics for describing a comparison's "type" to facilitate selection of relevant reference comparisons. For example, when selecting which reference comparisons should be used to provide context for the outsole pattern scores, it is important to select those comparisons that reflect the general conditions of the currently considered case. These general conditions may include, for example, the completeness and clarity of the outsole pattern in Q and the complexity of the outsole pattern in K. Clarity could be numerically represented using the manual annotations discussed in the previous bullet point. In particular, after alignment one can identify which regions of Q correspond to the shoe outsole as seen in K. Using the pixels inside this region, one could compute an average clarity value after assigning each clarity level a numeric value. Alternatively, one could simply report the proportion of the identified region each clarity level (assigned in Q) occupies. Outsole complexity could be represented by the proportion of pixels in K that fall on feature edges (i.e., the concentration of edge pixels). For RAC comparisons, one might consider the size of the RAC as observed in K and local clarity in the corresponding region of Q as comparison "type" features. For overall outsole comparisons and RAC comparisons, respectively, the corresponding comparison "type" features would be evaluated for any incoming case comparison and a distance would be computed between the case features and those from every available reference comparison. The reference comparisons for which the "comparison type" metrics are close to those of the considered case would be deemed most relevant and used to form the reference distributions that provide context for the similarity scores obtained in the case comparison.

- Real crime-scene impressions involve many deformations due to the flexibility of outsoles. The alignment method we use in this paper is rigid alignment (i.e., rigid translation and rotation). When a RAC in K is projected onto Q, we search around the local neighborhood of the projected patch in Q to account for local distortions. We plan to develop approaches for performing a more accurate *flexible alignment* for optimally matching contact regions in the crime scene impression with contact regions in the test impression.

- For future versions, we plan to explore the annotation of potential RACs visible in the questioned impression, annotation of RACs by tracing their boundaries, annotation of wear regions in the test impression (allowing wear to be explicitly analyzed), and metrics for RAC comparisons that take into account the difficulty of finding RACs in most crime scene impressions.

- Deploying our workflow in casework will require a more complete database of reference comparisons that provides a reasonable coverage of the *factor space* (by which we mean combinations of quality levels, levels of partialness, complexity of outsole design patterns, lifting methods, substrates, etc.)

- In the future, we plan to map output scores to the prevalent examiner conclusion scale (e.g., SWGTREAD conclusions (21) or OSAC conclusions) by calibrating the scores from our workflow with examiner judgements when such data become available.

This paper has presented a feasible implementation of a workflow for an end-to-end system that can provide quantitative support for evaluation of footwear evidence by examiners. We have demonstrated the workflow on two scenarios, a pristine comparison set and a more realistic mock crime scene comparison set. We are currently developing a new version of the system that will implement many of the desired improvements enumerated above. With future refinements, it is envisioned that the system will evolve into one that can be deployed

in routine casework.

References

1. National Research Council. Strengthening forensic science in the United States: a path forward. Committee on Identifying the Needs of the Forensic Sciences Community. Washington, DC: National Academies Press, 2009. Document No.: 228091

2. President's Council of Advisors on Science and Technology (PCAST). Report to the President – Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Washington, DC: Executive Office of the President of the United States, 2016 Sep.

3. Richetelli N, Hammer L, Speir JA. Forensic footwear reliability: Part III – positive predictive value, error rates, and inter-rater reliability. J Forensic Sci 2020;65(6):1883-93. doi: 10.1111/1556-4029.14552.

4. Hammer L, Duffy K, Fraser J, Daeid NN. A study of the variability in footwear impression comparison conclusions. J Forensic Identif 2013;63(2):205–18.

5. Richetelli N, Lee MC, Lasky CA, Gump ME, Speir JA. Classification of footwear outsole patterns using Fourier transform and local interest points. Forensic Sci Int 2017;275:102-9. doi: 10.1016/j.forsciint.2017.02.030.

6. Kortylewski A, Vetter T. Probabilistic compositional active basis models for robust pattern recognition. In: Wilson RC, Hancock ER, Smith WAP, editors. Proceedings of the British Machine Vision Conference (BMVC); 2016 Sep 19-22; York, UK. BMVA Press, 30.1-30.12. doi: 10.5244/C.30.30.

7. Kong B, Supancic J, Ramanan D, Fowlkes CC. Cross-domain image matching with deep feature maps. Int J Comput Vis 2019;127(11):1738-50. doi: 10.1007/s11263-018-01143-3.

8. Cui J, Zhao X, Liu N, Morgachev S, Li D. Robust shoeprint retrieval method based on local-to-global feature matching for real crime scenes. J Forensic Sci 2019;64(2):422-30. doi: 10.1111/1556-4029.13894.

9. Wu Y, Wang X, Nankabirwa NL, Zhang T. LOSGSR: Learned opinion score guided shoeprint retrieval. IEEE Access 2019;7:55073-89. doi: 10.1109/ACCESS.2019.2912585.

10. Yekutieli Y, Shor Y, Wiesner S, Tsach T. Expert assisting computerized system for evaluating the degree of certainty in 2D shoeprints. Washington, DC: US Department of Justice, 2012 Sep. NCJ Number 250336.

11. Wiesner S, Shor Y, Tsach T, Kaplan-Damary N, Yekutieli Y. Dataset of digitized RACs and their rarity score analysis for strengthening shoeprint evidence. J Forensic Sci 2010;65(3):762-74. doi: 10.1111/1556-4029.14239.

12. Shor Y, Wiesner S, Tsach T, Gurel R, Yekutieli Y. Inherent variation in multiple shoe-sole test impressions. Forensic Sci Int 2018;285:189-203. doi: 10.1016/j.forsciint.2017.10.030.

13. Petraco NDK, Gambino C, Kubic TA, Olivio D, Petraco N, Statistical discrimination of footwear: A method for the comparison of accidentals on shoe outsoles inspired by facial recognition techniques. J Forensic Sci 2010;55(1):34-41. doi: 10.1111/j.1556-4029.2009.01209.x.

14. Speir JA, Richetelli N, Fagert M, Hite M, Bodziak WJ. Quantifying randomly acquired characteristics on outsoles in terms of shape and position. Forensic Sci Int 2016;266:399-411. doi: 10.1016/j.forsciint.2016.06.012.

15. Richetelli N, Bodziak WJ, Speir JA. Empirically observed and predicted estimates of chance association: Estimating the chance association of randomly acquired characteristics in footwear comparisons. Forensic Sci Int 2019;302:109833. doi: 10.1016/j.forsciint.2019.05.049.

16. Richetelli N, Nobel M, Bodziak WJ, Speir JA. Quantitative assessment of similarity between randomly acquired characteristics on high quality exemplars and crime scene impressions via analysis of feature size and shape. Forensic Sci Int 2017;270: 211-22. doi: 10.1016/j.forsciint.2016.10.008.

17. Spencer NA, Murray JS. A Bayesian hierarchical model for evaluating forensic footwear evidence. Ann Appl Stat 2020;14(3):1449-70.doi: 10.1214/20-AOAS1334.

18. Damary NK, Mandel M, Yekutieli Y, Wiesner S, Shor Y. Spatial modeling of randomly acquired characteristics on outsoles with application to forensic shoeprint analysis. arXiv:1912.08272v2 [Stat.AP] 2020.

19. Evett IW, Lambert JA, Buckleton JS. A Bayesian approach to interpreting footwear marks in forensic casework. Sci Justice 1998;38(4):241-7. doi:10.1016/S1355-0306(98)72118-5.

20. Skerrett J, Neumann C, Mateos-Garcia I. A Bayesian approach for interpreting shoemark evidence in forensic casework: Accounting for wear features. Forensic Sci Int 2011;2010(1-3):26-30. doi: 10.1016/j.forsciint.2011.01.030.

21. SWGTREAD. Range of conclusions standard for footwear and tire impression examinations. Scientific Working Group for Shoeprint and Tire Tread Evidence, 2013 Mar. `https://www.nist.gov/system/files/documents/2016/10/26/swgtread_10_range_of_conclusions_standard_for_footwear_and_tire_impression_examinations_201303.pdf` (accessed May 31,2021).

22. Champod C, Evett IW, Jackson G. Establishing the most appropriate databases for addressing source level propositions. Sci Justice 2004;44(3):153-64, 2004. doi: 10.1016/S1355-0306(04)71708-6.

23. Dror IE, Thompson WC, Meissner CA, Kornfeld I, Krane D, Saks M et al. Context

management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. J Forensic Sci 2015;60(4):1111-2. doi: 10.1111/1556-4029.12805.

24. Venkatasubramanian G, Hegde V, Padi S, Iyer H, Herman M. Comparing footwear impressions that are close non-matches using correlation-based approaches. J Forensic Sci 2021;66(3):890-909. doi: 10.1111/1556-4029.14658.

25. Rosten E, Drummond T. Machine learning for high-speed corner detection. In: Leonardis A, Bischof H, Pinz A, editors. Computer Vision – ECCV 2006. 2006 May 7-13; Graz, Austria. Lecture Notes in Computer Science, vol 3951. Springer, Berlin, Heidelberg, 430-43. doi: 10.1007/11744023_34.

26. Bomze IM, Budinich M, Pardalos PM, Pelillo M. The maximum clique problem. In: Du DZ, Pardalos PM, editors. Handbook of Combinatorial Optimization, Suppl Vol A, 1999. Kluwer Academic Publishers, Boston, 1-74. doi: 10.1.1.56.6221

27. Tu PH, Hartley RI, inventors. Lockheed Martin Corporation, assignee. Fingerprint matching by estimation of a maximum clique. US Patent 5,933,516. 1999 Aug 3.

28. Venkatasubramanian G. Cliquematch: Finding correspondence via cliques in large graphs. 2020. https://doi.org/10.5281/zenodo.4277288 (accessed April 21, 2021).

29. Rida I, Fei L, Proenca H, Nait-Ali A, Hadid A. Forensic shoe-print identification: a brief survey. arXiv:1901.01431v3 [cs.CV] 2020.

30. Park S, Carriquiry A. An algorithm to compare two-dimensional footwear outsole images using maximum cliques and speeded-up robust feature. Statistical Analysis and Data Mining: The ASA Data Science Journal 2020 Apr;13(2):188-99. doi: 10.1002/sam.11449.

31. Kong B, Supancic J, Ramanan D, Fowlkes C. Cross-domain forensic shoeprint matching. In: Proceedings of the British Machine Vision Conference (BMVC); 2017 Sep 4-7; London, UK. BMVA Press.

32. Stigler SM. Francis Galton's account of the invention of correlation. Statist Sci 1989;4(2):73-9. doi: 10.1214/ss/1177012580.

33. Dogan G, Bernal J, Hagwood CR. A fast algorithm for elastic shape distances between closed planar curves. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7-12; Boston, MA. IEEE Computer Society, 2015;4222-30. doi: 10.1109/CVPR.2015.7299050.

34. EverOS V1.0. Dalian Everspry Sci & Tech Co., Ltd. http://www.everspry.com/en/products/products_03.htm (accessed April 21, 2021).

35. Neumann C. Defence against the modern arts: The curse of statistics: Part I – FRStat. Law Probab Risk 2020;19(1):1-20. doi: 10.1093/lpr/mgaa004.

36. Parzen E. On estimation of a probability density function and mode. Ann Math Statistic 1962;33(3):1065-76. doi: 10.1214/aoms/1177704472.

37. Gramacki A. Nonparametric kernel density estimation and its computational aspects. 1st ed. Switzerland: Springer, 2018. doi: 10.1007/978-3-319-71688-6.

38. Cheng Y, Kashyap RL. An axiomatic approach for combining evidence from a variety of sources. J Intell Robot Syst 1988;1(1):17-33. doi: 10.1007/BF00437318.

39. Sentz K., Ferson S. Combination of evidence in Dempster-Shafer theory. Albuquerque, NM: Sandia National Laboratories, 2002 Apr. Report No.: SAND2002-083. doi:10.2172/800792.

40. Lund SP, Iyer H. Likelihood ratio as weight of forensic evidence: A closer look. J Res Nat Inst Stand Technol 2017;122:27. doi: 10.6028/jres.122.027.

Appendix A - Sensitivity of Score-Based Likelihood Ratios (SLRs)

The density plots in Figure 16 are shown to illustrate how the curve heights that are used to compute a score-based likelihood ratio (SLR) can change with different bandwidth choices (top panel of figure) or when a single additional score is added to the existing set of KCNM scores (bottom panel of figure). We chose to illustrate the effect of adding a value to the set of KCNM scores instead of the KM scores because the observed score of 0.5591 from the considered example is higher than any of the observed KCNM values and density values in distribution tails are particularly unstable when using kernel density estimation.

During kernel density estimation each observed value is replaced by a small bell curve (normal distribution) whose peak is located at the observed value. The curve in the bottom panel reaches its minimum at 0.5591 because adding a score of 0.5591 to the existing set of KCNM scores would cause the largest increase to the estimated KCNM density for a score of 0.5591. The curve is roughly symmetric around 0.5591 because the contribution of the added point to the estimated density at 0.5591 falls off as the small bell curve for the added point slides away to either side of 0.5591. Looking carefully at the middle set of density plots in the bottom panel, you can see the bell curve for the added value create a shoulder in the overall blue density at 0.5591. As the score added to the KCNM set increases, the small bell curve for the added score pulls away to the right, with its distribution eventually appearing almost entirely above the score 0.5591.

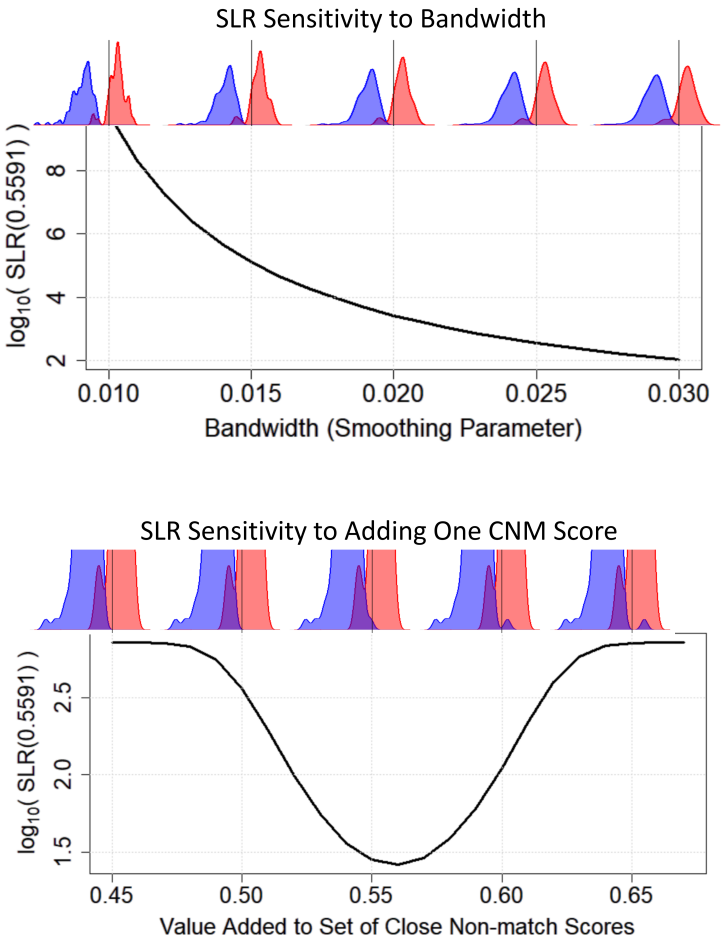

Figure 16: *(Top) Relationship between the bandwidth (i.e., smoothing parameter) applied to the known close non-match (KCNM) and known match (KM) scores during kernel density estimation and the resulting* $log_{10}(SLR)$ *value for the observed outsole pattern score of 0.5591. Mated and close non-match kernel density estimates are shown for bandwidth choices of 0.01, 0.015, 0.02, 0.025, and 0.03, respectively. The vertical lines in the density plots show the position of the observed score, 0.5591. (Bottom) Relationship between the value of a single score added to the set of KCNM scores and the resulting* $log_{10}(SLR)$ *value for the observed outsole pattern score of 0.5591, evaluated using a bandwidth of 0.023. Mated and close non-match kernel density estimates are shown for added values of 0.45, 0.5, 0.55, 0.6, and 0.65, respectively. These density plots are cropped to remove the top portions of the density curves to increase the visibility of the effect of the added value on the KCNM distribution. The vertical lines in the density plots show the position of the observed score, 0.5591.*

Appendix B - Reference Distributions for Final Scores

The reference distributions for final scores are based on the number of RACs being compared. We simulate such distributions using the reference distributions of outsole pattern and RAC scores. The simulated distributions are constructed as follows:

1. Assume we need to simulate a reference distribution of final scores that are from pairs with k RACs compared.

2. We have separate reference distributions of outsole pattern scores and RAC scores.

3. Suppose we want to simulate N values representing the final score distribution for KM comparisons involving k RACs.

4. Repeat (a), (b), (c) N times.

   (a) We randomly select an outsole pattern score from the reference distribution for KM pattern scores.

   (b) We randomly select k RAC scores from the reference distribution for KM RAC scores.

   (c) We find the $\log_{10}$(SLR)s for the k RAC scores in step (b) and add them to the pattern score from step (a). This gives a simulated final score for a KM comparison.

   Thus we generate the N final scores to represent the reference distribution for KM comparisons involving k RACs.

5. Suppose we want to simulate N values representing the final score distribution for KCNM comparisons involving k RACs.

6. Repeat (d), (e), (f) N times.

(d) We randomly select an outsole pattern score from the reference distribution for KCNM pattern scores.

(e) We randomly select k RAC scores from the reference distribution for KCNM RAC scores.

(f) We find the $\log_{10}$(SLR)s for the k RAC scores in step (e) and add them to the pattern score from step (d). This gives a simulated final score for a KCNM comparison.

Thus we generate the N final scores to represent the reference distribution for KCNM comparisons involving k RACs.

For pristine comparisons, we considered values of k equal to 0 through 20 since this range should cover all situations in our experiments. These distributions are shown in Figure 17.

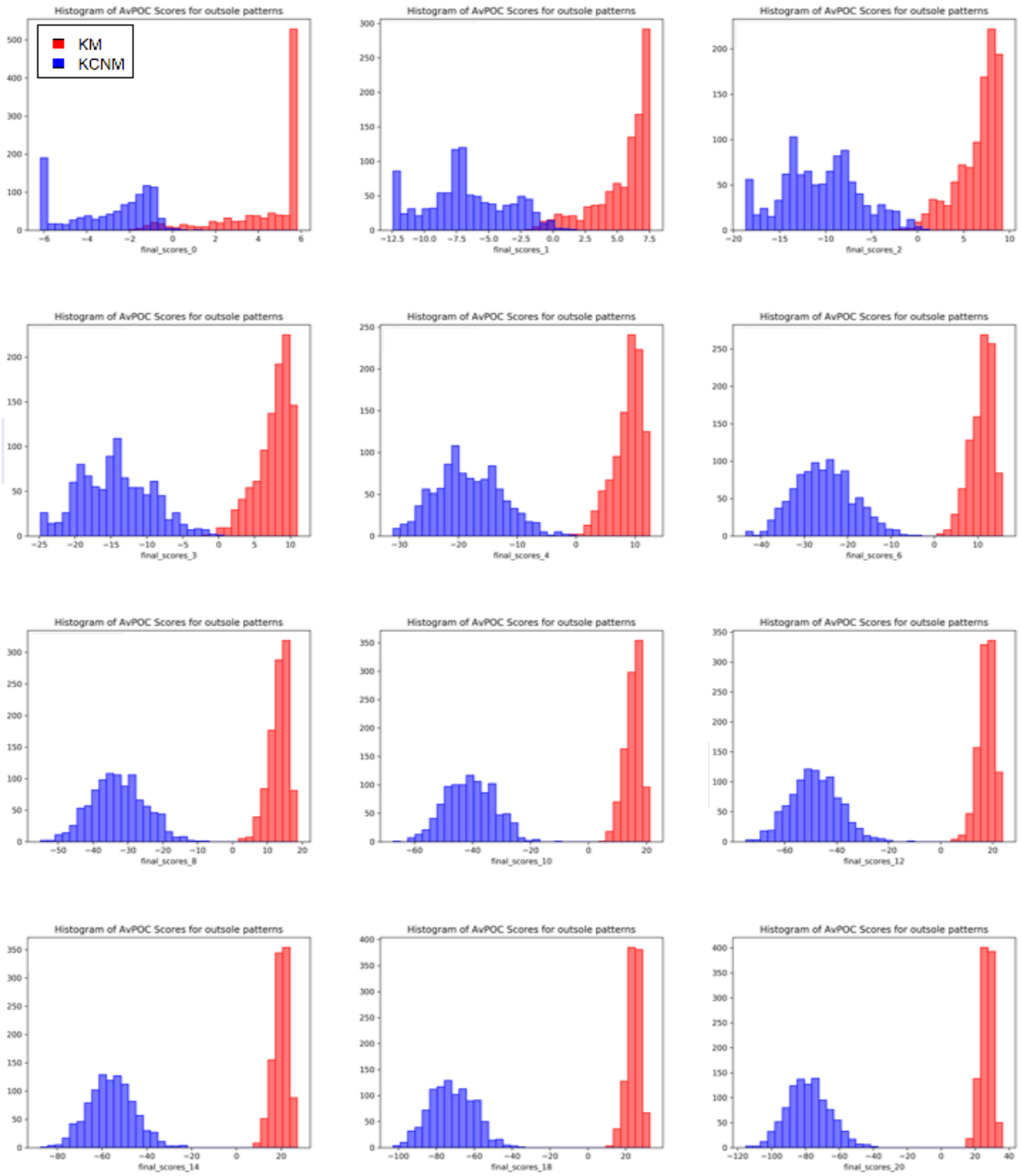

Figure 17: *Reference distributions for final scores of pristine comparisons based on the number of RACs marked up in K. Distributions for 0 to 20 RACs are precomputed. Here only 12 of those 21 distributions are displayed.*