Comparing Footwear Impressions that are Close Non-Matches Using Correlation-Based Approaches

Gautham Venkatasubramanian, MSc, (Former) Research Associate,
Vighnesh Hegde, MSc, (Former) Research Associate,
Sarala Padi, PhD, Research Associate, 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 thank Brian McVicker and Brian Eckenrode from the FBI for providing the boot data used in our study and for general guidance on the subject of forensic footwear impression comparisons. We thank Jacqueline Speir for making the West Virginia University mock crime-scene data publicly available and for discussions and feedback. We also thank Steve Lund, Weiqing Chen, Yooyoung Lee, Günay Doğan and Adam Pintar, current or former members of the NIST Forensic Footwear Research Team, for discussions and feedback. 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

Forensic activities related to footwear evidence may be broadly classified into the following two categories: (1) intelligence gathering and (2) evidential value assessment. Intelligence gathering includes identifying the make and model of a shoe impression by comparing its design elements with a database of outsole designs of known make and model, linking footwear impressions from one crime scene to those from a different crime scene, linking suspects to crime scenes, and other activities that provide leads for investigators. Assessment of evidential value, as practiced in the United States, involves a trained footwear examiner evaluating the degree of similarity between a known shoe of interest (together with its test impressions) and footwear impressions obtained from a crime scene, by performing side-by-side visual comparisons. However, the need for developing quantitative approaches for expressing similarities during such comparisons is being increasingly recognized by the forensic science community. In this paper, we explore the ability of similarity metrics to discriminate between impressions made by a shoe of interest and impressions made by close non-matching shoes. Close non-matching shoes largely share the same design and size. Therefore the ability to effectively discriminate between them requires considering, either explicitly or implicitly, not only design and size, but also wear patterns and, to some extent, individual characteristics. This type of discrimination is necessary for assessment of evidential value. The similarity metrics examined in this paper are correlation-based metrics, including Normalized Cross Correlation, Phase-Only Correlation, $AvNCC$ and $AvPOC$. The latter two metrics are based on features obtained from a convolutional neural network. Experiments are performed using Everspry impressions, FBI boot data that consist of impressions of boots worn by FBI trainees, and the West Virginia University footwear impression collection. The results show that Phase-Only Correlation performs as well as or better than the other metrics in all cases for the data sets we considered.

KEYWORDS

footwear impressions, footwear evidence, shoeprints, similarity metrics, correlation matching, deep learning, convolutional neural network

HIGHLIGHTS

- Quantitative approaches for expressing similarity during footwear impression comparisons.
- We consider evidence assessment as opposed to intelligence gathering.
- Effective discrimination between matches and close non-matches is needed for this purpose.
- Most other studies do not consider non-mates that are only close non-matches.
- Results indicate that Phase-Only Correlation performs as well as or better than the other metrics we considered.

Footwear marks are the most abundant form of evidence present at crime scenes. Yet, according to a U.S. Department of Justice publication (see Table-4 in [1]) less than 0.2% of requests for services received and completed in 2014 by publicly funded forensic crime labs in the U.S. were for footwear or tiremark impressions. Such under-use of footwear evidence is perhaps due to the increased focus on DNA and fingerprints as these often allow extremely strong associations to be made between crime scene evidence and a particular person of interest. Another contributing factor may be the views expressed by the U.S. National Academy of Sciences [2] and the President's Council of Advisors on Science and Technology [3] regarding scientific validity of pattern evidence disciplines. Forensic footwear examiners provide expert opinions and interpretations that arise from personal knowledge, training, and experience. However, the basis for examiner conclusions cannot be empirically demonstrated within the current framework. Recognizing this, the forensic science community is actively pursuing development of methods that can provide quantitative support for examiner conclusions.

Forensic activities related to footwear evidence may be broadly classified into two categories: (1) intelligence gathering and (2) evidential value assessment. Intelligence gathering involves activities such as identifying the make and model of a shoe impression by comparing its design elements with a database of outsole designs of known make and model, linking footwear impressions from one crime scene to those from a different crime scene, linking suspects to crime scenes, and various other activities that provide leads for investigators. Assessment of evidential value, as practiced currently in the United States, involves a trained footwear examiner evaluating the degree of similarity/dissimilarity between a known shoe of interest (together with its test impressions) and footwear impressions obtained from a crime scene, by performing side-by-side comparisons. Typically, four areas of comparison are considered for assessing similarities: outsole design, size, wear and individual characteristics (also called Randomly Acquired Characteristics (RACs)). Aided by his/her training and experience, an examiner makes an assessment of the likelihood of the observed degree of

similarity if the crime scene impression was made by the shoe of interest and its likelihood if some shoe other than the shoe of interest made the crime scene impression. The examiner chooses one of the levels from the SWGTREAD [4] (or a similar) conclusion scale to express his/her assessment in a report or during testimony.

Evidential value assessment requires (1) comparison of a questioned crime scene impression (Q) with one or more test impressions (T) produced from the shoe of interest in the current case and, (2) a large collection of ground-truth-known, mated and non-mated impressions that adequately represent scenarios likely to be encountered in case work. From this collection, comparisons would be chosen in which the crime scene impressions are similar (e.g., similar quality, complexity) to the impression in the current case. The collection of chosen comparisons (or "reference comparisons") would be used to place the current case similarity score in the context of previous observations of the system's behavior in these similar scenarios, both when comparing true mates and when comparing close non-matches, and facilitate weight of evidence assessments. It is worth noting that a higher similarity score, by itself, does not imply greater support for the shoe of interest being the source. It is the frequency of occurrence of the score in a casework comparison among reference mated comparisons relative to the frequency of occurrence of that score among reference close non-match comparisons that speak to the weight of evidence. Additional discussion of this approach may be found in ([5], [6]).

The need for developing quantitative approaches for expressing similarities/dissimilarities during such comparisons is being increasingly recognized by the forensic science community. A metric that can discriminate between arbitrary Qs and Ts is not as useful as one that can discriminate between close non-matches. Close non-match pairs largely share the same design and size. Therefore, to discriminate between such pairs, the discriminating information used by the metric would have to include, either explicitly or implicitly, wear characteristics and, possibly, individual characteristics, in addition to size and design information. (Note that we don't consider RAC comparison metrics explicitly in this paper; see ([5], [7], [6]) for

additional discussion of this topic.) This is important for evidence evaluation because it can be used to differentiate the higher levels of an interpretation scale that involve support for the proposition that the crime-scene impression was created by the shoe of interest. It is for this reason that, in this paper, we consider only non-mates that are close non-matches. In addition, it is close non-matches that are usually the most challenging for forensic examiners (and for similarity metrics as well), as other types of comparisons are generally easier for examiners to interpret.

This paper focuses on the performance of correlation-based similarity metrics, including Normalized Cross Correlation (NCC), Phase-Only Correlation (POC), $AvNCC$ and $AvPOC$. The latter two metrics are based on features obtained from a pretrained convolutional neural network. All of these metrics are tested for their power to discriminate between impressions made by a shoe of interest and other shoes that are considered to be close non-matches to the shoe of interest. For this paper, we define comparisons of close non-matches to be comparisons between impressions from two different shoes of the same make, model and size. To assist in this effort we performed experiments on three different data sets:

1. **Everspry Scanner Impression Data:** We acquired this data in our lab from a subset of the shoes in our possession using the Everspry EverOS scanner [8]. Five replicate impressions were made from each of five different pairs of shoes (10 shoes in total). We investigated the performance of the similarity metrics by focusing on how well the scores could discriminate comparisons between a questioned impression Q and a test impression T that are both from the same shoe versus comparisons between a Q and a T where T is from a shoe that is a close non-match of the shoe that produced Q.

2. **FBI Boot Data:** We acquired a collection of impressions of boots worn by FBI trainees for a period of several months. All the trainees performed more or less the same set of activities during their training period. All the boots were of the same make and model though there are several different men's and women's sizes in the collection.

Two replicate impressions were made from each of 36 different pairs of boots (72 boots in total). As with the Everspry data, we investigated the performance of the similarity metrics by focusing on how well the scores could discriminate comparisons between a Q and a T that are both from the same boot versus comparisons between a Q and a T where T is from a boot that is a close non-match of the boot that produced Q.

3. **West Virginia University (WVU) Footwear Impression Data:** We selected pairs of impressions Q and T from the West Virginia University footwear impression data [9] consisting of true matches and close non-matches where the Q impressions are mock crime-scene impressions. We used impressions from 18 different pairs of shoes (36 shoes in total). Unlike the FBI boot data, which consist of a single make and model outsole design, the subset of impressions from the WVU dataset (and also the Everspry data set) chosen for our experiments include several different outsole designs.

The questioned impressions used in the first two experiments are of much higher quality than what might be encountered in actual casework. The rationale for conducting these experiments is that they serve as *controls* in the sense that similarity metrics that are unable to discriminate between the shoe that is the source of Q from other shoes of the same make and model are not worthy of further consideration. As it turns out, the considered similarity metrics do seem to have good discriminating power in situations with high quality Q's. The third experiment, using the WVU footwear impression data, assesses the discriminating performance of these metrics in situations where the Q's are more representative of crime scene conditions.

The next section gives a brief overview of some recent research in developing similarity measures for footwear impression comparisons.

Recent Research Related to Similarity Measures for Footwear Impression Comparisons

There has been increasing interest in developing quantitative methods for footwear impression comparisons as evidenced by several recent publications. Richetelli et al. [9] investigated the classification performance of methods based on (1) the Fourier-Mellin transform (FMT), (2) the Phase-Only Correlation (POC), and (3) local interest points using the scale-invariant feature transform (SIFT) and compared using the random sample consensus (RANSAC) approach. They found that POC outperforms the other methods regardless of image type or image quality. They also point out the need to explore deep learning methods to solve the classification problem.

Kortylewski et al. [10, 11] developed similarity metrics based on generative models using Gabor filters and compared this method to those discussed in earlier publications. Kong et al. [12] investigated the performance of similarity metrics based on the idea of transfer learning from pretrained deep neural networks. They report that their methods outperformed earlier methods for the database retrieval problem.

Park et al. [13] developed new similarity scores based on feature engineering and machine learning using random forests and tested their performance using high quality impressions. They used maximum cliques applied to points obtained by the method of speeded-up robust features (SURF). The metrics they examined include, along with the random forest-SURF approach, percent overlap of SURF points after alignment, and Phase-Only Correlation (POC). Cui et al. [14] used 'deep belief networks' and 'spatial pyramid matching' to develop a local-to-global feature matching score for the database retrieval problem.

All of these articles develop scoring systems that, in theory, can either be used for database retrieval or for computing score-based likelihood ratios (SLRs) for weight of evidence assessments. Note that scoring systems that work well for make and model discrimination (database retrieval) may not necessarily work well for comparing Q and T for

evidence evaluation. Evidential value assessment is a more challenging problem than make and model identification since the metric has to incorporate, when possible, features specific to a given impression due to wear and usage that are unlikely to be present in impressions made by other shoes of the same make, model and size. Park et al. [13] is one of the few studies that specifically considered this by performing experiments to discriminate between matches and close non-matches. Their study used impressions obtained from two different make/models each at two different sizes. On the other hand, the WVU footwear impression data allowed us to consider a greater variety of make/model shoes. Furthermore, in our view, the degraded impressions considered in Park et al. [13] are not as representative of crime-scene-like conditions as are the mock crime-scene impressions available from the WVU footwear impression data. Therefore, we believe that results from our study add to the existing body of knowledge about the ability of similarity scores to discriminate between the real source shoe and other shoes of the same make and model.

We also believe that the final word is not out yet as to which of the similarity metrics should be recommended for use in casework. There are many promising similarity metrics and a separate, large scale study using a common data base of ground truth known, case work like, impressions needs to be conducted to fully understand the discriminating ability of competing metrics. Therefore, the focus of this paper is to demonstrate the feasibility of using quantitative methods for evidential value assessment and not to recommend a particular similarity metric for casework considerations. Undoubtedly, further research will lead to improved methods that are suitable for casework.

Materials and Methods

This section describes the data as well as the similarity metrics we used in our experiments.

Everspry Impression Data Details

From a collection of shoes in our possession, we chose 5 pairs of shoes with different

levels of wear (10 shoes in total) and generated 5 impressions per shoe using the Everspry EverOS scanner [8], which generates impression images at 300 PPI. This resulted in 50 total impressions. To obtain a set of mated pairs of impressions, we selected 5 of the 10 shoes and, for each of the 5 shoes, considered all possible pairs from the 5 impressions for the shoe, resulting in 10 combinations, or 10 comparisons, for each shoe. That gave us a total of 50 mated comparisons.

To obtain a set of close non-match pairs of impressions, we considered pairs of impressions where one impression is from the left shoe and the other impression is from the corresponding right shoe of the pair. Each left shoe has 5 impressions and each right shoe has five impressions. Considering all possible combinations of these impressions results in 25 comparisons. For each comparison, one of the shoe impressions is "flipped" on the vertical axis to create a close-non-matching impression. This results in 25 close non-match pairs for each pair of shoes. We used 5 pairs of shoes, resulting in 125 pairs of close non-matches. Figure 1 shows an example of Everspry impressions of a mated pair and a close non-match pair.

Many of the Everspry shoe impressions contain manufactured artifacts such as lettering or the Nike symbol. When such an impression is flipped, the appearance of these artifacts may differ from those in the original unflipped versions; for example lettering will appear as in a mirror image or the Nike logo will be reversed. These artifacts therefore provide image features that can help in discriminating an impression from its flipped close non-match, which is undesirable. We want to make certain that such artifacts in the images are not used in this way. So the artifacts are erased in all impressions before computation of the similarity metrics for computing scores of both mates and non-mates (see Figure 2).

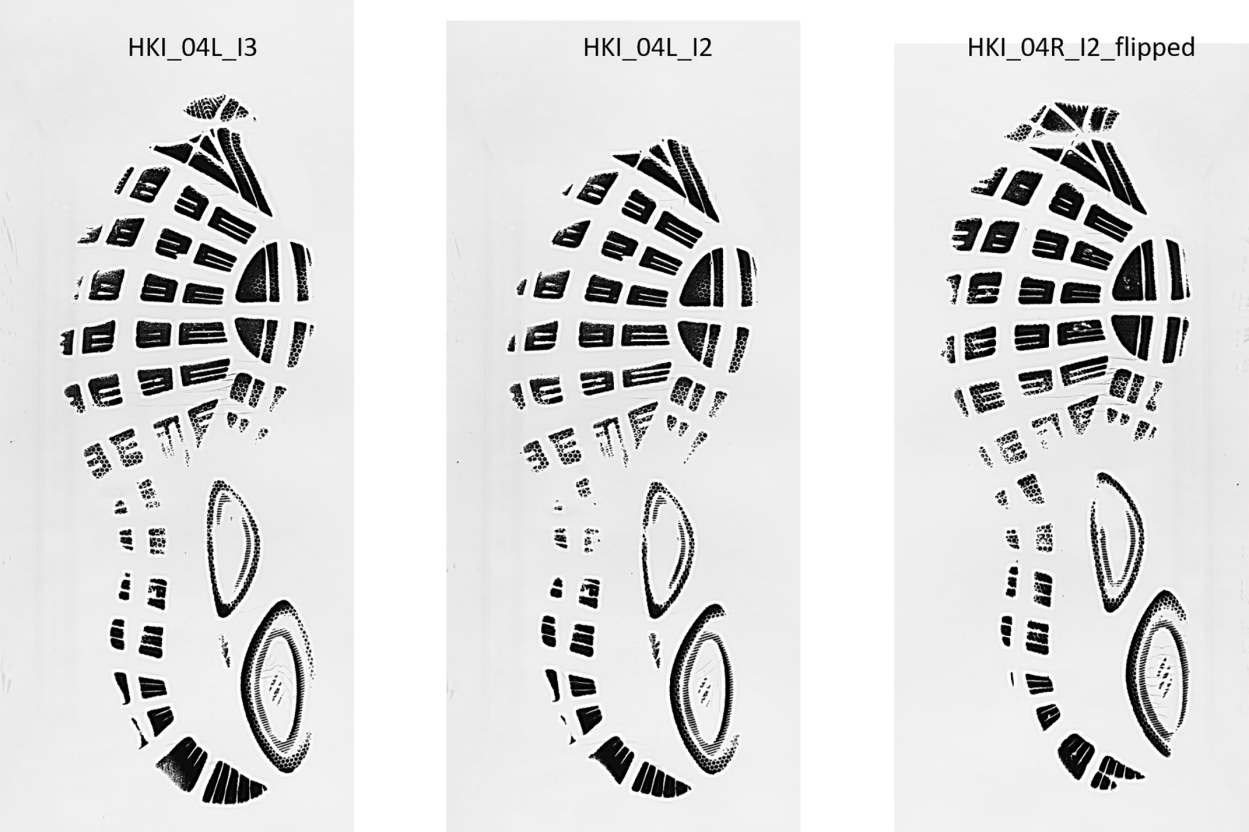

Figure 1: Comparisons using Everspry impressions: The impression on the left (from the left shoe) is compared with the mated impression in the center (a different impression from the left shoe) and the close non-match impression on the right (taken from the right shoe and flipped). These impressions are from the shoe pair HKI_04.

FBI Boot Data Details

We acquired a total of 36 pairs of boots from the FBI of which 18 pairs are from men and 18 pairs are from women. Several sizes are represented. For men, the boot sizes represented are

8.5–9.0–9.5–10.0–10.5–11.0–11.5–12.0–13.0–14.0

and for women the available sizes are

6.0–6.5–7.0–7.5–8.0–8.5–9.0–9.5–10.0–11.0

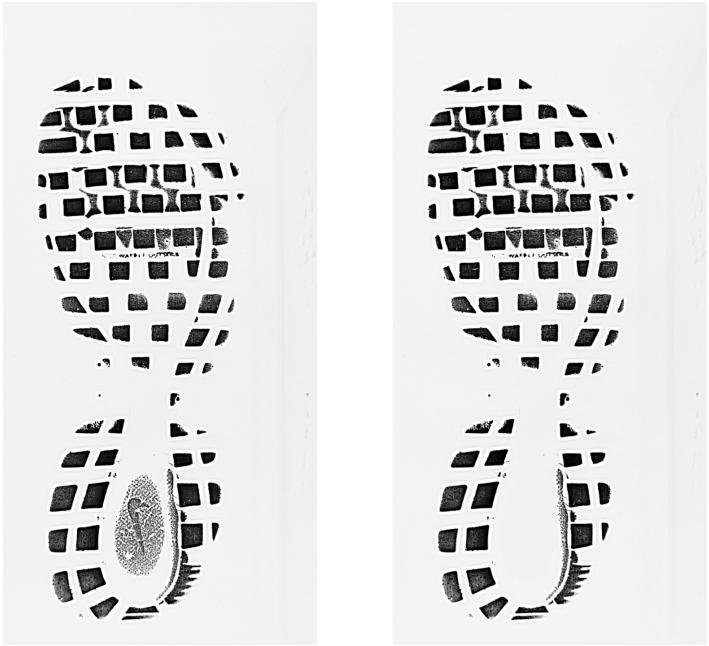

Figure 2: Eversprv images: The Nike logo on the bottom of the impression on the left is erased. The result is shown in the impression on the right.

Thus, there are a total of 72 boots (counting left and right boots separately). All the boots are of the same make and model. They were all used for 16 weeks under similar conditions. Test impressions were made from these boots using fingerprint powder. Two replicate test impressions were prepared from each boot in succession without re-applying fingerprint powder. Therefore, the first impression is normally darker than the second.

Each impression was obtained as follows:

1. Brush light coating of black fingerprint powder on outsole of boot.
2. Place boot on foot.
3. Place transparent adhesive sheet on floor.
4. Walk (with powdered boot) onto transparent adhesive sheet.
5. Remove adhesive sheet from boot.

6. Cover impressed adhesive sheet with acetate.

7. Repeat steps 3-6 to make a second test impression from same boot (without re-powdering).

Following this, each test impression was scanned at 600 ppi in gray scale. The image files were named as follows:

`Shoenumber_(M/W)_Size_(R/L)_(A/B)_T`

where:

M,W - Men/Women

R,L - Right,Left

T - Test impression

A, B - Two replicate test impressions for the same shoe where

A = replicate 1 and B = replicate 2

Table 1 summarizes our inventory of boot impressions.

Men's boots:	10 different sizes 72 men's impressions in total - 36 are left and 36 are right Out of 36, 18 are of type "A" and 18 are of type "B"
Women's boots:	10 different sizes 72 women's impressions in total - 36 are left and 36 are right Out of 36, 18 are of type "A" and 18 are of type "B"

Table 1: FBI boot impressions used in our experiments

The top of Figure 3 shows the FBI boot impressions obtained from boot pair #20. The reader may note that these boot impression images have a label in the region separating the toe region from the heel region that identifies the source of the impression, whether the impression is from a left boot or a right boot, and whether the impression is the first one or the second one after coating the outsole with fingerprint powder. There is also a black

rectangle that appears below the heel region of the impression that was placed there to hide some information and anonymize the data. The label and the rectangular region provide image features that can help in discriminating one boot impression from another, which is undesirable. We want to make certain that artificial artifacts in the images are not used in this way. So we use only the toe region from each boot to illustrate our methods. See bottom of Figure 3.

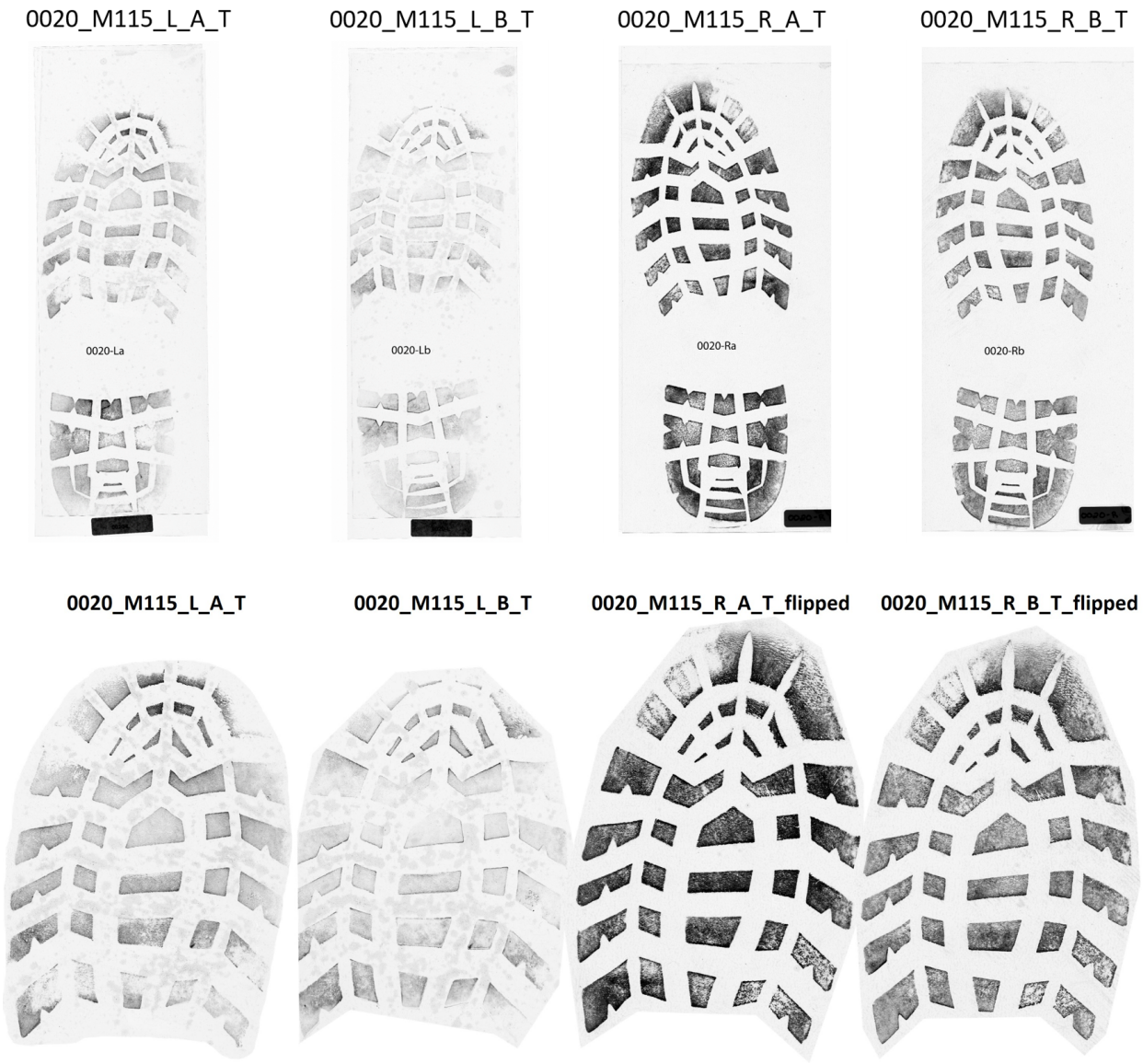

Figure 3: (Top) Two replicates of the left boot (`0020_M115_L_A_T` and `0020_M115_L_B_T`) and two replicates of the right boot (`0020_M115_R_A_T` and `0020_M115_R_B_T`) from boot pair #20. (Bottom) Boot impression `0020_M115_L_A_T` is compared with `0020_M115_L_B_T` (true match) and two close non-matches `0020_M115_R_A_T_flipped` and `0020_M115_R_B_T_flipped` from boot pair #20. Only toe regions are used.

Since we are interested in performance of our metrics for discriminating between actual matches and close non-matches, we use the "flipped" version of an impression to create a

close-non-matching impression (as is done with the Everpry impressions discussed above). For instance, consider the impression `0020_M115_L_A_T` (see bottom of Figure 3). The actual match for this impression is `0020_M115_L_B_T` whereas the close non-matches for this impression are `0020_M115_R_A_T_flipped` and `0020_M115_R_B_T_flipped`. Thus, each questioned impression is compared to three reference impressions of which exactly one is a match and the other two are close non-matches.

Ideally, the similarity score when comparing mated pairs would be at rank 1 position among the three scores. In reality, it could also be at rank 2 position or rank 3 position. We tally the number of times the score for a mated pair ends up in rank 1 position for each of the 144 impressions. Results are given in a later section.

WVU Footwear Impression Data Details

We obtained a public "crime scene" database created by researchers at West Virginia University [9] consisting of 18 pairs of shoes – 9 pairs used for making mock crime scene impressions with dust and 9 pairs used for impressions with blood. Four different substrates were used for each shoe in the case of dust (acetate, ceramic, paper, and vinyl) whereas 3 different substrates were used for each shoe in the case of blood (acetate, ceramic, and vinyl). For blood, there were two versions made, one with enhancement using leuco-crystal violet (LCV) and one without any enhancement. Thus, 72 impressions were made with dust (9 left shoes, 9 right shoes, each with 4 different substrates: $(9+9)\times 4 = 72$; 36 from left shoes and 36 from corresponding right shoes), 54 impressions with blood enhanced with LCV (9 left shoes, 9 right shoes, each with 3 different substrates: $(9+9)\times 3 = 54$; 27 from left shoes and 27 from right shoes), and 54 impressions with blood without any enhancement $(9+9)\times 3 = 54$; 27 from left shoes and 27 from right shoes). However, 6 of the 72 mock crime-scene impressions for dust (4 left shoe impressions and 2 right shoe impressions) were unusable resulting in 32 left-shoe impressions (3 impressions on ceramic and one on paper were unusable) and 34 right-shoe impressions (2 right-shoe impressions on ceramic were

unusable). Also, 2 of the mock crime-scene impressions for blood (shoe pair #045, right shoe on acetate, un-enhanced as well as enhanced) were unusable, resulting in a total of 53 mock crime-scene un-enhanced impressions and 53 enhanced impressions (54 left-shoe impressions and 52 right-shoe impressions). The reader is encouraged to refer to Richetelli et al. [9] for further details regarding the creation of these mock crime-scene impressions.

There are also exemplar test impressions in the database that consist of high quality impressions made from the 18 left shoes and from the 18 right shoes used to create the mock crime-scene impressions. We used these test impressions and the mock crime-scene impressions to evaluate the performance of our similarity metrics. The mated pair comparisons consisted of comparing the 18 left-shoe test impressions with the corresponding left-shoe mock crime-scene impressions. The nonmated comparisons consisted of comparing the test impressions with flipped versions of corresponding right-shoe impressions. In all, we have 172 comparisons – 86 mated comparisons and 86 comparisons of close non-matches. Table 2 summarizes the number of comparisons made in the different scenarios considered.

Substrate	Mated	Nonmated (Close Non-matches)
Blood (without LCV)	27	26
Blood (with LCV)	27	26
Dust	32	34
Total	86	86

Table 2: Counts for mated and nonmated comparisons for each substrate considered.

Figure 4 shows an example of a mated pair and a close non-match pair taken in dust. Additional example impressions similar to the ones used in our study are shown in Figure 2 in Richetelli et al. [9].

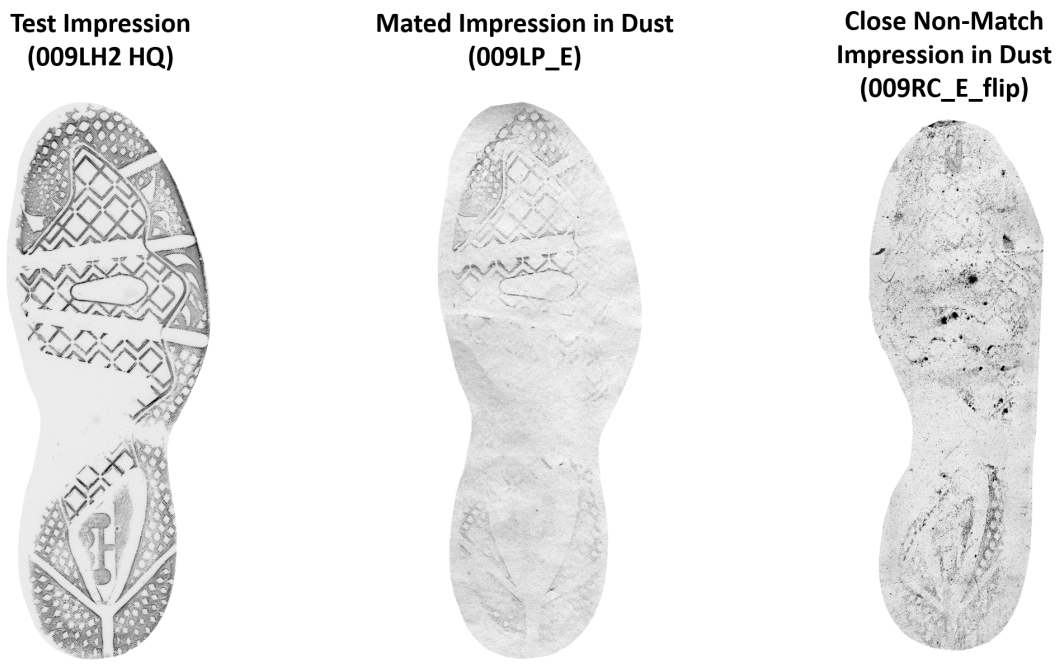

Figure 4: Comparisons using West Virginia University impressions taken in dust: The test impression on the left (from the left shoe) is compared with the mated dust impression on paper in the center (from the left shoe) and the close non-match dust impression on ceramic tile on the right (taken from the right shoe and flipped). These impressions are from the shoe pair #009.

Similarity Metrics

We started with two well-known correlation-based similarity metrics. These are Normalized Cross Correlation (NCC) and Phase-Only Correlation (POC). We also examined a deep learning approach proposed by Kong et al. [12].

Our approach requires that Q and T be aligned before computing the similarity score. Our alignment algorithm, discussed in greater detail in the next section, is therefore applied to each pair of impressions being compared. For Everspry and FBI boot impressions, we use automatic corner detection on each image followed by a registration process based on finding large cliques in a graph theoretic representation of pairs of corner points with one corner

point from one image and the other from the second image. We refer to this approach as the "Clique Matching" algorithm.

For WVU impressions, the informational content of many of the mock crime-scene impressions is so low that the corner detection algorithm finds too many non-corner spurious points and not enough actual corner points. We therefore manually select about a dozen corresponding "anchor points" in the two impressions, distributed around the impression. Ideally, there would be at least 2 points at the top of an impression, 2 at the bottom, 2 on the left side, and 2 on the right side. Zooming into the image is used to better localize each point selected. The user should have high confidence that corresponding points derive from a single point on the shoe. Since all impression pairs are either mates or close non-matches (implying they largely share the same outsole design), then ideally points selected would lie on corners or other unique features of design elements in the outsole. After the anchor points are chosen, a rigid transformation (rotation and translation) between the two sets of points is computed using the Kabsch algorithm [15]. A somewhat similar method for alignment is used in Richetelli et al. [9]. Figure 5 shows this method of alignment applied to a close non-match pair in the WVU data.

Once the two images of an impression pair are aligned, we apply NCC and POC as follows. Let x_{ij} and y_{ij} denote the values in pixel position (i, j) in Q and T, respectively. The following formula gives the definition of NCC when comparing Q and T.

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

What is referred to as the Normalized Cross Correlation in the field of computer vision ([12, 16]) is nothing more than the Pearson Correlation in statistics as one can recognize from Equation (1).

A variation of this approach uses a correlation value calculated from the phase component of the 2D Fourier Transforms of Q and T. This concept is discussed in Richetelli et al. [9]

and many earlier authors as well. Following notation similar to that in Richetelli et al. [9], suppose $G_1(u,v)$ and $G_2(u,v)$ denote the 2-dimensional Fourier transforms of the pixel intensity values in Q and T, respectively. These can be expressed using the amplitude components $A(u,v), B(u,v)$ and phase components $\sigma(u,v), \theta(u,v)$ as

$$G_1(u,v) = \mathcal{F}\left(A(u,v)e^{\boldsymbol{j}\sigma(u,v)}\right) \tag{2}$$

$$G_2(u,v) = \mathcal{F}\left(B(u,v)e^{\boldsymbol{j}\theta(u,v)}\right) \tag{3}$$

where $\mathcal{F}$ is the 2-dimensional Fourier-transform operator and $\boldsymbol{j} = \sqrt{-1}$. Then the Phase-Only Correlation (POC) between Q and T is defined as

$$POC(Q,T) = \mathcal{F}^{-1}\left(\frac{G_1(u,v)\,G_2(u,v)}{|G_1(u,v)\,G_2^*(u,v)|}\right) \tag{4}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier-transform operator, and $|z|$ and z^* denote the magnitude (norm) and the complex-conjugate of the complex number z.

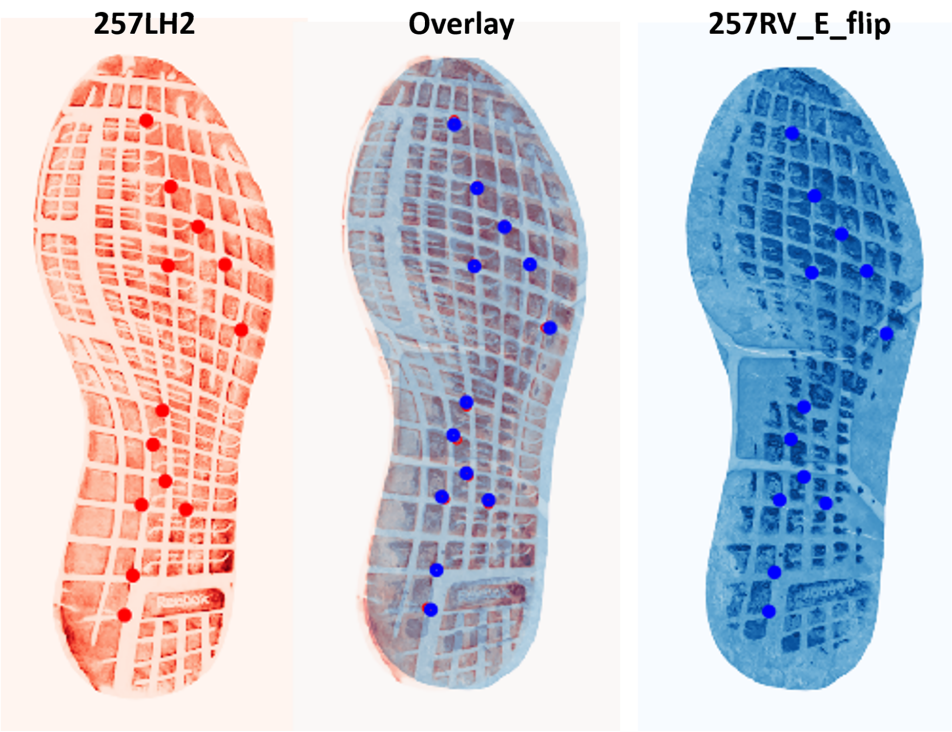

Figure 5: Example of alignment using West Virginia University impressions; the crime scene impression is taken in blood enhanced with LCV: The test impression in the left image (from the left shoe) is aligned with the close non-match blood+LCV impression on vinyl tile in the right image (from the right shoe and flipped). Thirteen manually selected corresponding anchor points are displayed in the two images. The overlay after alignment, along with the anchor points, is shown in the center image. These impressions are from the shoe pair #257.

In addition to computing NCC and POC, we can also apply the pre-trained ResNet-50 convolutional neural net model (layer ResNet-2bx, as done by Kong et al. [12]) for which the network weights are publicly available. This results in 256 "feature maps" for each image. Each feature map is a gray-scale array such that high values in the array correspond to the presence of a certain feature in the original image at a particular location. Each feature map serves as a feature extractor for a certain type of image feature. Such features may not necessarily be apparent to a human but are *seen* by the algorithm. More details regarding filtering and feature extraction in images can be found in [17].

In this approach, the input to the neural net is two images, Q and T. Each image is processed by the first few layers of the (same) Resnet-50 model (up to layer ResNet-2bx). The output of this processing is two sets of 256 feature maps, one set for each input image. Two separate similarity metrics are then computed for these two feature-map sets, $AvNCC$ and $AvPOC$. These metrics are described in more detail below. A detailed overview of ResNet is available in [18].

The filters needed to produce the feature maps were previously learned by training ResNet on the ImageNet database of 1.2 million images of all kinds (animals, birds, trees, pencils, etc., see [19]). The rationale is that filters used in the early layers of a network are tuned to extract "general features" (edges, corners, other atomic shapes) and therefore the same filters have the potential to extract informative, discriminating features from any image class (footwear impressions, fingerprints, etc). However, we emphasize that *we did not do any additional training in our study.* We simply use the pre-trained network with all its weights and apply it to our data for feature extraction. We analyzed the 256 individual similarity metrics computed from corresponding pairs of feature maps (also often called *channels)* from each impression, and determined that most or all of the discriminating power came from about 15 to 20 channels out of the 256 available channels. We could have used this subset of channels to define our similarity metric but we did not do so because we did not have enough data to use a part of it for *training* and a part of it for *testing* and *validation*. It is for this reason that we used a simple average of the scores from all of the 256 channels. When a much larger set of ground truth known pairs of impressions becomes available (either created by us or by other researchers) we will be able to fine-tune the deep learning based metrics which may perhaps result in more powerful discrimination.

If the two images being compared come from the same source then the feature map i ($i = 1, 2, \ldots, 256$) for Q may be expected to have a high correlation with the corresponding feature map i for T, whereas the correlation is not expected to be high if T and Q do not come from the same source shoe. This reasoning applies to correlations found for each

corresponding pair of feature maps. Let NCC_k represent the value of NCC when comparing the k^{th} feature map from Q to the k^{th} feature map from T. Also let $x_{ij}^{(k)}$ and $y_{ij}^{(k)}$ denote the values in pixel position (i, j) for the k^{th} feature map from Q and T, respectively. The following formula gives the definition of NCC_k when comparing Q and T.

$$NCC_k(Q,T) = \frac{\sum_{i,j}\left(x_{ij}^{(k)} - \overline{x}^{(k)}\right)\left(y_{ij}^{(k)} - \overline{y}^{(k)}\right)}{\sqrt{\sum_{i,j}\left(x_{ij}^{(k)} - \overline{x}^{(k)}\right)^2}\sqrt{\sum_{i,j}\left(y_{ij}^{(k)} - \overline{y}^{(k)}\right)^2}}. \tag{5}$$

This process results in 256 normalized cross correlation values for a given pair (Q,T) corresponding to $k = 1, 2, \ldots, 256$. We simply take the average of these 256 correlations and use the result as a similarity score. We use the notation $AvNCC$ to denote this score. Thus

$$AvNCC(Q,T) = \frac{1}{256}\sum_{k=1}^{256} NCC_k(Q,T). \tag{6}$$

In a similar manner, we compute the POC value (based on Equation (4)) using the k^{th} corresponding feature maps from Q and T for $k = 1, 2, \ldots, 256$, and denote the resulting value by POC_k. We take the average of these 256 phase-only correlations and use the result as a similarity score. We use the notation $AvPOC$ to denote this score. Thus

$$AvPOC(Q,T) = \frac{1}{256}\sum_{k=1}^{256} POC_k(Q,T). \tag{7}$$

Kong et. al. [12] introduced a similarity score that they refer to as multi-channel normalized cross correlation ($MCNCC$) which is based on a generalization of the concept of Pearson Correlation Coefficient. However, they did not consider phase-only correlations. It is readily seen that the $AvNCC$ score is a special case of $MCNCC$ as defined in [12] when the feature map elements are uncorrelated with each other.

As mentioned earlier, for each Everspry impression from a particular shoe, we regard it as the questioned impression Q and compare it with other impressions from the same shoe that serve as the test impressions T for true matches; for close non-matches, we use

each of the five *flipped* impressions from the corresponding opposite shoe of the pair as T. For each FBI boot impression, we regard it as the questioned impression Q and compare it with one true match (replicate impression) and two close non-matches (flipped versions of the boot for the other foot). For each WVU test impression T, we compare it with a set of mock crime-scene impressions, either true matches or close non-matches. The discrimination performances of the similarity scores are investigated by the use of appropriate receiver operating characteristic (ROC) plots and also by comparing the distribution (densities) of the mated and non-mated scores. Results of our analyses are discussed in a later section.

The Problem of Alignment

Given a crime scene impression Q and a test impression T, one first faces the task of aligning or registering the two impressions so that a meaningful assessment of correspondence can be made. The meaning of alignment is clear when Q and T are both from shoes of the same make, model and size. In such a case one expects to see an almost exact correspondence between the contact regions in the two impressions and also the noncontact regions in the two impressions. If the two impressions are well aligned then we may even expect a more or less exact correspondence between contact regions in the two impressions at the pixel level. The same is expected to hold for pixels in the noncontact regions.

The meaning of alignment is often unclear when comparing impressions from shoes of different make and model or shoes from the same make and model but different sizes. In these cases it is common practice to rotate each impression to a standard vertical orientation such that the longest line segment connecting two boundary points of the shoe impression, typically the tip of the toe and a point in the bottom of the heel, is vertically oriented. For the database retrieval problem it is generally adequate to rotate each impression to a standard vertical orientation since we expect to compare Q to a library of shoe impressions of mostly different makes and models. However, when comparing Q to a close non-match T

– for instance, (a) a flipped version of a test impression from the opposite shoe, or (b) an impression made by another shoe of the same make, model, and size of the same foot – a more accurate alignment is desired. This is a key component of our proposed approach. Many of the articles mentioned above avoid this careful alignment step by resorting to the standard vertical orientation approach since their focus is on database retrieval and not assessment of evidential value. To the best of our knowledge they do not consider comparisons of Q to close non-matches ([13] is an exception).

Image Alignment as a Point Pattern Problem

By alignment we mean that we find a transformation (rigid or non-rigid) to map each pixel from the crime scene image to a corresponding pixel in the test impression image. If we consider an outsole impression image as a large set of points in contact with a surface, the problem of aligning two images reduces to finding a transformation that maps a subset of the first point set (corresponding to Q) to a subset of the second point set (corresponding to T) in such a way that the spatial relationships between point pairs in the first point set are in agreement with the spatial relationships between corresponding pairs in the second point set. Once we find such a transformation, it can be applied to every pixel of Q to find the corresponding pixel in T. However, it is not computationally feasible to consider each pixel of the shoeprint image as a contact point. So we either have to consider a sample of contact points or choose those contact points that are expected to simplify the process of finding the correct alignment (often referred to as *interest points*). We pursue this latter approach.

We first identify a collection of points on the impression that are corner points, that is, they are points that lie on the corners of design features. While these points are a (small) subset of all contact points, they often provide an effective way of identifying points of correspondence, especially when the two impressions have the same source or come from sources with the same outsole design. Fortunately, many corner point detection algorithms have been developed in the field of computer vision and for the Everspry and FBI boot

impressions, we use a method based on one such algorithm. For reasons stated earlier, we manually select the corner points for the WVU mock crime-scene impressions.

Corner Point Detection

We found that the readily available corner detection algorithms like Harris Corner Detector [20], Shi-Tomasi Corner Detector [21], and FAST [22] do not provide the required amount of accuracy (as measured by false positive and false negative rates) when used for finding corners in the shoeprint images. We chose to use the FAST algorithm but with a minor modification. To declare a pixel to be a corner (see Figure 6), the conventional FAST algorithm considers 16 border pixels of a Bresenham circle [23] of radius 3 (in red) around the pixel. Our modification considers 8 additional pixels (in orange) in the neighborhood based on some experiments we conducted to improve the corner detection performance. All of these pixels are part of the neighborhood nbd around the pixel (in black).

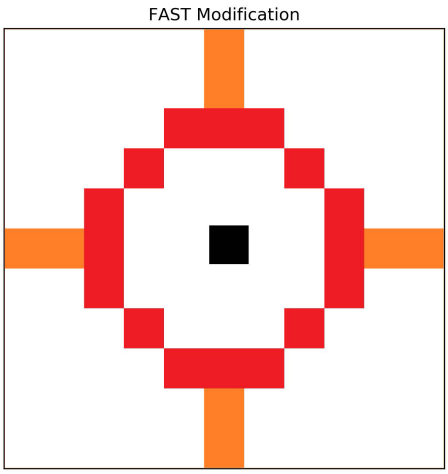

Figure 6: FAST pixels (red) and additional pixels (in orange) help find better corners.

Our modification computes the corner response $CR_{i,j}$ for a given pixel $I_{i,j}$ as follows:

- We pad the input image so that each pixel has a valid neighborhood.
- For each pixel $I_{i,j}$, we compute the relative intensities of the pixels in its neighborhood $nbd(I_{i,j})$ by subtracting the value $I_{i,j}$ from all the pixels in $nbd(I_{i,j})$.

- If a connected subset S of $nbd(I_{i,j})$
  - has more than half the points in the neighborhood (i.e., $|S| \geq 0.5 \times |nbd(I_{i,j})|$)
  - has all positive or all negative values for relative intensity (that is, all brighter pixels or all darker pixels than pixel $I_{i,j}$),
  - has relative intensity values exceeding a given threshold t in absolute value,

  then, the corner response $CR_{i,j}$ is the absolute value of the average of the relative intensities of the pixels in S.

- Otherwise, $CR_{i,j}$ is zero.

Further, we apply an averaging filter in order to reduce spurious maxima in the corner response image. This was consistently seen to result in improved corner detection performance. Figure 7 shows an example of the results obtained by the modified FAST and how it compares with three other corner detection methods. In our limited set of experiments we found that the modified FAST finds more correct corners and fewer false corners. While this does not prove that the modified FAST method is superior, it serves to explain why we chose the modified FAST for our study.

Alignment through Clique Matching

Once points of interest (in our work we used specifically corner points but they can be any collection of interest points) have been obtained from the shoeprint images, we consider the problem of aligning the two given point patterns. This problem, and graph-matching and maximum-clique based approaches to solve it, have a long history [24]. Recently, Park et al. [13] have used the maximum-clique approach for footwear impression alignment.

Let S1 be a shoeprint with M points of interest, and let S2 be a shoeprint with N points of interest. We write

$$S_1 = \{p_1, p_2, p_3, \ldots, p_M\}$$

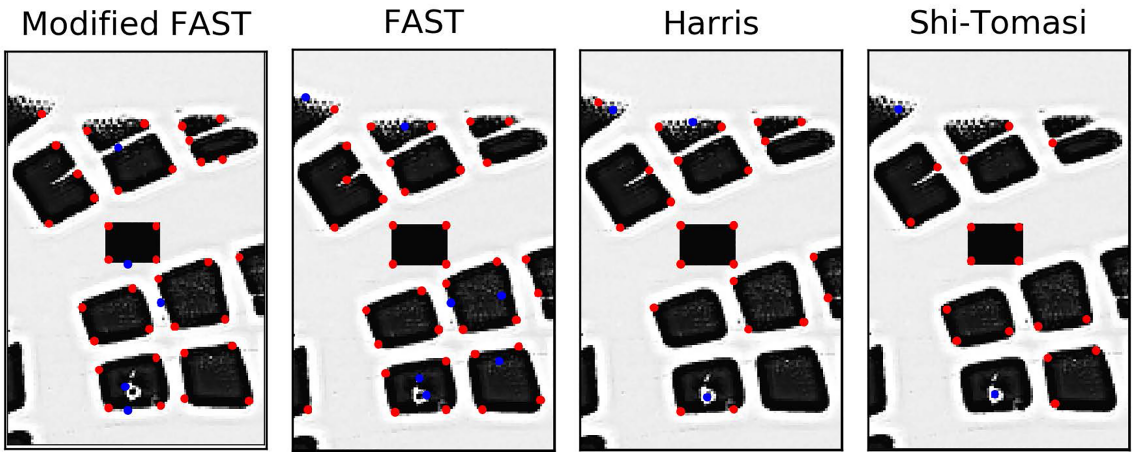

Figure 7: An illustration of results from applying different corner detectors. The modification to FAST resulted in more true corners and fewer false corners in our limited set of experiments. Points marked in blue are detected as *corner points* by the algorithm but appear to be false corners when examined visually. Points marked in red are detected by the algorithm and appear to be true corners.

and

$$S_2 = \{q_1, q_2, q_3, \ldots, q_N\}$$

where $p_i = (x_i, y_i),\ i = 1, 2, \ldots, M$ and $q_j = (u_j, v_j),\ j = 1, 2, \ldots, N$ are the interest points from S1 and from S2, respectively. For ease of notation, we use S1 to denote both the shoeprint and the set of its interest points. We now seek a transformation F that maps the maximum number of points from S1 to S2 while preserving spatial distances between corresponding pairs. We call these *corresponding points*. We restrict ourselves to rigid transformations (rotation and translation). However, allowing a small amount of distortion may lead to improved registration.

Finding the largest collection of corresponding points in S1 and S2 is well known to be equivalent to the problem of finding a maximum clique in a suitably constructed graph consisting of vertices and edges defined as follows (see [25]).

- Each pair of points $v_{ij} = (p_i, q_j)$ with $p_i \in S1$ and $q_j \in S2$ is a vertex of the graph G.

Since S1 has M points and S2 has N points, the graph G has $M \times N$ vertices.

- Suppose $v_{ij} = (p_i, q_j)$ and $v_{rs} = (p_r, q_s)$ are vertices of G such that $p_i \neq p_r$, $q_j \neq q_s$, then we draw an edge connecting v_{ij} and v_{rs} if and only if the Euclidean distance between p_i and p_r, denoted by $d(p_i, p_r)$, is nearly equal to the Euclidean distance between q_j and q_s, denoted by $d(q_j, q_s)$. That is,

$$|d(p_i, p_r) - d(q_j, q_s)| \leq \epsilon$$

for some sufficiently small $\epsilon > 0$.

Once such a graph G has been constructed, we seek the largest subset $\mathcal{C}$ of vertices in G such that every pair of vertices in $\mathcal{C}$ is connected by an edge. Such a subset of G forms a subgraph of G called a maximum *clique* of G. Each vertex $v_{ij} = (p_i, q_j)$ that is part of the clique implies that the point p_i of $S1$ and the point q_j of S2 correspond via a rigid transformation. Ideally, this same rigid transformation applied to the entire impression Q is expected to align it with the impression T. In practice, this approach has proved to be effective as long as the amount of distortion that occurred when the shoe produced the impression is relatively small. In such cases we can use the Kabsch algorithm [15] to obtain the required rigid transformation F once a set of corresponding points of Q and T have been identified. An example of maximum-clique-based alignment with Evespry impressions is shown in Figure 8.

It is well known that the maximum clique problem is in the NP-Hard class of problems; finding the solution to such problems takes exponential time complexity in the worst case. However, since our problem requires the discovery of just a single maximum clique, the computations can be sped up considerably in practice. In fact, all we really need is a large clique; it does not need to be an actual maximum clique. A large clique will find a large number of corresponding points in the two impressions which will facilitate alignment. The larger the clique the better we expect the alignment to be. Additionally, we can ensure

that the graph G is sufficiently sparse by using an appropriate value of ϵ while constructing the edges of G. This also helps reduce the computation time. We developed software that implements such speed-up techniques for use in our study. The code is written in Python and is available as a package ([26],[27]).

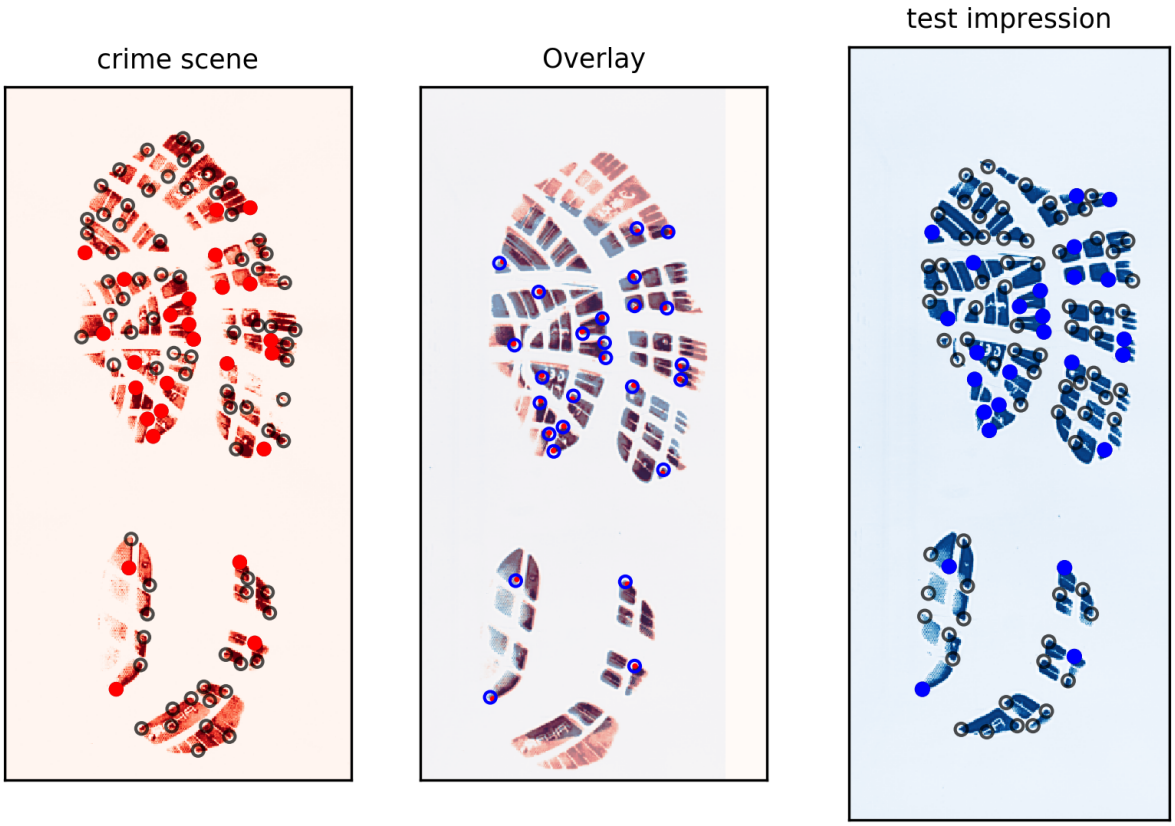

Figure 8: Maximum-clique-based alignment with Everspry impressions: The right-most impression is the test impression T. Corner points automatically found are marked by unfilled and filled blue circles. The left-most impression is the crime scene impression Q after applying the clique-based alignment algorithm. Corner points automatically found are marked by unfilled or filled red circles. The middle image is a superposition of the Q and T impressions. The points in the maximum clique are marked by filled circles in the left and right impressions. The middle image shows only the corresponding points using unfilled blue circles for points from T and filled red circles for points from Q.

Clique Size as a Discriminator

It is reasonable to expect that two impressions would have more corresponding corner points when they are made by the same shoe than when they are made by different shoes,

even if they are close non-matches. With this in mind we examined the sizes of the maximum cliques found by the clique matching algorithm for mated pairs and pairs that are close non-matches. Our findings are shown in Figure 9.

It is clear that maximum cliques tend to be larger for mated comparisons than for close non-match comparisons. One reasonable explanation for this behavior is that the wear patterns for the left and right shoes are different and hence, when either shoe impression is flipped and used as a close non-match for the impression from the other shoe, the clique-match algorithm finds a greater number of corresponding points, and hence larger cliques, for mated pairs. This also results in better alignment for mated impressions than for nonmated impressions which, in turn, affects measures based on correlations. As we will see in the next section, the correlation-based metrics are indeed able to discriminate very well between mates and close non-matches.

Clique size analysis is not performed for WVU data because the alignments for WVU data are based on a small number of manual markup of 'anchor points' (about a dozen points) for both mated and nonmated comparisons. The clique-match algorithm is not used for WVU data.

Results

Everspry Data

As described above, for each Everpsy impression from a particular shoe, we compare it with other impressions from the same shoe for true matches and flipped impressions from the opposite shoe for close non-matches.

Figure 10 shows ROC plots for NCC, POC, $AvNCC$, and $AvPOC$. The plots are based on scores from comparisons of 50 known mated pairs and 125 known close-non-match pairs. These plots suggest that NCC and POC may be better discriminators between matches and close non-matches than $AvNCC$ and $AvPOC$, as indicated by the Area Under the

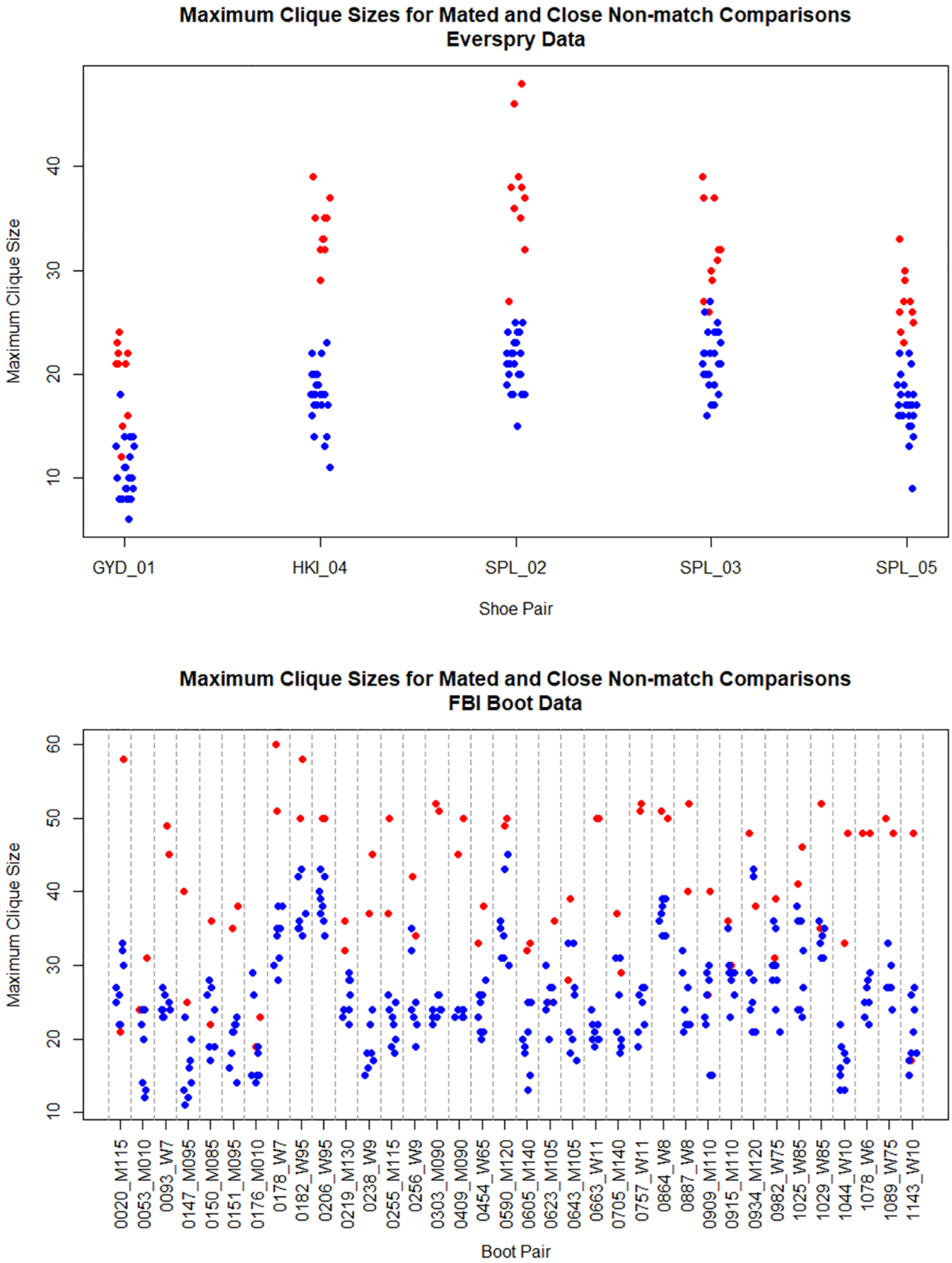

Figure 9: Maximum clique sizes for mated comparisons (red points) and close non-match comparisons (blue points) for Everspry data (top figure) and FBI boot data (bottom figure). Maximum clique sizes show a strong tendency to be larger for mated comparisons.

Curve (AUC) values. More experiments need to be conducted to confirm whether or not this pattern generalizes.

Figure 11 shows plots of the fitted probability densities (smoothed histograms) for all four scoring methods with mated score distributions shown in red and close-non-match score distributions in blue. The plots show the degree of separation between the mated and close-non-match densities for these similarity scores.

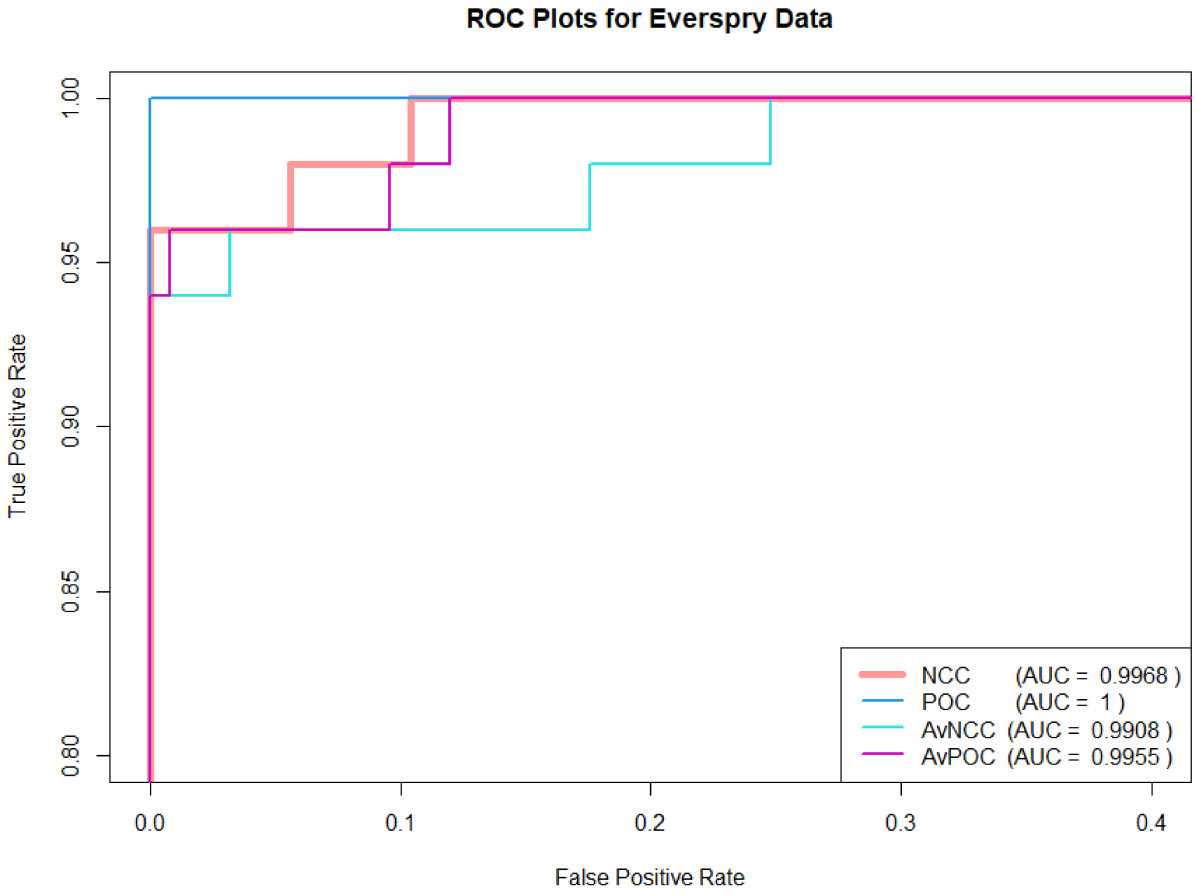

Figure 10: ROC plots for Everspry data: *NCC* (red), *POC* (blue), *AvNCC* (cyan) and *AvPOC* (magenta).

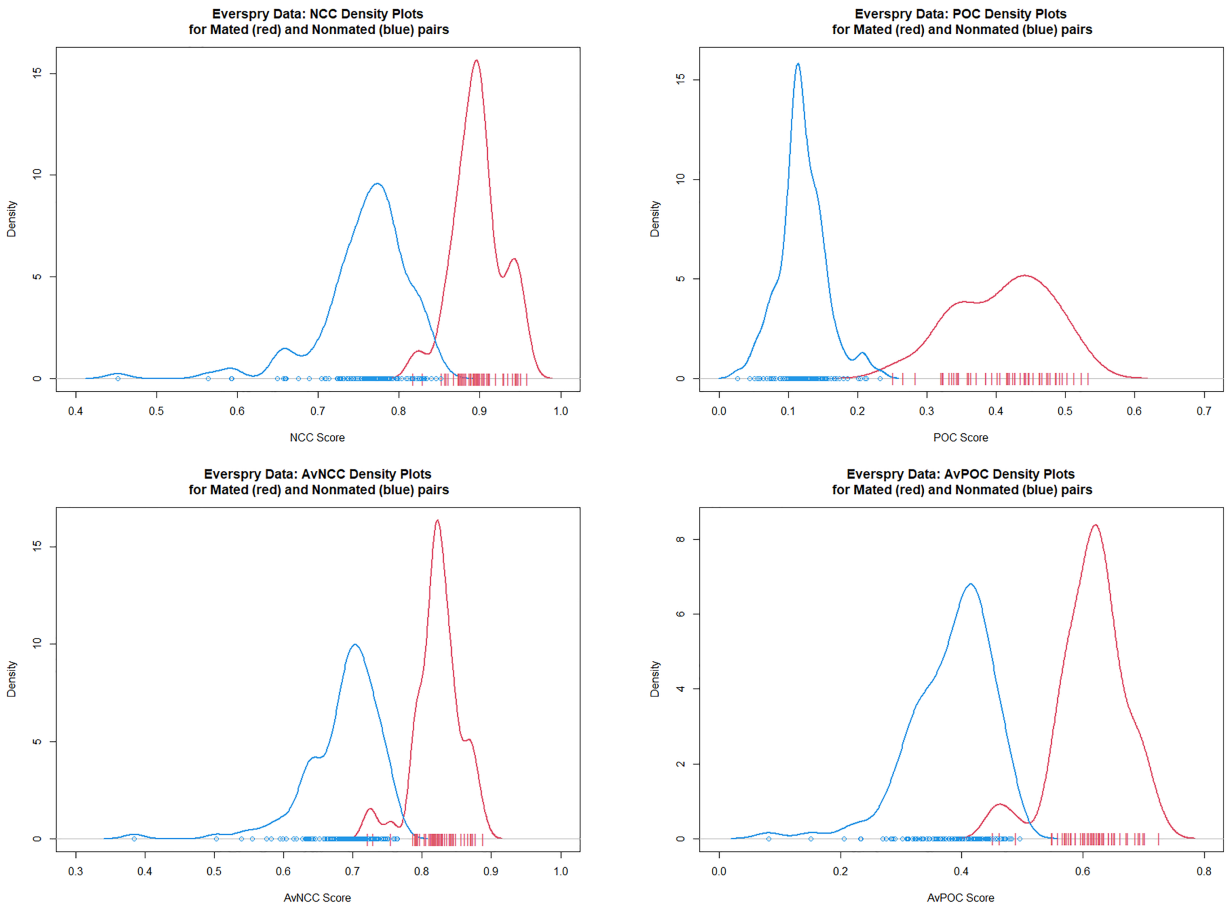

Figure 11: Density plots for mated (red) and nonmated (blue) scores for Everspry data: *NCC* (top-left), *POC* (top-right), *AvNCC* (bottom-left) and *AvPOC* (bottom-right). In each of the four figures, the red vertical tick marks on the horizontal axis represent the scores for mated comparisons and the blue open circles on the horizontal axis represent the scores for close non-match comparisons.

We also examined how often a comparison of an impression with its known mate produced a score that was higher than the scores when compared with close non-matches. Each impression was compared with a lineup of 6 impressions consisting of one mate and 5 close non-mates. As each shoe has 5 impressions (true mates) and 5 close non-mates (from the opposite shoe), this results in 20 lineups for each pair of shoes. For 5 pairs of shoes this gives us 100 total lineup comparisons. Table 3 provides summary information regarding the rank

of the match score for the true mate among all the scores from the lineups.

	Rank (largest to smallest)						
Scoring Method	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Total
NCC	99	1	0	0	0	0	100
POC	100	0	0	0	0	0	100
$AvNCC$	100	0	0	0	0	0	100
$AvPOC$	100	0	0	0	0	0	100

Table 3: Number of times a mated comparison yields a score higher than scores from close non-matches for Everspry data.

We note that, except for NCC, every scoring method correctly places the mated impression at rank 1 position (the possible ranks are 1, 2, 3, 4, 5, or 6) in all 100 lineup comparisons whereas NCC does so in 99 cases out of 100. In the one case where the mated score was not at rank 1 for NCC it was actually in rank 2 position.

FBI Boot Data

As described above, for each impression in our collection of FBI impressions, there are three comparisons done – one true match and two (flipped) close non-matches from opposite shoes. This is shown in the bottom of Figure 3.

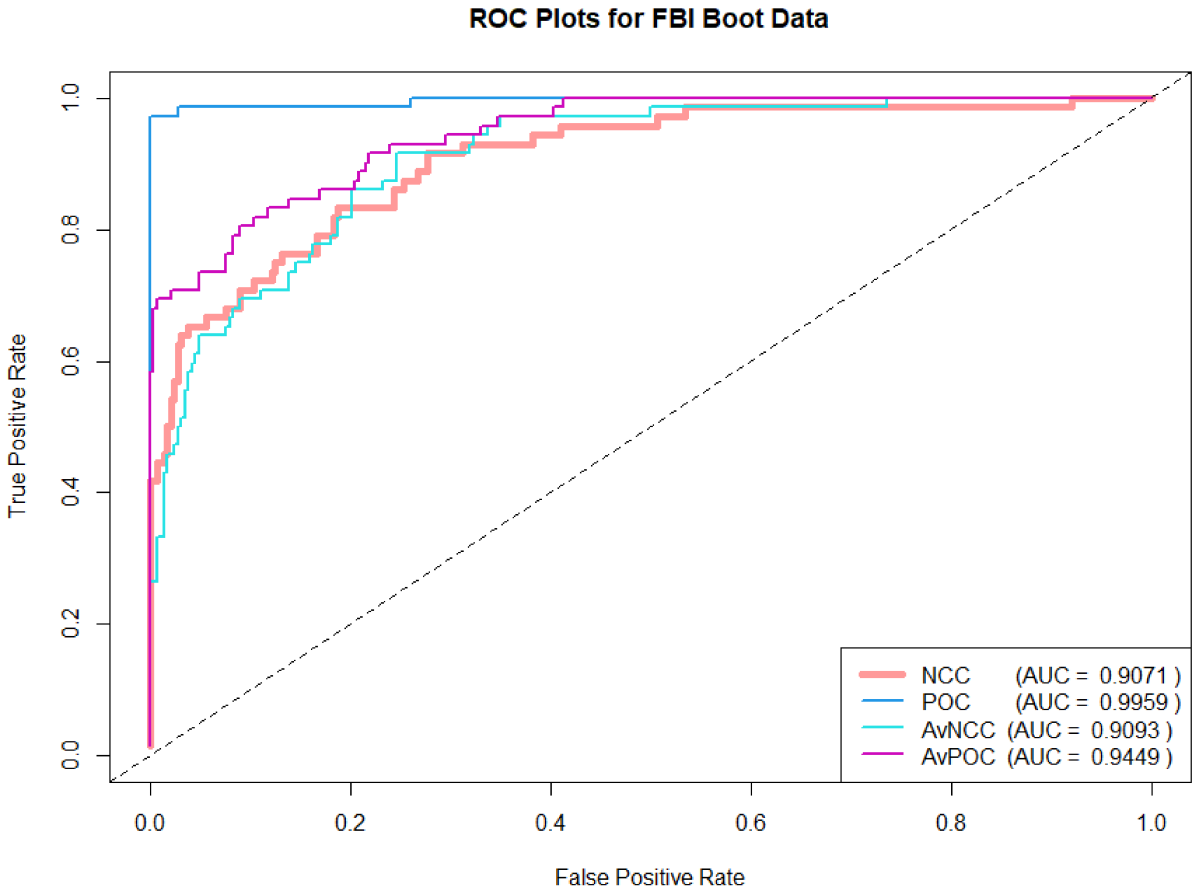

Figure 12: ROC plots for FBI boot data: *NCC* (red), *POC* (blue), *AvNCC* (cyan) and *AvPOC* (magenta).

Figure 12 shows ROC plots for the FBI boot data. The plots are based on scores from comparisons of 72 known mated pairs and 288 known close-non-match pairs. *POC* appears to have noticeably better discrimination power than the other scoring methods for this data, as indicated by the AUC values.

Figure 13 shows plots of the fitted probability densities (smoothed histograms) for all four scoring methods with mated score distributions shown in red and close-non-match score distributions in blue. The plots show the degree of separation between the mated and close-non-match densities for these similarity scores.

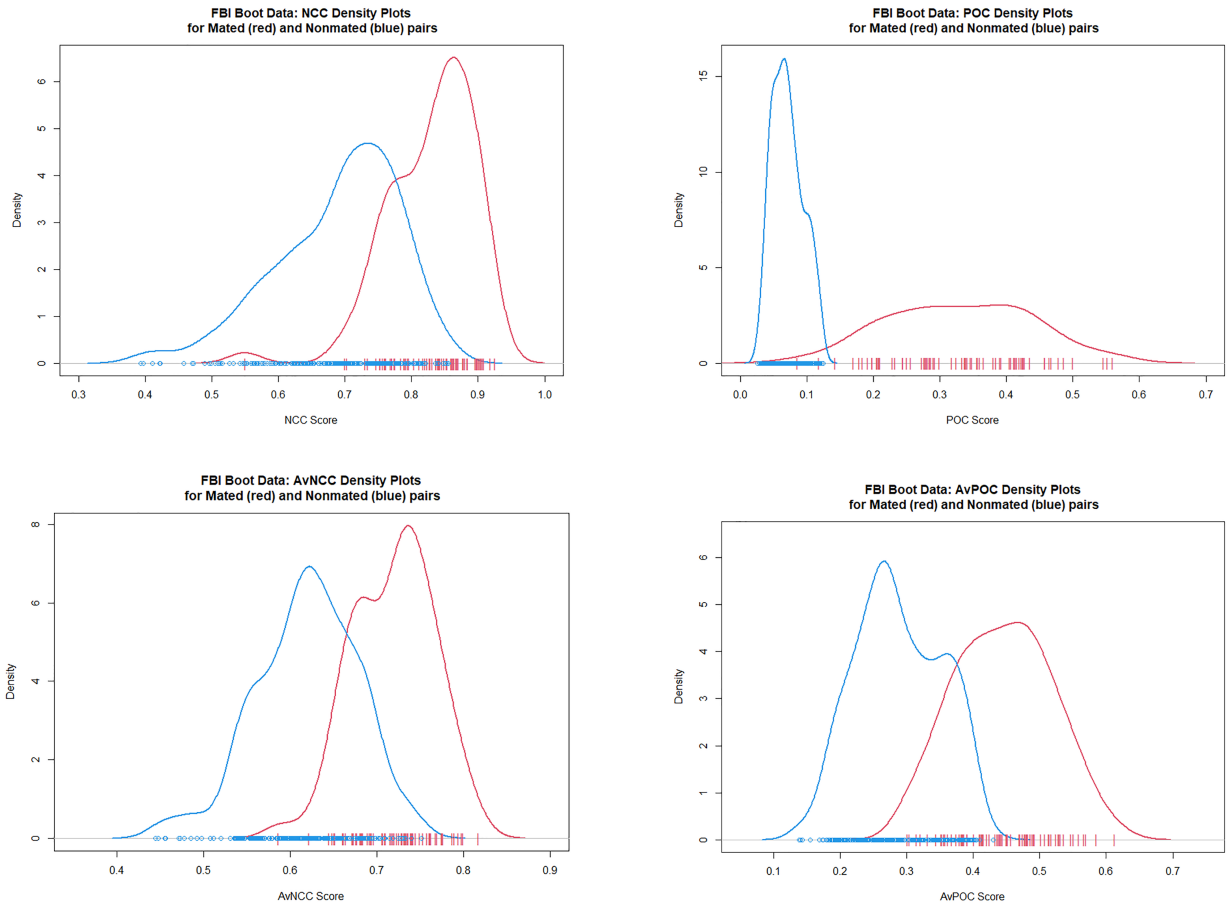

Figure 13: Density plots for mated (red) and nonmated (blue) scores for FBI Boot data: *NCC* (top-left), *POC* (top-right), *AvNCC* (bottom-left) and *AvPOC* (bottom-right). In each of the four figures, the red vertical tick marks on the horizontal axis represent the scores for mated comparisons and the blue open circles on the horizontal axis represent the scores for close non-match comparisons.

It is also of value to examine how often a comparison of an impression with its known mate produced a score that was higher than the scores when compared with close non-matches. As we use 72 mated pairs, we have a total of 144 lineup comparisons. Table 4 provides summary information pertaining to this question.

Scoring Method	Rank (largest to smallest)			Total
	Rank 1	Rank 2	Rank 3	
NCC	132	10	2	144
POC	143	1	0	144
$AvNCC$	135	8	1	144
$AvPOC$	143	1	0	144

Table 4: Number of times a mated comparison yields a score higher than scores from close non-matches for FBI boot data.

We note that NCC correctly places the mated impression at rank 1 position (the possible ranks are 1, 2, or 3) in 132 out of 144 comparisons, $AvNCC$ places the mated comparison at rank 1 position 135 out of the 144 comparisons, and POC as well as $AvPOC$ do so in 143 out of 144 comparisons.

WVU Data

As described above, for each test impression in our collection of WVU impressions, we compare it with a set of mock crime-scene impressions, either true matches or (flipped) close non-matches from opposite shoes.

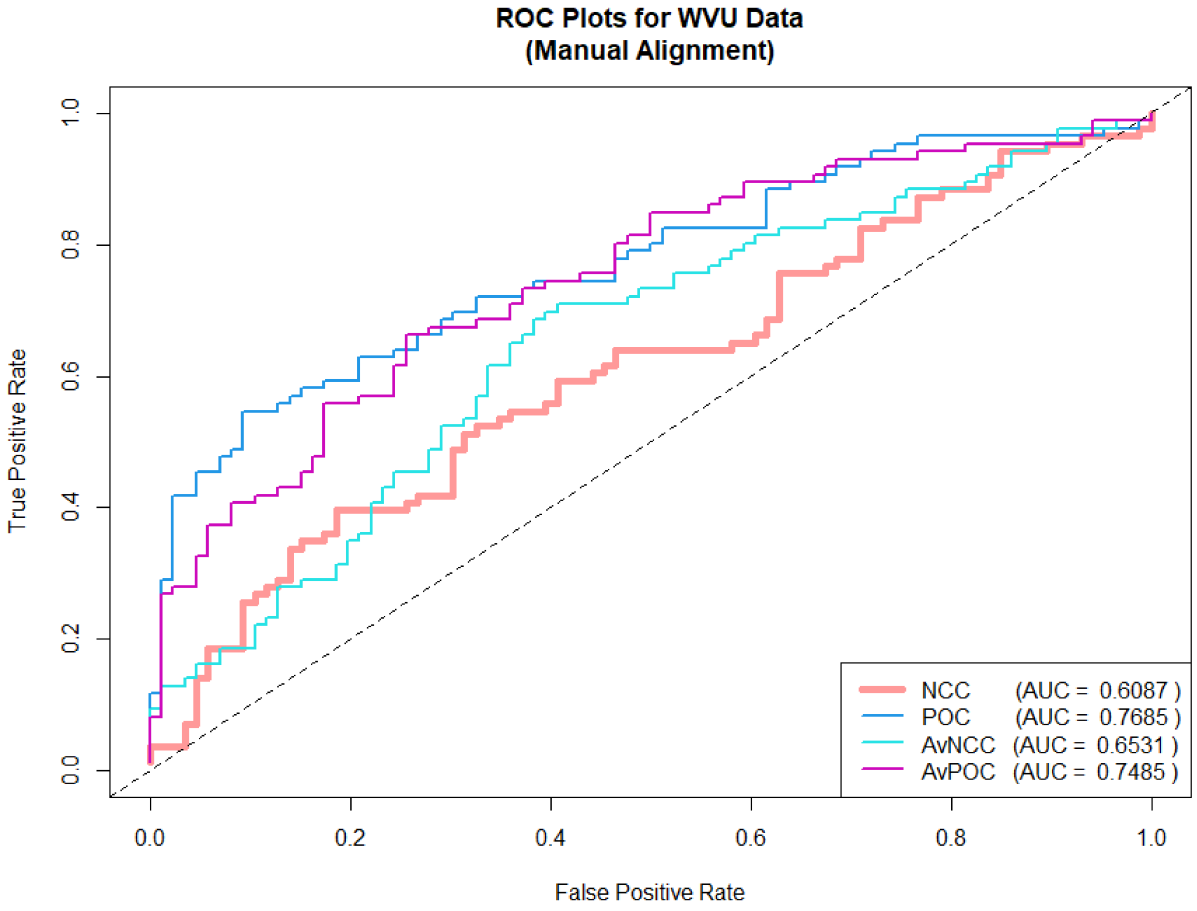

Figure 14: ROC plots for WVU mock crime-scene data: NCC (red), POC (blue), $AvNCC$ (cyan) and $AvPOC$ (magenta).

Figure 14 shows the ROC plots for the WVU data. The plots are based on scores from comparisons of 86 known mated pairs and 86 known close-non-match pairs. POC appears to have better discrimination power than the other scoring methods for this data, as indicated by the AUC values. The densities for mated scores and close non-match scores are compared in Figure 15.

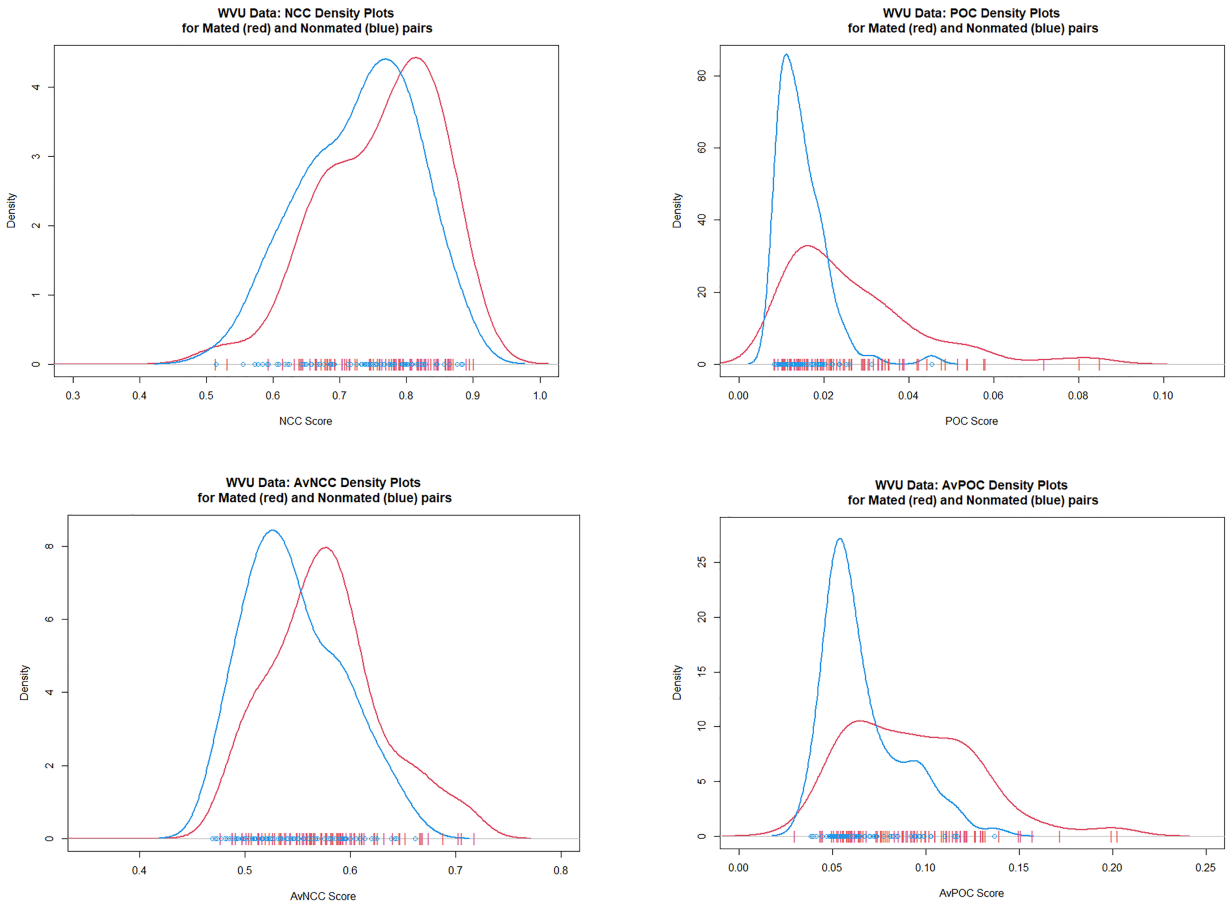

Figure 15: Density plots for mated (red) and nonmated (blue) scores for WVU data: *NCC* (top-left), *POC* (top-right), *AvNCC* (bottom-left) and *AvPOC* (bottom-right). In each of the four figures, the red vertical tick marks on the horizontal axis represent the scores for mated comparisons and the blue open circles on the horizontal axis represent the scores for close non-match comparisons.

Figure 16 shows ROC plots for mock crime-scene impressions involving dust, blood without LCV, and blood with LCV. The plots use mated pairs and close non-matches from the WVU data. For dust, the plot is based on scores from comparisons of 32 known mated pairs and 34 known close-non-match pairs; for blood without LCV, 27 known mated pairs and 26 known close-non-match pairs; and for blood with LCV, 27 known mated pairs and 26 known close-non-match pairs.

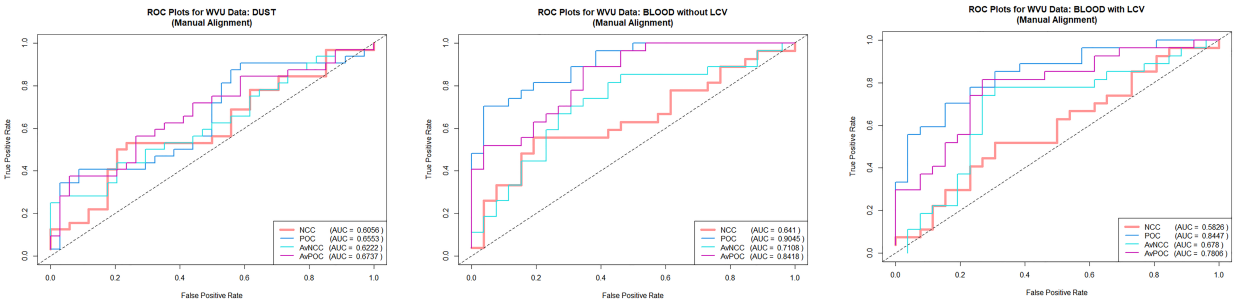

Figure 16: ROC plots for WVU mock crime-scene data. From left to right: dust, blood without LCV, blood with LCV.

In dust, $AvPOC$ is the best discriminator for these data with POC being a close second. In each of blood (without LCV) and blood with LCV, POC is the best discriminator for these data. The performance in blood and blood with LCV is significantly better than the performance in dust, where the discrimination is just slightly better than random.

Relative to what we see in the cases of the Everspry and FBI data, all of the scoring methods with WVU data show reduced performance. This is perhaps to be expected since the mock crime-scene images contain much less information than the Everspry and FBI images, where the questioned impressions have much higher quality.

For all three sets of data (Everspry, FBI, WVU), the scores do not have the benefit of extracting and using explicit wear and RAC information, as opposed to considering an overall pattern that implicitly contains wear and RACs. We speculate that, if wear and RAC information could be explicitly extracted in an accurate manner, either automatically or by a user performing appropriate annotations on the test and questioned impressions, this information combined with the scoring methods we use might result in much improved performance. This would particularly affect results using WVU data, as results with Everspry and FBI data are already quite good. This is a topic of ongoing research.

Conclusions

In this paper, we have explored the use of similarity metrics for footwear evidence evaluation, as opposed to the more commonly researched database retrieval application. We have therefore focused on data sets that include only known matches and known close non-matches. A metric that can discriminate between arbitrary Qs and Ts may not perform well for discriminating between matches and close non-matches. Close non-matching shoes largely share the same design and size. Therefore the ability to effectively discriminate between them requires considering, either explicitly or implicitly, wear patterns and possibly individual characteristics in addition to design and size. This type of discrimination is necessary for assessment of evidential value.

POC has been reported (Richetelli et al. [9] and others) to be an effective similarity metric for the data-base retrieval application. Recently, Park et al. [13] have explored the performance of a particular implementation of POC, which they call POC-R, along with other methods, for discriminating between matches and close non-mates using a collection of Everspry impressions they created in their laboratory. They found the POC-R metric to be much less effective than a metric using Random Forests on heuristically defined features. For the data sets we used in our experiments, we found that our implementation of POC performs extremely well for the Everspry data set as well as for the FBI boot data. For the WVU data, the performance of POC is encouraging for blood and blood with LCV impressions; however it performs quite poorly for dust impressions. The differences in our conclusions regarding Everspry impressions could be due to two factors – (1) we use different data sets, and (2) we use different alignment methods.

We also explored the performance of a deep learning approach, specifically, the use of features from a pre-trained ResNet-50 model and calculating the average correlation scores (NCC and POC) across the various feature channels. In our experiments, we did not find any improved performance from the deep learning approach. Neither $AvNCC$ nor $AvPOC$ provided any improvement to the performance of POC for the data sets we considered.

Results from our experiments also suggest that our metrics likely take into account some individual and wear characteristics of shoes as all comparisons are done with shoes that largely have the same outsole design. However, additional experiments are needed to decompose the scores into components associated with outsole design, size, wear, and RACs.

Fusing our scores with RAC-based scores is expected to provide a more powerful summary of the overall correspondence between Q and T. This will be investigated in a future study, although preliminary results supporting this possibility are demonstrated in [5].

References

1. Durose MR, Burch AM, Walsh K, and Tiry E. Publicly funded forensic crime laboratories: Resources and Services, 2014. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics 2016. NCJ250151

2. NAS Report. Strengthening Forensic Science in the United States: A Path Forward. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, Document No.: 228091 2009

3. PCAST Report. REPORT TO THE PRESIDENT: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. President's Council of Advisors on Science and Technology 2016

4. SWGTREAD. Range of Conclusions Standard for Footwear and Tire Impression Examinations. Scientific Working Group for Shoeprint and Tire Tread Evidence 2013

5. Venkatasubramanian G, Hegde V, Lund S, Iyer H, and Herman M. Quantitative evaluation of footwear evidence: initial workflow for an end-to-end system. In Review 2020

6. Lund S. Footwear Impression Comparison System (FICS): a workflow built while focusing on casework performance. In Review 2020

7. Hegde V and Lund S. Expanding the NIST footwear evidence framework: revised RAC scores, weighting reference comparisons, and mapping results to examiner conclusions. In Review 2020

8. Everspry. Dalian Everspry Sci & Tech Co., Ltd. (2020). Available from: http://www.everspry.com/en/products/products_03.htm

9. Richetelli N, Lee MC, Lasky CA, Gump ME, and Speir JA. Classification of footwear outsole patterns using Fourier transform and local interest points. Forensic Sci. Int. 2017; 275:102–9

10. Kortylewski A, Albrecht T, and Vetter T. Unsupervised footwear impression analysis and retrieval from crime scene data. Asian Conference on Computer Vision 2014 :644–58

11. Kortylewski A and Vetter T. Probabilistic compositional active basis models for robust pattern recognition. British Machine Vision Conference 2016

12. Kong B, Supancic J, Ramanan D, and Fowlkes C. Cross-domain forensic shoeprint matching. Proceedings of British Machine Vision Conference 2017

13. Park S and 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; 13:188–99

14. Cui J, Zhao X, Liu N, Morgachev S, and Li D. Robust shoeprint retrieval method based on local-to-global feature matching for real crime scenes. J. Forensic Sci. 2019; 64:422–30

15. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography 1978; 34:827–8

16. Kong B, Supancic J, Ramanan D, and Fowlkes CC. Cross-domain image matching with deep feature maps. International Journal of Computer Vision 2019; 127:1738–50

17. Burger W and Burge MJ. Digital Image Pocessing: An Algorithmic Introduction Using Java. 2016

18. He K, Zhang X, Ren S, and Sun J. Deep learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition 2016 :770–8

19. Deng J, Dong W, Socher R, Li LJ, Li K, and Fei-Fei L. Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009 :248–55

20. Derpanis KG. The Harris Corner Detector. York University 2004 :1–2

21. Shi J and Tomasi C. Good features to track. 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1994 :593–600

22. Rosten E and Drummond T. Machine learning for high-speed corner detection. European Conference on Computer Vision 2006 :430–43

23. Bresenham J. A linear algorithm for incremental digital display of circular arcs. Communications of the ACM 1977; 20:100–6

24. Conte D, Foggia P, Sansone C, and Vento M. Thirty years of graph matching in pattern recognition. International Journal of Pattern Recognition and Artificial Intelligence 2004; 18:265–98

25. Bomze IM, Budinich M, Pardalos PM, and Pelillo M. The maximum clique problem. Handbook of Combinatorial Optimization, Supplement Volume A 1999 :1–74

26. Venkatasubramanian G. cliquematch : finding correspondence via cliques in large graphs. In Review 2020

27. Venkatasubramanian G. Python package `cliquematch`. Available from: https://github.com/ahgamut/cliquematch