6-11-2020

Quantifying the similarity of 2D images using edge pixels: an application to the forensic comparison of footwear impressions

Soyoung Park  
_University of Iowa_, sypark@iastate.edu

Alicia L. Carriquiry  
_University of Iowa_, alicia@iastate.edu

Follow this and additional works at: [https://lib.dr.iastate.edu/csafe_pubs](https://lib.dr.iastate.edu/csafe_pubs)

Part of the Forensic Science and Technology Commons

**Recommended Citation**

Park, Soyoung and Carriquiry, Alicia L., "Quantifying the similarity of 2D images using edge pixels: an application to the forensic comparison of footwear impressions" (2020). _CSAFE Publications_. 53.  
[https://lib.dr.iastate.edu/csafe_pubs/53](https://lib.dr.iastate.edu/csafe_pubs/53)

This Article is brought to you for free and open access by the Center for Statistics and Applications in Forensic Evidence at Iowa State University Digital Repository. It has been accepted for inclusion in CSAFE Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Quantifying the similarity of 2D images using edge pixels: an application to the forensic comparison of footwear impressions

Abstract
We propose a novel method to quantify the similarity between an impression (Q) from an unknown source and a test impression (K) from a known source. Using the property of geometrical congruence in the impressions, the degree of correspondence is quantified using ideas from graph theory and maximum clique (MC). The algorithm uses the x and y coordinates of the edges in the images as the data. We focus on local areas in Q and the corresponding regions in K and extract features for comparison. Using pairs of images with known origin, we train a random forest to classify pairs into mates and non-mates. We collected impressions from 60 pairs of shoes of the same brand and model, worn over six months. Using a different set of very similar shoes, we evaluated the performance of the algorithm in terms of the accuracy with which it correctly classified images into source classes. Using classification error rates and ROC curves, we compare the proposed method to other algorithms in the literature and show that for these data, our method shows good classification performance relative to other methods. The algorithm can be implemented with the R package shoeprintr.

Keywords
Maximum clique, learning algorithms, shoe outsole comparison, pattern matching, image analysis

Disciplines
Forensic Science and Technology

Comments
This is a manuscript of an article published as Park, Soyoung, and Alicia Carriquiry. "Quantifying the similarity of 2D images using edge pixels: an application to the forensic comparison of footwear impressions." Journal of Applied Statistics (2020): 1-28. Posted with permission of CSAFE.

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/csafe_pubs/53
Quantifying the similarity of 2D images using edge pixels: An application to the forensic comparison of footwear impressions

Soyoung Park\textsuperscript{a} and Alicia Carriquiry\textsuperscript{a}

\textsuperscript{a}Department of Statistics, Iowa State University, U.S.A.

\textbf{ARTICLE HISTORY}
Compiled May 21, 2020

\textbf{Abstract}
We propose a novel method to quantify the similarity between an impression ($Q$) from an unknown source and a test impression ($K$) from a known source. Using the property of geometrical congruence in the impressions, the degree of correspondence is quantified using ideas from graph theory and maximum clique (MC). The algorithm uses the $x$ and $y$ coordinates of the edges in the images as the data. We focus on local areas in $Q$ and the corresponding regions in $K$ and extract features for comparison. Using pairs of images with known origin, we train a random forest to classify pairs into mates and non-mates. We collected impressions from 60 pairs of shoes of the same brand and model, worn over six months. Using a different set of very similar shoes, we evaluated the performance of the algorithm in terms of the accuracy with which it correctly classified images into source classes. Using classification error rates and ROC curves, we compare the proposed method to other algorithms in the literature and show that for these data, our method shows good classification performance relative to other methods. The algorithm can be implemented with the R package \textit{shoeprintr}.

\textbf{KEYWORDS}
Maximum clique; Learning algorithms; Shoe outsole comparison; Pattern matching; Image analysis

\section{1. Introduction}

The need to compare two or more two-dimensional images arises in many situations. In the forensics context, examiners are often asked to determine whether a latent impression at the crime scene could have been made by a putative object in possession of the subject. Examples include fingerprints, tire tread marks, footwear prints, and several others. Here we focus on quantifying the similarity between two shoe outsole impressions, but the method we propose can be used more widely.

Footwear impressions are ubiquitous in crime scenes and can be powerful evidence to link a suspect’s footwear to the crime. There are different approaches to collect shoe outsole prints from a scene that depend on how the print is deposited, but at least in terms of usage, high-resolution photography appears to be popular [3, 6]. For our work, we consider two-dimensional (2D) images, such as those obtained using photography.
For a forensic examiner, the question of interest is one of source: could the prints at the crime scene have been made by the defendant’s shoe?

Shoe outsole images can be compared using class, subclass, or individual characteristics. Class characteristics include size, make, model, and any other attribute that can be expected to be the same across a very large number of shoes. Sub-class characteristics are common across a smaller number of shoes and include small differences in the outsole pattern created by a specific mold, for example. Finally, individual characteristics (or randomly acquired characteristics, RACs) are believed to be unique to a shoe sole and arise from wear and tear. Stone [30] estimated that the probability of observing one matching RAC in two images is approximately 1 in 16,000; under the assumption of independence of the location of the RACs, the probability of a match decreases exponentially as the number of RACs used for comparison increases. Kaplan-Damary et al. [16], however, have argued that the locations of RACs on a shoe outsole follow a non-uniform process. In current practice, forensic examiners focus on RACs when assessing whether the correspondences observed between the putative shoe and the print at the crime scene warrants an identification conclusion. In most cases, examiners carry out a visual comparison and form their subjective expert opinion on whether two impressions originated from the same shoe into one of the seven-scale decisions thorough the guideline by SWGTREAD [8]. One drawback of the focus on RACs is that when the crime scene print is of low quality, small RACs are often not visible.

In what follows, we use $Q$ to denote the questioned outsole impression(s) found at the crime scene, and $K$ to denote the reference or known impression obtained from a test impression from the putative shoe source. Figure 1 is an example of images of a partial shoe print (or *latent*) found at the crime scene and the putative source shoe from a suspect or from a reference database. As mentioned above, current practice consists of visually comparing the two impressions to make a decision regarding source. A report by the National Research Council [20] was critical of this practice because the decision about source heavily relies on the examiner’s experience and subjective assessment. In addition to the issue of subjectivity, little is known about the accuracy and repeatability of shoe outsole comparisons by forensic examiners. Majamaa and Ytti [19] and Shor and Weisner [27] question whether examiners are likely to reach
An advisory committee to the Obama administration on matters of science and technology pointed to a potential lack of reliability and accuracy among footwear examiners when addressing questions of source (PCAST, [14]). The PCAST report also addresses the question of probative value of shoe print evidence; even if two impressions are indistinguishable, what does the high degree of similarity indicate regarding source?

In practice, comparing $Q$ and $K$ is challenging. First, images obtained from a crime scene impression are sometimes partial or smudged. Second, to compare two impressions of $Q$ and $K$, the images first need to be rotated and aligned and sometimes re-scaled. Third, images are subject to noise and background effects, and therefore whatever we conclude from the comparison of $Q$ and $K$ is affected by the quality of images. Finally, to be able to accurately identify and characterize RACs, the resolution of the image needs to be very high, which rarely occurs in real casework.

The goal of our work is two-fold. First, we wish to develop a semi-automated approach to compare $Q$ and $K$ and calculate a score to quantify the degree of similarity (or correspondence) between the images, that overcomes some of the challenges listed above. The score should rely on features associated with class and sub-class characteristics, wear patterns, and RACs, without being prohibitively expensive to compute. The decision to exclude a putative shoe from consideration (or decisions of different sources) can depend on class characteristics alone, but to conclude that the impressions have a common source, the score must include information provided by the RACs. Consequently, including RACs into the score is a necessary, but not sufficient, condition for identification. Second, we propose an approach to obtain a similarity score for a pair of impressions, which can be used to assess the probative value of the evidence.

There have been no studies designed to test the performance of comparison algorithms on footwear impressions sharing the same outsole design but used by different people. However, this is a relevant issue when attempting to quantify the probative value of the evidence. If the examiner cannot distinguish between different shoes with the same class characteristics and degree of wear, then the footwear evidence has low probative value unless the crime scene impression was made by a shoe of uncommon make and model or rarely observed size.

This paper is organized as follows. In Section 2, we review the literature on methods to quantify the similarity between 2D images, with emphases on approaches used to compare two outsole impressions. In Section 3, we propose an algorithm to compare any two 2D images that rely on the concept of maximum clique and show how the algorithm that we propose works on several examples when two images are very similar. We describe an experimental dataset of images, introduce an R-package to implement the proposed algorithm, and develop an illustrative example in Section 4. We compare the performance of the algorithm we propose with other existing methods reviewed by Richetelli et al. [26]. Finally, Section 5 summarizes our findings, provides an interpretation of results, and discusses potentially useful variations for our algorithm. An Appendix A provides additional details regarding the location and number of the selected circular regions in $Q$ and their effect on the performance of the algorithm.

2. Earlier work

The problem that consists of comparing two images is not new and has arisen in many different disciplines, including footwear examination. Bouridane et al. [4] proposed a fully automatic system for matching and retrieval of 2D images that uses a fractal

Phase-only correlation (POC) is used for comparing two outsole images (especially when one of them is obtained from a partial print) in [10, 11]. A potential drawback of the POC is its lack of invariance to rotation and scale. In response, Gueham et al. [12] proposed using the Fourier-Mellin transform – a version of POC that is rotation and scale invariant – to automatically pair similar images. The Fourier-Mellin transform correlation (FMTC) can be computed by calculating a POC using images that have been transformed to the log polar domain.

Recently, Richetelli et al. [26] showed that POC and FMTC, exhibit good classification ability for sorting shoe outsole images into mated and non-mated categories. Using high-quality images, they showed that POC results in the highest area under the ROC curve (AUC) when classifying shoe impressions by source. Gueham et al. [9] note that POC is robust to translation, brightness, and noise of images when comparing two or more of them. The fact that POC is not robust to rotation and scale, however, means that before implementing POC on a pair of images, the images must be aligned and scaled as accurately as possible. In their examples, Richetelli et al. [26] re-scaled all shoe images to have the same size and manually adjusted rotation angles between mated and non-mated pairs of comparisons. We revisit the usefulness of POC and FMTC in Section 4.3 and compare their performance to that of other methods for image comparison. Details about POC and FMTC can be found in [26].

Several authors have proposed comparison algorithms that make use of characteristics observed in specific regions of the outsole. Tang et al. [31, 32] use Attribute Relational Graphs (ARG), where nodes represent features of the outsole and edges reflect relationships between the nodes. Kortylewski et al. [18] proposed focusing on periodicity on the outsole pattern to develop a system for data retrieval which is invariant to rotation and translation and is robust to noise. The images with which they worked are available to the public in the University of Basel website (https://fid.dmi.unibas.ch/). Wang et al. [33] propose a fully automatic retrieval system for crime scene prints that relies on a Wavelet-Fourier transform. A recent paper, Alizadeh and Kose [2] proposes a method for retrieving shoe outsole impressions by using a blocking sparse representation technique which is resistant to distortions in rotation and scale.

The evaluation of the performance of the various algorithms has been carried out in an ad-hoc manner, often on datasets for which we do not know ground truth (i.e., which pairs of images were deposited by the same shoe and which pairs were not), and furthermore, the datasets have not been available to the general scientific community (a notable exception is [18]). Except for commercial vendors, scientists have invested limited effort to create a database of outsole patterns. Recently, Kong et al. [17] described the creation of such a database, perhaps as a first step in the construction of a national reference database. Shor et al. [28] are in the process of creating a database of controlled, replicated impressions from which one can estimate within-source variability. They highlight the importance of understanding variability when creating test impressions under different conditions. Regardless of the abundance of algorithms that have been described in the literature, forensic practice continues to rely on subjective
assessments by footwear experts.

3. A novel algorithm to quantify the similarity between 2D images

3.1. The signature of a shoe outsole impression

We consider the coordinates of all edges detected by the Prewitt operator [25] in the shoe outsole image as the starting points of interest to compare two images \( Q \) and \( K \). We consider all edge points in the outsole image because they represent the outsole boundaries. Some of these will correspond to class or sub-class characteristics, but some might include RACs that appear uniquely on each shoe outsole. Therefore, the algorithm we propose directly considers the outsole pattern and indirectly takes into account RACs when extracting a signature of the shoe impression. The disadvantage of using edges is that the comparison is high-dimensional. At 300 dpi, a typical outsole image contains about 10,000 edge points, and therefore measuring the similarity between two images is computer and time-intensive. In practice, we down-sample the image at a 20% rate.

We extract a signature from an outsole image \( Q \) by focusing on three semi-arbitrary circular regions in the impression, and the edge points within them. One advantage of limiting the comparison to the three circular areas on the image \( Q \) is that we can select regions of the image that are both interesting and less contaminated by noise. We choose to define circular target areas because circles are invariant to rotation. The number of circles we propose was selected by trial and error and represent a compromise between accuracy and computational burden. Given that choice, we then construct triangles, denoted by \( \Delta \), with edges connecting the centroids of those circles in both \( Q \) (\( \Delta_Q \)) and \( K \) (\( \Delta_K \)).

Once the three circles have been selected, we compare the two triangles formed by connecting the three centers of the circles in \( Q \) (\( \Delta_Q \)) and the three centers of the circles that the algorithm found in \( K \) (\( \Delta_K \)). To quantify the similarity between two triangles, we rely on the concept of congruence of two triangles – when all corresponding sides and interior angles are congruent. By one of the conditions of congruence of two triangles, if the sides of \( \Delta_Q \) and \( \Delta_K \) are of similar length (SSS), then we say that the two triangles are congruent. We use the degree of congruence as an additional feature useful to measure the degree of similarity between two shoe outsole impressions. Three is the minimum number of circles to build the signature of an outsole; additional circles may improve accuracy but at the cost of increased computational burden.

The first step that consists in selecting the three circular regions in \( Q \) is not automated and enables the forensic practitioner to select the portions of the latent that are most promising. Manual selection of the areas of interest in the questioned impression is also useful when the latent print is partially observed. The coordinate values of all edge points within the selected circles are the data with which we work.

Let \( q_1, q_2, q_3 \), represent three circular areas in \( Q \) selected by the examiner. Using the edge points and other attributes, we compute the value of a set of features from which we will eventually construct a similarity score. For example, the distances between the centers of the three circles are informative. Figure 2 is an example of a known mated pair, obtained by imaging the same shoe twice and assigning one image to \( Q \) and one to \( K \). By applying the comparison algorithm we propose (see Section 3.2) we identify the three best corresponding circles \( k_1^*, k_2^*, k_3^* \) in \( K \). When two impressions have a common source, we expect to see a high degree of similarity between features
of the circular areas and between the triangles formed by circle centroids in $Q$ and $K$. By the congruence of both triangles formed by joining the centroids, the three pairwise distances between centers in $Q$ and in $K$ should be close enough to suggest that the two impressions were made by the same shoe.

We define six similarity features using the differences $q_1 - k_1^*$, $q_2 - k_2^*$, $q_3 - k_3^*$. These are: (1) Average number of points in the maximum clique (Clique size) (2) Average percentage of points in the $K$ circles that overlap points in $Q$ circles (% Overlap in $K$), (3) Average percentage of points in the three circles of $Q$ that overlap the corresponding circles in $K$ (% Overlap in $Q$), (4) Standard deviation of the estimated rotation angles computed when aligning the three pairs of regions (SD of rot. angle), (5) Average of the median distance between overlapping points (Med. distance of OP) and (6) Average of the absolute difference in the length of the sides of the triangle formed by joining the centroids in $Q, K$ (Diff. in triangle $Q-K$). We argue below that none of the six features individually has sufficient discriminatory power to correctly identify mated and non-mated pairs of images. Therefore, we propose combining features into a score using a supervised learning algorithm such as a random forest [5].

### 3.2. Maximum clique to align and overlay two images

To compute the value of the features described earlier, we must first align the images and then find subsets of edge points in both images that are congruent. To do so, we use the idea of maximum clique (MC). A clique in an undirected graph is a subset of its vertices where every vertex in the subset is connected by an edge to all other vertices in the subset, i.e., the subgraph induced by the clique is complete. The maximal clique is a clique that cannot be extended by including one more adjacent vertex. The maximum clique

---

Figure 2.: Circular areas in $Q$ (selected by the examiner) are in the left panel, and corresponding circles in $K$ (identified by the algorithm) are in the right panel when $Q$ and $K$ are known mated pairs. The numbers in the boxes indicate the length of sides of the triangles in $Q$ and $K$ formed by uniting the circle centroids.
clique is a clique of the largest possible dimension in a given graph. As an example, in the graph Figure 3, the maximal cliques are \{1,2,3\}, \{2,3,4\}, \{2,4,5\}, \{3,4,6\}, \{4,5,6,7\} and the maximum clique is \{4,5,6,7\}.

The algorithms we develop to quantify the similarity between two outsole images rely on the concept of the maximum clique. A maximum clique is invariant to rotation and translation because it depends on the pairwise distances between nodes in the graph. Thus, the main idea is the following: although outsole pattern images may be translated, rotated and subjected to noise and other loss of information, the geometrical relationships between the points that constitute the pattern will not change much. Thus, local maximum cliques can be used to find corresponding positions in the two images so that we can align them.

The first step in the algorithm is manual. An examiner who wishes to compare the two images first marks at least three circular areas \(q_1, q_2, q_3\) in image \(Q\), in regions of interest to the examiner. There are many ways how to choose interesting areas, but in Section 4.3, we manually find three circular areas \((q_1, q_2, q_3)\) in \(Q\) that are most likely to be in contact with the floor, such as upper left and right areas, lower left area in the outsole. We consider first the circle \(q_1\) which has center coordinates \((c_{x,q_1}, c_{y,q_1})\) and radius \(r_{q_1}\). There is a set of \(n_{q_1}\) edge points within \(q_1\) which we denote by \(S_{q_1}\), a \(2 \times n_{q_1}\) matrix of \(x\) and \(y\) coordinate values, where

\[
S_{q_1} = \begin{pmatrix} x_1 & \ldots & x_{n_{q_1}} \\ y_1 & \ldots & y_{n_{q_1}} \end{pmatrix}.
\]

The next step is to find the closest circular region to \(q_1\) in \(K\) from among a set of \(n_K\) candidate circles. Candidate circle \(k_i\) in \(K\) has center \((c_{x,k_i}, c_{y,k_i})\) and radius \(r_{k_i}\) and the edge points within are \(S_{k_i}\) (\(2 \times n_{k_i}\) matrix). The goal is to select the circle \(k_{c_i}\) that is closest to \(q_1\) by computing the maximum clique, or the subset of edge points in \(S_{k_i}\) that is congruent to a subset of edge points in \(S_{q_1}\). There are \(m\) points in the maximum clique, \(M_{q_1,k_i}\), and we denote the \(j^{th}\) point in \(M_{q_1,k_i}\) by \(p_{j,q_1}\) and \(p_{j,k_i}\).

More formally:

\[
M_{q_1,k_i} = \{(p_{1,q_1}, p_{1,k_i}), \ldots, (p_{m,q_1}, p_{m,k_i})\}^T
\]

\[
p_{j,q_1} = (x_{j,q_1}, y_{j,q_1}), j = 1, \ldots, m
\]

\[
p_{j,k_i} = (x_{j,k_i}, y_{j,k_i}), j = 1, \ldots, m.
\]

In practice, finding the maximum clique in a pair of images is computationally
expensive. Therefore, rather than considering all edge points in \( q_1 \) (which can be in the hundreds), we divide circle \( q_1 \) into a moderate number of equally sized bins (we used 30) and select one edge point randomly from each non-empty bin. Denote a set of 30 random edge points from \( S_{q_1} \) as \( S_{q_1,30} \). We then find the maximum clique \( M_{q_1,k_i} \) between the 30 edge points in \( S_{q_1,30} \), and all edge points in \( S_{k_i} \). Given \( M_{q_1,k_i} \), we can now align a circle \( q_1 \) to the same coordinate system as \( K \) by multiplying all edge points \( S_{q_1} \) by a \( 2 \times 2 \) rotation \((\Sigma_A)\) with rotation angle \((\theta)\) and \(2 \times n_{q_1}\) translation matrix \((T_A)\) so that:

\[
S^K_{q_1} = \Sigma_A \times S_{q_1} + T_A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \times S_{q_1} + \begin{pmatrix} T_x & \cdots & T_x \\ T_y & \cdots & T_y \end{pmatrix}
\]

We have now produced an initial map of the points in circle \( q_1 \) onto the coordinate system for points in circle \( k_i \). We refer to the set of points mapped from circle \( q_1 \) to the coordinate system of shoe \( K \) as \( S^K_{q_1} \) in the rotated and translated circular area \( q^K_1 \).

Finally, for each point in \( S^K_{q_1} \), we find the closest point in \( S_{k_i} \). When the Euclidean distance between two points, one in \( S^K_{q_1} \) and one in \( S_{k_i} \), is less than 2 pixels, we say that points overlap. Using the set of overlapping points in \( S_{q_1} \), we re-calculate a center and a radius for the circular area in \( K \) that is the most closely aligned with circle \( q_1 \) or \( q^K_1 \). To ensure that every point in \( q_1 \) is covered by the circle in \( K \), we increase the value of the estimated radius of the circle in \( K \) by an arbitrary amount. In the examples we discuss later, we increased the estimated radius of the circle in \( K \) by about 20%.

Section 3.3.1 illustrates how the algorithm proceeds when finding a single matching circle (one in \( Q \) and the other in \( K \)), as an example.

The process is repeated for the two additional circles \( q_2, q_3 \) to find the corresponding mated regions \( k^*_2, k^*_3 \) in \( K \) that show the best correspondence with \( q_2, q_3 \). In the end, we determine the best matching circular areas \( k^*_1, k^*_2, k^*_3 \) in \( K \) that correspond to the examiner-selected areas \( q_1, q_2, q_3 \) in \( Q \).

To select candidate circles \( k_i, i = 1, \ldots, n_K \) in the known shoe image \( K \), we proceed in a systematic way. The centers of the candidate circles are placed on a set of fixed coordinate values \((c_{x,k_i}, c_{y,k_i})\) in \( K \) that are separated in the vertical and horizontal directions by a distance equal to the radius of the circles. In this way, the union of the \( n_K \) candidate circles covers the entire shoe outsole image \( K \). To improve computational efficiency, we limit the search area to the sections in \( K \) in which we expect to find the best match for each \( q_i \) (see Section 4.3).

Once the pairs \((q_1, k^*_1), (q_2, k^*_2)\) and \((q_3, k^*_3)\) have been identified, we can obtain the value of the six quantitative features defined in Section 3.1. Section 3.3.2 shows the process of how the algorithm performs a comparison between two images. Even with highly optimized code, searching for similar circular areas in \( K \) is computationally intensive.

There are variations to this base algorithm that have the potential of increasing computational efficiency without sacrificing accuracy. For example, if the two images are roughly aligned initially, we could limit the search for \( k^*_1 \) to a neighborhood in \( K \) that roughly corresponds to the area in \( Q \) that contains \( q_1 \). Or we could use information about the distances between \( q_1 \) and \( q_2 \) and \( q_1 \) and \( q_3 \) to project the location of \( k^*_2, k^*_3 \) once \( k^*_1 \) has been located. We discuss the pros and cons of these and other approaches in Section 5.
3.3. Example implementation of the proposed algorithm

Here we illustrate the implementation of the algorithm we propose when comparing images from known mated pairs (KM) and also address more challenging problems such as comparing images of outsoles of different shoes of the same brand, model and size, and comparing a reference image with a partial impression. Known mated (KM) pairs are constructed by selecting images produced by the same shoe scanned multiple times. The known non-mated pairs (KNM) are built by pairing images that are known to come from different shoes. In Section 4.3, we show results obtained when the algorithms are tested on a large set of pairs of images for which we know the source.

3.3.1. A single circle

We proceed as in Section 3.1, and select one circular area in Q which we denote \( q_1 \). In Figure 4, the points in \( q_1 \) are shown in blue. The circle \( q_1 \) is selected manually, and in this case, it has a radius equal to 50-pixel units. The candidate matching circle \( k_i \) in K is shown in red in the figure, and has a larger radius, to ensure that it covers the entire set \( S_{q_1} \). To quantify the similarity between \( (q_1, k_i) \) we record the value of eight features (see Table 1).

![Diagram](image)

Figure 4.: Left top panel shows pairwise distances between points in the maximum clique. Right top panel shows points in the maximum clique, geometrically congruent subset points in \( q_1 \) and \( k_i \). Bottom left panel: Red points are all points in \( k_i \), and blue points are in \( q_1^K \) that is mapped \( q_1 \). Bottom right panel: the area in K that is most aligned with circle \( q_1 \)

Figure 4 shows some of the features. In the figure, the top left panel shows the pairwise Euclidean distances between \( m \) points in \( q_1 \) and \( k_i \) which form the maximum
The pairwise Euclidean distances between points in $M_{q,k}$ in $q_1$ and $k_i$ are almost the same, and all lie approximately on the 45° line. This is what we expect when points in $M_{q,k}$ are the maximum clique. The panel on the top right of the figure shows the points in the maximum cliques that overlap, after translation and rotation. Recall that the maximum clique was obtained from the 30 points selected from circle $q_1$ and the corresponding points in circle $k_i$. Those points are geometrically congruent regardless of their absolute coordinate values. The bottom left panel shows the entire sets of points in $q_i^K$ and $k_i$ (left panel) after rotation and translation applied, and the subset of points that are defined as overlapping (OP), meaning that their distance is less than 2-pixel units. The bottom right plot shows the area in $K$ that results in the closest match with circle $q_1$.

<table>
<thead>
<tr>
<th>Clique size</th>
<th>Rotation angle</th>
<th>Overlap on $k_i$</th>
<th>Overlap on $q_i^K(q_1)$</th>
<th>Med-pairwise distance</th>
<th>$c_x$</th>
<th>$c_y$</th>
<th>Final radius</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>12.05</td>
<td>0.75</td>
<td>0.97</td>
<td>0.3</td>
<td>54.5</td>
<td>688.5</td>
<td>50.28</td>
</tr>
</tbody>
</table>

In Table 1, column headings denote the following:

1. **Clique size**: The number of points in the maximum clique. We expect to see a larger number when impressions have a common source.
2. **Rotation angle**: For circle $q_1$, this is the rotation angle that results in the best alignment with circle $k_i$ when using the set of points in the maximum clique.
3. **Overlap on $k_i$**: The proportion of points $k_i$ that overlaps points in $q_i^K$.
4. **Overlap on $q_i^K(q_1)$**: The proportion of points in $q_1$ that overlaps points in $k_i$.
5. **Med-pairwise distance of OP**: The median of pairwise Euclidean distances between overlapping points (OP) in $k_i$ and $q_i^K$.
6. **$c_x$**: The estimated $x$ coordinate of circular area in $K$ that is the most closely aligned with $q_1$.
7. **$c_y$**: The estimated $y$ coordinate of circular area in $K$ that is the most closely aligned with $q_1$.
8. **Final radius**: The estimated radius of the circular area in $K$ that is the most closely aligned with $q_1$.

### 3.3.2. Comparing KM

The prints shown on the left half of Figure 5 are replicate images of the same shoe outsole obtained by the same operator one after the other. The two impressions have a different orientation and also differ in the quality of the image.

In both images, we begin by cropping pixels with coordinate values in the bottom 1st percentile or the top 99th percentile of all coordinate values. This cropping is to remove extraneous signals outside the boundary of the shoe outsole and to focus on the inner portion of the outsole impression—two images on the right in Figure 5 result after cropping the outer-most pixel values.

We manually select circle $q_1$ centered on $(75.25, 600.4)$ and with a radius equal to 50-pixel units. The left panel of Figure 6 shows the points included in $q_1$ in blue. The goal is to find the best matching circle in $K$.

Consider several candidate circles, $k_i, i = 1, ..., n_k$ with radius $r_k$ to compare with circle $q_1$. The centers of the candidate circles are placed at fixed $x,y$ coordinate values. This is illustrated in the right panel of Figure 6; in the figure, we select locations on
Figure 5.: Upper two panels: the entire set of edge points in $Q$ (left) and $K$ (right). Bottom two panels: Edge points after cropping points with coordinate values below 1% and above 99% in $Q$ (left) and $K$ (right).

Figure 6.: Left panel: Circle $q_1$ is fixed in the questioned shoe, $Q$. Right panel: Candidate circles will be compared in the reference shoe, $K$.

20%, 40%, 60%, and 80% of the range of $x$-coordinate values in the reference shoe. In the figure, the $x$-coordinates are (49, 98, 147, 196). The $y$-coordinate values are placed along the vertical axes so that circles of the radius ($r_k$) 65 pixels overlap along those axes. In the illustrative example in the right panel of Figure 6, with fixed $x$-axis of 49, there are nine circles where centers have $y$-coordinates (246, 311, 376, 441, 506, 571, 636, 701, 766). Table 2 shows the results obtained when comparing circle $q_1$ and nine candidate circles with center $x$-coordinate values at 49, as in Figure 6. We number the candidate circles by $k_1, ... k_{36}$ starting from the bottom left of the impression. The four circles at the bottom of the left-most vertical line share few points (or none) with $q_1$, so they are excluded from the list of the candidates. Overall, the list of candidate circles includes 36 that are placed along the vertical lines defined by the intersection with the $x$-axis. Among the candidate circles in $K$, we select the one with the highest overlap with $q_1$. We do not show the results from all possible comparisons here, but we do display results from five of the comparisons in Table 2. Circle $k_8$ in Table 2 has the highest overlap with $q_1$ among all candidate circles in $K$, after some rotation and
translation. The rotation angle that resulted in the best alignment between $k_8$ and $q_1$ is 12.05 degrees.

Table 2.: Comparisons of $q_1$ and a sample of candidate circles in $K$

<table>
<thead>
<tr>
<th>Circle</th>
<th>Clique</th>
<th>Rot. size</th>
<th>Overlap</th>
<th>Overlap</th>
<th>Med-pairwise</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_5$</td>
<td>12</td>
<td>52.61</td>
<td>0.49</td>
<td>0.27</td>
<td>0.64 69</td>
</tr>
<tr>
<td>$k_6$</td>
<td>14</td>
<td>33.13</td>
<td>0.41</td>
<td>0.39</td>
<td>0.86 48.5</td>
</tr>
<tr>
<td>$k_7$</td>
<td>15</td>
<td>30.86</td>
<td>0.43</td>
<td>0.48</td>
<td>0.87 56.5</td>
</tr>
<tr>
<td>$k_8$</td>
<td>18</td>
<td>12.05</td>
<td>0.75</td>
<td>0.97</td>
<td>0.30 54.5</td>
</tr>
<tr>
<td>$k_9$</td>
<td>14</td>
<td>90.00</td>
<td>0.65</td>
<td>0.46</td>
<td>1.00 59</td>
</tr>
</tbody>
</table>

Figure 7.: Circle matching between circle $q_1$ and circle $k_8^{\text{adjusted}}$; (1) Before the alignment is adjusted, (2) After the alignment, (3) Close points after the alignment has been adjusted

Once we have identified the closest candidate circle to $q_1$, we increase its radius so that all of $q_1$ is covered by $k_8$. In this example, the adjusted candidate circle $k_8^{\text{adjusted}}$ has center coordinates equal to (54.5, 688.5), and its radius is 15 units larger value the estimated radius, so that the new radius is 50.28 + 15 = 65.28. Figure 7 shows the overlap of circles $q_1$ and $k_8^{\text{adjusted}}$ before and after we improve the alignment, we found $k_1^*$. Note that after alignment, circle $q_1$ (blue) is fully included within circle $k_1^*$ (red). The similarity features are shown in Table 3.

Table 3.: Step3: Similarity features from comparison of circles $q_1$ and $k_8^{\text{adjusted}}$

<table>
<thead>
<tr>
<th>Clique</th>
<th>Rot. size</th>
<th>Overlap</th>
<th>Overlap</th>
<th>Med-pairwise</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1^*$</td>
<td>18</td>
<td>12.13</td>
<td>0.73</td>
<td>0.97 0.29</td>
</tr>
</tbody>
</table>

We find that $k_1^*$ in $K$ has a 97% overlap with $q_1$. The overlapping points are shown in Figure 7, panel (3) and were identified as described earlier. The final matching circle has a radius of 50.28 and center at (54.5, 688.5) in $K$. The left pair of images in Figure
8 shows fixed circle $q_1$ in the questioned print $Q$ and the most similar circle $k_1^*$ in the known print, $K$.

![Figure 8.](image)

Figure 8.: Left panel: images $Q$ and $K$ with most similar circles. Right panel: final results after matching three circles, $q_1$ (red), $q_2$ (yellow), $q_3$ (green) in $Q$, to most similar circles $k_1^*$ (red), $k_2^*$ (yellow), $k_3^*$ (green) in $K$.

We repeat the process for two additional circles in $Q$, $q_2$, and $q_3$, and find the most similar circles in $K$, which we denote $k_2^*$, $k_3^*$. The similarity features are shown in Tables 4 and 5 and illustrated in Figure 8. All three similar circles in $K$ show a degree of overlap that exceeds 90% and similar rotation angles around 12 degrees. In addition, the pairwise Euclidean distances between centers of $q_1$ and $q_2$ in image $Q$ is 451.74, similar to the distance between the centers of $k_1$ and $k_2$ in image $K$. This is also true of the two other possible pairwise distances between circle centers. (In Table 5, $\Delta$ means ‘triangle’ from centers of circles in each image) In the right two panels in Figure 8, we use the same color pairs of circles in $Q$ and in $K$ that are most similar.

Table 4.: Information of fixed circle in $Q$ and corresponding matched circle in $K$ in Section 3.3.2

<table>
<thead>
<tr>
<th>Fixed Center in $q_i$</th>
<th>Found Center in $k_i^*$</th>
<th>Radius in $k_i^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle $q_1$ (75.25, 600.40)</td>
<td>Circle $k_1^*$ (54.5,688.5)</td>
<td>50.28</td>
</tr>
<tr>
<td>Circle $q_2$ (110,50)</td>
<td>Circle $k_2^*$ (179.5, 255)</td>
<td>50.57</td>
</tr>
<tr>
<td>Circle $q_3$ (170,470)</td>
<td>Circle $k_3^*$ (174,579.5)</td>
<td>50.57</td>
</tr>
</tbody>
</table>

Table 5.: The matching results with similarity features in Result Section 3.3.2

<table>
<thead>
<tr>
<th>Matching $q_i - k_i^*$</th>
<th>Clique size</th>
<th>Rot. angle</th>
<th>Overlap on $k_i^*$</th>
<th>Overlap on $q_i$</th>
<th>Med p.wise distance</th>
<th>Circle comp. $\Delta$ in $q_i$'s $\Delta$ in $k_i^*$'s</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1 - k_1^*$</td>
<td>18</td>
<td>12.13</td>
<td>0.73</td>
<td>0.97</td>
<td>0.29</td>
<td>1-2 451.74 451.16</td>
</tr>
<tr>
<td>$q_2 - k_2^*$</td>
<td>17</td>
<td>10.57</td>
<td>0.53</td>
<td>0.91</td>
<td>0.43</td>
<td>1-3 161.19 161.74</td>
</tr>
<tr>
<td>$q_3 - k_3^*$</td>
<td>20</td>
<td>12.14</td>
<td>0.63</td>
<td>1.00</td>
<td>0.24</td>
<td>2-3 325.58 324.55</td>
</tr>
</tbody>
</table>

3.3.3. Performance on very similar images

Here we compare images produced by different shoes that belong to the same individual, are of the same size, brand, and model, and have a slightly different degree of wear. This is arguably one of the more challenging comparisons. The degree of wear and any
RACs are reflected on the impression, as one would expect. For example, in Figure 9, we see repeating zigzag patterns in the middle and right part of the impression from shoe $Q$ (left), but those repeating zigzag patterns are not captured well in shoe $K$ (right). None of the five replicate images obtained from shoe $K$, showed the repeating zigzag patterns in the middle of the image.

![Figure 9: Q and K images, with most similar circles.](image)

In Figure 9, we again use the same color to represent pairs of most similar circles. Note that the algorithm failed, and found circles $k^*_1$, $k^*_2$, $k^*_3$ that are clearly placed in different positions in the print than $q_1$, $q_2$, $q_3$. The summary of results shown in Tables 6 and 7 also show that the results from the comparison of images are not indicating a strong conclusion to the same source, compared to result in Table 5. For example, the rotation angles that were estimated when finding the best alignments are high, which is an indication that similar circles can only be found after much adjustment of $K$.

Table 6.: Similarity between fixed circles in $Q$ and best matching circles in $K$ in Section 3.3.3

<table>
<thead>
<tr>
<th>Fixed Circle</th>
<th>Center in $q_i$</th>
<th>Found Circle</th>
<th>Center in $k^*_i$</th>
<th>Radius in $k^*_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle $q_1$</td>
<td>(55.25, 580.4)</td>
<td>Circle $k^*_1$</td>
<td>(100.5,112.5)</td>
<td>51.91</td>
</tr>
<tr>
<td>Circle $q_2$</td>
<td>(75, 220)</td>
<td>Circle $k^*_2$</td>
<td>(105, 508)</td>
<td>52.8</td>
</tr>
<tr>
<td>Circle $q_3$</td>
<td>(170,450)</td>
<td>Circle $k^*_3$</td>
<td>(64.5,586)</td>
<td>52.03</td>
</tr>
</tbody>
</table>

Table 7.: Matching results between circle $q_i$ and $k^*_i$ in Section 3.3.3

<table>
<thead>
<tr>
<th>Matching $q_i - k^*_i$</th>
<th>Clique size</th>
<th>Rot. angle</th>
<th>Overlap on $k^*_i$</th>
<th>Overlap on $q_i$</th>
<th>Med p.wise distance</th>
<th>Circle Comp.</th>
<th>Dist. of $\Delta$ in $q_i$’s</th>
<th>Dist. of $\Delta$ in $k^*_i$’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1 - k^*_1$</td>
<td>15</td>
<td>31.32</td>
<td>0.40</td>
<td>0.63</td>
<td>0.92</td>
<td>1-2</td>
<td>360.94</td>
<td>395.53</td>
</tr>
<tr>
<td>$q_2 - k^*_2$</td>
<td>15</td>
<td>4.07</td>
<td>0.13</td>
<td>0.61</td>
<td>0.43</td>
<td>1-3</td>
<td>173.70</td>
<td>474.87</td>
</tr>
<tr>
<td>$q_3 - k^*_3$</td>
<td>15</td>
<td>67.61</td>
<td>0.28</td>
<td>0.60</td>
<td>0.87</td>
<td>2-3</td>
<td>248.85</td>
<td>87.89</td>
</tr>
</tbody>
</table>

As an additional test of the power of the algorithm to tell apart impressions from two shoes with the same pattern, we compared impressions from the left and the right shoes of the same pair. To do so, we flipped the image of the left outsole so that it looked like the impression from the right shoe. These two impressions (e.g., one from
the left foot and one from the right foot) have the same outsole design worn during the same time by the same person.

Figure 10.: Three example comparison of two known mated pairs between two replicates of a left (KM-L1L2), between two replicates of a right (KM-R1R2) and a known non-mated pair between the left and the right shoe of the same pair (KNM-L1R1).

As before, we selected three circular areas in each Q and found the most similar circles in the K images. The values of the similarity features in each of the three comparisons are shown in Table 8.

Table 8.: Similarity features in comparisons KM-L1L2, KM-R1R2 and KNM-L1R1

<table>
<thead>
<tr>
<th>Q - K</th>
<th>Comp.</th>
<th>Clique size</th>
<th>Rot. angle</th>
<th>Overlap on $k_i^*$</th>
<th>Overlap on $q_i$</th>
<th>Median pairwise distance in $\Delta$ for $q - k^*$</th>
<th>Abs. diff. in length for $q - k^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM-L1L2</td>
<td>$q_1 - k_1^*$</td>
<td>17</td>
<td>3.03</td>
<td>0.62</td>
<td>0.98</td>
<td>0.43</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>$q_2 - k_2^*$</td>
<td>15</td>
<td>1.43</td>
<td>0.54</td>
<td>0.97</td>
<td>0.40</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>$q_3 - k_3^*$</td>
<td>18</td>
<td>1.93</td>
<td>0.65</td>
<td>0.97</td>
<td>0.28</td>
<td>2.16</td>
</tr>
<tr>
<td>KM-R1R2</td>
<td>$q_1 - k_1^*$</td>
<td>16</td>
<td>11.15</td>
<td>0.62</td>
<td>0.89</td>
<td>0.26</td>
<td>3.30</td>
</tr>
<tr>
<td></td>
<td>$q_2 - k_2^*$</td>
<td>12</td>
<td>12.32</td>
<td>0.34</td>
<td>0.72</td>
<td>0.67</td>
<td>3.37</td>
</tr>
<tr>
<td></td>
<td>$q_3 - k_3^*$</td>
<td>18</td>
<td>11.36</td>
<td>0.56</td>
<td>0.99</td>
<td>0.29</td>
<td>0.38</td>
</tr>
<tr>
<td>KNM-L1R1</td>
<td>$q_1 - k_1^*$</td>
<td>15</td>
<td>17.14</td>
<td>0.38</td>
<td>0.51</td>
<td>0.94</td>
<td>24.37</td>
</tr>
<tr>
<td></td>
<td>$q_2 - k_2^*$</td>
<td>12</td>
<td>10.54</td>
<td>0.33</td>
<td>0.48</td>
<td>0.50</td>
<td>11.55</td>
</tr>
<tr>
<td></td>
<td>$q_3 - k_3^*$</td>
<td>13</td>
<td>8.92</td>
<td>0.29</td>
<td>0.41</td>
<td>0.88</td>
<td>18.77</td>
</tr>
</tbody>
</table>

4. Data analysis and results

We initially implemented the comparison algorithm on a database with 300 mated pairs and 300 non-mated pairs of impressions obtained from shoes of different brands
and models. Since the outsole from different shoe brands has different design elements, the algorithm does a good job of telling mated and non-mated pairs apart. Results can be found in Chapter 3 in [21]. For the comparison algorithm, comparing two very different outsole patterns is easier than comparing two impressions that share class characteristics. In real casework, interesting comparisons are between images obtained from similar objects. In the case of footwear evidence, examiners carry out a comparison only if the suspect’s shoe and the print at the crime scene have the same overall pattern indicating the same brand and model. In this section, we evaluate the performance of the proposed comparison method using a database of images obtained from a collection of shoes of the same brand, model, and size.

4.1. Data

Few datasets of shoe outsole impressions are publicly available and are usable for research purposes. As part of our project, we constructed a longitudinal database of 2D outsole impressions that is available in https://forensicstats.org a public repository. One hundred and sixty participants, about half of them female, were recruited to participate in a study to collect outsole impression information. Participants had to wear shoe sizes 8, 8.5, 10, or 10.5 to qualify. Each was allocated a brand new pair of shoes, either Adidas or Nike, equipped with a step counter. Participants were asked to return the shoes every eight weeks, during a period of about six months, so that we could obtain new impressions. Each measurement on each shoe made during the study period was replicated four times. In one case, one of the replicate measurements was discarded, so only three replicate observations are available for that shoe. Here, we use a subset of the data consisting of 60 pairs of Nike Winflow shoe size 8.5 (38 pairs) and 10.5 (22 pairs).

![Figure 11. Scanned images of two right shoes from different pairs of shoe size 10.5 obtained with the EverOS scanner. Two shoes were used for six months by two participants. The two panels on the left are replicated images of the same shoe from the pair with ID 01. The two panels on the right correspond to two replicate images of the right shoe from the pair with ID 10.](http://www.shopevident.com)

We collected shoe impressions using an EverOS footwear scanner (http://www.shopevident.com) that produces 2D images of the outsole. The instrument scans the shoe outsole as the person steps onto the scanner by capturing the weight distribution of the wearer and the corresponding areas of the outsole that are in touch with the scanning surface. The scanned images have a resolution of 300 dpi. In addition to the outsole scan, output images include a ruled border that helps with the determination of shoe size. Figure 11 shows two replicates of the scans of two right shoes of size 10.5 from different pairs.
We focus on the images obtained from the subset of 60 pairs of shoes during the fourth data collection occasion when shoes had been worn for about six months. Because the impressions were obtained from shoes of the same brand and model, the differences between the outsoles are due only, or at least primarily, to wear and to RACs.

4.2. R-package shoeprintr

To carry out the comparisons using the proposed method, we developed an R package called shoeprintr in collaboration with Omni Analytics (https://omnianalytics.io/). The package is available on GitHub, at https://github.com/CSAFE-ISU/shoeprintr. The package uses a parallelized maximum clique algorithm for speed and efficiency.

Because shoeprintr is provided as an R package, documentation for all exported functions is available and includes descriptions of all parameters, as well as example code for the routines. The primary function is boosted_clique, which takes two sets of points, and based on the given set of parameters, performs a sampling routine to enhance the speed of the matching using MC, and uses statistical techniques to calculate the best overlay between them. Verbose output and resulting plots are available to assess the results as well. Finally, the GitHub file entitled README, available at https://github.com/CSAFE-ISU/shoeprintr, provides a working example on how to install shoeprintr and run an example comparison. Also, Chapter 6: Shoe Outsole Impression Evidence in online book Open Forensic Science in R [23] shows more details how to use R-package shoeprintr on shoe impressions.

There is still an opportunity to improve some of the routines. For example, speed is still limiting when the number of pairwise comparisons between edge points to form a graph is higher than a few hundred. Furthermore, execution of the package requires the compilation of the pmc binary as described in the README file, which means that changes to the output format of the binary image could impact the results of shoeprintr. We continue to improve the package and will upload new versions and the corresponding documentation as they become available.

4.3. Implementation of a learning algorithm and classification results

We randomly selected 70% of the pairs of shoes of each size to train a random forest and set aside the other 30% for testing and estimation for cross-validation. Therefore, out of 60 pairs of shoes, 27 pairs of size 8.5 and 15 pairs of size 10.5 were included in the training set. To construct the mated pairs (KM) of images in the training set, we compared replicate images of the same shoe. There were four replicates for each shoe, so for each, we can obtain 6 (= 4 \times (4 - 1)/2) KM pairs. For one shoe, there were only three replicate images available so that we could construct only 3 KM pairs. This resulted in 501 = 83 \times 6 + 3 KM pairs. To construct the non-mated (KNM) pairs of images, we paired the first replicate of each shoe with five randomly selected shoes from a different pair of the same size. Altogether, we constructed 420 = 42 \times 2 \times 5 KNM pairs of images. We proceeded in the same way when constructing the KM and the KNM pairs of images in the testing dataset, and produced 216 and 180 test KM and KNM pairs, respectively.

We followed the steps in Section 3.2 to implement the algorithm and to then extract six similarity features from the local circular matching areas. Features are combined
into a similarity score using random forest on the training dataset.

Suppose that the toe and/or heel regions are visible for $Q$ and $K$ so that the examiner can coarsely align the two images, for example by making the toe area the top and heel area the bottom of the images. If the toe or the heel areas of $Q$ are not visible, then the examiner can still select a circular area in $Q$ and plausibly matching areas in $K$ to initiate the comparison. Here, we suppose that $Q$ and $K$ are roughly aligned toe to toe at the top of the image. Selection of the three circular areas of interest in $Q$ can be performed manually (for example, by an examiner looking at the image on a computer) or automatically. If the latter, it may be reasonable to focus on regions of the outsole that are most likely to be in contact with the floor, and where we would expect to find the most wear and tear. Those regions include the ball of the foot and the heel, in a typical complete print. Consider the impression on the left panel of Figure 12. The three circles we select in $Q$ are located on the left and right upper half of the outsole image and on the heel area as follows: In image $Q$, define $q_1$ as a circle with center located on the 30th quantile and 75th quantile of the range of $x$ and $y$ coordinates, and with radius equal to 50-pixel units. The other two circles are placed in the upper right quadrant and on the bottom left quadrant of the impression. Circle $q_2$ on the bottom left quadrant has center at the 20th and 25th quantiles of the $(x,y)$ coordinates, and the third circle $q_3$ is located in the upper right quadrant and has a center on the 75th and 65th quantiles of the $x$ and $y$ coordinate ranges. Circles $q_2, q_3$ have the same radius as circle $q_1$. When $Q$ is a partial impression, the location of the three circular areas can be adjusted to fit within the latent impression.

To find the closest matching circles in $K$, in principle, we would need to consider 30 to 40 candidate circular areas that cover the entire impression, when the radius for $Q$ is 50 units and for $K$ is 65 units as in our study. Since each circle-to-circle comparison takes between 1 and 2 minutes of computing time where circles in $K$ contain around 400-500 pixels, comparing two impressions can take up to an hour on a server with ten cores running Ubuntu. By confining the search for the best matching circles in $K$ to quadrants, as shown in Figure 12, we speed up the comparison significantly, by over 75%, to 10-15 minutes per comparison. Even if we confine searches to quadrants, the union of candidate circles should cover the entire target areas on impression $K$.

The proposed method produces results such as those displayed in Tables 4 and 5 and
Some similarity features are associated with each of the circle-to-circle comparisons, and some are associated with the geometric arrangement of the three circular areas. Features such as clique size, proportion of overlapping points in \( k^* \) and \( q \), median of Euclidean distance between OP after \( q \) and \( k^* \) are aligned, are extracted for each of the three circle-to-circle comparisons. To these, we add a feature including the distances between the centroids of the circular areas that are computed from the geometric arrangement of the circles on the impressions. In all, we consider six features that can be combined to quantify the degree of similarity between \( Q \) and \( K \). Ideally, these features will take on different values when the comparison involves known mated and known non-mated pairs of shoes.

![Figure 13.: Values of the six similarity features among KM and KNM pairs of images in the training dataset. With the exception of SD of rotation angle, other values are averaged over three circles.](image)

Figure 13 shows the distribution of the values of the six similarity features (five of them averaged over the three circular areas and the standard deviation over three rotation angle estimations) in pairs of mated and non-mated impressions. Although the distributions of the features among mated and non-mated impressions overlap, they have different median and range. For example, the values of the standard deviation of rotation angle among KM impressions tend to be smaller than the SDs observed among KNM pairs, as we would expect.

![Figure 14.: Variable importance from the trained random forest algorithm with 10-fold cross-validation.](image)

The goal is to quantify the similarity between two outsole impressions and to develop a classification algorithm that can accurately determine whether two impressions
may have been produced by the same shoe. Since no single feature is discriminating enough, we combine the six similarity features computed into a single score via learning algorithms. We trained several learning algorithms including a random forest (RF), Bayesian additive regression trees (BART), a support vector machine (SVM), and a k-nearest neighbor (KNN). We then tested the predictive performance of each algorithm on the same test set of pairs of images with known source. Most of the algorithms exhibited reasonably low false positive and false negative rates when determining the source of a pair of images. We decided to use the RF in the analyses that follow because it had good predictive power, but is also fast and results in a similarity score that is interpretable in the context of the problem.

Figure 15.: Predicted random forest score (class probability for class KM) on the training comparisons (left panel) and on the test comparisons (right panel)

In the training phase, we implemented 10-fold cross-validation and computed an internal (to the training data) classification error. The RF algorithm produces a ranking of importance among the six similarity features, shown in Figure 14. The most important variable is the median distance among OP after alignment of the images has been accomplished. The second most important feature is clique size. After that, % Overlap in $Q$ and $K$, and SD of the rotation angle follow in terms of discriminatory power. The least important variable is the difference in distances between centers of circular regions in $Q$ and $K$. This is because we confine the search areas at the beginning of the comparison. When we search for matching areas with no restrictions, the feature we denote Diff. in triangle $Q$ and $K$ plays an important role.

Figure 16.: Credible intervals (lines) of the predicted score (black dots) from the BART in the test comparisons

We only report the out-of-bag classification error estimate, obtained by applying the
algorithm to the independent testing dataset. The class probability for KM produced
by the RF is used as the final similarity score based on the six similarity features.
Figure 15 shows the estimated similarity scores for the KM and KNM for both 10-fold
CV results from the training set (left panel) and from the test dataset. The RF was
able to learn how to distinguish KM and KNM pairs of images in the training set, so
the distributions of score values in the two groups do not overlap, as shown in the left
panel of Figure 15. When the trained RF is tested in the hold-out test set, then the
two score distributions are less well separated, as should be expected. The RF score
ranges between 0 and 1, and the empirical densities of scores in test set obtained for
the KM and KNM pairs of images have modes around 0.85 and 0.1, respectively. An
alternative approach is to compute a similarity score the BART (Bayesian Additive
Regression Trees) algorithm. The advantage of BART is that it permits calculating a
credible interval of the predicted similarity score. We randomly picked 50 KM and 50
KNM in the test comparisons and showed a credible interval of the predicted score
to show the variability of the score in Figure 16. The variability of the score from the
mated comparisons gets wider as the computed score is close to zero. This also happens
when we consider KNM pairs of images and the score approaches 1.

To evaluate the classification performance of the algorithm we propose, we also
implemented two other methods, phase-only correlation (POC) and Fourier-Mellin
transform correlation (FMTC), which showed good performance when comparing high-
quality outsole images in [26]. POC is highly sensitive to rotation, so when we use it
to compare two outsole images, we have to adjust the rotation angle between the two
impressions. We used two approaches to compute the rotation angle between the two
impressions: (1) using intensity-based image registration with MATLAB (POC-R), and
(2) by detection of the principal axis of the shoe outsole impression (POC-P). Given
the good discrimination performance of POC as reported by Richetelli et al. [26], we
explored including POC-R as an additional feature in the RF. This version of the RF
that includes POC-R as an additional feature is denoted RF-POC-R. We note that
Richetelli et al. [26] carried out a manual alignment of images and that non-mated
shoes had different outsole patterns [29]. This may explain why POC exhibits such
good performance in their case.

Since FMTC is invariant to scale, rotation, and translation, we would expect the
method to work well when comparing outsole impressions, but that was not the case
here. POC-P does not fare much better, probably because the estimate of the rotation
angle obtained by comparing the tilt of the two major axes in the impressions is not
precise enough. Of the three, the most discriminating approach appears to be POC-
R; we conclude that POC-R is an effective classifier when two impressions are of high
quality and correspond to the same shoe. POC is also sensitive to scale – this is another
limitation when one of the outsole impressions in the comparison is a partial print from
which it is difficult to estimate size accurately.

The assessment of the performance of each method is based only on the comparisons
among pairs of impressions in the test set, none of which were used to train the RF.
We calculated the ROC curves, shown in Figure 17, for each of the five classifiers
introduced earlier, plus a simple classifier that relies exclusively on the proportion
of overlapping points in Q. The ROC curve of POC-R increases rapidly toward the
upper left corner of the plot but then plateaus at a moderate sensitivity value. The
RF-6 behaves similarly, but for the same false positive rate (FPR), it reaches a higher
sensitivity. When we use POC-R as an additional feature in the RF, we achieve the
best classification performance, as shown in Figure 17, at least when both images are
of high quality.
Figure 17.: ROC curve on test set using six classification methods: Random forest with seven features, (RF-POC-R), random forest with six features (RF-6), percent of overlapping points on Q (% Overlap on Q), phase-only correlation with image registration (POC-R), phase-only correlation with principal axes alignment (POC-P) and Fourier-Mellin transform correlation (FMTC).

Table 9.: Diagnostic values from the ROC curve

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>EER</th>
<th>Opt. threshold</th>
<th>FPR + FNR</th>
<th>FPR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF-POC-R</td>
<td>0.970</td>
<td>0.089</td>
<td>0.540</td>
<td>0.157</td>
<td>0.050</td>
<td>0.107</td>
</tr>
<tr>
<td>RF-6</td>
<td>0.913</td>
<td>0.189</td>
<td>0.600</td>
<td>0.328</td>
<td>0.078</td>
<td>0.250</td>
</tr>
<tr>
<td>% Overlap on Q</td>
<td>0.820</td>
<td>0.250</td>
<td>0.870</td>
<td>0.477</td>
<td>0.222</td>
<td>0.255</td>
</tr>
<tr>
<td>POC-R</td>
<td>0.775</td>
<td>0.255</td>
<td>0.094</td>
<td>0.368</td>
<td>0.039</td>
<td>0.329</td>
</tr>
<tr>
<td>POC-P</td>
<td>0.728</td>
<td>0.371</td>
<td>0.060</td>
<td>0.649</td>
<td>0.089</td>
<td>0.560</td>
</tr>
<tr>
<td>FMTC</td>
<td>0.680</td>
<td>0.395</td>
<td>0.056</td>
<td>0.733</td>
<td>0.094</td>
<td>0.639</td>
</tr>
</tbody>
</table>

Table 9 includes additional information regarding the ROC curve of the six classifiers we compared. The first two columns in the table show area under the ROC curve (AUC), and equal error rate (EER), which is the error rate when FPR and false negative rate (FNR) are the same. The best classifier will have the highest AUC and lowest EER. The optimal threshold is the value of the similarity score that results in a minimum of the sum of FPR and FNR in the ROC space. Overall, in this particular application, RF-6 and RF-POC-R exhibit the best classification performance, at least in terms of AUC and EER. POC-R shows the lowest FPR, but at the expense of a high FNR. This means that the POC-R classifier tends to conclude different sources too often for pairs of impressions left by the same shoe. The classifier that shows the worst performance is FMTC, a surprising finding given that FMTC is invariant to translation, scale, and rotation angle.
4.4. Varying the number of circles

In Section 3.1, we defined three circular regions from which we extract the signature of $Q$. In this Section, we investigate the impact of the number of circular regions in $Q$ on the accuracy of the classifier. We consider two, three, four, five, and six circular regions in $Q$, and search for the corresponding areas in $K$. Additional details are found in the Appendix A.

When we define only one circular region in $Q$, the SD of rotation angle can not be calculated, so we did not consider this case. For the cases where two to six circular regions were selected in $Q$, we carried out the same comparison steps described in Section 4.3. We trained the random forest with 10-fold cross-validation using the features extracted from two, three, four, five, and six circular regions. Using the trained RFs, we then predicted whether pairs of KM and KNM pairs of images in the test set had the same or a different source. For example, if the RF is trained with data from four circular areas, then predictions are also based on four circular region comparisons in the test set.

Figure 18.: Prediction error rates are obtained with the cut-off of 0.5 in test comparisons when varying the number of circles.

Figure 18 shows the false negative rate (FNR), false positive rate (FPR), and the overall prediction error in the test comparisons from the RF trained on features extracted from two to six circles. The threshold we used was 0.5. Both FNR, FPR, and the overall prediction error decrease as we add more circular regions to determine the signature of $Q$, but all plateau in this example as we reach five.

Figure 19.: ROC curves show the prediction performance of the RF trained on two, three, four, five, and six circular regions.

The diagnostic values from the ROC curves shown in Figure 19 are in Table 10. As the number of circular regions in $Q$ increases, the AUC of the ROC curves also increases. Interestingly, the optimal threshold is also decreases and the surface area in $Q$ used in the comparison increases. When the number of circles in $Q$ reaches six,
the optimal threshold was 0.329, lower than the cut-off of 0.5. This suggests that by increasing the area in $Q$ used in the comparison we also increase the risk of over-fitting.

Table 10.: Diagnostic values from the ROC curves displayed in Fig. 19. Sensitivity and specificity are calculated with the optimal threshold as a cut-off in test comparisons.

<table>
<thead>
<tr>
<th># of circles</th>
<th>AUC</th>
<th>Opt.threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.919</td>
<td>0.640</td>
<td>0.900</td>
<td>0.782</td>
</tr>
<tr>
<td>3</td>
<td>0.913</td>
<td>0.600</td>
<td>0.922</td>
<td>0.750</td>
</tr>
<tr>
<td>4</td>
<td>0.940</td>
<td>0.460</td>
<td>0.906</td>
<td>0.884</td>
</tr>
<tr>
<td>5</td>
<td>0.960</td>
<td>0.442</td>
<td>0.911</td>
<td>0.907</td>
</tr>
<tr>
<td>6</td>
<td>0.960</td>
<td>0.329</td>
<td>0.883</td>
<td>0.921</td>
</tr>
</tbody>
</table>

More circular regions included in the signature of $Q$ appear to improve accuracy and robustness of the classifier. However, the quality of image in the sub-regions matters. For the set of images we used in the analyses we present here, three or four circular regions strike a good balance between the bias and the variance of the classifier.

5. Discussion and future work

We have developed a method to compare two footwear impressions, that uses graph theoretic ideas and machine learning. The algorithm relies on the coordinates of the edge pixels of the outsole images. These pixels may represent a class or a sub-class characteristic, a unique RAC and any other marks that result from wear and tear. In this light, the similarity score produced by the random forest can discriminate between pairs of impressions that have a common or a different source effectively, even though all impressions were made by shoes that share class characteristics such as brand, model, and (approximately) wear time.

The comparison algorithm we propose has a wide range of applications, in terms of pattern matching and similarity of two 2D or 3D images. The alignment step, which we carry out using the idea of a maximum clique, is critical. Once images are aligned, it is possible to select features to then quantify the similarity between the two images. To extend the method and compare 3D images, one plausible approach is to first align the images using the $x, y$ coordinates corresponding to the plane and then extract features by incorporating the depth coordinate $z$. The current guidelines for evaluation of shoeprint evidence rely on the visual comparison carried out by an experienced human examiner who then makes a categorical conclusion on the seven-number scale adopted by SWGTREAD [8]. The algorithm we propose could be useful to align images using features chosen by the examiner and to calculate a similarity score more objectively. While we have focused here on footwear evidence, there are many other situations in forensic practice that require comparisons of patterns. These include the analysis of evidence such as fingerprints, surveillance photos, handwriting, tire treads, and many others. In most of these cases, the evidence consists of one or more 2D images.

While the similarity score we propose obtaining appears to have an excellent discriminating ability, there are some potentially challenging issues that deserve further investigation. The challenge arises when the pattern includes a small number of elements that repeat throughout the outsole. An example is shown on the left panel of images in Figure 20. One other challenge is the case of two non-mated impressions where one image is oversaturated, as the pair of images in the right of Figure 20. This
can occur when the outsole is smooth, with no apparent patterns in the outsole, or when the degree of wear and tear is extreme, or when the shoeprint at the crime scene is not well recorded. The solution to this problem might require a careful initial alignment step to precede detailed comparisons, something which we plan to investigate in the future.

In this manuscript, we have used Nike Winflow athletic shoes to both develop and test the algorithms we propose. An obvious question is whether the same fitted model would perform well when the pair of impressions to be compared were produced by shoes of different brand and model. We do not have images from a sufficiently broad variety of footwear brands to address this question but were able to make use of the images collected from Adidas shoes included in the study carried out in CSafe. Adidas shoes, like the Nike shoes we have used in this work, share class and wear characteristics. Recall that the overall incorrect classification rate on Nike test impressions of the RF-6 trained on the same type of shoes was about 18%. When the Nike-trained RF was used to classify Adidas shoes, the error rate almost doubled, to 30%. However, when we re-trained the RF using a mix of Nike and Adidas shoes, performance on the testing samples improved to 85% accuracy. The tentative conclusion we draw is that a model trained on one or a limited variety of outsole patterns is unlikely to perform well when implemented to classify a pair of images of a different pattern. This then begs the question of whether the most accurate approach to use in practice will involve a large number of specialized classifiers or a small number of classifiers that can apply more broadly. As mentioned earlier, the databases to allow this type of investigation are not yet available.

Supervised learning algorithms perform better when the combination of feature values is different in items that belong to different classes. In many applications, features are suggested by the nature of the problem. For example, Park and Carriquiry [22] classify glass fragments into the same and different pane classes and use the measured concentration of 18 chemical elements as the features in the learning algorithm. In other applications, however, the set of features must be defined by the user. In almost all applications, it is difficult to assemble the training dataset that will result in an algorithm with good out-of-sample classification properties. In the forensics context, additional difficulties include the fact that the images to be compared may have been obtained at different times in the life of an object, and that at least one of the images is sometimes of poor quality. Therefore, before we can determine whether either one
of the RF classifiers will be useful in real casework, much work remains to be done.

Acknowledgements

We are grateful to two anonymous reviewers and an editor for their extensive and constructive comments, which greatly improved our manuscript. Dr. Hariharan K. Iyer at the National Institute of Standards and Technology (NIST) has generously shared his ideas and suggestions and we have learned much from him. We also thank Dr. Eric Hare from Omni Analytics for his contribution to the computing problems in the analysis. Lastly, we are grateful to the footwear team in the Center for Statistics and Applications in Forensic Evidence (CSAFE) who collected the database of outsole impressions we use in the paper.

References


