Forensic Science International 331 (2022) 111126

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

The effect of image descriptors on the performance of classifiers of footwear outsole image pairs

Soyoung Park[a,1], Alicia Carriquiry[b,*,2]

[a] Department of Statistics, Pusan National University, Busan, South Korea
[b] Center for Statistics and Applications in Forensic Evidence (CSAFE), Iowa State University, USA

ARTICLE INFO

Article history:
Received 22 June 2021
Received in revised form 12 November 2021
Accepted 26 November 2021
Available online 29 November 2021

Keywords:
Pattern evidence
Degraded impressions
Footwear forensics
Image analysis
Random forests

ABSTRACT

Shoe prints are commonly found at the scene of a crime and can sometimes help link a suspect to the scene. Because prints tend to be partially observed or smudgy, comparing crime scene prints with reference images from a putative shoe can be challenging. Footwear examiners rely on guidelines such as those published by SWGTREAD [1] to visually assess the similarity between two or more footwear impressions, one reason being that reliable, quantitative methods have yet to be validated for use in real cases. To help in the development of such methods, we created a study dataset of images of outsole impressions that shared class characteristics and degree of wear and that were subject to a specific type of degradation. We also propose a method to quantify the similarity between two outsole images that extends the capabilities of MC-COMP [2]. The proposed method is composed of three steps; (1) extracting image descriptors, (2) aligning images using the maximum clique, (3) calculating similarity values using two different classifiers; (a) degree of overlap between the two images, and (b) a score produced by a random forest. To explore the performance of the algorithm we propose, we compared degraded, crime scene-like images to high-quality reference images produced by the same or by different shoes. Even though comparisons involved matches or very close non-matches, and one of the images was blurry, the algorithm shows good source classification performance.

1. Background

In the forensic pattern comparison disciplines including footwear impression analysis, examiners typically rely on visual, subjective assessment of the similarities between two items. In recent years, however, there has been a push to develop methods to quantify similarities and differences between a questioned item found at a crime scene (e.g., a fired bullet or a print from a shoe) to a reference item (bullet fired with the suspect's gun, the suspect's shoe).

Prints left by the shoes of the perpetrator are commonly found at the scene, but the footwear evidence is not always useful, either because it was not recorded, or because the information is not enough to lead to an identification of a specific shoe. In many cases, the information that can be obtained includes the brand, the model, and the size of the shoes that left the print, but unless one or more of those characteristics is rare, it is typically not possible to conclude that it was the perpetrator's shoe in particular that was present at the scene. To add to the challenge, shoe prints are subject to background effects such as the pattern or other attributes of the substrate (tile designs, blood consistency, dirt, snow) and can be more or less deformed and smudged depending on the activity of the perpetrator at the time the print was made and on whether the integrity of the scene was well preserved.

The current practice of evaluating shoeprint evidence consists in the visual comparison of two or more prints by a trained human examiner, who can rely on the guidelines published by SWGTREAD [1]. The conclusion scale that examiners often use includes seven different levels to represent the strength of the evidence in favor of "same source". The report from the National Research Council (2009)

* Corresponding author.
E-mail addresses: soyoung@pusan.ac.kr (S. Park), alicia@iastate.edu (A. Carriquiry).
[1] Park's work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C100711111).
[2] Carriquiry's work was partially funded by Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

https://doi.org/10.1016/j.forsciint.2021.111126
0379-0738/

[3] and the 2016 report from the President's Council of Advisors in Science and Technology (PCAST) [4] both highlight the need to develop objective and reliable methods for the quantitative analysis of pattern evidence. In the last several years, research groups in academia, industry and government have proposed new algorithmic approaches that address those needs [5–7].

The analysis of footwear impressions is particularly challenging for several reasons. First, impressions from the same shoe can vary depending on the gait and the weight of the person wearing the shoe, and also on the activity (walking, running, jumping, etc.) in which the wearer was engaged when the print was made. While class characteristics such as make and model might be readily obtainable from a crime scene impression, small, individual characteristics known as randomly acquired characteristics (or RACs) may not be visible in impressions that are blurry or contaminated by background. Researchers wanting to develop algorithms to analyze footwear evidence have limited access to databases with crime scene-like images that are large enough and for which ground truth is known. Finally, crime scene investigators are not always well-trained in the collection of footwear impressions from the scene, and therefore, the evidence suffers from the effects of poor illumination, less than ideal camera angles and other limitations.

Without a well designed, reasonably large and accessible database of images that mimic real evidence, and for which we know ground truth, it is difficult for researchers to develop good methods for the comparison of footwear images, and for evaluators to carry out black-box studies to assess the performance of human examiners. Notable exceptions are the data sets constructed by Kortylewski et al. [8] who uploaded a database of crime scene-like images to the public domain, and Richetelli et al. [9] who collected outsole impressions made on blood and dirt. Cervelli et al. [10] compare footwear retrieval systems on synthetic and real shoe marks from the crime scene and propose a matching method based on the Mahalanobis distance. Park and Carriquiry [2] produced artificially degraded images of outsoles to test the performance of an algorithm to quantify the similarity between two impressions. They found that even with a single image descriptor, the algorithm resulted in the correct identification of the source of a print, at least in the conditions under which it was tested.

Fig. 1 shows an example of a crime scene-like image reproduced from the database in [9]. The impression was made on acetate with a dusty outsole, and was then scanned. As is often the case when the print is made by a dusty outsole, elements of the outsole pattern are likely to be smudged and no longer distinguishable. Dust may cause the expansion of pattern elements, making their edges blurry.

There are multiple different approaches for recording footwear images [11]. In our lab, we rely on a step-on scanner that produces two-dimensional (2D) gray-scale images of the shoe outsole and that is manufactured by Everspry (https://www.shopevident.com/product/everos-laboratory-footwear-scanner). Several other labs in the US rely on the EverOS scanner as well. When an individual wearing the study shoe steps on the clear surface of the scanner, the scanner detects the weight distribution of the wearer and produces an image of the portions of the outsole that make contact with the scanning surface. The image is then stored as a 2D gray-scale image that can be analyzed using standard image analysis software.

The objectives of the study we discuss here are two fold. First, we construct an experimental dataset that includes a sequence of 2D images of the same sample of shoes, that were captured with increasing levels of blurriness. In every case, the source of the impression is known. Second, we extract several different image descriptors and use them – either individually or in combination – to classify pairs of images into same or different source categories. The shoes in the database share class characteristics including brand, model, size and degree of wear. Therefore, our non-mated comparisons are made using only close non-matches.

Fig. 1. An example of a crime scene-like print ('280RA_E_Background_Subtracted_flip.tif') with dust scanned on the surface of acetate from Richetelli et al. (2017) ([9]).

2. A study database of blurred footwear impressions

We randomly selected 12 pairs each of Nike Winflow 4, sizes 8 and 10, from a collection of about 80 pairs of shoes of the same make, model and sizes that were purchased in 2018. The 80 pairs of Nike shoes were worn by volunteers who participated in an earlier study, and who were asked to walk at least 10,000 steps per week in them, over a period of approximately six months.

Using the 24 pairs of Nike shoes, we created a database of increasingly blurry 2D images with the EverOS scanner. To decrease the sharpness of the scanned impressions, we interposed sheets of paper between the outsole and the scanning surface of the step-on scanner. By increasing the number of sheets of paper between the shoe and the scanner, we control the degree of blurriness in the scanned impressions. For each shoe, we obtained scans using zero, two, four, six, eight, and ten sheets of paper placed on top of the scanning area. We use the term *level of degradation* to refer to the six different "treatments". At all levels of degradation, we obtained three replicate images of the shoe. In all, the database includes 18 images obtained from each of 48 shoes for a total of 864 images. When scanning the shoes, there was only one person who wore all shoes.

Fig. 2 shows a sequence of impressions obtained from the same shoe, but with increasing degree of blurriness. As more sheets of paper are placed between the scanner and the shoe, the pattern of the outsole loses sharpness and edges of pattern elements become blurry. When the image is scanned with ten sheets of papers on top of the scanning surface, the elements of the pattern (pentagons, lines) are more likely to be expanded, smudged, and less well defined. It becomes difficult to distinguish the striped lines on the right

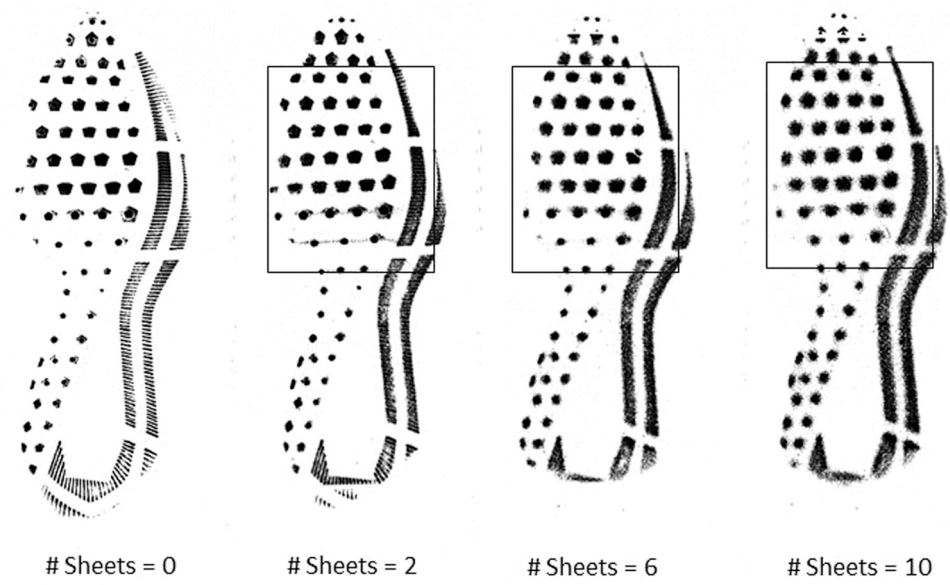

Fig. 2. Increasingly blurred impressions from a right side Nike Winflow 4 shoe. The same shoe was scanned with 0, 2, 6 or 10 sheets of paper on the scanning surface of the step-on EverOS scanner.In the analyses that follow, we ignore anything in the images outside of the marked rectangle, to simulate the situation where the impression is partially observed.

of the outsole or the shape of the polygons in the level-10 images. These artificially blurred impressions resemble the impressions scanned on dust in Fig. 1. The rectangles mark the area of the impression that is actually used in the comparisons; to simulate the case where the crime scene image is observed only partially, we ignore the data outside of the rectangle.

3. An algorithm to aid in source determination

3.1. Study design

In real case work, footwear evidence found at the scene of a crime is often degraded in some way, or is partially observed. Reference impressions obtained from the suspect's shoe or from some other known source, however, are typically high-quality images that include enough detail to permit identifying small elements such as RACs in the outsole. To ensure that the study we describe is somewhat realistic, we construct pairs of images for comparison that include one blurry and partially observed impression and one good-quality impression. We let Q denote the questioned or low-quality image and K denote the known-source or good-quality image. We use the 864 images in the database described in Section 2 to construct pairs of *mated* images, where both images are obtained from the same shoe, and pairs of *non-mated* images, where two close non-matching shoes are involved. Every comparison, whether involving mated or non-mated images, involves a partially observed, blurry image of the outsole and a clear, good-quality image.

Fig. 3 shows two pairs of impressions scanned at level 0 (images labeled (a) and (c)) and at level 10 (images labeled (b) and (d)) using

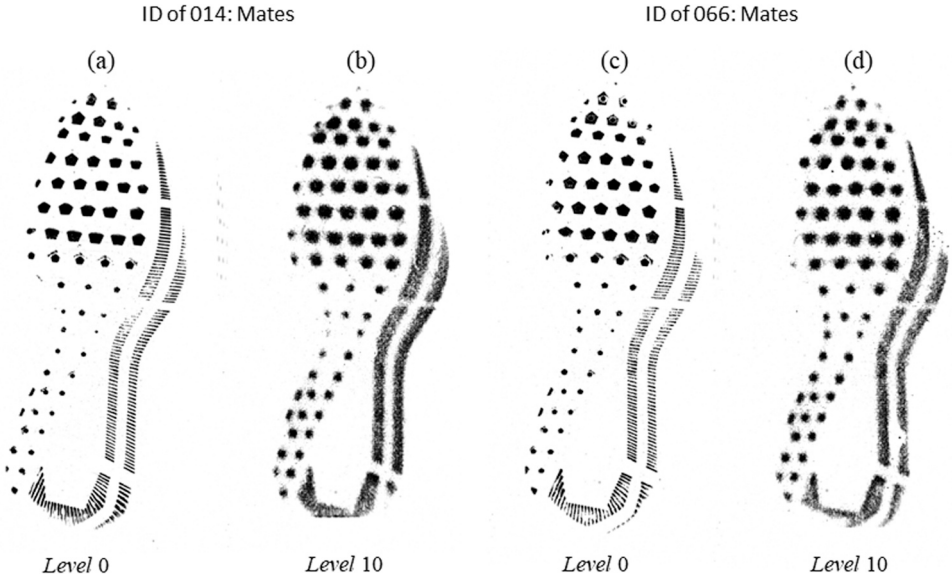

Fig. 3. Two pairs of mated pairs of images. The left two images were obtained using the shoe with ID number 014 and the right two images were obtained with shoe ID 066. In both pairs, the left-most impression corresponds to K and the right most impression corresponds to Q. The Q images were made with 10 sheets of paper placed between the shoe and the surface of the scanner.

Table 1
Number of comparisons of each type included in the study.

	Degradation level for Q					
Class	0	2	4	6	8	10
Mated comparisons	144	144	144	144	144	144
Non-mated comparisons	144	144	144	144	144	144

two shoes from the database (IDs 014 and 066). Images (a) and (b) were made with the same shoe ID 014, and both images (c) and (d) were made with shoe ID 066, meaning that each of the two pairs of images is mated. Images (a) and (c) are good-quality images and play the role of *K*, whereas images (b) and (d), which reflect only a portion of the outsole and have blurry edges, play the role of *Q*. A non-mated comparison would include, for example, images (a) and (d), or (b) and (c). By design, as mentioned earlier, the non-mated comparisons include only close non-matches with the same class characteristics such as brand, model, size and degree of wear.

We constructed a total of 1728 pairs of comparisons, half of which were between the same shoe, and half of which were between different shoes. Table 1 shows the number of mated (first row) and non-mated (second row) comparisons included in the study.

3.2. Feature descriptors

In image analysis, *features* (also called *points of interest*) are well defined, robustly extracted points in an image. Image descriptors play an important role in image registration (or alignment) and image comparison, and typically consist of edge, corner, blob, line and other types of groups of pixels. These features can be transformed to be more efficient and robust and to be scale and rotation invariant. Examples of such features and the type of descriptors on which they rely include SIFT (*blob*) [12], SURF (*blob*) [13], KAZE (*blob*) [14], ORB (*corner*) [15] and more.

The *SIFT* feature (scale-invariant feature transform) was introduced by Lowe (2004) [12], and is among the most popular image descriptors, at least in terms of usage. SIFT is robust to rotation and scale transformations, and to the illumination of an image. However, extracting the SIFT is computationally intensive.

The speeded-up robust feature *SURF* introduced by Bay et al. (2006) [13] is also robust to rotation and scale for image registration. While SIFT relies on Gaussian differences to approximate a Laplacian, SURF speeds up extraction time by using a box filter on an integral image (cumulative sums of pixel values to the left and above current location). Panchal et al. (2013) [16] showed that implementation of SURF is typically faster than implementation of SIFT, with no noticeable differences in performance.

The feature descriptor called KAZE was introduced by Alcantarilla et al. (2012) [14]. KAZE means “wind” in Japanese, and the name is meant to evoke the non-linear way in which the wind moves. As its name suggests, KAZE operates in a non-linear scale space, rather than the Gaussian scale space in which SIFT and SURF operate. KAZE features are extracted using nonlinear diffusion filtering, an adaptive filtering approach that differentially preserves the sharpness of pattern and noise. Because implementing the nonlinear filters requires solving sets of nonlinear differential equations, KAZE is more computationally demanding than SURF and comparable to SIFT [14].

The oriented FAST and Rotated BRIEF (ORB) [15] is a blended version of two feature extraction methods: FAST (Features from Accelerated Segment Test) [17] and BRIEF (Binary Robust Independent Elementary Features) [18]. Kulkarni et al. [19] argued that ORB outperforms SIFT and SURF when comparing outsole images.

In recent years, several research teams have carried out studies to compare the performance of the various feature descriptors that have been proposed. Some of the studies include [16,20,21].

Here, we focus on the comparison of pairs of images obtained either from the same shoe, or from different – but very similar – shoes. The goal was to determine whether one or more of SURF, KAZE, and ORB is robust, fast and efficient enough for use in forensic footwear comparisons, where at least one of the images in the comparison is degraded. In real applications, it is important to minimize the computational burden without compromising performance. Thus, to run our comparisons we chose the number of features that would strike a good balance between computational efficiency and prediction accuracy. We first carried out the comparisons using 500 features extracted using SURF, KAZE and ORB individually. In a second set of comparisons, we extracted 200 or 300 strong features using each of the three approaches, and then combined them into a single set of 500 or 600 features. The hypothesis we wished to test is whether there might be a gain in accuracy when combining approaches that exploit different attributes of the images. In summary, the six different approaches for feature extraction we tested were:

1. SURF(500)
2. KAZE(500)
3. ORB(500)
4. KAZE(200) & SURF(200) & ORB(200); *comb200*
5. KAZE(300) & SURF(300); *K-S-300*
6. KAZE(200) & SURF(200); *K-S-200*

For illustration, Fig. 4 shows the features extracted from the outsole of shoe ID 014 when the shoe is scanned at the best possible resolution. Each panel corresponds to one of the six methods described above. Fig. 5 shows the features extracted from an image of the same shoe, but when the shoe is scanned with ten sheets of paper placed on the surface of the scanner. Each type of descriptor has its own distribution of features, so there may be some advantages in combining features from different descriptors.

3.3. Alignment of two images

Features extracted using any of the algorithms described above can be represented by their coordinate values (x, y) in the images that we label *Q* (for questioned) and *K* (for known). To compare the two images, one plausible approach is to overlay one on the other and measure differences. We use an idea from graph theory, called the *maximum clique* [22,23]. The maximum clique is defined as the largest subsets of points in two images that share the same geometric arrangement. Consequently, for images *K* and *Q*, the maximum clique corresponds to the set of points that have the same pairwise distances. An advantage of the maximum clique is its invariance to rotation and scale.

One limitation of the approach is that identifying the maximum clique is an NP-complete problem that requires intensive calculations. If we have extracted 100 features from each of *Q* and *K*, there will be $\binom{100}{2} = 4950$ pairwise distances to compute in each image. Using those pairwise distances we then need to find the largest subset with the same set of distances between pairs of features (e.g., [24]).

3.4. Quantifying similarity between two aligned images

Once two sets of features are aligned, we measure the degree of similarity between the two sets.Suppose that we are able to identify variables or metrics that take on different values when the two images in the comparison are mated and when they are non-mated. If so, and if for a specific pair *Q* and *K* we obtain values of the metrics that indicate that the outsoles are “similar”, then it may be plausible to think that *Q* and *K* were made by the same shoe. This is the

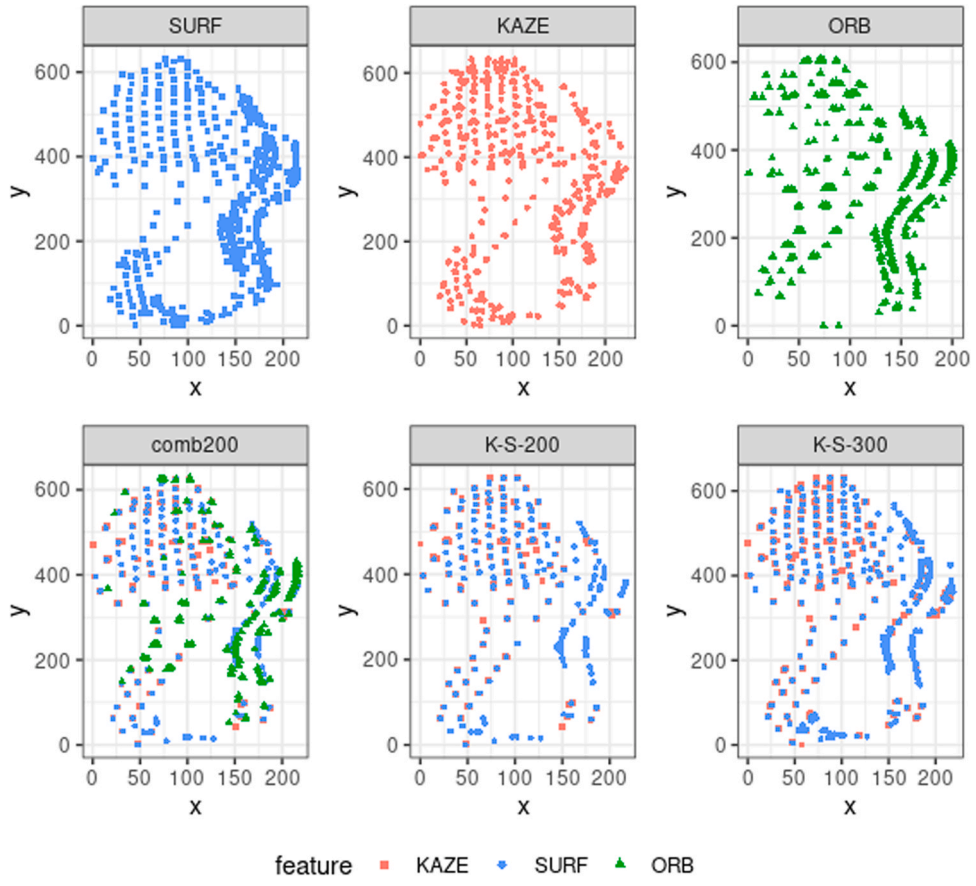

Fig. 4. Features extracted using six approaches using the same image of the outsole of the left shoe from the pair labeled ID 014. The six panels correspond to the case where *Q* was scanned at the best possible resolution.

approach we follow. To quantify similarity, we define three metrics that can be measured on the aligned images: .

1. The proportion of points (features) in *Q* and *K* that overlap, meaning that the difference in coordinates of the two points is within a pre-determined threshold (*% overlap*).
2. The size of the maximum clique (*clique size*).
3. The median value of the location discrepancy among overlapping pairs of points in *Q* and *K* (*median distance*).

A more detailed discussion can be found in [2].

3.5. R-package: Shoeprintr

To speed up calculations, rather than comparing the entire set of features in *Q* to those in *K*, we propose a down-sampling approach to reduce the number of interesting points in *Q*. We first divided *Q* into a regular grid with 100 sub-areas, and from each sub-area we randomly sampled one of the points of interest in the sub-area. This reduces the number of features in *Q* to 100 features that are well distributed on the entire image *Q*. We use the down-sampled set of features on *Q* and the entire set of features in *K* to calculate the maximum clique.

All of the calculations described in this manuscript can be implemented using the freeware `Shoeprintr`, an R-package developed by us. This package contains a function called *boosted_clique*, which implements a parallelized version of the maximum clique algorithm, and also includes functions to down-sample *Q*. Depending on the number of cores available, parallelizing the maximum clique calculations can reduce computing time by about 10-fold. With 500 features extracted from *Q* and *K*, and with down-sampling of *Q*, *boosted_clique* extracts the maximum clique in about 1–2 min when implemented on a Linux server with 16 cores running Ubuntu.

4. Results

4.1. Classifier based on %overlap

Using the same set of the mated and non-mated comparisons described in Section 3.1, we extracted features from *Q* and *K* to summarize the images. The number and type of points of interest were obtained using six different methods or combination of methods that were described in Section 3.2.

We first aligned the set of features extracted from *Q* and *K*, using the maximum clique approach. From each of the aligned pairs of images, we compute the value of the three similarity statistics described earlier. Initially, we focus on %overlap, and compute its value using all mated and non-mated pairs of images where *Q* is degraded at level 10. We repeated the same calculations for each of the six methods we list above and for an additional, independent approach that consists in using a phase-only correlation (POC) to quantify the similarity between the two images in a pair. The goal was to include a "standard" method to quantify similarity between two images that only relies on existing software.The POC-R is a POC value after two images are aligned using a built-in registration function called *imregform* in Matlab. The POC approach was discussed in [9] as a possible method to classify crime scene-like images.

Fig. 6 summarizes the results. The panels along the main diagonal show estimated densities for %overlap scores calculated on mated (orange) and non-mated (teal) pairs of images with *Q* degraded at level 10. Little or no overlap observed between the orange and teal

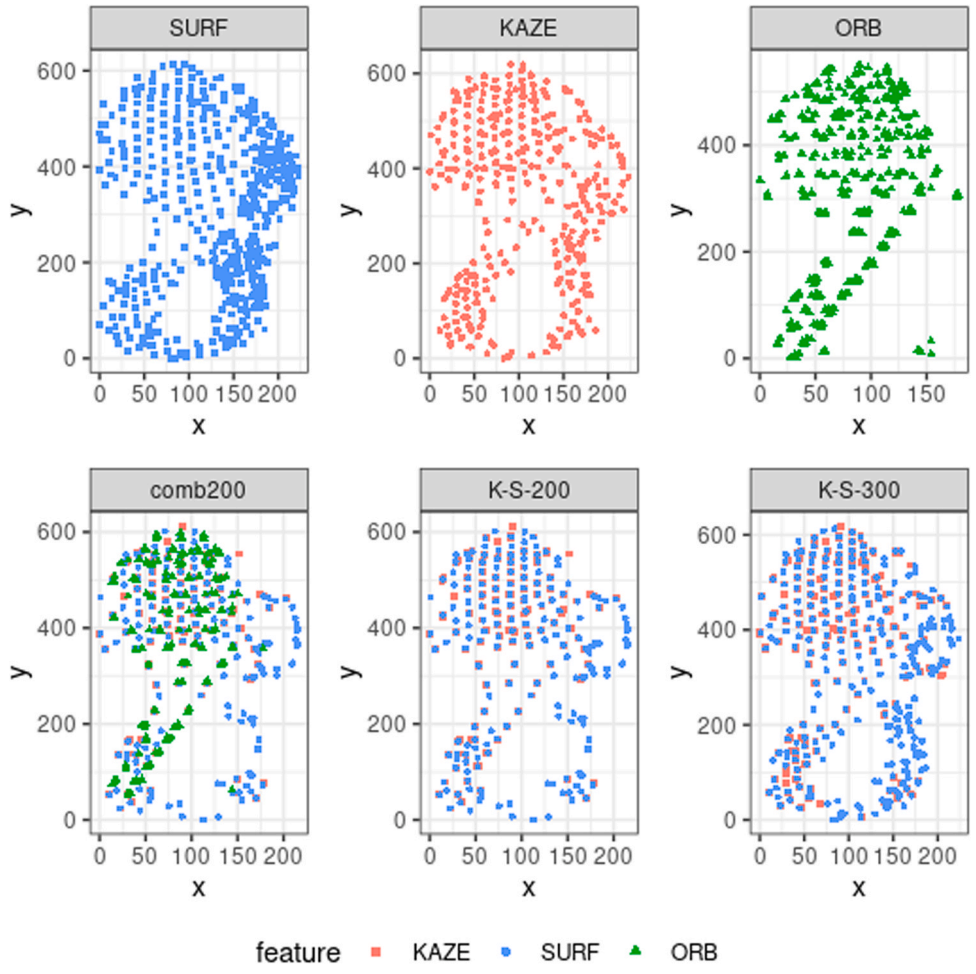

Fig. 5. Features extracted using six approaches using the same image of the outsole of the left shoe from the pair labeled ID 014. The six panels correspond to the case where the Q image was scanned at degradation level 10.

distributions is an indicator that%overlap can be used to estimate the probability that a pair of impressions was made by the same or by different shoes. In the ideal case of no overlap, %overlap would be able to discriminate perfectly between mated and non-mated pairs of images. We do observe some separation between the two distributions when features extracted using SURF, KAZE or the combinations of descriptors were used for alignment. When the % overlap scores relied on ORB features however, the discrimination power of the score disappeared. This is most likely due to the fact that ORB produces descriptors that are based on corners in the outsole pattern, which are blurred in the Q images we degraded. We included POC-R for comparison, and found that in this particular situation, POC-R fails to distinguish between mated and non-mated pairs of images. A reason for this poor performance is that POC is not invariant to rotation of one of the images, so to implement POC, we first need to estimate a rotation angle required to align the images. When Q is blurred, it is difficult to accurately estimate the rotation angle.

The off-diagonal panels below and above the densities in Fig. 6 provide information about the relationships between the %overlap scores that are obtained using different sets of features. As would be expected, there is a high positive correlation among the scores based on SURF and KAZE features (and on the corresponding combinations) both for mated and non-mated pairs of images. The pairwise correlations that involve ORB features or POC-R are lower, as we might have anticipated given that ORB features and POC-R rely on different attributes of the outsole pattern.

Fig. 7 shows the receiver operating characteristic (ROC) curves restricted to the subset where sensitivity of the %overlap statistic as a classifier exceeds 0.75. Each panel corresponds to a different level of blurriness of Q. As would be expected, the performance of the classifier decreases as Q becomes more degraded, regardless of the image descriptor (or combination) used in the analysis. Fig. 8 shows the corresponding areas under the ROC curves. Consistent with the score distributions shown in Fig. 6, SURF and KAZE features appear to have better discrimination ability than ORB. In fact, even at Level 10 of degradation of Q, the classifiers based on SURF and KAZE (or a combination) appear to distinguish between mated and non-mated images reasonably well. The combinations of features extracted by SURF, KAZE and ORB also lead to a good classifier, but this is probably due to the contribution of the SURF and KAZE features.

Table 2 shows the equal error rates (EER), the point on the ROC curve where the probability of a false positive and the probability of a false negative are the same. At Level 0, the %overlap classifier that relies on SURF has the lowest EER of about 0.04. The classifiers based on the combinations of SURF and KAZE features, as well the classifier based on KAZE features alone, all exhibit EERs below 10%. These are also the best performing classifiers when the Q image is blurry, except that the best performer in that case is the classifier based on the combination of SURF, KAZE and ORB features.

4.2. Alternative classifier based on three similarity metrics

So far we have presented the results we obtained when only the %overlap metric was used as a classifier. Here, we re-introduce the other two metrics – size of the maximum clique (MC) and median distance between overlapping pixels – and construct a similarity

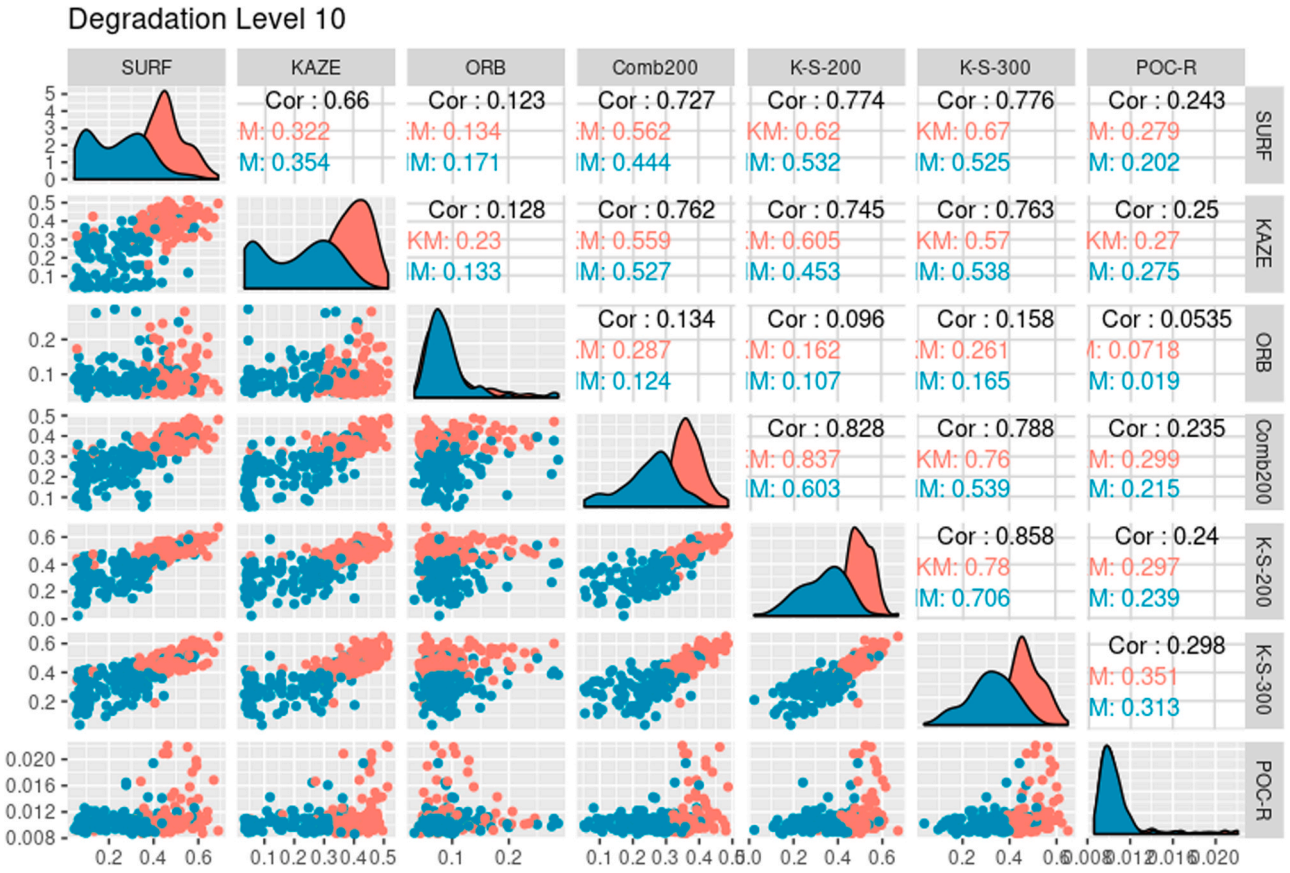

Fig. 6. Density estimation and correlation of the %overlap statistic computed using different features.

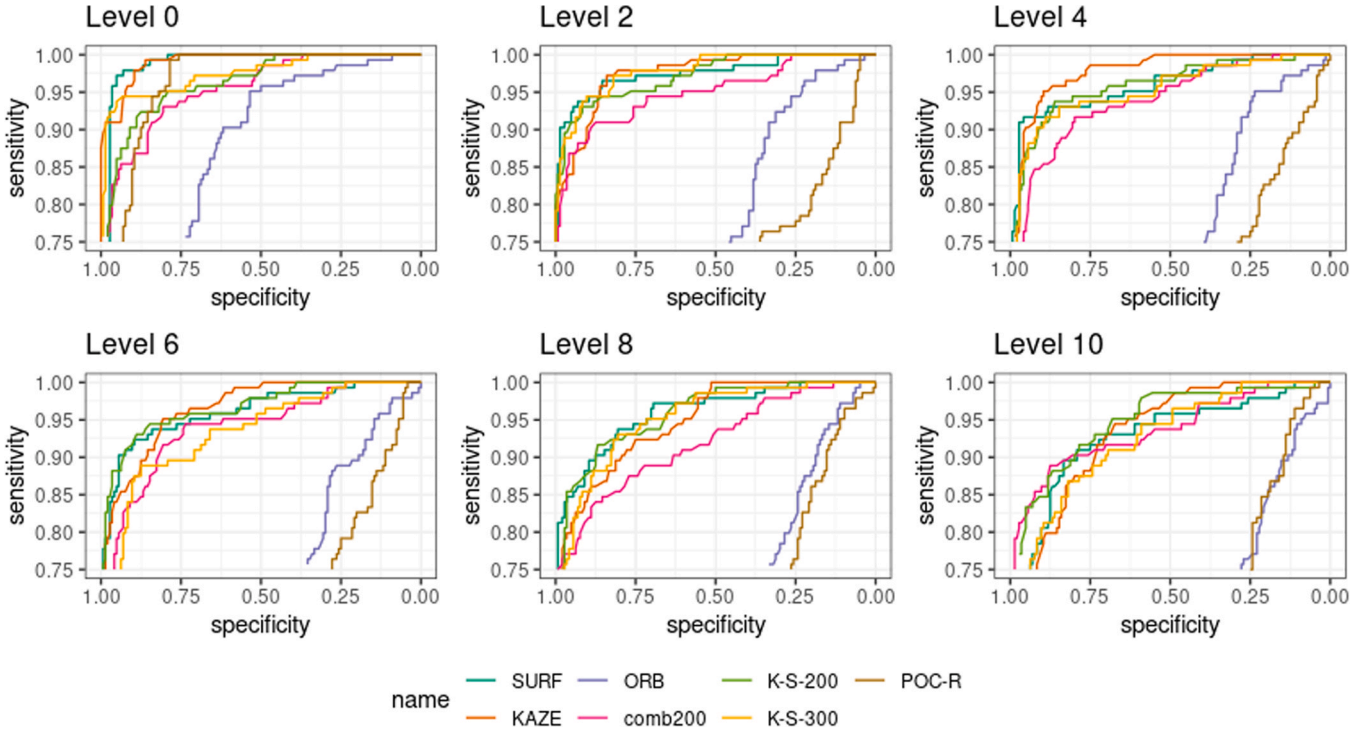

Fig. 7. Area under the ROC curve for each level of blurriness of *Q*.

score using a random forest to combine the three similarity statistics. As before, we use the six feature extraction methods described in Section 3.2 to construct six sets of similarity statistics. For each set, we re-train a RF using a subset equal to 70% of the mated and of non-mated pairs of images, and test the classifier using the rest (30%) of the paired images. In this analyses, we excluded the POC-R

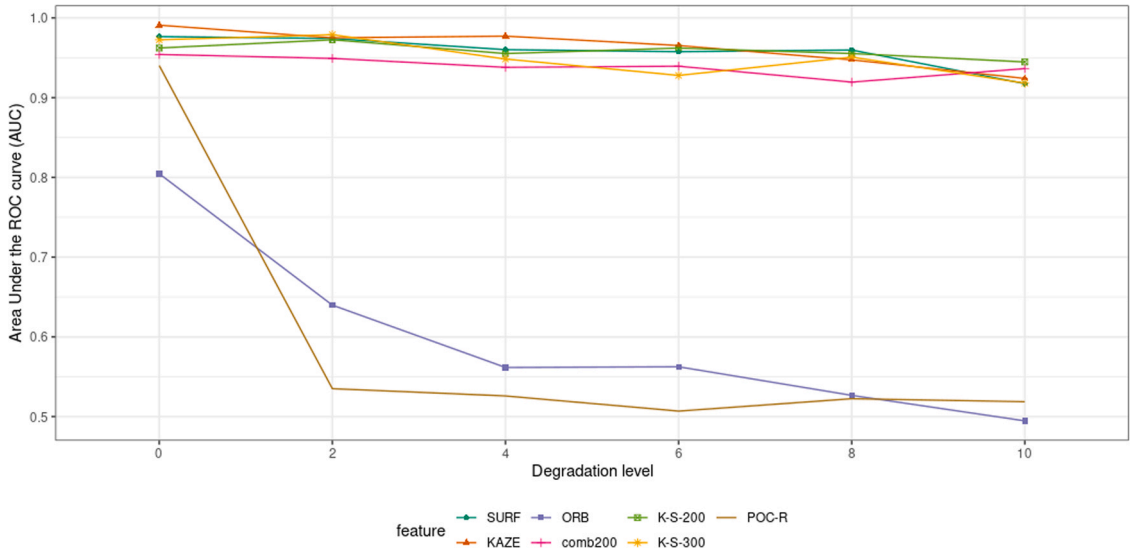

Fig. 8. Area under the ROC curve for each level of blurriness of *Q*.

Table 2
Equal error rates for the classifiers when the image of *Q* is of high quality (Level 0) or is blurry (Level 10).

Method	Level 0	Level 10
SURF	0.042	0.139
K-S-300	0.061	0.160
KAZE	0.067	0.170
K-S-200	0.097	0.125
comb200	0.132	0.119
ORB	0.264	0.486
POC-R	0.504	0.504

classifier because we did not expect it to perform any better than earlier.

Table 3 shows the EER computed for the RF classifiers based on each of the six different sets of features, at the two extreme degradation levels for *Q*. The first line in the table, labeled *RF-all* shows the EER for the classifier obtained by combining 18 similarity metrics (three metrics computed from each of the six sets of features), and the remaining six lines show the realized EERs obtained for the six image descriptors described earlier. When both *Q* and *K* in the test set are good-quality images, the RF score that combines 18 similarity features makes no classification errors. The other six RF scores, each combining the three similarity metrics based on the six sets of features also exhibit good accuracy regardless of the method used to extract features. An EER equal to 2.5% corresponds to one or two mis-classified pairs of images. When the *Q* image is degraded at Level 10, the accuracy of the RF classifiers ranges from 2.4% to 13%. Compared to the results shown in Table 2 in Section 4.1, combining all similarity metrics into a single score appears to noticeably decrease the EER in all cases. These results are encouraging, but need to be interpreted with some caution. In our study, the training set and the testing set of paired images were not independent, because some of the same shoes (in different combinations) were included in both sets. This is likely to have resulted in somewhat optimistic error rates.

Table 3
Equal error rates for the RF classifiers based on three similarity values when the image of *Q* is of high quality (Level 0) or is blurry (Level 10).

Method	Level 0	Level 10
RF-all	0.000	0.024
K-S-200	0.025	0.111
comb200	0.025	0.111
SURF	0.025	0.111
KAZE	0.025	0.119
K-S-300	0.025	0.119
ORB	0.025	0.133

Fig. 9 shows the relative variable importance for the performance of the RF classifiers based on different sets of features. The left panel shows results for the case where both the *Q* and *K* images are obtained at good resolution. The right panel corresponds to the case where *Q* was blurred (Level 10). The horizontal bars show the relative importance of each of the three similarity metrics computed from each of the six sets of features. There are, therefore, 18 horizontal bars in each of the two panels. To interpret the results shown in the figure, we can either focus on the feature extraction method (SURF, KAZE, ORB and combinations) or on the similarity metric (clique size, %overlap, or median distance).

Overall, similarity metrics obtained from the combined version of descriptors show the higher discrimination values as shown in Fig. 9. The three similarity metrics based on the set of ORB descriptors are lowest in terms of discrimination ability, a result that is consistent with the rest of our findings. One other conclusion we draw is that there appears to be value in combining image descriptors to construct a classifier. As the image of the questioned outsole becomes blurrier, similarity metrics based on sets of combined descriptors such as comb200 and K-S-200 play more important roles in the classifiers.

5. Discussion

Footwear evidence is commonly found in crime scenes. Yet, few crime labs in the United States have dedicated footwear evidence examiners and prints are often not even recovered by crime scene investigators. There are many explanations, but one reason appears to be the lack of objective, reproducible, and robust methods to compare questioned and reference impressions. The quantitative assessment of the similarity between two footwear images is challenging in realistic scenarios because crime scene images are typically smudgy, partially observed, and subject to background noise. While promising algorithms to carry out comparisons have been proposed (e.g., [2,8,25]), testing of their performance focused on mostly high quality images of crime scene impressions. For algorithmic tools to be useful for footwear examiners, they must be

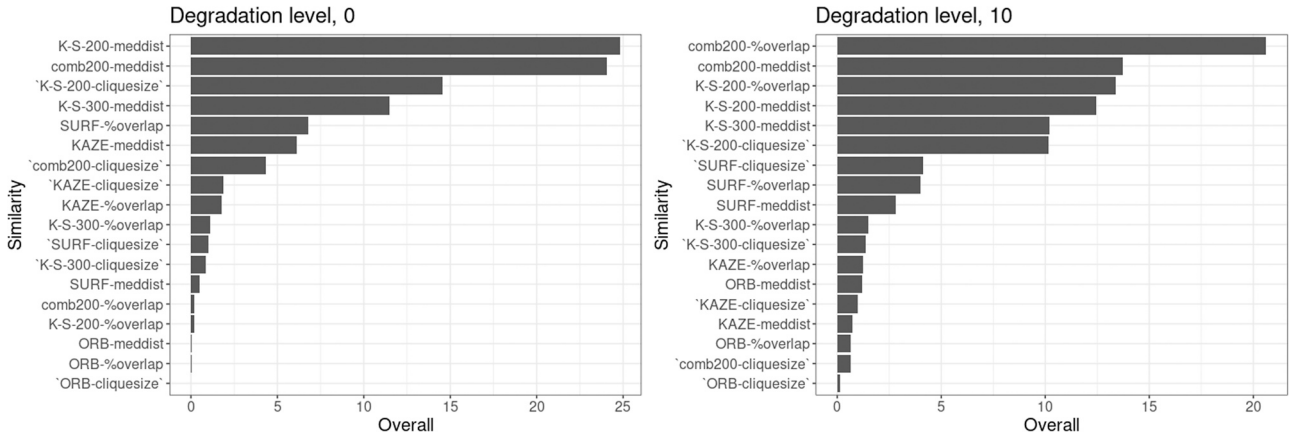

Fig. 9. Variable importance from the random forest training with all of similarity features.

shown to perform well even when one of the images in the comparison is degraded or partially observed.

In this manuscript, we revisit the approach proposed by [2] but pursue two different objectives. First, we explore whether different image descriptors or combinations of image descriptors, produce classifiers with comparable performance. Second, we test the performance of those classifiers on pairs of images where one of the images has been increasingly blurred. In all cases, we consider only the most challenging comparisons, where shoes share class characteristics (brand, model, size) and degree of wear. In this scenario, non-mated comparisons are carried out among close non-matching outsole impressions. We note that the shoes used in this study were purchased new and then assigned to volunteers, who walked in them for a period of about six months. To attempt to measure wear, at least roughly, we attached a step counter to each pair of shoes. We realize that this is an imperfect way to quantify wear, but it provided some idea of the variability between volunteers in terms of use. Two participants were outliers in that they exceeded the median number of steps in the group by tens of thousands of steps which resulted in shoes with a significantly higher degree of wear. We did not include those two pairs of shoes in the subset of shoes we used in this manuscript. The rest of the participants exhibited some variability, but overall, the degree of wear of the shoes we included in this analyses was comparable. In this light, we did not consider differences in wear when constructing the non-mated pairs of shoes. Information on RACs could also be used to construct the non-mated pairs of images. In this study, we did not attempt to identify or count RACs, however.

Outsole images can be degraded in many different ways. In our study, we focused on the degree of blurriness that one might encounter in an image obtained from a crime scene. To control the degree of blurriness of an image, we interposed sheets of paper between the outsole and the scanning surface of an EverOS scanner in our lab. When there were 10 sheets of paper between the shoe bottom and the scanner, much of the detail of the outsole pattern was lost, and edges of the pattern elements were blurred.

Our findings are encouraging. As argued by [2], when both images being compared are of good quality, it is possible to reach accuracy of about 95% when deciding whether a pair of images were produced by the same or by different shoes, even when using a simple classifier such as the %overlap. At present, these results hold only for the two models of athletic shoes in our study. When one of the two images is blurred, the algorithms predictably lose discriminating power, but the overall accuracy of at least some of them is still about 85–88%.

When three similarity metrics are combined into a single score using a random forest, accuracy increases to 97% and higher. In particular, the random forest score that combined the values of 18 similarity metrics based on six sets of features makes no classification errors in the test set. These results are to be interpreted cautiously, since in this study we only had 100 pairs of images in the training set, and 44 pairs in the test set. Thus, some over-fitting is likely. Further, there is overlap between the 18 features and it is unlikely that they all contribute noticeably to the performance of the classifier.

The approach we use to extract features from outsole images has an effect on the performance of algorithms in terms of classification error. We considered three image descriptors and constructed six sets of features using descriptors individually or in combination. We found that classifiers that rely on SURF and KAZE features tend to outperform those that rely on ORB features, regardless of the degree of blurriness of one of the images in a comparison pair. ORB uses pixels associated with corners of pattern design elements, and perhaps in the two models of athletic shoes we used, the corners are not well defined or are not important attributes of the outsole design.

There is much work still to do. A limitation to carry out this type of research is the lack of large databases with realistic crime scene-like impressions where ground truth is available. Ideally, we would like to have many different brands and models of shoes represented in the database, as well as a variety of substrates on which the print was made at the crime scene. Such an extensive database would permit selecting large enough subsets for training of random forests or other algorithms that are appropriate in different scenarios. At present, the limited scope of the databases that are available (including ours) tend to favor the use of simple classifiers such as the single %overlap, that require no training and can be used more broadly. One important take-home message is that we do not know whether any of the methods we discuss and that show promise in our set-up will generalize to other scenarios. So much more experimentation needs to occur before any of the proposed approaches can be implemented in real case work. In this work, we have tried to control computational effort without sacrificing accuracy. As computing power continues to increase, even computationally demanding algorithms will scale with the size of the databases of images that might need to be searched.

We are in the process of planning the construction of a more extensive database, which will be placed in the public domain to empower other researchers working in this area. Once those data become available, it will be possible to refine and adapt algorithmic approaches to quantitatively assess the similarity between two images. We believe that algorithms will never substitute well trained footwear examiners, but accurate, reliable algorithms can become a useful tool to aid in their analyses, at least in those cases where the quality of the crime scene image is sufficient to allow a comparison.

CRediT authorship contribution statement

Soyoung Park: Conceptualization, Formal analysis, Software, Investigation, Writing – original draft, Visualization. **Alicia Carriquiry:** Conceptualization, Formal analysis, Investigation, Writing – original draft, Supervision, Funding acquisition.

Acknowledgments

Dr. Hari Iyer has provided valuable guidance to us as we developed our methods and we are grateful to him. We are also grateful to our CSAFE colleagues for many productive discussion and for ideas to improve our work. As statisticians, we rely on the knowledge of practitioners, who own the problems and have the subject matter expertise. We wish to thank Ms. Leslie Hammer in particular for always keeping us on our toes, and Mr. David Kanaris and Brian McVicker for their insights and encouragement. Finally, the two anonymous reviewers of this manuscript provided helpful, constructive comments and we thank them for their time and effort.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.forsciint.2021.111126.

References

[1] Scientific Working Group for Shoeprint and Tire Tread Evidence (SWGTREAD), Standard for terminology used for forensic footwear and tire impression evidence, 2013. ⟨https://www.nist.gov/system/files/documents/2016/10/26/swgtread_15_standard_for_terminology_used_for_forensic_footwear_and_tire_impression_evidence_201303.pdf⟩.
[2] S. Park, A. Carriquiry, An algorithm to compare two-dimensional footwear outsole images using maximum cliques and speeded-up robust feature, Stat. Anal. Data Min.: ASA Data Sci. J. 13 (2020) 188–199.
[3] NRC, Strengthening forensic science in the United States: a path forward, National Academies Press, 2009.
[4] J., Holdren, E., Lander, W., Press, M., Savitz, W., Austin, C., Chyba et al., Report to the president forensic science in criminal courts: ensuring scientific validity of feature-comparison methods, Subcommittee on the Social and Behavioral Sciences Team: United States Government, 2016.
[5] E. Hare, H. Hofmann, A. Carriquiry, Algorithmic approaches to match degraded land impressions, Law, Probab. Risk 16 (2017) 203–221.
[6] A. Carriquiry, H. Hofmann, X.H. Tai, S. VanderPlas, Machine learning in forensic applications, Significance 16 (2019) 29–35.
[7] H.K. Iyer, S.P. Lund, Likelihood ratio as weight of forensic evidence: a closer look, J. Res. (NIST JRES) 122 (2017).
[8] A., Kortylewski, T., Vetter, Probabilistic compositional active basis models for robust pattern recognition, in: BMVC, 2016.
[9] N. Richetelli, M.C. Lee, C.A. Lasky, M.E. Gump, J.A. Speir, Classification of footwear outsole patterns using fourier transform and local interest points, Forensic Sci. Int. 275 (2017) 102–109.
[10] F. Cervelli, F. Dardi, S. Carrato, Comparison of footwear retrieval systems for synthetic and real shoe marks, in: 2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis, 2009, pp. 684–689.
[11] W.J. Bodziak, Footwear Impression Evidence Detection, Recovery, and Examination, second ed., CRC Press, Taylor and Francis, 2000.
[12] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[13] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Underst. 110 (2008) 346–359.
[14] P.F. Alcantarilla, A. Bartoli, A.J. Davison, Kaze features, in: European Conference on Computer Vision, Springer, 2012, pp. 214–227.
[15] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: an efficient alternative to sift or surf, in: 2011 International conference on computer vision, IEEE, 2011, pp. 2564–2571.
[16] P. Panchal, S. Panchal, S. Shah, A comparison of sift and surf, Int. J. Innov. Res. Comput. Commun. Eng. 1 (2013) 323–327.
[17] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: European conference on computer vision, Springer, 2006, pp. 430–443.
[18] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: European conference on computer vision, Springer, 2010, pp. 778–792.
[19] A. Kulkarni, J. Jagtap, V. Harpale, Object recognition with orb and its implementation on fpga, Int. J. Adv. Comput. Res. 3 (2013) 164.
[20] S.A.K. Tareen, Z. Saleem, A comparative analysis of sift, surf, kaze, akaze, orb, and brisk, in: 2018 International conference on computing, mathematics and engineering technologies (iCoMET), IEEE, 2018, pp. 1–10.
[21] H.-J. Chien, C.-C. Chuang, C.-Y. Chen, R. Klette, When to use what feature? sift, surf, orb, or a-kaze features for monocular visual odometry, in: 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), IEEE, 2016, pp. 1–6.
[22] I.M. Bomze, M. Budinich, P.M. Pardalos, M. Pelillo, The maximum clique problem, in: Handbook of combinatorial optimization, Springer, 1999, pp. 1–74.
[23] P.R. Östergård, A fast algorithm for the maximum clique problem, Discret. Appl. Math. 120 (2002) 197–207.
[24] L. Change, Efficient maximum clique computation and enumeration over large sparse graphs, Int. J. Very Large Databases (VLDB) 29 (2020) 999–1022.
[25] S.N., Srihari, Analysis of Footwear Impression Evidence – Final Technical Report to the U.S. Department of Justice, Technical Report, State University of New York at Buffalo, 2010.