
An efficient clustering-based retrieval framework for real crime scene footwear marks

Yi Tang*, Harish Kasiviswanathan and
Sargur N. Srihari

Department of Computer Science and Engineering,
Centre of Excellence for Document Analysis and Recognition,
University at Buffalo,
The State University of New York,
Amherst, New York, 14228, USA
E-mail: yitang@buffalo.edu
E-mail: harishka@buffalo.edu
E-mail: srihari@cedar.buffalo.edu
*Corresponding author

Abstract: As the most common type of evidence at crime scenes, footwear marks are found more often than fingerprints, and yet left largely unused due to lack of efficient and reliable tools. While the central task is stated simply – retrieve the closest matches among a database of known outsole prints – the difficulty is the poor quality of the marks and a very large and increasing number of outsole patterns. Since grouping the database into clusters can dramatically speed-up retrieval, we propose clustering based on recurring outsole patterns. The clustered database is used to retrieve similar prints for a given crime scene mark. Geometric shapes like line segments, circles and ellipses are proposed as features for crime scene marks. Then these features are structurally represented in the form of an attributed relational graph (ARG). Robust ARG matching is achieved with the introduced footwear print distance (FPD), a similarity measure for footwear prints. Sensitivity analysis of FPD is performed to show its robustness. The proposed system is invariant to scale, translation, rotation and insensitive to noise and degradations of the prints. Experiments show that the approach outperforms other state-of-the-art footwear print retrieval systems.

Keywords: real crime scenes; footwear marks; similarity measure; image retrieval; clustering; Hough transform; attributed relational graph; ARG; earth mover's distance; EMD; shoe prints; ellipse detection; footwear print distance; FPD.

Reference to this paper should be made as follows: Tang, Y., Kasiviswanathan, H. and Srihari, S.N. (xxxx) 'An efficient clustering-based retrieval framework for real crime scene footwear marks', *Int. J. Granular Computing, Rough Sets and Intelligent Systems*, Vol. X, No. Y, pp.000–000.

Biographical notes: Yi Tang received his BE in Computer Science from Beijing University of Posts and Telecommunications, China in 2006. He is currently a PhD candidate in Computer Science at University at Buffalo, The State University of New York. His research interests include pattern recognition, machine learning, data mining and computational forensics.

Harish Kasiviswanathan received his BE (honours) in Electrical and Electronics from Birla Institute of Technology and Science (BITS), Pilani, India. He worked as an Aerospace Engineer at Honeywell Technology Solutions Lab, Bangalore, India. He is currently pursuing his MS in Computer Science at University at Buffalo, The State University of New York. His interests include computer vision, machine learning and computational forensics.

Sargur N. Srihari is a SUNY Distinguished Professor in the Department of Computer Science and Engineering at University at Buffalo, the State University of New York. He is the Founding Director of CEDAR, the Centre of Excellence for Document Analysis and Recognition, which was recognised as the First US Postal Service Centre of Excellence in 1991. He received his BSc in Physics and Mathematics from the Bangalore University (National College) in 1967, BE in Electrical Communication Engineering from the Indian Institute of Science, Bangalore in 1970, and PhD in Computer and Information Science from the Ohio State University, Columbus in 1976. He is a Fellow of IEEE and IAPR. His research interests include pattern recognition, machine learning and computational forensics.

1 Introduction

As evidence in crime scenes, footwear marks can serve to narrow-down a list of suspects as well as provide a clue to link multiple crimes. Despite being found more frequently than fingerprints, footwear marks remain largely ignored as evidence. This state of affairs is largely because crime scene prints tend to be partial, noisy or highly degraded and the number of existing outsole patterns is very large. Manual comparison of crime scene marks with a large database of outsole patterns is very tedious and inefficient. At present there is no fully automated footwear print retrieval system to assist the forensic examiner, largely because most features to characterise footwear patterns fail with crime scene marks.

There is very little history of research on clustering footwear prints and its use in retrieving the closest print for a given crime scene mark. Retrieval based on a query consisting of a crime scene footwear mark is difficult as there are no known robust features that suit both partial crime scene marks and degraded database prints.

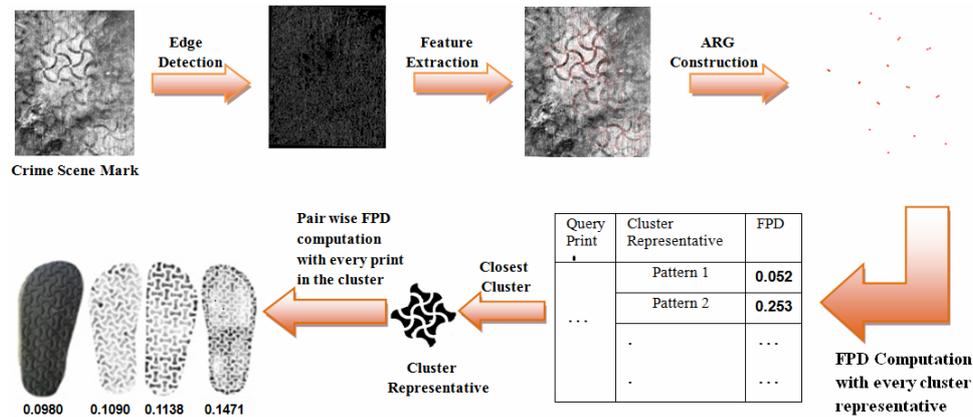
Geometric shapes like line segments, circles and ellipses are commonly found in footwear patterns. They withstand wear-and-tear for a long time, their shapes are preserved and commonly found in crime scene marks. Based on this observation, they are proposed here as features. Their relationship structure is represented in the form of an attributed relational graph (ARG). Attributes of nodes and edges in the ARG are defined to capture spatial relationships in terms of relative distance, position, dimension and orientation and also to be invariant to scale, rotation and translation.

Prints represented in the form of ARGs are compared by defining a *footwear print distance (FPD)*, which is built upon the nested earth mover's distance (EMD) (Kim et al., 2010). As a similarity measure it promises robustness in the matching of partial crime scene marks. Sensitivity analysis indicates that the FPD changes very little when node and edge attributes change marginally; thereby demonstrating insensitivity to small errors in the extraction of line segments, circles and ellipses.

Clustering and retrieval are two *different* but *related* problems. Clustering (Bishop, 2006) is the problem of grouping a set of objects into *clusters* based on similarity, so that objects within a cluster are as similar as possible, while objects of different clusters are as dissimilar as possible. The problem of retrieval is to find the most relevant objects in a database for the given query object. Similarity and relevance become *equivalent* if they are based on the same distance measure. Clustering has been used in information retrieval (Liu and Croft, 2004) for improving speed and accuracy.

Due to the nested structure of FPD, it is computationally intensive and slow in retrieval. This is alleviated by clustering, or grouping, the prints of similar outsole patterns. Domain knowledge (of the dataset) is used to determine the cluster centres prior to clustering. This speeds up clustering by assigning each object to its closest pre-determined cluster centre and avoiding iterative re-computation of cluster centres.

Figure 1 Proposed system for retrieving similar prints for a crime scene mark (see online version for colours)



Note: FPD is shown below retrieved similar prints.

Retrieval speed is increased by comparing the crime scene mark with a cluster representative rather than with every print in the database. Once the closest cluster is determined the crime scene mark is compared with prints in the cluster to determine the closest print. The main contributions of this paper are:

- Geometric shapes (line segments, circles and ellipses) are proposed as features, and new algorithms to extract features are presented.
- The features are represented as an ARG. Node and edge attributes are defined in a way to make it invariant to scale, translation, rotation and insensitive to noise and degradation of the prints.
- FPD, a distance measure built upon nested EMD is introduced as a similarity measure for footwear prints. Robustness of FPD is shown with sensitivity analysis.
- A system (shown in Figure 1) is proposed to retrieve similar prints with queries of crime scene footwear marks.

- The proposed retrieval system is compared with the state-of-the-art image matching algorithm such as scale invariant feature transform (SIFT) (Lowe, 2004), shape context-based object recognition algorithm (Belongie et al., 2002) and existing image retrieval systems specifically designed for footwear prints.

The rest of the paper is organised as follows: Section 2 presents the existing footwear print retrieval systems. Section 3 describes the algorithms for feature extraction. Section 4 deals with the structural representation of the extracted features in the form of an ARG. Section 5 introduces the FPD, a similarity measure for footwear prints. Section 6 deals with the clustering of footwear prints using recurring patterns. Section 7 presents the performance of the proposed method on a dataset of 2,660 prints and 300 crime scene marks, followed by sensitivity analysis of FPD in Section 8. Finally, Section 9 concludes the paper.

2 Related work

While there are many existing footwear print retrieval systems that work with synthetic and clean footwear prints, there is no published literature on clustering known footwear prints or retrieving them with real degraded queries. No one has reported good performance of their system with real crime scene footwear marks. Mikkonen and Astikainen (1994) proposed a classification system for shoeprints in which classification codes based on basic patterns are used to identify and classify the partial footwear impressions. Geradts and Keijzer (1996) described an automatic classification for shoe outsole designs. Here, different shapes in shoes are recognised using Fourier features and then these features are used in a neural network to classify the footwear. Alexander et al. (1999) presented a fractal pattern matching technique with mean square noise error as a matching criteria to match the collected impression against database prints.

Chazal et al. (2005) proposed a fully automated shoe print classification system which uses power spectral density (PSD) of the print as a pattern descriptor. Here, crucial information of the print is preserved by removing the low and high frequency components and 2D correlation coefficient is used as similarity measure. Zhang and Allinson (2005) proposed an automated shoe print retrieval system in which edge direction histogram (EDH) is used to represent the shapes in shoes. 1-D discrete Fourier transform on the normalised EDH is used as features and Euclidean distance is used as similarity measure. Pavlou and Allinson (2006) presented an automatic footwear classification system where maximally stable external region (MSER) feature detectors encoded with SIFT descriptors are used as features and then Gaussian feature similarity matrix and Gaussian proximity matrix are used as similarity measure.

Ghouti et al. (2006) described a ShoeHash-based approach to classify the shoeprints. Directional filter banks (DFBs) is used to capture the local and global details of the shoeprints and the energy dominant blocks are used as a feature vector. Finally, normalised Euclidean distance is used as similarity measure. Su et al. (2007) proposed a shoeprint retrieval system based on topological and pattern spectra. Here, pattern spectrum is constructed using the area measure of granulometry and topological spectrum is constructed using the Euler number. Then, normalised hybrid measure of these two is used to match the shoeprints. Crookes et al. (2007) described two ways to classify shoeprints:

- 1 in the spatial domain, modification of the existing techniques: Harris-Laplace detector and SIFT descriptor is proposed; the Harris corner detector is used to find the local features; Laplace-based automatic scale selection is used to decide the final local features and nearest neighbour is used as similarity measure
- 2 in the transform domain, phase-only correlation (POC) is used to match the shoeprints.

Sun et al. (2008) used expectation-maximisation (Dempster et al., 1977) and *K*-means (Bishop (2006)) clustering to group colour prints of shoe. Here, RGB information is used as features but most of the database prints and all the crime scene marks are grey-scale images hence their method is of limited use in computational forensics. Gueham et al. (2008) evaluated the performance of optimum trade-off synthetic discriminant function (OTSDF) filter and unconstrained OTSDF filter in classifying partial shoeprints.

AlGarni and Hamiane (2008) proposed an automatic shoeprint retrieval system in which Hu's moment invariants are used as features. Then results from standard similarity measures like Euclidean, city block, Canberra and correlation distances are compared. Xiao and Shi (2008) presented a computerised shoeprint matching using PSD and Zernike moments. Jingl et al. (2009) presented a new feature directionality to match shoeprints. Here, features extracted from cooccurrence matrix, Fourier transform and directional mask are matched using sum-of-absolute-difference. Nibouche et al. (2009) proposed a solution for matching rotated partial shoeprints. Harris points encoded with SIFT descriptors are used as features and they are matched using random sample consensus (RANSAC). Dardi et al. (2009) described a texture-based retrieval system for shoeprints. A Mahalanobis map is used to capture the texture and then matched using a correlation co-efficient measure. They have reported cumulative match characteristic (CMC) with 87 known prints and 30 real crime scene marks. Wang et al. (2009) presented a wavelet and fuzzy neural network-based approach to recognise footprints. Patil and Kulkarni (2009) used the Gabor transform to extract the multi-resolution features from shoeprints and then the Euclidean distance to match these features.

3 Feature extraction

Features being the crucial component of the retrieval system, the chosen features should be reliable. Colour, texture and shape can be used to distinguish images (Rui et al., 1999). Colour components are missing in crime scene marks while textures are susceptible to wears and hard to be captured. Geometric shapes can be easily captured and they are the most durable and reliable features of the outsole especially, in crime scene footwear marks. Shape features are also robust against occlusion and incompleteness, i.e., the variation or missing local region on the outsole will affect shape features in other regions to the minimal extent.

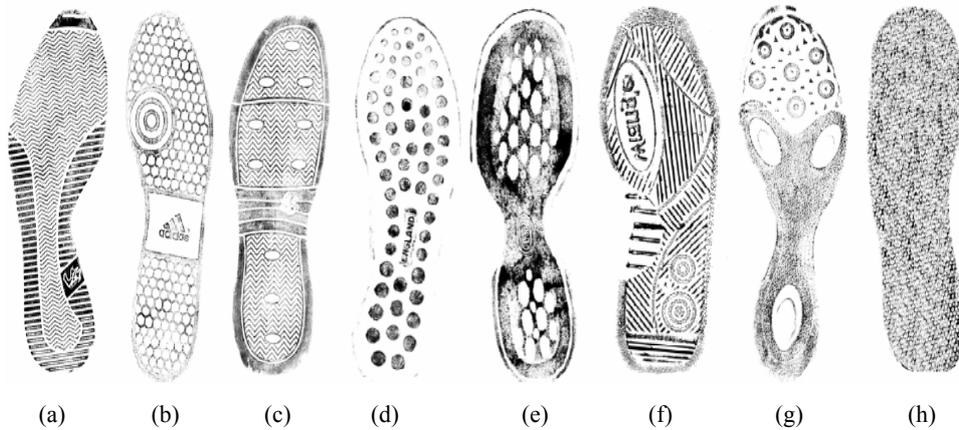
Visual inspection of 5,034 footwear prints from Foster and Freeman (<http://www.fosterfreeman.com/>) Dataset revealed that 91.8% of footwear prints can be represented by three basic shapes: line segments, circles and ellipses. Based on these shapes, footwear prints can be classified into eight types: piecewise lines, only

circles/arcs, only ellipses, circles and ellipses, lines and circles, lines and ellipses, lines, circles and ellipses and only texture. Any shapes other than circles and ellipses are approximated by piecewise lines. Combinations of these shapes can be used to distinctively identify the pattern of the footwear print. The distribution of fundamental shapes is given in Table 1, and sample prints are shown in Figure 2.

Table 1 Distribution of fundamental shapes in footwear prints

<i>Fundamental shapes</i>	<i>Number of prints</i>
Piecewise lines only	3,397
Circles/arcs only	73
Ellipses only	15
Lines and circles	812
Lines and ellipses	285
Circles and ellipses	5
Lines, circles and ellipses	37
Texture only	410
Total	5,034 prints

Figure 2 Eight types of footwear prints, (a) piecewise lines (b) lines and circles (c) lines and ellipses (d) only circles/arcs (e) only ellipses (f) lines, circles and ellipses (g) circles and ellipses (h) only texture



The standard Hough transform (SHT) (Hough, 1962; Nixon and Aguado, 2002) detects shapes in an image by mapping foreground pixels into a parameter space represented by an n -dimensional accumulator array, where n is the number of parameters used to describe the shape of interest in Hough space. Each significant pixel from the shape of interest would cast a vote in the same cell of an accumulator array, hence all pixels of a shape gets accumulated in a single cell. The number of valid peaks in the accumulator array would correspond to the number of shapes in the image.

The foremost step in the feature extraction is to perform morphological operations such as dilation and erosion. This makes the interior region of the boundary *uniform* and hence the Canny (1986) edge detector does not detect any edges inside the

boundary. This helps to enhance the quality of the edge image. In the proposed retrieval system (Figure 1), three basic shapes are detected: *straight line segments*, *circles* and *ellipses*.

Circle detection: SHT with a three-dimension accumulator array is used to detect circles. Gradient orientation (Goulermas and Liatsis, 1999) and spatial constraints (e.g., constraints on the distance between the centres of two candidate circles) are used to eliminate spurious circles. Pixels of detected circles are *removed* from the edge image and fed as input for ellipse detection. In the subsequent subsections we present two algorithms to detect ellipses and line segments respectively.

3.1 Ellipse detection

An ellipse in a Cartesian plane is described by the centre (p, q) , length of the semi-major/minor axes a, b and the orientation θ . In case of SHT, these five parameters demand a five-dimensional accumulator which is computationally expensive but randomised Hough transform (RHT) (Xu and Oja, 1993; McLaughlin, 1996) is computationally advantageous. In case of ellipse detection in footwear prints, RHT cannot be used directly¹. Hence, in RHT, we incorporate ideas:

- 1 decomposition of footwear prints into connected components
- 2 elimination of unwanted components based on their eccentricity²
- 3 smart selection of three random points based on local smoothness indicator (LSI) (explained below) at each pixel
- 4 elimination of spurious ellipses using gradient orientation.

We introduce a measure of local smoothness on image edges, *LSI*. The LSI at a *foreground* pixel z is defined by the standard deviation of gradient orientation of *foreground* pixels within the neighbourhood of z , i.e., $LSI(z) = std(\{\theta_{z'} | z' \in N_z\})$, where *std* stands for standard deviation, $\theta_{z'}$ is the gradient orientation at pixel z' , N_z is a 7×7 neighbourhood window centering at z . The LSI of good ellipse pixels should be within a proper interval $[s_l, s_u]$ to ensure that they are locally smooth. Algorithm 1 summarises ellipse detection. Detected ellipses before and after validation are shown in Figure 3. True ellipse pixels are removed from the edge image and the output is fed as input for line detection.

Algorithm 1 Ellipse detection

Input: Original image of footwear print, \mathbf{I} and its edge map \mathbf{I}_e .

Output: Detected ellipses and their parameters.

1. Compute the gradient orientation of \mathbf{I} .
2. Decompose \mathbf{I}_e into components, find the eccentricity e of each component.
3. Eliminate connected components with $e < 0.3$ and *size* < 20 pixels.
4. **For each** connected component \hat{C} **Do**
5. Pick three pixels randomly
6. Compute LSI at each pixel to get s_1, s_2 and s_3

7. **If** s_1, s_2, s_3 are all within interval $[s_l, s_u]$

8. **Then** Apply RHT and find parameters of the ellipse (p, q, a, b, θ)

9. Find candidate pixels that satisfy ellipse equation

$$\frac{((x-p)\cos\theta + (y-q)\sin\theta)^2}{a^2} + \frac{((y-q)\cos\theta + (x-p)\sin\theta)^2}{b^2} = 1$$

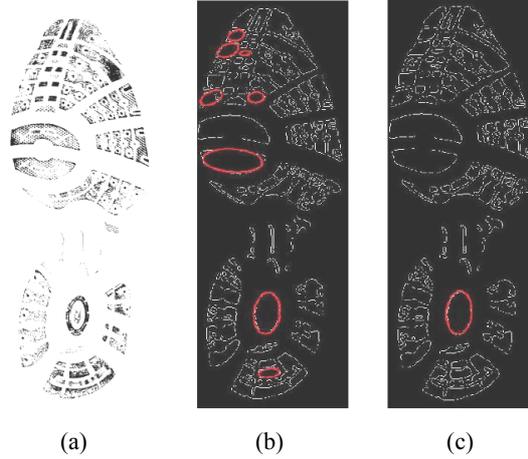
10. Find analytical derivative d at each candidate pixel using

$$d = \frac{\left(\frac{-2}{a^2}\right)(x-p)\cos\theta\sin\theta + (y-q)\sin^2\theta + \left(\frac{2}{b^2}\right)(y-q)\cos^2\theta - (x-p)\cos\theta\sin\theta}{\left(\frac{-2}{a^2}\right)(x-p)\cos^2\theta + (y-q)\cos\theta\sin\theta + \left(\frac{2}{b^2}\right)(y-q)\cos\theta\sin\theta - (x-p)\sin^2\theta}$$

11. Accept the pixel as an ellipse pixel if the difference between d and tangent of gradient orientation is below T_1

12. Declare \hat{C} as an ellipse if $\frac{\text{number of ellipse pixels}}{\text{circumference of ellipse}} \geq T_2$

Figure 3 Validate ellipses extracted using gradient orientation, (a) input (b) detected ellipses before validation (c) true ellipse after validation (see online version for colours)



How to choose threshold T_1, T_2, s_l and s_u : In Canny (1986) edge detection errors exist in the gradient orientation computation due to following steps:

- 1 increase in the size of sobel mask to make it less sensitive to noise leads to poor localisation
- 2 discrete differentiation operation in the sobel mask leads to approximation error in gradient.

In an ellipse E , the analytical derivative at each ellipse pixel P represents the ideal value of the slope of the tangent at P , while the tangent of gradient orientation at P is an *estimated* value of the slope of the actual tangent to the ellipse E at P . Ideally, these two values should be identical but because of the errors introduced in the gradient orientation

computation, there is *always* some difference between them. For spurious ellipse pixels, this difference is large while it is small for true ellipse pixels. Hence, threshold T_1 is introduced to distinguish true ellipse pixels from spurious ones. If the value of T_1 is set too small, then some true ellipse pixels might be eliminated; if it is set too large, then spurious ellipse pixels might be accepted as true ones. In our experiments we found that T_1 falls in the range of [0 0.26] hence we set T_1 to 0.26.

For complete ellipses, the ratio $\frac{\text{number of ellipse pixels}}{\text{circumference of ellipse}}$ equals 1. But, ellipses in the degraded footwear prints are mostly *partial* making the ratio less than 1. In experiments we found that ratio falls in the range [0.4 1] so we set T_2 to 0.4. From multiple trials of experiments we found that, when s_l and s_u are set to 0.03 and 0.25 respectively, selected pixels are good for determining ellipse parameters.

3.2 Line detection

On average, line segments in a footwear print ranges from 200–300. Each group of co-linear pixels generates a peak in Hough matrix. Numerous line segments make it difficult to differentiate true peaks from spurious ones. Further, useful but short line segments sometimes get missed. Hence, SHT alone is unlikely to give satisfying results for complex images like shoeprints. We propose a new algorithm that is able to detect line segments accurately and completely. First, connected components are labelled in the edge image. Then, for each component, SHT is applied and peaks are detected. When a peak in Hough matrix is identified and the line segments are extracted, the pixels contributing to those line segments are eliminated from the edge image. Then, an updated Hough matrix is obtained by applying SHT on the *modified* edge image. This procedure repeats until all line segments have been detected. Algorithm 2 summarises line detection in footwear prints. Sample results of feature extraction are shown in Figure 4.

Algorithm 2 Line segment detection

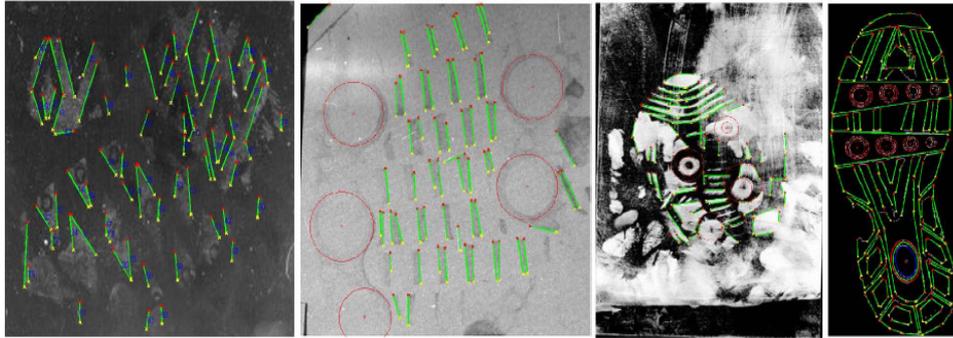
Input: Original image \mathbf{I} and edge map \mathbf{I}_e after removing all circle/ellipse pixels

Output: Line segments and their parameters

1. Find the connected components in \mathbf{I}_e
2. **For each** connected component \hat{C} **Do** /*outer loop*/
3. Form a sub-image \mathbf{I}_{es} of the same size as \mathbf{I}_e
4. Compute Hough matrix \mathbf{M} using SHT
5. **For each** peak P in Hough matrix \mathbf{M} **Do** /*inner loop*/
6. Initialise counter c to 0
7. Compute row-column indices of pixels in \mathbf{I}_{es} that map to P
8. Extract line segments in \mathbf{I}_{es} associated with P
9. Merge two line segments if their distance is below threshold d_{\min}
10. Discard line segments if ratio $\frac{\text{number of online pixels}}{\text{length of line segment}} < R_{\min}$
11. **If** at least one new line segment associated with P is accepted

12. **Then** suppress pixels that are mapped to accepted line segment(s)
 13. Apply SHT on I_{es} to update Hough matrix \mathbf{M}
 14. Reset counter c to 0
 15. **Else** suppress neighbouring cells of peak P in \mathbf{M}
 16. Increment counter c by 1
 17. **If** $c \geq N_{\max}$ **then** break
-

Figure 4 Results of feature extraction for a few crime scene footwear marks and a database print (see online version for colours)



Notes: Detected lines, circles and ellipses are shown in green, red and blue respectively.
Best viewed in colour.

How to choose threshold d_{\min} , R_{\min} and N_{\max} : threshold d_{\min} is used to decide whether to merge two line segments based on their distance. R_{\min} plays a similar role as T_2 in Algorithm 1. Threshold N_{\max} is the maximum number of consecutive failures in retrieving true line segments from each component \hat{C} . When counter c reaches N_{\max} , which reliably indicates that all true line segments in \hat{C} have been detected, the execution breaks from the inner loop and continues searching for line segments from the next component. These parameters have been set as follows and used throughout the experiments without the need of extra tuning: $d_{\min} = 16$ pixels, $R_{\min} = 0.48$, and $N_{\max} = 20$.

4 Attributed relational graph

Relational structures have been used to represent complex objects and scenes (Haralick and Shapiro, 1992). Such graphical representation (Bunke et al., 2008) has great advantages over feature vector because of its ability to explicitly model relationships between individual parts and feature points.

After feature extraction, a footwear print has been decomposed into a set of primitives. To obtain a structural representation of these primitives, an ARG (Sanfeliu and Fu, 1983; Bunke and Messmer, 1995) is built for each print. An ARG is a three-tuple $(V; E; A)$ where V is the set of nodes, E is the set of edges and A is the set of attributes. Lines, circles and ellipses are defined as nodes. Each edge describes the spatial relationship between nodes. The attributes include node attributes (unary) and edge attributes (binary).

There are three types of nodes (lines, circles and ellipses) and nine types of edges (line-to-line, line-to-circle, line-to-ellipse, circle-to-line, circle-to-circle, circle-to-ellipse, ellipse-to-line, ellipse-to-circle and ellipse-to-ellipse denoted as L2L, L2C, L2E, C2L, C2C, C2E, E2L, E2C, E2E edges). To tackle the case of nodes being missing or incorrectly detected due to noise, occlusion and incompleteness, a *fully-connected directed graph* is adopted. This means that there is a directed edge from each node to all other nodes. To distinguish one node from the other or one pair of nodes from another, node and edge attributes have been carefully defined to quantify the spatial relationships between each pair of nodes in terms of distance, position, size and orientation, which are tabulated in Table 2, where $N-\alpha$, $N-rs$, rd , pd , rp , $N-ro$ represent normalised relative angle, normalised relative size, relative distance, perpendicular distance, relative position, and normalised relative orientation respectively. All these attributes have been normalised in the range [0 1]. The weights shown for each edge attribute are determined using the sensitivity analysis described in Section 5.2.2. This rich feature descriptor is *invariant to scale, rotation, and translation*, as the value of each attribute does not depend on the scale, orientation or position of the footwear print. It is also *insensitive to noise and degradations*. It mimics the way humans distinguish one collection of shapes from the other.

Table 2 Node and edge attributes

Edge type	Att	Definition	Normalisation	Weight
L2L	$N-\alpha$	$\frac{ L1.\theta - L2.\theta }{180}$	-	0.4472
	$N-rs$	$\frac{L1.len}{L1.len + L2.len}$	-	0.4472
	rd	$\frac{dist(L1.m, L2.m)}{L1.len + L2.len}$	$\frac{rd}{\sqrt{1 + rd^2}}$	0.4472
	pd	$\frac{dist(L1.m, L2)}{L1.len + L2.len}$	$\frac{pd}{\sqrt{1 + pd^2}}$	0.4472
	rp_1	$\frac{\min(OA, OB)}{\max(OA, OB)}$	$\frac{rp_1 + 1}{2}$	0.4472
	rp_2	$\frac{\min(OA , OB)}{\max(OA , OB)}$	$\frac{rp_2 + 1}{2}$	0.4472
C2C	$N-rs$	$\frac{C1.r}{C1.r + C2.r}$	-	0.7071
	rd_1	$\frac{dist(C1.cen, C2.cen)}{C1.r + C2.r}$	$\frac{rd_1}{\sqrt{1 + rd_1^2}}$	0.7071
	rd_2	$\frac{dist(C1.cen, C2.cen)}{C1.r - C2.r}$	$\frac{rd_2}{\sqrt{1 + rd_2^2}}$	0.0

Table 2 Node and edge attributes (continued)

<i>Edge type</i>	<i>Att</i>	<i>Definition</i>	<i>Normalisation</i>	<i>Weight</i>
L2C	<i>N-rs</i>	$\frac{L.len}{C.r + L.len}$	-	0.5774
	<i>rd</i>	$\frac{dist(C.cen,L)}{C.r}$	$\frac{rd}{\sqrt{1+rd^2}}$	0.5774
	<i>rp</i>	$\frac{\min(S1,S2)}{\max(S1,S2)}$	$\frac{rp+1}{2}$	0.5774
L2E	<i>N-rs</i>	$\frac{L.len}{L.len + E.ER}$	-	0.5
	<i>rd</i>	$\frac{dist(L.m,E.cen)}{L.len + E.ER}$	$\frac{rd}{\sqrt{1+rd^2}}$	0.5
	<i>rp₁</i>	$\frac{\min(OA,OB)}{\max(OA,OB)}$	$\frac{rp_1+1}{2}$	0.5
	<i>rp₂</i>	$\frac{\min(OA , OB)}{\max(OA , OB)}$	$\frac{rp_2+1}{2}$	0.5
	<i>N-ro</i>	$\frac{ L.\theta - E.\theta }{180}$	-	0.5
C2L	<i>N-rs</i>	$\frac{C.r}{C.r + L.len}$	-	0.5774
	<i>rd</i>	$\frac{dist(C.cen,L)}{C.r}$	$\frac{rd}{\sqrt{1+rd^2}}$	0.5774
	<i>rp</i>	$\frac{\min(S1,S2)}{\max(S1,S2)}$	$\frac{rp+1}{2}$	0.5774
C2E	<i>N-rs</i>	$\frac{C.r}{C.r + E.ER}$	-	0.5774
	<i>rd</i>	$\frac{dist(C.cen,E.cen)}{C.r + E.ER}$	$\frac{rd}{\sqrt{1+rd^2}}$	0.5774
	<i>rp</i>	$\frac{\min(OA,OB)}{\max(OA,OB)}$	$\frac{rp+1}{2}$	0.5774
E2E	<i>e_ratio</i>	$\frac{E1.e}{E1.e + E2.e}$	-	0.2236
	<i>f(Δe)</i>	$\frac{E1.e - E2.e + 1}{2}$	-	0.2236
	<i>rd</i>	$\frac{dist(E1.cen,E2.cen)}{E1.ER - E2.ER}$	$\frac{rd}{\sqrt{1+rd^2}}$	0.4472

Table 2 Node and edge attributes (continued)

<i>Edge type</i>	<i>Att</i>	<i>Definition</i>	<i>Normalisation</i>	<i>Weight</i>	
E2E	<i>N-rs</i>	$\frac{E1.ER}{E1.ER + E2.ER}$	-	0.4472	
	<i>N-ro</i>	$\frac{ E1.\theta - E2.\theta }{90}$	-	0.4472	
	<i>rp</i>	rp (E1.major-axis, E2.major-axis)	$\frac{rp+1}{2}$	0.4472	
<i>Node attributes</i>					
<i>Node</i>	<i>Attributes</i>	<i>Definition</i>			
Circle	Quality	$\frac{\text{Number of pixels on circle}}{\text{Circumference of circle}}$			
Circle	Completeness	Standard deviation of the angle that all on-circle pixels make with respect to the centre			
Ellipse	Eccentricity	$\sqrt{1 - \frac{b^2}{a^2}}$			
<i>Symbols and its definition</i>					
<i>L</i>	Line segment	<i>E</i>	Ellipse	<i>dist</i>	Euclidean distance
<i>r</i>	Radius	<i>mid</i>	Mid-point	<i>rp</i>	Relative position
<i>e</i>	Eccentricity	<i>cen</i>	Centre	<i>rd</i>	Relative distance
<i>C</i>	Circle	<i>pd</i>	Perpendicular distance	<i>rs</i>	Relative size
<i>len</i>	Length	<i>ER</i>	$\sqrt{a * b}$	θ	Orientation
<i>att</i>	Attributes	<i>N</i>	Normalised	<i>a, b</i>	Semi-major axis and semi-minor axis of the ellipse respectively
<i>max</i>	Maximum	<i> </i>	Absolute value	<i>p, q</i>	Centre of the ellipse

5 Similarity measure

Image retrieval applications typically employ histogram (or probability density) distance measures. Bin-by-bin distance measures such as Euclidean distance (or its generalisation known as the Minkowski distance) and Kullback-Leibler divergence are perceptually unsatisfactory. Kolmogorov-Smirnov is only applicable to one-dimensional problems. EMD, a cross bin distance metric has become most popular in content-based image retrieval (Rubner et al., 2000). Advantages of EMD include: allows partial matches, ability to efficiently handle high-dimensional feature spaces and closeness to perceptual similarity when applied to image histograms.

5.1 Earth mover's distance

EMD evaluates the least amount of work that is needed to transform one distribution into the other. Consider the evaluation of the distance between two signatures (histograms)

$P_1 = \{P_{1i} | 1 \leq i \leq n_1\}$ and $P_2 = \{P_{2j} | 1 \leq j \leq n_2\}$. The bins $[P_{1i}]$ have corresponding weights $\mathbf{w}_1 = [w_{1i}]$ and similarly $[P_{2j}]$ have weights $\mathbf{w}_2 = [w_{2j}]$. The ground distance matrix $\mathbf{C} = [c_{ij}]$ specifies *ground distance* between all pairs of bins, c_{ij} . The flow matrix $\mathbf{F} = [f_{ij}]$, where f_{ij} is the amount of ‘supplies’ transferred from bin P_{1i} to bin P_{2j} . The goal is to find proper values of \mathbf{F} in order to *minimise* the overall work given by

$$WORK(\mathbf{w}_1, \mathbf{w}_2, \mathbf{C}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{ij} f_{ij} \quad (1)$$

which is subject to the following constraints:

$$f_{ij} \geq 0, \quad \forall 1 \leq i \leq n_1, 1 \leq j \leq n_2, \quad (2)$$

$$\sum_{j=1}^{n_2} f_{ij} \geq w_{1i}, \quad \forall 1 \leq i \leq n_1, \quad (3)$$

$$\sum_{i=1}^{n_1} f_{ij} \leq w_{2j}, \quad \forall 1 \leq j \leq n_2, \quad (4)$$

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_{ij} = \min \left(\sum_{i=1}^{n_1} w_{1i}, \sum_{j=1}^{n_2} w_{2j} \right). \quad (5)$$

Constraint 2 allows moving ‘supplies’ from P_1 to P_2 and not vice versa. Constraint 3 limits the amount of ‘supplies’ that can be sent by the bins in P_1 to their weights. Constraint 4 limits the bins in P_2 to receive no more ‘supplies’ than their weights. Constraint 5 forces to move the maximum amount of ‘supplies’ possible. This amount is referred to as the total flow in the transportation problem.

This is a linear programming problem which is solved efficiently by the transportation simplex algorithm (Hillier and Liebermann, 1995). Once the flow matrix \mathbf{F} is found, the EMD is defined as the overall work normalised by the total flow

$$EMD(P_1, P_2) = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{ij} f_{ij}}{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_{ij}}. \quad (6)$$

The computation of EMD assumes that there exists a proper distance measure to compute ground distance matrix \mathbf{C} , where the element c_{ij} is the unit distance between a pair of bins P_{1i} and P_{2j} , i.e., the work required to move one unit of ‘supplies’ from the source bin P_{1i} to the destination bin P_{2j} . Such distance can be straightforwardly defined between histogram bins because of their strict relative order.

5.2 Footwear print distance

Robust ARG matching requires an assignment algorithm that yields not only a correspondence between two sets of vertices but also the similarity between them. Popular assignment algorithms are nearest neighbour search, Hausdorff distance, bipartite matching and EMD. EMD is the state-of-the-art assignment algorithm (Kim et al., 2010;

Rubner et al., 2000). In EMD, the bins are replaced by vertices and relations between them. Both vertices (nodes) and relations (edges) have attributes associated with them. The vertices also have associated weights with them, which are useful in performing assignment. However, when matching two ARGs, the ground distance between two vertices depends not only on the two vertices themselves, but also is related to their incident edges. Therefore, computing the ground distance between two vertices, involves a combinatorial optimisation procedure to establish correspondence as consistently as possible between the *attributed trees* rooted at vertices. Hence, *direct application of the basic EMD algorithm cannot solve the ARG matching problem* and it needs to be augmented with a method for computing the ground distance matrix between all pairs of nodes.

Recently, Kim et al. (2010) have used nested structure of EMD to achieve robust ARG matching in computer vision. This kind of technique has never been analysed in the forensics domain until now. However, their method will not work well when two graphs to be matched have multiple attributes of different scales, and the difference in each attribute between two ARGs contribute unequally to the resulting overall distance. In this case, we need to apply appropriate weights on different attributes to balance their contributions to the overall distance, so that the difference in one feature/attribute will not dominate the overall distance. This step is essential as crime scene marks are created in an uncontrolled environment and they are highly degraded and partial, too.

5.2.1 Computation of FPD

We propose a method to learn the weights for different attributes using sensitivity analysis and incorporate the learned weights into the framework of EMD to finally derive a distance measure (FPD) for footwear prints. First, we elaborate how FPD is derived from EMD, and how learned weights are incorporated, followed by how to learn the weight vector.

A completely connected ARG is formally defined as $P = (V, R, n)$ where $V = \{V_i | 1 \leq i \leq n\}$ is the set of nodes and $R = \{R_{ij} | 1 \leq i, j \leq n\}$ is the set of relations between nodes. Each node has a weight and an attribute vector, $V_i = (w_i, \mathbf{v}_i)$ and each relation R_{ij} has an attribute vector \mathbf{r}_{ij} .

Let ARG of 1st and 2nd footwear prints be $F|P_1 = (V_1, R_1, n_1)$ and $F|P_2 = (V_2, R_2, n_2)$ respectively. To compute the FPD between FP_1 and FP_2 , an appropriate mapping M between the two sets of nodes is needed. The cost or ground distance matrix is $\mathbf{C} = [c_{ij}]$ where $c_{ij} = c(V_{1i}, V_{2j} | V_{1i} \in V_1, V_{2j} \in V_2)$. The unit cost or distance between V_{1i} and V_{2j} is evaluated based on the similarity of the spatial configurations at the two nodes, which is explained later in this section.

By providing *identical* weights for all nodes the nested structure of EMD can handle the case of subgraph matching, i.e.,

$$w_{1i} = w_{2j} = \frac{1}{\max(n_1, n_2)}, \quad 1 \leq i \leq n_1, \quad 1 \leq j \leq n_2. \quad (7)$$

Unlike EMD, a node of FP_1 can transfer its weight to only one node of FP_2 . This is known as *uniqueness constraint*. To enforce one-to-one correspondence, each node i in the first ARG can match only one node j in the second ARG or left unmatched, i.e., f_{ij}

may take the value of either $\frac{1}{\max(n_1, n_2)}$ or 0, $\forall i \in \{1, \dots, n_1\}, j \in \{1, \dots, n_2\}$.

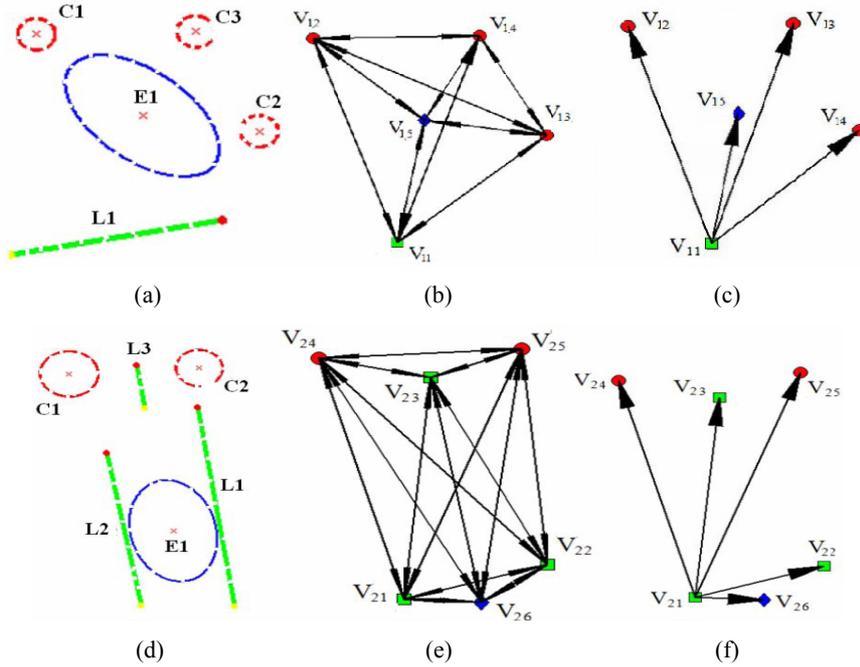
Therefore, we rewrite equation (6) as

$$\text{FPD}(FP_1, FP_2) = \frac{\frac{1}{\max(n_1, n_2)} \sum_{\{(i,j)|f_{ij}>0\}} c_{ij}}{\sum_{\{(i,j)|f_{ij}>0\}} f_{ij}} \quad (8)$$

The total number of correspondence pairs between the two ARGs is $\min(n_1, n_2)$ so the total amount of flow transferred from FP_1 to FP_2 is $\frac{\min(n_1, n_2)}{\max(n_1, n_2)}$. Substituting this term for the denominator in equation (8) we get,

$$\text{FPD}(FP_1, FP_2) = \frac{\sum_{\{(i,j)|f_{ij}>0\}} c_{ij}}{\min(n_1, n_2)} \quad (9)$$

Figure 5 (a) Footwear print FP_1 (b) ARG of FP_1 (c) Attributed tree of FP_1 rooted at V_{11} (d) Sample print FP_2 (e) ARG of FP_2 (f) Attributed tree of FP_2 rooted at V_{21} (see online version for colours)



Note: Green squares, red circles and blue diamonds represent lines, circles and ellipses respectively.

Cost determination between two nodes: For a given pair of nodes in two graphs, say V_{1i} and V_{2j} , how one node is different from the other depends not only on the nodes, but also on how they relate to their respective neighbours in terms of distance, orientation, position, etc. This means that the distance c_{ij} between the two nodes should be evaluated based on the distance between two attributed relational sub-graphs, rooted at V_{1i} and V_{2j} respectively. Each attributed relational sub-graph is an *attributed tree* (AT) (Pelillo et al., 2001). ARG and Attributed tree for two sample prints are shown in Figure 5. This leads to a *nested structure of ARG matching*, which consists of inner and outer steps. For the outer step, the unit cost or distance between V_{1i} and V_{2j} , is defined as

$$c(V_{1i}, V_{2j}) = \text{EMD}(AT_{V_{1i}}, AT_{V_{2j}}), \quad (10)$$

where $AT_{V_{1i}}$ and $AT_{V_{2j}}$ are attributed trees rooted at V_{1i} and V_{2j} in the two ARGs. The tree $AT_{V_{1i}}$ consists of the root vertex V_{1i} and its connection to the rest of the $n_1 - 1$ vertices.

To calculate the distance between the two trees $AT_{V_{1i}}$ and $AT_{V_{2j}}$ using EMD framework, we build the inner cost matrix $\hat{\mathbf{C}} = [c_{ij}^*]$ whose elements correspond to pairwise node-to-node (V_{1i} to V_{2j}) distances *in the two trees*. The *inner cost* between V_{1i} and V_{2j} takes into account not only the unary attributes of the nodes but also their edges attributes and is calculated by

$$c(V_{1i}, V_{2j}) = \alpha d_E(\mathbf{v}_{1i}, \mathbf{v}_{2j}) + (1 - \alpha) d_E(\mathbf{Q} * \mathbf{r}_{1i}, \mathbf{Q} * \mathbf{r}_{2i}) \quad (11)$$

where α is a weight co-efficient in the interval $[0, 1]$, d_E is the Euclidean distance, \mathbf{r}_{1i} is the attribute vector of the edge between V_{1i} and V_{1i} , \mathbf{Q} is the weight vector and the operator '*' denotes the element-wise product between two vectors. Parameter α reflects the relative importance of the difference of node attributes and the difference of edge attributes in the evaluation of inner cost between two nodes, and is set to 0.5 assuming *equal importance*. Weight vector \mathbf{Q} for all edge attributes is derived using sensitivity analysis described in Section 5.2.2.

Nodes V_{1i} and V_{2j} may have one of three possible labels: 'L', 'C' and 'E' corresponding to lines, circles, or ellipses respectively. Thus there are 9 combinations of labels for (V_{1i} , V_{2j}). A line cannot match with a circle or an ellipse regardless of their attributes and neighbours; while a circle and ellipse can match to some degree. Thus the unit matching cost for *non-matching* label pairs is $c('L', 'C') = c('L', 'E') = 1$. For other label pairs, the node-to-node inner costs are determined using equation (11). Algorithm 3 summarises FPD computation.

Algorithm 3 FPD between prints FP_1 and FP_2

Input: ARGs of $FP_1 = (V_1, R_1, n_1)$, $FP_2 = (V_2, R_2, n_2)$,

where $V_1 = \{(V_{1i}, w_{1is}, \mathbf{v}_{1i}) \mid 1 \leq i \leq n_1\}$, $V_2 = \{(V_{2j}, w_{2jt}, \mathbf{v}_{2j}) \mid 1 \leq j \leq n_2\}$

and $R_1 = \{(R_{1is}, r_{1is}) \mid 1 \leq i, s \leq n_1\}$, $R_2 = \{(R_{2jt}, r_{2jt}) \mid 1 \leq j, t \leq n_2\}$

Output: FPD(FP_1, FP_2)

1. Set each component $w_{1i} = w_{2j} = \frac{1}{\max(n_1, n_2)}$, $1 \leq i \leq n_1$, $1 \leq j \leq n_2$
2. Compute outer cost matrix **C**

For $k_1 = 1$ to n_1 **Do**

For $k_2 = 1$ to n_2 **Do**

If (is Compatible (V_{1k_1}, V_{2k_2})) == false)

Then $c(k_1, k_2) = 1$

Else build attributed trees rooted at V_{1k_1} and V_{2k_2} , viz., $AT_{V_{1k_1}}$ and $AT_{V_{2k_2}}$

For $i = 1$ to n_1

For $j = 1$ to n_2

If (is Compatible(V_{1i}, V_{2j})) == false)

Then $c'(i, j) = 1$

Else $c'(i, j) = \alpha * d_E(\mathbf{v}_{1i}, \mathbf{v}_{2j}) + (1 - \alpha) * d_E(\mathbf{Q} * \mathbf{r}_{1k_1}, \mathbf{Q} * \mathbf{r}_{2k_2})$

$c(k_1, k_2) = \text{EMD}(AT_{V_{1k_1}}, AT_{V_{2k_2}})$
3. Compute flow matrix **F** using simplex algorithm to minimise the overall cost n1

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{ij} f_{ij}$$
4. **return** FPD(FP_1, FP_2) = $\sum_{\{(i,j) | f_{ij} > 0\}} c_{ij} / \min(n_1, n_2)$

Function is Compatible(Node V_1 , Node V_2)

/* to check if node types are compatible, e.g., a line and a circle are incompatible. */

If ($V_1.label == 'L'$) & ($V_2.label == 'C'$ || $V_2.label == 'E'$)

return false;

Else if ($V_2.label == 'L'$) & ($V_1.label == 'C'$ || $V_1.label == 'E'$)

return false;

Else return true;

Analysis of time complexity: The theoretical complexity of EMD used to compute $c(k_1, k_2)$ in Algorithm 3 is $O(n_1 n_2 n_{iter})$, where n_{iter} is the number of iterations used by the simplex algorithm, which can be approximated by $O(n_1 n_2 \max(n_1, n_2))$ (Kim et al., 2010). Step 2 of Algorithm 3 may involve at most $n_1 n_2$ times of EMD computation in the worst case [when function call is compatible(V_{1k_1}, V_{2k_2}) returns true each time], so the time complexity of FPD is $O(n_1^2 n_2^2 \max(n_1, n_2))$.

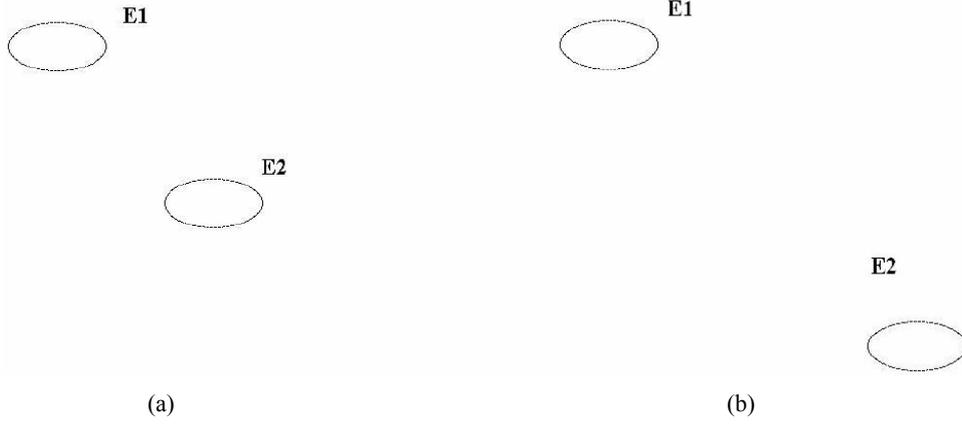
5.2.2 Computing weight vector using sensitivity analysis

Distance between ARGs has different sensitivities for different attributes. Weight vector **Q** in equation (11) takes care of the difference in sensitivities. For clarity, the same equation is shown below:

$$c(V_{1i}, V_{2j}) = \alpha d_E(\mathbf{v}_{1i}, \mathbf{v}_{2j}) + (1-\alpha) d_E(\mathbf{Q} * \mathbf{r}_{1i}, \mathbf{Q} * \mathbf{r}_{2i}) \quad (12)$$

where α is 0.5 as explained before, d_E is the Euclidean distance, \mathbf{r}_{1i} is the attribute vector of edge between V_{1i} and V_{1i} , \mathbf{Q} is the weight vector and $*$ is the element-wise product between two vectors.

Figure 6 Example of two two-node prints used in sensitivity analysis, (a) P_1 (b) P_2



Consider two synthetic prints shown in Figure 6, each of which contains two identical ellipses, i.e., $n_1 = n_2 = 2$. We assume that α equals 0.5, which means that node attributes and edge attributes are weighed equally. Using Algorithm 3 we arrive at the following distance between the two prints

$$\text{FPD}(P_1, P_2) = 0.25(d_E(\mathbf{v}_{11}, \mathbf{v}_{21}) + d_E(\mathbf{v}_{12}, \mathbf{v}_{22}) + d_E(\mathbf{Q} * \mathbf{r}_{112}, \mathbf{Q} * \mathbf{r}_{212})) \quad (13)$$

Similarly, for two prints with arbitrary n_1 and n_2 nodes, with $n_1 \leq n_2$, we have

$$\text{FPD}(P_1, P_2) = \frac{\alpha \left\{ \sum_{i=1}^{n_1} n_1 d_E(\mathbf{v}_{1n_i}, \mathbf{v}_{2n_i}) + 2 \sum_{\{(i,j)|i<j\}} d_E(\mathbf{Q} * \mathbf{r}_{1ij}, \mathbf{Q} * \mathbf{r}_{2ij}) \right\}}{n_1^2} \quad (14)$$

We assume that we can make a print become totally different from itself by changing every value of both node and edge attributes from one extreme to the other. This means that $\Delta r = 1$, where r is an edge attribute. Suppose that P_1 and P_2 are such two prints whose distance $\text{FPD}(P_1, P_2)$ equals the maximum of 1. Assuming that the difference of node attributes in each correspondence pair, as well as the difference of edge attributes in each correspondence pair, take *equal* responsibility for the distance $\text{FPD}(P_1, P_2)$. The number of shares of such responsibility is $n_1^2 + n_1(n_1 - 1) = n_1(2n_1 - 1)$.

$$d_E(\mathbf{Q} * \mathbf{r}_{1is}, \mathbf{Q} * \mathbf{r}_{2jt}) = \frac{n_1}{\alpha(2n_1 - 1)}, \quad 1 \leq i, s \leq n_1, \quad 1 \leq j, t \leq n_2. \quad (15)$$

We can rewrite the above equality as follows.

$$\sqrt{\sum_{k=1}^m Q_k (r_{1isk} - r_{2jik})^2} = \frac{n_1}{\alpha(2n_1 - 1)}, \quad (16)$$

where m is the dimension of the edge attribute vector.

Finally, we arrive at the values of weight vector as follows.

$$Q_k = \frac{\sqrt{\left[\frac{n_1^2}{\alpha n_1 (2n_1 - 1)} \right]^2}}{m} = \frac{2n_1}{(2n_1 - 1)\sqrt{m}}, \quad \forall k \in \{1, 2, \dots, m\} \quad (17)$$

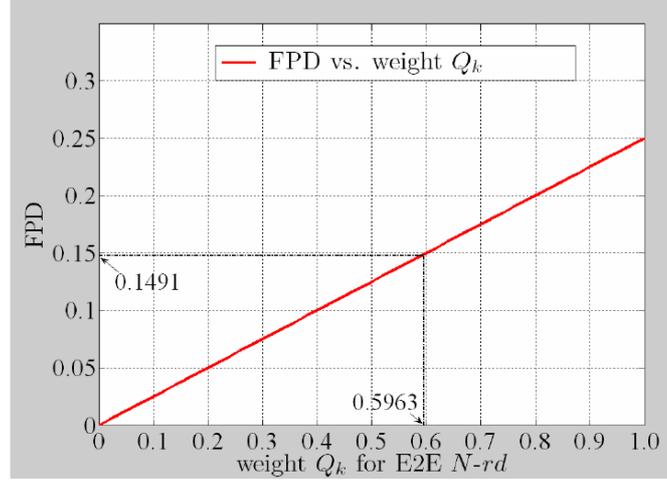
For large n_1 , $\frac{2n_1}{(2n_1 - 1)} \approx 1$, thus we have $Q_k \approx \frac{1}{\sqrt{m}}$. When $n_1 = 2$, $Q_k = \frac{4}{3\sqrt{m}}$. This indicates that we can determine the weights $\{Q_k, k = \{1, \dots, m\}\}$ by first deriving the value of Q_k in the case of two-nodes, then multiplying it by $\frac{3}{4}$. The contribution of each edge attribute for all pairs of nodes to distance can be calculated as

$$\frac{\frac{2n_1}{(2n_1 - 1)\sqrt{n}} * 1 * n_1 (n_1 - 1) * \alpha}{n_1^2} = \frac{n_1 - 1}{(2n_1 - 1)\sqrt{n}}.$$

From the definition of edge attributes (shown in Table 2), $E2E$ edge has six attributes. Since the attributes eccentricity ratio and Δe bear apparent dependency between each other, the number of independent attributes for $E2E$ edge is 5, i.e., $m = 5$. We chose the attribute *normalised relative distance* $N-rd$ in Table 2 to conduct the experiments on prints with *only two* nodes to determine the relationship between the distance FPD and $N-rd$ as well as the weight Q_k . This attribute is supposed to contribute to the final distance an amount of $\frac{2-1}{(2*2-1)\sqrt{5}} = 0.1491$.

In the experiments, the maximum change of $N-rd$ is 1. Our goal is to find the value of Q_k such that the distance reaches 0.1491 when the $\Delta N-rd$ is 1 (we set the $N-rd$ of P_1 to be 0, so $\Delta N-rd = N-rd(P_2) = 1$). From Figure 7, we get the value of Q_k as 0.5963, which is the same as what we would get if we substitute 5 for m , and 2 for n_1 in equation (17). Finally, we obtain weight Q_k for attribute $E2E$ $N-rd$ for prints with n_1 nodes by multiplying 0.5963 by $\frac{3}{4}$ to get 0.4472, which is shown in subtable of $E2E$ in Table 2.

Similarly, the weights for other attributes have been determined by means of both experiments and derivations. The weights are shown in the last columns in the definitions of edge attributes in Table 2.

Figure 7 Plot of distance FPD against weight Q_k (see online version for colours)

Notes: $N-rd(P_1) = 0$ and $N-rd(P_2) = 1$

6 Clustering using recurring patterns

As the time complexity of FPD is $O(n_1^2 n_2^2 \max(n_1, n_2))$ it is a computationally intensive graph distance measure despite being robust and accurate hence to overcome the trade-off between its speed and accuracy we clustered the known prints in database beforehand to aid the retrieval.

Clustering algorithms can be divided into partition-based, density-based and hierarchical-based methods (Aldenderfer and Blashfield, 1984). Popular algorithms like K -means (partition-based) and hierarchical clustering require huge amount of computation of distance between data points. For instance, assignment/reassignment of data points to the nearest cluster in K -means requires a huge number of distance computations. Similarly, building similarity matrix in hierarchical clustering is computationally expensive for a large dataset. In other words, an efficient distance measure (*usu.* computed in linear time) is required for a practical application of K -means or hierarchical clustering, for which FPD is *not* efficient. Moreover, K -means clustering assumes that data are represented by real vectors with *equal* dimension; Gaussian mixture model (GMM) (Bishop, 2006) assumes normally distributed data. Both assumptions are *not valid* for footwear prints represented by ARGs. Hence the existing clustering algorithms are not readily applicable to footwear prints, and so we decided to incorporate domain knowledge to aid the clustering.

Recurring patterns (shown in Figure 8) such as wavy pattern, concentric circles (Girod, 1982; Foster and Freeman, <http://www.fosterfreeman.com/>; Mikkonen et al., 1996) are typically found in footwear prints and each of them can represent a group of similar prints. Each pattern is simple and its graph structure has a small number of nodes. Further, the ARG representing a footwear print has 200–300 nodes on average and nodes can vary considerably in terms of relative size, position, etc. This makes the feature space very *sparse* and therefore similar footwear prints tend to stay close to each other and

dissimilar ones stay apart. Hence, to cluster the huge dataset we use *recurring patterns* as cluster representatives, which serve as *initial seed clusters* (Basu et al., 2002). This clustering approach is outlined in Figure 9 and is summarised as Algorithm 4.

Figure 8 33 cluster representatives determined from a database of 2,660 prints

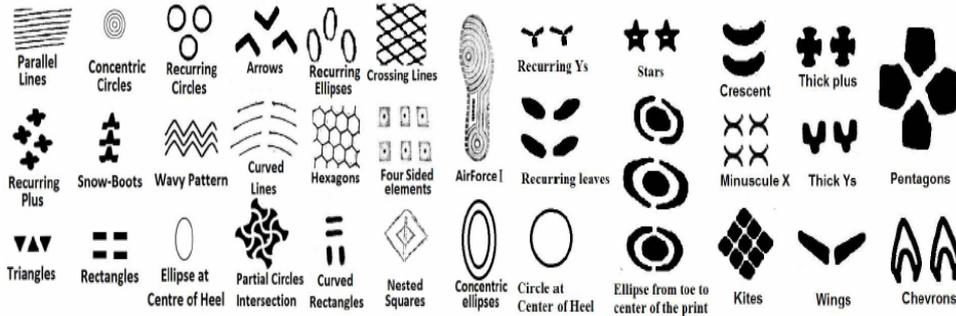
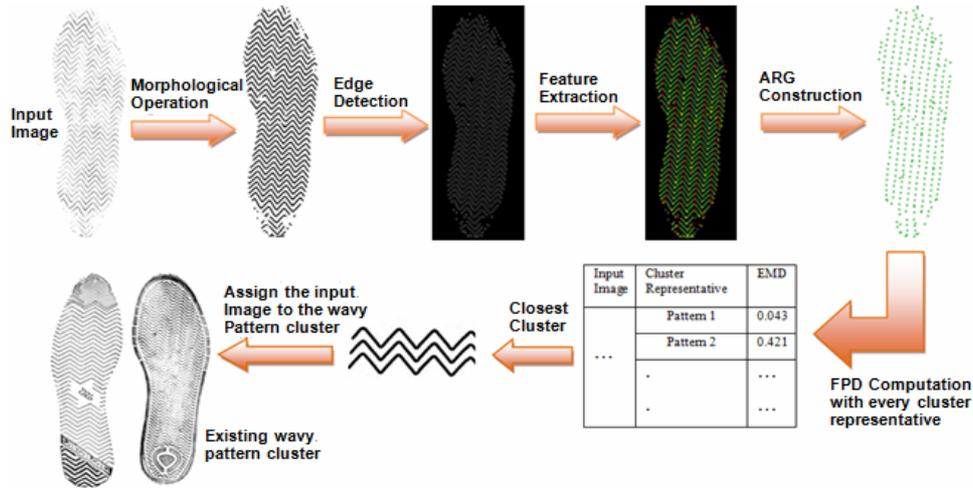


Figure 9 Clustering Footwear prints in database (see online version for colours)



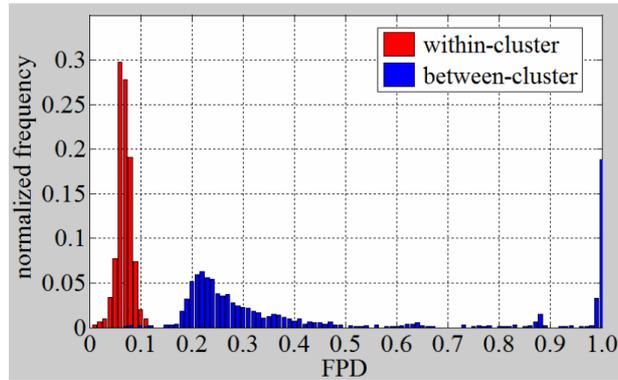
Algorithm 4 Clustering database prints

Input: Set D of footwear prints and recurring patterns P_1, P_2, \dots, P_m
Output: Clusters of footwear prints

1. Set # of clusters K to m , and set each recurring pattern as a cluster representative
2. **For each** footwear print $p \in D$ **Do**
3. Compute $FPD(p, P_i), \forall i = 1, \dots, K$.
4. $j = \arg \min_i (FPD(p, P_i))$
5. **If** $FPD(p, P_j) \leq T$
6. **Then** assign p to cluster P_j
7. **Else** increment K by 1
8. Create a new cluster P_K and assign p to P_K

How to choose threshold T : T is an important parameter as it decides whether to assign a print p to its closest cluster P_j (existing) or let it be a new cluster. An appropriate value of T should be located between average within-cluster FPD and average between-cluster FPD. On the other hand, a footwear print usually contains one dominant recurring pattern and one or more non-dominant pattern(s). For instance, the print in Figure 2(a) has wavy pattern as a dominant pattern and rest of the parallel lines as a non-dominant pattern. We randomly selected 200 prints and manually assigned them to a cluster representative by recognising the dominant pattern of each print. If no recurring pattern was found in them, then they form a new cluster. After the clusters were formed, we computed pair-wise FPDs and visualised the distributions of within/between-cluster FPD (shown in Figure 10). From Figure 10, it is evident that $T = 0.15$ separates the two distributions well hence 0.15 is an optimal threshold value.

Figure 10 Distributions of within/between-cluster FPD (see online version for colours)



6.1 Upper bound of FPD comparisons

Let the database consists of N prints and m initial clusters. We assume that it is *equally likely* that a print p is to be clustered into any of m existing clusters, or becomes a new cluster, i.e., the probability of creating a new cluster is $\frac{1}{m+1}$. The 1st print needs to be compared with m clusters. For the 2nd print, the *expected* number of clusters to be compared is $m \cdot \frac{m}{m+1} + (m+1) \cdot \frac{1}{m+1} < m + \frac{1}{m}$. For the 3rd print, this expectation³ will be $m \cdot \frac{m^2}{(m+1)^2} + (m+1) \cdot [\frac{1}{m+2} + \frac{1}{(m+1)^2}] + \frac{m+2}{(m+1)(m+2)} < m + \frac{2}{m}$. In general, the expected number of clusters with which the k^{th} ($k = 1, \dots, N$) print needs to be compared is less than $m + \frac{k-1}{m}$. Then the total number of comparisons with existing clusters for N prints⁴ will not exceed $Nm + \frac{N(N-1)}{2m}$. In our case, $N = 2,660$ and $m = 33$ hence the number of FPD comparisons will not exceed 194,180. However, hierarchical clustering would require ${}_{2,660}C_2 = 3,536,470$ FPD computations to build the similarity matrix. So the

proposed clustering method reduces computation load by 94.5%. This reduction speeds up the retrieval and it is evident from Table 3.

Table 3 Retrieval speed

<i># of scene marks used in experiments</i>	<i>Average time per query without clustering</i>	<i>Average time per query with clustering</i>
300	120 minutes	10 minutes

7 Experiments and results

The dataset we used contains 2,660 known prints and 300 test images from Foster and Freeman. The known prints (shown in Figure 11) were created by taking impressions of footwear outsoles provided by footwear vendors, while the test images are real crime scene marks (shown in Figure 12), which are mostly partial, were taken at different illumination/orientation. Each known print has meta-data information such as the brand and model of the footwear. The resolution of database images and crime scene images varies from 72 dpi to 150 dpi and 72 dpi to 240 dpi respectively. Known prints and crime scene marks come from *disjoint* sources. Forensic experts provided the ground truth of matching prints for every crime scene mark.

Figure 11 Sample known footwear prints

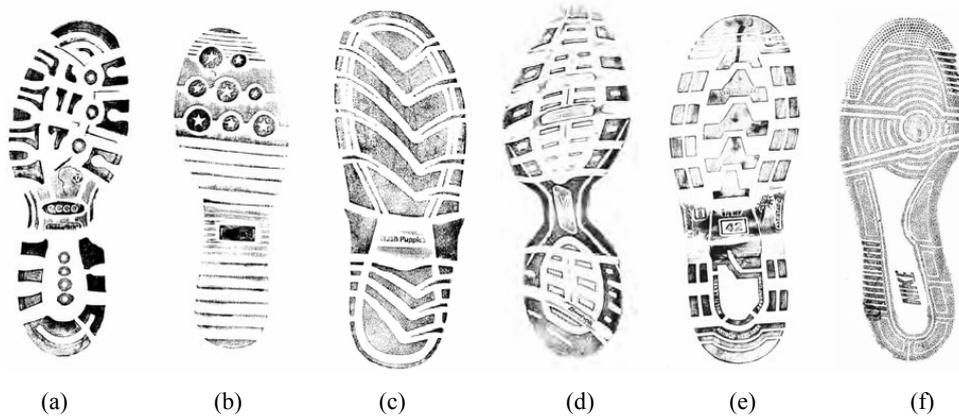
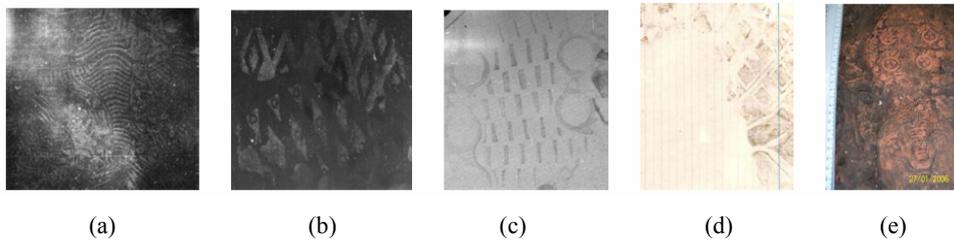


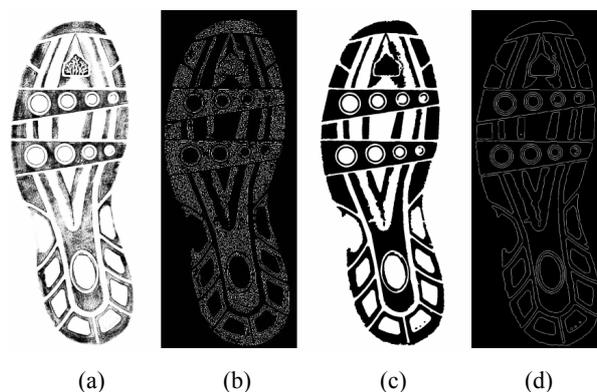
Figure 12 Sample crime scene images (see online version for colours)



We conducted the experiments in the following steps:

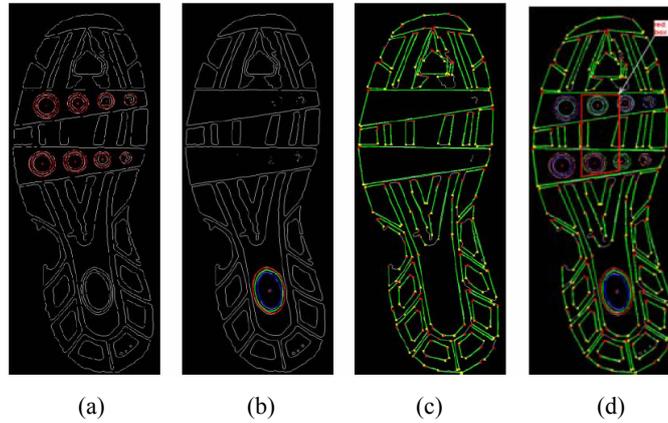
- Step 1 The first step in the feature extraction is to perform morphological operations such as dilation and erosion. This helps to enhance the quality of the edge image, which is used as input for circle detection. Result of Step 1 for a sample footwear print is shown in Figure 13.
- Step 2 SHT is used to detect circles in footwear prints. Pixels of detected circles shown in Figure 14(a) are removed from the edge image and fed as input for ellipse detection. Pixels of detected ellipses shown in Figure 14(b) are removed from the edge image and the output is fed as input for line detection. Figure 14(c) shows the detected line segments. Features are extracted in the order of circle, ellipse and line. This is because circles are special case of ellipses and arbitrary shapes in footwear print are approximated by piecewise lines. Figure 14(d) sums up all the features.
- Step 3 For each detected feature, node and edge attributes shown in Table 2 are calculated and finally an ARG is constructed. One such ARG is shown in Figure 15.
- Step 4 From visual inspection, 33 recurring patterns (shown in Figure 8) were determined. Using Algorithm 4, 1,460 footwear prints were associated with one of the 33 clusters whereas the remaining 1,200 prints were so unique that each of them was a cluster by itself. Sample clusters from the clustered database are shown in Figure 16.
- Step 5 In the clustered database, real crime scene marks were used as queries in the retrieval system (shown in Figure 1) and the closest prints were retrieved. The crime scene mark was *first* matched against every cluster representative to find the closest cluster. *Then within the closest cluster*, crime scene mark was matched against each of the prints and the top k matches from the cluster were retrieved. Sample retrieval results for four crime scene marks are shown in Figure 17.

Figure 13 Results of morphological operations on sample image, (a) original grey-scale image (b) edge image of (a) (c) result of morphological operations on (a) (d) edge image of (c)



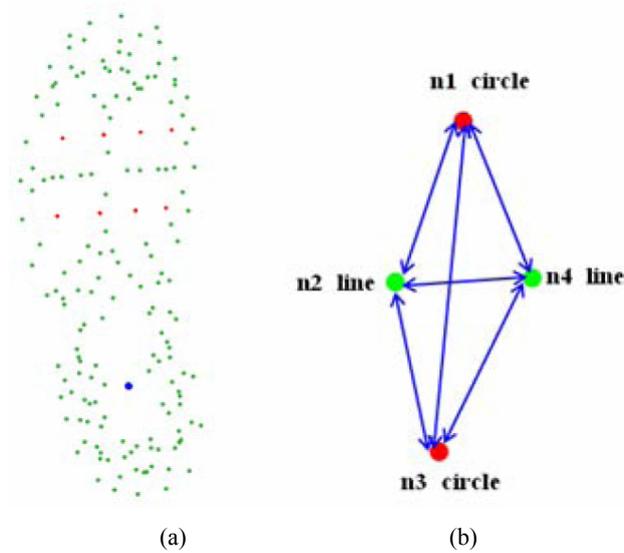
Note: Morphological operations improve the quality of edge image significantly.

Figure 14 Results of extracting shape features from edge image in Figure 13(d) in the order of circle → ellipse → line, (a) circles (b) ellipses (c) line segments (d) all features (see online version for colours)



Notes: Red box in (d) indicates a small region in the footwear print. Best viewed in colour.

Figure 15 ARG construction, (a) ARG for footwear print shown in Figure 13(a) (b) subgraph of graph (a) for the region enclosed within the red box of Figure 14(d) (see online version for colours)

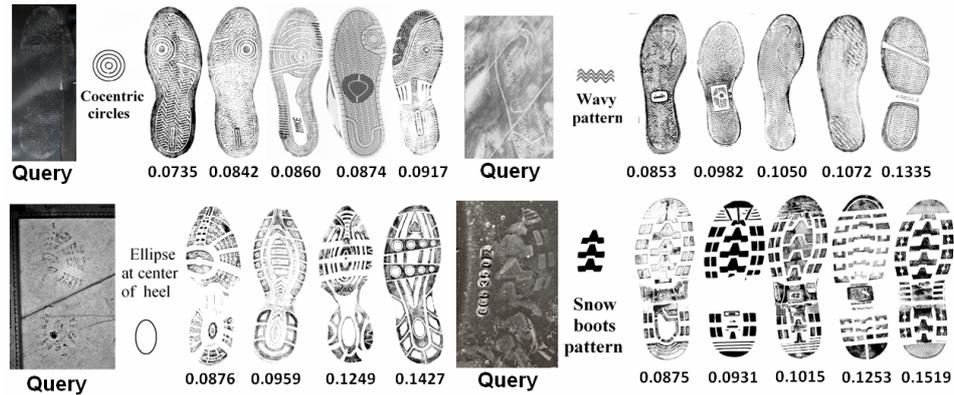


Notes: Green and red dots represent lines and circles respectively. Please note this ARG is fully connected and edges are omitted for clarity. Best viewed in colour.

Figure 16 Sample clusters from clustered database (see online version for colours)

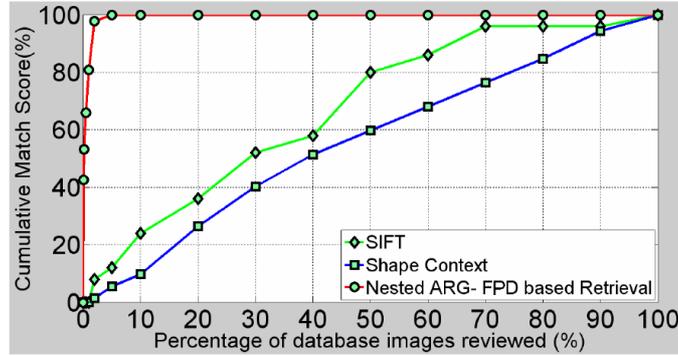
Cluster Name	Cluster Representative	Sample prints in the cluster for the cluster representative	Cluster Name	Cluster Representative	Sample prints in the cluster for the cluster representative
Wavy Pattern			Parallel Lines		
Recurring Plus			Circles Dominated		
Partial Circles Interleaves			Triangles Dominated		

Figure 17 Retrieval results (*top matches*) for sample crime scene marks



Note: For each query, the matched cluster representative and the top four or five matches along with the corresponding FPDs are listed.

To evaluate the system’s performance, the retrieval system was tested on 300 real crime scene marks. CMC and mean average precision (MAP) were used as performance metrics. CMC is the probability of finding a match in the first n percent of the sorted database. The probability of a match [cumulative match score (CMS)] is estimated by the proportion of times a matching footwear print appears in the first n percent of the sorted database. SIFT (Lowe, 2004) is the state-of-the-art image matching algorithm and shape context-based method (Belongie et al., 2002) has been widely used in matching shape patterns. Hence, we compare our approach against SIFT and shape context. The CMC curve of the three methods is shown in Figure 18. CMC curve of our approach remains the same before and after clustering, but it is evident from Table 3 that clustering gives 12 times of improvement in retrieval speed. Further, CMS of our system is much higher than the other two approaches.

Figure 18 CMC of proposed approach, SIFT and shape context-based approach (see online version for colours)

MAP provides a single measure of quality across recall levels. Let the set of relevant items for a query $q_j \in Q$ be $\{i_1, \dots, i_{n_j}\}$ and R_{jk} be the set of ranked retrieval results from the top result until relevant item i_k , then *MAP* is defined as

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{n_j} \sum_{k=1}^{n_j} Precision(R_{jk}) \quad (18)$$

where $Precision(R_{jk}) = \frac{\# \text{ of relevant items retrieved}}{\# \text{ of retrieved items}}$. *MAP* of our approach, SIFT and shape context is shown in Table 4.

Table 4 *MAP* comparison with SIFT and shape context

<i>Method</i>	<i>SIFT</i>	<i>Shape context</i>	<i>Proposed approach</i>
<i>MAP</i>	0.17	0.10	0.56

Since there is no benchmark dataset of real crime scene marks, direct comparison with the existing footwear print retrieval system is impractical but most of the existing systems (summarised in Table 5) work with only synthetic and clean (i.e., almost noise-free) prints. None of them have designed their system to capture features from crime scene marks. Dardi et al. (2009) have used three techniques (luminance image, images processed by Laplacian edge detector and Canny edge detector) to retrieve closest print(s) for 30 crime scene marks. *CMS* of the three methods varies for a given query and the *CMS* of the best match in at least one of the three methods (considering highest ranking technique for each mark) is reported as 74% at 10% of the sorted database. But their method is tested only with a very small dataset (87 known prints and 30 real crime scene marks). Table 5 compares our approach with existing systems in terms of performance and method of conducting experiments.

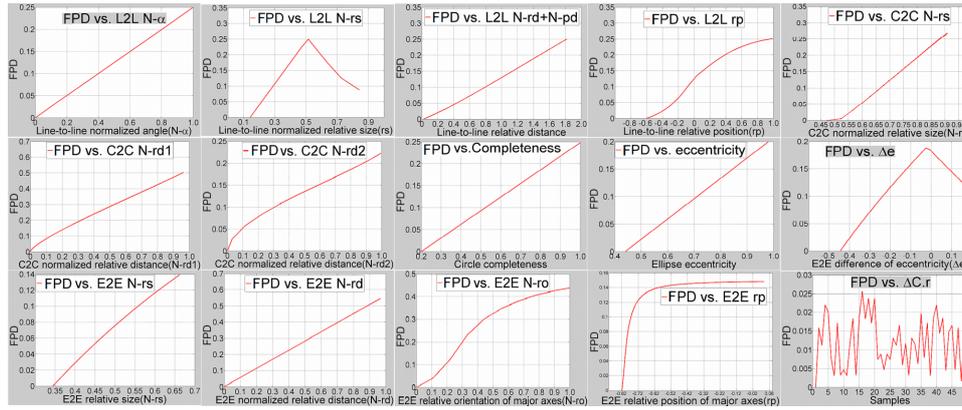
Table 5 Comparison of our nested ARG-FPD-based approach with the state-of-the-art footwear print retrieval systems

Footwear print retrieval systems	Full print		Partial print		Experiments with crime scene marks	Short-coming	Dataset used
	CMS at 1%	CMS at 10%	CMS at 1%	CMS at 10%			
Chazal et al. (2005)	64	90	50	77	-	Lacks scaling invariance	475 prints from dataset of Forensic Science Laboratory, Dublin, Ireland
Zhang et al. (2005)	85.4	97.44	-	-	-	Not tested with partial prints	512 prints from Foster and Freeman Dataset
Pavlou and Allinson (2006)	86	93	85	92	-	Not tested with real crime scene marks	368 prints provided by Forensic Science Services dataset, UK
Crookes et al. (2007)	100	100	100	100	-	Tested only with synthesised SoCs	500 clean prints and 50 degraded prints
Crookes et al. (2007)	100	100	100	100	-	Lacks rotational invariance	100 clean prints and 64 synthetic scene images
Gueham et al. (2008)	-	-	-	95.68	-	Tested only with 100 prints	100 prints from Foster and Freeman Dataset
Dardi et al. (2009)	-	-	-	-	CMS at 10%:73% CMS at 5%:40% CMS at 1%:10%	87 known prints and 30 SoCs	87 known prints and 30 real crime scenes from ENSFI group
Nested ARG-FPD	100	100	100	100	CMS at 10%:100% CMS at 5%:100% CMS at 1%:70%	-	2,660 known prints and 300 real crime scene marks from Foster and Freeman Dataset

8 Sensitivity analysis

Sensitivity analysis (Smith et al., 2008) is a powerful system validation technique, so we conducted extensive experiments in a subset of 1,400 prints to investigate how sensitive FPD is to changes in attributes defined in Table 2. Results are shown in Figure 19.

Figure 19 Sensitivity of FPD to variation in each attribute shows that FPD is *insensitive* to small errors in the feature extraction (see online version for colours)



Notes: There are 15 plots – from top to bottom, left to right: labelled (a)-(e), (f)-(j), (k)-(o). A linear correlation is seen between FPD and most attributes indicating that FPD *consistently* approximates human perceptual distance.

There are 15 graphs shown (labelled *a-o*), each of which is a plot of distance with respect to one of the attributes. The first four graphs correspond to how the distance between two lines (L2L) changes as attributes $N-\alpha$, $N-rs$, $N-rd$, $N-pd$, and rp are varied. The next four graphs correspond to how the distance between two circles (C2C) changes as the attributes $N-rs$, $N-rd_1$, $N-rd_2$, and completeness of circle are varied. Similarly, the attributes eccentricity ratio, Δe , $N-rs$, $N-rd$, $N-ro$, and rp are varied to find the change in distance between two ellipses (E2E) in the subsequent six graphs. The final graph varies the radius of one of the two circles to find the change in distance between two circles (C2C).

A linear correlation is seen between distance and most attributes which clearly indicate that FPD measures human perceptual distance. The four plots (*b, j, n, o*) show non-linear behaviour. The reason for their behaviour are explained as follows:

- (b) L2L $N-rs$: for instance, let the two lines be l_1 and l_2 . Initially, $l_1.len > l_2.len$, as l_1 becomes shorter than l_2 ($l_1.len < l_2.len$) the algorithm will start to minimise FPD hence FPD drops after reaching a peak.
- (j) E2E Δe : similar reason as (b).
- (n) E2E rp : when the major axes of two ellipses are far, the rate of change in rp becomes very small hence FPD initially increases fast and then saturates.
- (o) C2C $\Delta C.r$: when the radius r of one of the two circles vary randomly within 15%, the change of FPD is always below 0.025.

9 Conclusions

This paper proposes geometric shapes like line segments, circles and ellipses as features for crime scene footwear marks and presents algorithm to detect line segments and ellipses in footwear prints. These features provides high distinctiveness and are very effective in retrieving similar prints for a partial and noisy crime scene mark. This distinctiveness is achieved by structurally representing the geometric shapes in the form of an ARG which captures the spatial relationships such as relative dimension, relative size, relative distance and orientation between every geometric shapes in a print. The attributes for every shape are defined in a way to provide scaling, rotation and translation invariance. Further, FPD, a perceptual similarity measure for matching degraded footwear prints has been introduced. The trade-off between accuracy and speed of FPD is overcome by clustering the database prints beforehand. A database of known prints is clustered based on recurring outsole patterns.

Eventually, we have proposed a system to retrieve similar footwear prints for real crime scene marks, an unsolved task in the field of computational forensics. From experimental results it is evident that the retrieval speed improves significantly with the clustered database and FPD plays a crucial role in both retrieval and clustering of footwear prints by ensuring robust matching of prints. Further, sensitivity analysis of FPD showed that it is very robust to the small changes in the attributes of ARG. CMCs and MAP of our approach were compared with the state-of-the-art image matching algorithm SIFT and widely used shape matching algorithm shape context to show the efficacy of our system. Results from Foster and Freeman Dataset were compared with the existing footwear print retrieval systems to show that our system outperforms the state-of-the-art footwear print retrieval systems.

Future direction is to speed up the FPD computation further by using approximation methods (Huet and Hancock, 1999; Pavlou and Allinson, 2009) and pre-filtering the whole database using histogram-based indexing (Grauman and Darell, 2004; Berg et al., 2005). Another direction is to use the proposed system to get the best match for the crime scene mark and use that to interpret the result in terms of the strength of the evidence it can support to find if the suspect is the offender. One can use the Likelihood ratio approach (Everett et al., 1998) to interpret the strength of the evidence. As future work, this approach will be investigated to come up with a probabilistic framework for evidence interpretation of footwear print retrieval and identification. Other potential application is to combine multiple marks from crime scene and use it to enhance the evidence strength.

Acknowledgements

This work was funded in part by award 2007-DN-BX-K135 of the National Institute of Justice, Office of Justice Programmes, and US Department of Justice. Points of view expressed are those of the authors and do not necessarily represent the official position or policies of the US DoJ. We would like to thank Foster and Freeman, Ltd. for providing us datasets.

References

- Aldenderfer, M.S. and Blashfield, R.K. (1984) *Cluster Analysis*, SAGE Publications, Beverly Hills, CA, USA.
- Alexander, A., Bouridane, A. and Crookes, D. (1999) 'Automatic classification and recognition of shoeprints', *Proceedings of the Seventh International Conference on Image Processing and its Applications*, Manchester, Vol. 2, pp.638–641.
- AlGarni, G. and Hamiane, M. (2008) 'A novel technique for automatic shoeprint image retrieval', *Forensic Science International*, Vol. 181, No. 1, pp.10–14.
- Basu, S., Banerjee, A. and Mooney, R.J. (2002) 'Semi-supervised clustering by seeding', *International Conference on Machine Learning*, pp.27–34.
- Belongie, S., Malik, J. and Puzicha, J. (2002) 'Shape matching and object recognition using shape contexts', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 4, pp.509–522.
- Berg, A.C., Berg, T.L. and Malik, J. (2005) 'Shape matching and object recognition using low distortion correspondences', *CVPR 2005: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp.26–33.
- Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, Springer, Secaucus, NJ, USA.
- Bunke, H. and Messmer, B.T. (1995) 'efficient attributed graph matching and its application to image analysis', *Proceedings of the 8th International Conference on Image Analysis and Processing*, pp.45–55.
- Bunke, H., Irniger, C. and Neuhaus, M. (2008) 'Graph matching challenges and potential solutions', *International Conference on Image Analysis and Processing*, Vol. 3617, pp.1–10.
- Canny, J. (1986) 'A computational approach to edge detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 6, pp.679–698.
- Chazal, P.D., Flynn, J. and Reilly, R.B. (2005) 'Automated processing of shoeprint images based on the Fourier transform for use in forensic science', *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 3, pp.341–350.
- Crookes, D., Bouridane, A., Su, H. and Gueham, M. (2007) 'Following the footsteps of others: techniques for automatic shoeprint classification', *Second NASA/ESA Conference on Adaptive Hardware and Systems*, pp.67–74.
- Dardi, F., Cervelli, F. and Carrato, S. (2009) 'A texture based shoe retrieval system for shoe marks of real crime scenes', *International Conference on Image Analysis and Processing*, Vol. 5716, pp.384–393.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society*, Vol. 39, No. 1, pp.1–38.
- Everett, I.W., Lambert, J.A. and Buckleton, J.S.(1998) 'A Bayesian approach to interpreting footwear marks in forensic casework', *Science & Justice*, Vol. 38, No. 4, pp.241–247.
- Foster and Freeman, 'Solemate', (online) available at <http://www.fosterfreeman.com/>.
- Geradts, Z. and Keijzer, J. (1996) 'The image database REBEZO for shoeprints with developments on automatic classification of shoe outsole designs', *Forensic Science International*, Vol. 82, No. 1, pp.21–31.
- Ghouthi, L., Bouridane, A. and Crookes, D. (2006) 'Classification of shoeprint images using directional filter banks', *International Conference on Visual Information Engineering*, pp.167–173.
- Girod, A. (1982) 'Computerized classification of the shoeprint of Burglars shoes', *Forensic Science Int.*, Vol. 1, No. 1, pp.59–65.
- Goulermas, J. and Liatsis, P. (1999) 'Incorporating gradient estimations in a circle-finding probabilistic Hough transform', *Pattern Analysis and Applications*, Vol. 2, No. 3, pp.239–250.

- Grauman, K. and Darrell, T. (2004) 'Fast contour matching using approximate earth mover's distance', *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp.220–227.
- Gueham, M., Bouridane, A. and Crookes, D. (2008) 'Automatic classification of partial shoeprints using advanced correlation filters for use in forensic science', *International Conference on Pattern Recognition*, pp.1–4.
- Haralick, R.M. and Shapiro, L.G. (1992) *Computer and Robot Vision*, Addison Wesley, Boston, MA, USA.
- Hillier, F.S. and Liebermann, G.J. (1995) *Introduction to Mathematical Programming*, McGraw-Hill, New York.
- Hough, P.V. (1962) 'Method and means for recognizing complex patterns', US Patent 3069654.
- Huet, B. and Hancock, E.R. (1999) 'Line pattern retrieval using relational histograms', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 12, pp.1363–1370.
- Jingl, M.Q., Ho, W.J. and Chen, L.H. (2009) 'A novel method for shoeprints recognition and classification', *International Conference on Machine Learning and Cybernetics*, Vol. 5, pp.2846–2851.
- Kim, D.H., Yun, I.D. and Lee, S.U. (2010) 'Attributed relational graph matching algorithm based on nested assignment structure', *Pattern Recognition*, Vol. 43, No. 3, pp.914–928.
- Liu, X. and Croft, W.B. (2004) 'Cluster-based retrieval using language models', *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.186–193.
- Lowe, D.G. (2004) 'Distinctive image features from scale-invariant keypoints', *International Journal of Computer Vision*, Vol. 60, No. 2, pp.91–110.
- McLaughlin, R. (1996) 'Randomized Hough transform: better ellipse detection', *TENCON 1996: IEEE Digital Signal Processing Applications*, Perth, Vol. 1, pp.409–414.
- Mikkonen, S. and Astikainen, T. (1994) 'Databased classification system for shoe sole patterns – identification of partial footwear impression found at a scene of crime', *Journal of Forensic Science*, Vol. 39, No. 5, pp.1227–1236.
- Mikkonen, S., Suominen, V. and Heinonen, P. (1996) 'Use of footwear impressions in crime scene investigations assisted by computerised footwear collection system', *Forensic Science Int.*, Vol. 82, No. 1, pp.67–79.
- Nibouche, O., Bouridane, A., Gueham, M. and Laadjel, M. (2009) 'Rotation invariant matching of partial shoeprints', *International Machine Vision and Image Processing Conference*, pp.94–98.
- Nixon, M. and Aguado, A. (2002) *Feature Extraction and Image Processing*, Academic Press, Elsevier.
- Patil, P.M. and Kulkarni, J.V. (2009) 'Rotation and intensity invariant shoeprint matching using Gabor transform with application to forensic science', *Pattern Recognition*, Vol. 42, No. 7, pp.1308–1317.
- Pavlou, M. and Allinson, N.M. (2006) 'Automatic extraction and classification of footwear patterns', *Lecture Notes in Computer Science, Proc. Intelligent Data Engineering and Automated Learning*, pp.721–728.
- Pavlou, M. and Allinson, N.M. (2009) 'Automated encoding of footwear patterns for fast indexing', *Image Vision Computing*, Vol. 27, No. 4, pp.402–409.
- Pelillo, M., Siddiqi, K. and Zucker, S.W. (2001) 'Many-to-many matching of attributed trees using association graphs and game dynamics', *Proceedings of 4th International Workshop on Visual Form*, pp.583–593.
- Rubner, Y., Tomasi, C. and Guibas, L.J. (2000) 'The earth mover's distance as a metric for image retrieval', *International Journal of Computer Vision*, Vol. 40, No. 2, pp.99–121.
- Rui, Y., Huang, T.S. and Chang, S-F. (1999) 'Image retrieval: current techniques, promising directions, and open issues', *Journal of Visual Communication and Image Representation*, Vol. 10, No. 1, pp.39–62.

- Sanfeliu, A. and Fu, K.S. (1983) 'A distance measure between attributed relational graphs for pattern recognition', *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 13, No. 3, pp.353–362.
- Smith, E.D., Szidarovszky, F., Karnavas, W.J. and Bahill, A.T. (2008) 'Sensitivity analysis, a powerful system validation technique', *The Open Cybernetics and Systemics Journal*, Vol. 2, No. 1, pp.39–56.
- Su, H., Crookes, D. and Bouridane, A. (2007) 'Shoeprint Image retrieval by topological and pattern spectra', *International Machine Vision and Image Processing Conference*, Vol. 27, pp.341–350.
- Sun, W., Taniar, D. and Torabi, T. (2008) 'Image mining: a case for clustering shoe prints', *International Journal of Information Technology and Web Engineering*, Vol. 3, No. 1, pp.70–84.
- The MathWorks, Inc., 'Measure properties of image regions', available at <http://www.mathworks.com/help/toolbox/images/ref/regionprops.html>.
- Wang, R., Hong, W. and Yang, N. (2009) 'The research on footprint recognition method based on wavelet and fuzzy neural network', *International Conference on Hybrid Intelligent Systems*, pp.428–432.
- Xiao, R. and Shi, P. (2008) 'Computerized matching of shoeprints based on sole pattern', *Lecture Notes in Computer Science; Proceedings of the 2nd International Workshop on Computational Forensics*, Vol. 5158, pp.96–104.
- Xu, L. and Oja, E. (1993) 'Randomized Hough transform (RHT): basic mechanisms, algorithms, and computational complexities', *Computer Vision Graphics and Image Processing: Image Understanding*, Vol. 57, No. 2, pp.131–154.
- Zhang, L. and Allinson, N. (2005) 'Automatic shoeprint retrieval system for use in forensic investigations', *5th Annual UK Workshop on Computational Intelligence*.

Notes

- 1 There are around 50,000 foreground pixels in a footwear print of typical size 600×800 and picking three foreground pixels from them in random will never narrow down to the right ellipse.
- 2 The eccentricity (The MathWorks, Inc., <http://www.mathworks.com/help/toolbox/images/ref/regionprops.html>) of one component/region is defined as the eccentricity of the ellipse that has the same second-moments as the region.
- 3 Let x be a random variable that denotes the number of clusters which need to be compared with the 3rd print. Then

$$\Pr(x = m) = \frac{m}{m+1} \cdot \frac{m}{m+1} = \frac{m^2}{(m+1)^2},$$

$$\Pr(x = m + 1) = \frac{1}{m+1} \cdot \frac{m+1}{m+2} + \frac{m}{m+1} \cdot \frac{1}{m+1} = \frac{1}{m+2} + \frac{m}{(m+1)^2},$$

$$\Pr(x = m + 2) = \frac{1}{m+1} \cdot \frac{1}{m+2}.$$

- 4 This total is less than

$$m + \left(m + \frac{1}{m}\right) + \dots + \left(m + \frac{k-1}{m}\right) + \dots + \left(m + \frac{N-1}{m}\right) = Nm + \frac{N(N-1)}{2m}.$$