Expert Assisting Computerized System for Evaluating the Degree of Certainty in 2D Shoeprints

Final Technical Report

TP-3211

SUBMITTED TO:
U.S. Department of Justice
Office of Justice Programs
National Institute of Justice
810 Seventh Street N.W.
Washington, DC 20531

AUTHORS:

Yoram Yekutieli
Hadassah Academic College
Dept. of Computer Science
37 Hanevi'im St. P.O.Box 1114,
Jerusalem, Israel 9101001
Email: yoramye@hadassah.ac.il

Yaron Shor, Sarena Wiesner & Tsadok Tsach
Israel National Police
Division of Identification and Forensic Science (DIFS)
1 Bar-Lev Road
Jerusalem, Israel 91906
Email: yaronshor@gmail.com
sarenawiz@gmail.com
ttsach@gmail.com

September 2016[1]

[1] This report describes the work done in a 2-years project (with half a year amendment) funded by the NIJ (Task plan 3211) that took place from 2010 to 2012. The report was first submitted on December 2012. Different aspects of this work were presented since 2011 (see chapter 4 Dissemination of research findings).

Abstract

In recent years there is a growing demand to fortify the scientific basis of forensic methodology. The 2009 NAS report: "Strengthening Forensic Science in the United States: A Path Forward" focused on the need for a more profound scientific base for forensic science in general, and specifically for pattern comparison.

The demand for numerical and statistical presentation of the comparison results is seldom fulfilled in the pattern comparison field. Since meaningful databases that can support such calculations do not exist yet, even the best expert's work is based solely on expertise and experience. Many argue, with much justification, that an expert, even with 20 years of experience, cannot merely rely on his memory when trying to evaluate an examined phenomenon's weight and frequency of appearance. Such comparison results will not yield consistent numbers. Without a good methodology to compare a feature to many others of its kind, the investigator cannot reliably calculate the criteria to estimate what level of certainty applies in a particular case. Thus any expert's assertion about-the level of certainty in a given case will lack a sufficient numerical and objective base.

This study answers how to evaluate the rareness of acquired accidentals, also known as randomly acquired characteristics (RACs) such as scratches, nicks, tears and holes, as they appear on shoe sole test impressions, and how to address the presence of an array of accidentals on one sole. To achieve this goal, three tools were developed:

1. A large, consistent and adequate database of accidentals. All the accidentals were scanned and their contours digitized.
2. Statistical models to determine chances of finding similar accidentals on other soles in three aspects: shape, location and orientation.
3. Practical tools enabling the shoeprint expert to mark new accidentals and quickly evaluate their evidential value in a more scientifically reliable manner than previously done.

The database was created in the following manner: Test impressions were prepared from worn shoe soles. Accidentals appearing on the sole, such as cuts and scratches, were semi-automatically marked by a qualified examiner using the specially designed "MarkAccidentals" software, developed in this study. Each characteristic was automatically stored in the database, with its digital representation, including its location and calculated orientation. Nearly 9,000 accidentals were recorded and stored.

Since calculating the probability of the accidental's location required many more accidentals, the locations of 20,000 more accidentals were recorded using dedicated software we developed (FaDeMa) that enables rapid location marking. Our database contains test impressions made from shoes of many different designs, and in order to compensate for the effect of different sole designs – the locations were normalized to a standard (universal) sole.

The orientation of each accidental and its accuracy were calculated, and saved in the database.

The statistical algorithm developed, enables the **SESA** (***S**tatistic **E**valuation of **S**hoeprint **A**ccidentals*) system to calculate the probability of finding another feature similar to a particular feature of a scanned and digitized accidental. The system can calculate the chance to get another accidental in the same location, or the same orientation as the examined one. Combining all the probabilities together (by multiplication, assuming independence) reveals the result of the comparison. The number received at the end of the process serves the expert as a guiding number, allowing more objective and accurate results and conclusions.

The system works so far only on test impression, and the research to allow the evaluation of real shoeprint is yet to follow. This work appeared in preliminary form in conferences (see chapter 4 *Dissemination of research findings*).

Contents

Executive Summary

In recent years demand has grown to fortify the scientific basis of forensic methodology. The 2009 report of the NAS: "Strengthening Forensic Science in the United States: A Path Forward" focused on the need for a more profound scientific base of forensic science in general and specifically of pattern comparison. Several branches of the forensic science had already begun the efforts to calculate and present their findings in a more scientific way.

The existing pattern comparison field seldom fulfills the demand for quantitative and statistical presentation of results. Since there are no meaningful databases that can support such calculations as yet even the best expert's comparative work is based solely on expertise and experience. Without a good methodology of comparing a feature against many others of its kind, investigators lack a reliable method of calculating the criteria for determining what level of certainty to address in a particular case. Furthermore, the level of certainty given by the expert will lack sufficient quantitative or objective data.

When an expert determines a level of certainty that a specific mark (e.g. a hand writing sample, or a striation mark of accidental on shoe sole) was created by a specific object (or man), he must know (or knowledgeably estimate) the possibility that another mark will look like the examined one. The degree of resemblance between the "correct" match and the "wrong" match influences the chance of the expert to err.

Currently in most areas of forensic science, the databases and procedures for determining such probabilities do not exist. The problem is inherent to pattern comparison methods: comparing two visual features does not yield a numerical result. This project overcame the shortcoming by creating of a computerized database of digital representations of the examined phenomenon- in our case- accidentals on shoe soles.

Accidentals, also known as randomly Acquired Characteristics (RACs) are features on a shoe outsole resulting from random events of removing or adding of material to the outsole. These include, but not limited to: cuts, scratches, tears, holes, stone holds, abrasions and the acquisition of debris. The location, orientation and shape of these characteristics contribute to the uniqueness of a shoe outsole. The location is defined as the location of the accidental on the shoe-sole. The orientation is determined as the angle between the long axes of the accidental in relation to the long axes of the shoe-sole. The shape is the contour of the accidental. Accidentals change over time as the shoe is further used but for comparison purposes, the accidental's features at a set time are examined. Our mission was to build a system for assisting shoeprint examiners in assessing the

evidential value, or rarity, of acquired accidental defects on shoe soles. Evidential value is highly dependent on the rarity of the accidentals.

The question of rarity of an accidental was dealt with by estimating the probability of occurrence of the different features of an accidental: its location, orientation and shape. Estimating those probabilities is a nontrivial process composed of the following steps:

(1) Extracting the features and building the database. This in itself is a multi-step process.

(a) The spatial features (location and orientation of an accidental) must be defined with respect to a coordinate system. To compare numerous accidentals on many test impressions of various models, we developed a method for aligning the test impressions. This was done by defining a *standard (universal) coordinate system* of test impressions. The different test impressions were each assigned a specific shoe coordinate system by using consistent anchor points and finding the long axis. This coordinate system was later used both as a reference frame for the spatial features of the specific shoe and in the joint alignment of many shoes.

(b) Methods were developed for measuring each of the features. The shape of an accidental was extracted semi automatically by segmenting the image to foreground and background. The extracted shape was corrected using human intervention when needed. The result was a digital representation of the accidental's contour. The location of an accidental was calculated as the centroid of the contour, with respect to the shoe alignment. The orientation of an accidental was extracted from the shape by using PCA[2] to find the major and minor axis of the shape. Orientation was defined as the angle of the major axis with respect to the X axis of the shoe-specific coordinate system.

(c) Tools for marking numerous accidentals were used either for extracting the shape and location, or for marking accidentals merely by their locations:

(1) The CONTOURS data set composed of nearly 9,000 accidentals from 300 test impressions. For each accidental we have its contour and location (other attributes such as orientation are derived later from the contour). The CONTOURS data set is delivered with the SESA DVD, and is used directly by the delivered software tools.

(2) The FAST data set composed of 20,000 accidentals from 600 test impressions, only the location marked for each accidental.

[2] Principal component analysis performed by Eigen decomposition (see section 2.5.2).

(2) Deriving the statistical model for the rarity of accidentals. The following steps were taken:

(a) Error estimation: Having the database, we estimated the different components of the model: the distribution of each feature, and the error in estimating the values.

- The error in location was found to be largely composed of the error in assigning the shoe-aligned coordinate system, and to a lesser extent on the error in estimating the centroid of the accidentals. The magnitude of location error was estimated to be 0.5 cm.
- The error in estimating orientation was found to be mainly related to the shape of the accidental. Elongated shapes had much lower orientation errors compared to condensed, rounded shapes. We therefore developed the *orientation score* – the degree of elongation of a shape. Orientation scores ranged from a little above 1 for circular shapes to 100 or higher for the longest, narrowest shapes. We described orientation error as a function of orientation score. Error values ranged from very high errors of 60 degrees or more for the rounded shapes, to 1-2 degrees, for the longest shapes.
- The error in shape estimation was mainly dependent on the variability among different test impressions of the same shoe, and to a lesser extent on the variability between the human operators that operated the software. To estimate shape error we developed a shape dissimilarity measure. We described the distribution of dissimilarity values of two groups: The first is the population of correct matches, i.e. pairs of contours that originated from the same accidental (and are different because of differences between the test impressions and the marking process). The second population is of the non-matches, i.e. pairs of contours that originated from different accidentals. As expected, we found that the correct matches had lower dissimilarity values than the non-matches. The two distributions overlapped, meaning that, in some cases, different accidentals resemble each other even more than two impressions of the same accidental. This happened mainly for small accidentals (few millimeters in diameter).

(b) Describing the distributions of features: For each of the two sets of accidentals (n = ~8,900 and ~20,000) we superimposed the locations on one universal coordinate system.

The distribution of locations of accidental was estimated using a 2D kernel estimation technique[3]. We proposed a method to compensate for the non-equal contribution of different shoe types and wear patterns by developing the notion of *accumulated contact areas* of the shoes. Using those methods we showed that the distribution of accidental location is approximately uniform. Since we were not able to estimate non-uniformity accurately, we used a uniform assumption for the statistical model of locations.

The distribution of orientation was found to be uniform. We separated the accidentals into groups according to their orientation score and found that only the longest, narrowest shapes had any clear divergence from uniformity of orientations. The distribution of the longest accidentals showed peaks in the direction of the long axis of the shoe.

Since shape is a high dimensional feature (as opposed to the two dimensional location and one dimensional orientation) we did not estimate its distribution directly. Instead we described the distribution of shape dissimilarity values (as noted above). We found an optimal threshold between the distributions of the two populations of correct matches and the non-matches. Two shapes that have a dissimilarity value below the threshold have a higher probability of originating from the same accidental. If their dissimilarity value is above the threshold, they probably belong to different accidentals. This was used to estimate rarity of a new shape in the following way. Each new shape was compared to all (~8,900) shapes in our database. The number of comparisons with dissimilarity values below the threshold was found. This is the number of shapes that were similar enough to the target shape, to be considered as originating from the same accidental. The proportion of this number relative to the size of the database is an estimate of the new shape's rarity. We found that small accidentals (of few millimeters in length) had more similar results than larger accidentals. This is related to the uniqueness of shapes. A large accidental has enough information to be identified in spite of the noise (the error in estimating its shape).

(c) Combining the estimated errors and distribution of features to estimate the rarity:

As noted above, the error in estimating the location of accidental was found to be ~0.5 cm in both dimensions. A shoe has an area of approximately 300 cm^2. Adopting a grid metaphor for the locations of accidentals, we estimated the number

[3] Kernel density estimation – a non-parametric way to estimate the probability density function. See section 2.4.7 for details.

of grid cells by dividing the shoe area by the squared error, yielding 1,200 cells. We also found that each shoe has, on average, 30 noticeable (to the naked eye) accidentals. Assuming that the locations are uniformly distributed we concluded that the probability of locating an accidental in one of the grid cells is 30/1,200 = 1/40.

The probability of finding an accidental with a specific orientation was calculated in the following manner. First we used its shape to calculate its orientation score. Using the score we found its orientation error. Since the distribution of orientations was found to be uniform, we concluded that the probability of finding an accidental with a specific orientation is its orientation error divided by 180^0 (180^0 is the full range of possible orientations). The implication is that round accidentals are not unique – the error in estimating their orientation is so big that they give little or no information. On the other hand, finding the orientation of longer, narrower accidentals is more accurate, thus these accidentals are rarer. For example, such an accidental with an orientation score of 150 has an orientation error of 2.25^0 and a probability of occurrence of 1/80.

The probability of finding an accidental with a specific shape was already described above.

(d) Producing one number: The probability of finding an accidental similar to a target one is product of multiplying the three probabilities (finding similar features of location, orientation and shape). This calculation is based on the assumption that the probabilities of the three features are independent. This assumption was not validated during this research.

(3) Building software tools for shoeprint comparison experts, which enable marking new accidentals and assessing their rarity:

Dedicated software tools were built during the different phases of research. Those tools include modules for collecting the data, marking the location and shape of accidentals, establishing a common coordinate system on test impressions, extracting the different features, measuring the errors, analyzing the statistics and presenting the results in a clear, didactical way. A complete software package for the aid of the shoeprint comparison expert was developed – the SESA[4] package.

[4] The SESA (*Statistic Evaluation of Shoeprint Accidentals*) DVD version 1.0b1 contains the *CONTOURS* data set, the two software tools (*MarkAccidentals* and *CompareAccidentals*) and a comprehensive user manual that covers installing and operating the tools.

As a result of the procedure described above, this project solves, partially, the problem of calculating the results of pattern comparison for test impressions of shoe soles in a scientific and mathematically sound manner.

The methodological concept developed simultaneously with the advance of this project marks a big step towards transforming the entire area of pattern comparison to an agreed, scientifically based field, and leads to founding large databases for the various comparison fields, and algorithms to calculate the chance to get other features, similar (under restrictions) to an examined one. This research project had to overcome several obstacles in order to achieve this goal:

1. The existing accidental databases in the world were composed of actual photographs of accidentals and not their digital representation. Today the database of computerized accidentals is accessible to all experts, allowing them to compute actual probabilities of the chance a similar accidental. This was achieved by assembly of a large digital database of accidentals.

2. Development of software tools to assist the footwear impression examiners in assessing the rarity of accidentals: The tools enable marking of new accidentals in a consistent way, and comparing them to a large database. These tools contain the following:

    a) A universal shoe-aligned coordinate system allowing the accumulation of accidentals from many shoes onto one coordinate system.

    b) The distribution of locations of accidentals, developing the notion of accumulated contact areas.

3. Presenting the problem of variability of test impressions in a measurable way: One instance of an accidental is bound to be slightly different from another. This phenomenon, known for years, received no quantitative treatment before. We calculated the rate of dissimilarity between multiple instances of the same accidental.

4. The discovery that locations of accidentals are (as a first approximation) evenly distributed on the sole of the shoe, when they are normalized to a universal shoe sole.

5. The discovery of a uniform distribution of orientations of accidentals on the shoe sole. In contrast to several studies done on smaller populations, for most shapes, the distribution is uniform. A deviation from uniformity appeared only for the most

elongated shapes which had higher probability of occurrence in the direction of the shoe's long axis.

6. Creating two populations of matching values: correct matches (known match) and non-correct matches (known no-match).

The tools and methods developed during this project compose a comprehensive and methodological system for collecting data, assembling a database, and using the database to estimate the statistics of accidentals. It is our belief and hope that the use of SESA system can lead the pattern comparison field in general, and shoeprint comparison in particular, to the new, scientific era.

1. Introduction

1.1. Statement of the problem

The scientific attack on the forensic science has intensified during the last decade. In 2005, Saks and Kohler [1] claimed that a "paradigm shift" must come and shake the forensic science from its base, since forensics was never really based on science. The emerging of DNA analysis emphasizes this need, since unlike other forensic fields; DNA is soundly based on statistics and scientific methods.

Today, shoeprints revealed at crime scenes are compared manually against suspect shoes. The first step is to determine whether the class characteristics match. These include the sole pattern, size and wear. Matching these features narrows down the potential population of shoes that could have left the shoeprint, but their discriminative value is limited. Once a match in all class characteristics is achieved, the shoeprint comparison examiner searches for accidental characteristics that appear both on the sole of the shoe and on the shoeprint from the crime scene. Accidentals, also known as Randomly Acquired Characteristics (RACs) are features on a shoe outsole resulting from random events of removing or adding of material to the outsole. These include, but not limited to: cuts, scratches, tears, holes, stone holds, abrasions and the acquisition of debris. The location, orientation and shape of these characteristics contribute to the uniqueness of a shoe outsole. The location is defined as the location of the accidental on the shoe-sole. The orientation is determined as the angle between the long axes of the accidental in relation to the long axes of the shoe-sole. The shape is the contour of the accidental. If such accidentals are identified, the examiner evaluates their rarity based on his knowledge and experience, and the level of confidence derived from the combination of all the accidentals present on the shoeprint. This is less reliable than automated methods of calculating rarity.

When testifying in court on shoeprint comparison, experts state their opinion about the chances of getting another shoeprint bearing the same characteristics. This expert opinion lacks the potential or known error rate requested by the Daubert ruling [2].

In brief, the Daubert criteria consist of four independent principles:

1. Whether the theory or technique in question can be (and has been) tested
2. Whether it was published in peer-reviewed journals gaining widespread acceptance within the relevant scientific community
3. Whether standards exist controlling the method operation
4. Is there a known or potential error rate of the method.

In order to meet the Daubert challenge, which has surfaced in shoeprint cases in the U.S and the U.K, one must show that shoeprint comparison, and its interpretation, fulfill all of the Daubert criteria. However, this includes defining the practiced technique, and reporting its potential error rate.

The NAS report [3], in 2009, was called "Strengthening Forensic Science in the United states, A Path Forward". Among the common issues that arose in this report was the demand to fortify the scientific basis of the different degrees of certainty, including the amount of information needed to determine "uniqueness" or "individuality". Concerning shoeprint comparison, the NAS report stated:

> …But there is no defined threshold that must be surpassed, nor are there any studies that associate the number of matching characteristics with the probability that the impressions were made by a common source. It is generally accepted that the specific number of characteristics needed to assign a definite positive identification depends on the quality and quantity of these accidental characteristics and the criteria established by individual laboratories. [3, p. 147]

This project's main goal is to assist forensic science and specifically shoeprint comparison in quantifying these probabilities. Our project leads the way to a more "scientific" shoeprint examination and evaluation. The challenge of evaluating the chance to get repetition of a "unique" accidental characteristic was the heart of this research project. Because of the random nature of the accidentals, their appearance is supposed to be unrepeatable; hence "predicting the unrepeatable" is a major problem. As presented by Petaco et. al. [4]

> While it is a strongly held belief by many footwear examiners that the patterns of accidental marks on shoes are unique, this is an inductive conclusion that has not been thoroughly studied using controlled experiments.

Comparing two shapes to each other is challenging in the sense that neither a mathematical representation of the shapes nor the choice of an appropriate algorithm to measure to degree of similarity between two shapes is obvious.

Another complication in seeking a pure mathematical representation of a feature found on a sole of a shoe is that, due to the dynamic process in which shoeprints are embedded, every step produces a slightly different print. This phenomenon is the reason for making

several test impressions from the suspect's shoe in order to compare it to the scene of crime shoeprint [5].

The phenomenon of slightly distorted shapes in an unrepeatable manner leads to the problem of how to determine when two shapes are two versions of the same accidental, or two different shapes that happen to look alike. This project developed tools to solve such questions.

1.2.Literature citation and review

Several attempts were made during past years to evaluate shoeprints in a scientific and mathematical way. The chance to find another shape similar to a known accidental and the chance to have an accidental in a specific place on the sole were examined by forensic experts.

In the Netherlands, all experts routinely use a guideline for shoeprint analysis [6,7,8]. In short, this entails the expert's attaching a numerical value to each acquired feature in order to calculate the overall evidential value of the comparison between a shoeprint and the shoe. This procedure allows for an objective evaluation and quantification of accidental characteristics by different experts. The main drawback of these guidelines is that the value set for an accidental is arbitrary and is not derived from the uniqueness of the accidental. Another disadvantage of the Dutch Guidelines is the fact that it deals only with the shape of the accidental and not with its location and orientation.

Several countries obtain databases of photographed accidentals, and the expert can visually compare the defect in debate to the "standard" accidentals and derive the "correct" weight it should bear. Obviously, this search can be very arduous and time consuming if the database is large.

Many efforts were made in recent years to make the confidence level of evaluation in shoeprint comparison less expert-dependent than common today, creating a model of objective accidentals evaluation [7,8,9,10,11,12].

Some efforts have been made to automatically sort pattern of shoes, in order to facilitate finding the shoe type involved in a crime. However, as of now, no available commercial system is capable of automatic classification and identification.

Several groups are actively working on automating the process of recognition and classification of shoeprints. Their main goal is to replace human experts in the tedious work of shoeprints classification, and to help in matching a given shoeprint to a known database.

Some image processing and pattern recognition techniques have been developed for automatic comparison of tool marks [13,14,15,16] and they may be applied to the domain of shoeprint analysis.

Alexander et al [17,18] suggested replacing the process of human classification of shoeprints with automatic pattern matching. They built a system capable of searching through a small database of shoeprint images and identifying potential matches.

Huynh et al [19] presented a fully automatic shoeprint recognition system. Their main concern was to classify shoeprints into groups based on the sole pattern. They used several pre-processing steps to enhance the images, remove noise and obtain a scale and rotation normalized image. They extracted the 2D direct-Fourier-transform coefficients, measured the similarity between these features and ranked the images accordingly. The system was compared to classification by human experts and achieved 54% correct matches for the matches ranked on the top of the list. The system was later improved and achieved 65% correct matches for the matches ranked on the top of the list [20] and was also able to deal with partial prints.

Pavlou and Allinson [21] used local feature detection and description in their system for automatic extraction and classification of shoeprints and achieved 85% correct matches for the matches ranked on the top of the list.

Gueham et al [22] developed a method for matching shoeprint images using phase-only correlations – a specific feature of the Fourier coefficients. This enabled an improvement in the system's ability to match low quality and partial shoeprints images.

Su et al [23,24,25] improved several aspects of the system developed by the same group [17,18], including a better threshholding of noisy shoeprint images [23] and better local feature detection and analysis [24, 25].

Finally, Srihari's group (Tang et al [26,27 28] and Srihari [29]) developed a comprehensive methodology and system for matching shoeprint evidence by identifying geometrical patterns (mainly of class characteristics). This work dealt with crime scene impressions and presented methods for image enhancement, feature extraction, similarity computation between evidence and known impression, retrieval of closest match in a database and computation of match uncertainty.

Only the last group tried to estimate statistics of matching, but it was done on the level of geometric patterns and not using specific accidentals. None of the works described above tried to evaluate the probability of occurrence for specific accidentals.

1.3. Statement of rational for the research

In this project we address some of the problems stated above and present a large step towards making shoeprint evidence scientifically sound. First, we created a large database of accidentals, a necessary component for research and standardization of the field. Second, we developed algorithmic methods for evaluating the chances of finding a combination of similar accidentals on a different shoe sole. And last, software tools for marking accidentals and analyzing their statistics were developed; tools for the aid of the shoe expert community.

1.4 Basic assumptions

The research and development on this project followed several assumptions that were taken in order to meet the challenges. The results presented here are dependent on those assumptions. Checking the assumptions is one of our major goals in future research.

- Creating an outsole-pattern database and statistic calculations of pattern rarity is time consuming, yet useful for a limited period of time because of the rapidly changing shoe market. Therefore, this research focused on accidentals, assuming that the behavior of accidentals is persistent for a much longer period of time.
- Most shoe-soles are made from similar material and therefor will tear in a similar manner. This enabled us to ignore the outsole pattern and compare accidentals on various sole patterns and to develop a database that can withstand the frequent changes in shoe style and patterns and contains many more shoes than a pattern dependent database.
- The three features of each accidental: shape, location and orientation are independent [30]. This enabled us to multiply their probabilities. Further investigation is required in the future in order to characterize the dependencies, if any exist, and take them into account in calculating the probability of occurrence of an accidental.
- Each accidental on the sole occurred during an individual incident and therefor the probabilities are independent. This enabled us to multiply the probabilities of multiple accidentals and so calculate probabilities of configurations of accidentals. This too should be further investigated.

2. Methods and Results

2.1. Data sets

2.1.1. Raw data: Images

We have collected approximately 900 scans (600 DPI, scanned using Microtek i800 scanner) of test impressions of many different shoe types.

For some shoes we have the three different image types: *laboratory test impression*, *shoe sole photograph* and *scene of crime photographs* (Fig. 1).

Fig. 1. Examples of the three types of images for three different shoes.

Most of the work was done on test impressions, with shoe sole photographs as a reference.

2.1.2. Data sets

Major

Using methods described below, we have collected two major data sets for accidental characteristics (accidentals):

- **CONTOURS** data set, composed of ~8,900 accidentals from 300 test impressions. For each accidental we have its contour and location (other attributes such as orientation are derived later from the marked contour).
- **FAST** data set, composed of 20,000 accidentals from 600 test impressions. For each accidental we only have its location.

Both data sets were used in building the statistical models, as is described below.

The CONTOURS data set is delivered with the accompanying DVD, and is used directly by the delivered software tools.

Minor - Repeated impressions and marking data sets

We prepared five datasets of repeated markings. In each such dataset, k accidentals were marked on d different test impressions of the same shoe m times (k = 10 to 18, d = 18 to 20, m = 3 to 4). The purpose was to estimate the different sources of variability between

instances of the same accidental: the variability due to the nature of the impression process, the variability due to the marking process, and the variability within and between the markings of individual human operators. The outcome was a methodological evaluation of the error in estimation of each of the attributes of an accidental: its location, orientation and shape. The process of error estimation and the role of errors in developing the statistical models are described below (section 2.3-2.6).

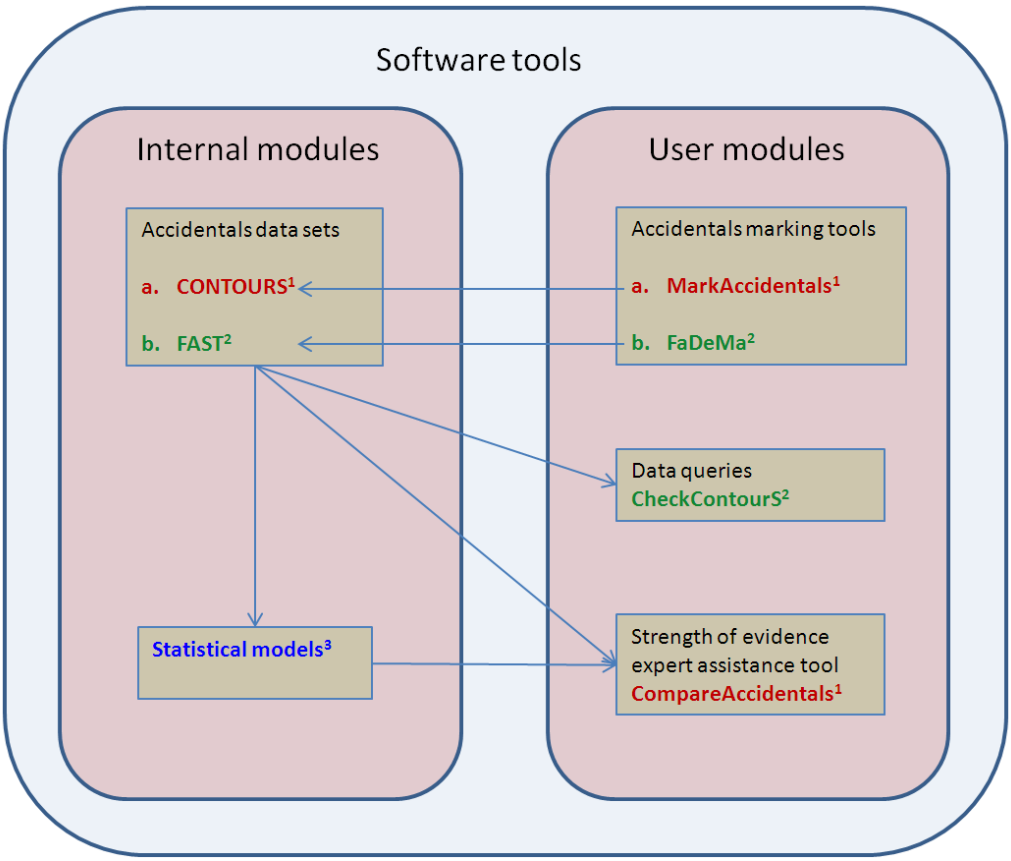
Collecting the data sets

The CONTOURS data set was collected by marking the contours of accidentals. This was achieved by building a tool (MarkAccidentals) used by trained human operators. Two major types of information are extracted from this data set: accidentals contours and accidentals locations.

The FAST data set was collected using another tool - **FaDeMa** (Fast Defects [accidentals] Marking) developed especially to collect larger numbers of accidentals locations. This tool allows the human operators to mark approximately 60 accidentals locations per hour, which is 10 times faster than the rate using MarkAccidentals. Each accidental is marked only by one point (and not by its entire contour).

2.2. The software tools

In this project we developed software tools to collect data, analyze data, build the statistical models and make statistical inference using these models (Fig. 2). The deliverable package contains two expert-assistant-tools for assessing the probability of accidentals in test impressions:



1. Delivered (SESA DVD). 2. Used internally during R&D. 3. Used by the SESA tools; explained below.

Fig. 2. The different software tools developed in the project and their relations. Some tools are delivered in the SESA DVD while others are not, and were only used by us in the research and development done in this project. The statistical models component is used by the SESA DVD tools and explained in detail below.

Using the SESA package, the shoeprint expert can mark new accidentals and find out how rare they are.

Examples of using the software tools

Here we use excerpts from the SESA user manual to demonstrate some of the abilities of the tools. For a complete explanation on installing and using the tools, you are referred to the SESA user manual (available on the SESA DVD).

A. MarkAccidentals

This tool is used to mark new accidentals on test impressions, or edit existing accidentals.

The MarkAccidentals tool has various functions. The main functions are: viewing two images of the same shoe (e.g. a test impression image and a shoe sole photograph) side by side (Fig. 3), viewing already marked accidentals (Fig. 4, 5), marking a new accidental (Fig. 6, 7) and saving marked accidentals (Fig. 8). The user can also change already marked accidentals by cleaning excess markings, add manual markings, and change the attributes of an accidental.

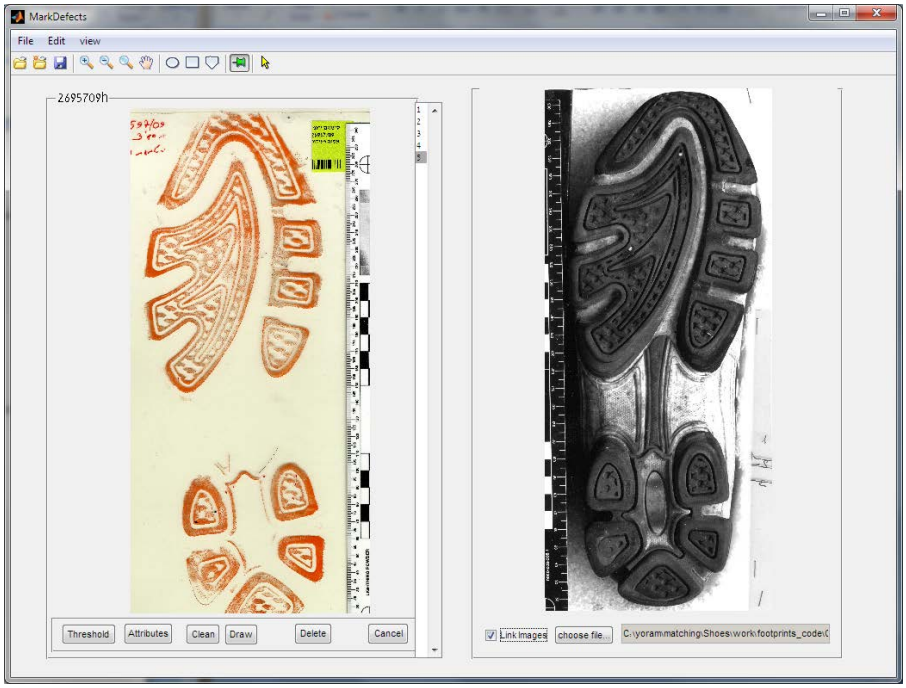


Fig. 3. Viewing two images side by side. *Left*: test impression. *Right*: shoe sole of the same shoe.

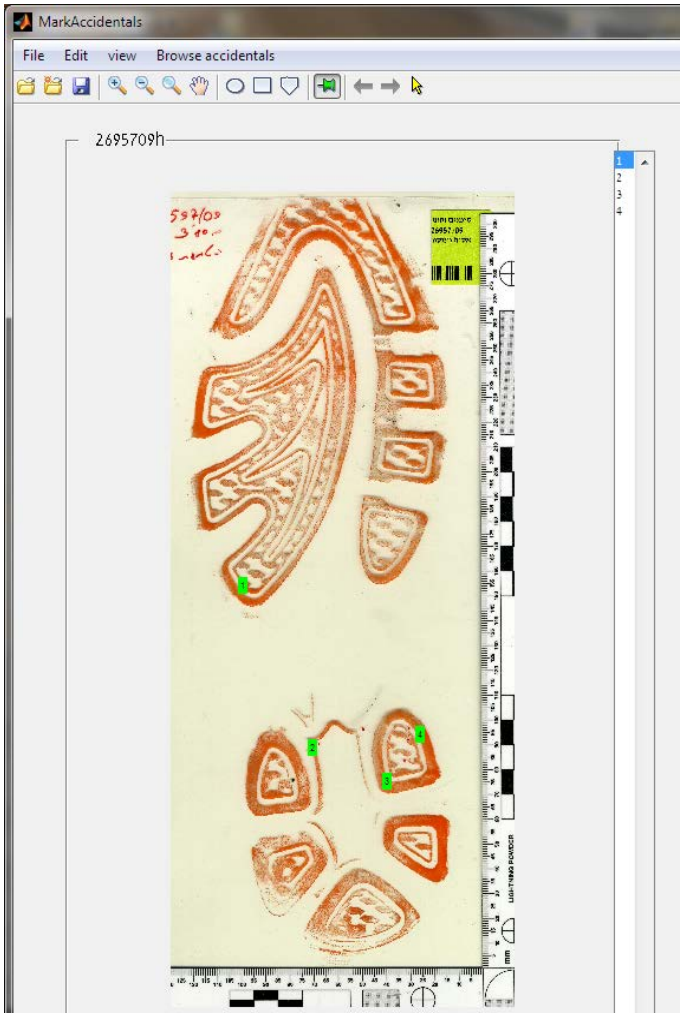


Fig. 4. Location of already marked accidentals and their ID numbers shown over the test impression.

Fig. 5. Zoomed in view of accidental #3 from Fig. 4.

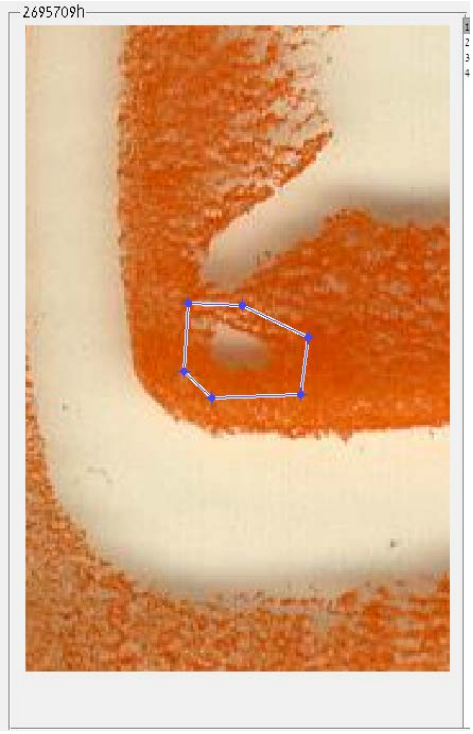


Fig. 6. Choosing a new accidental.

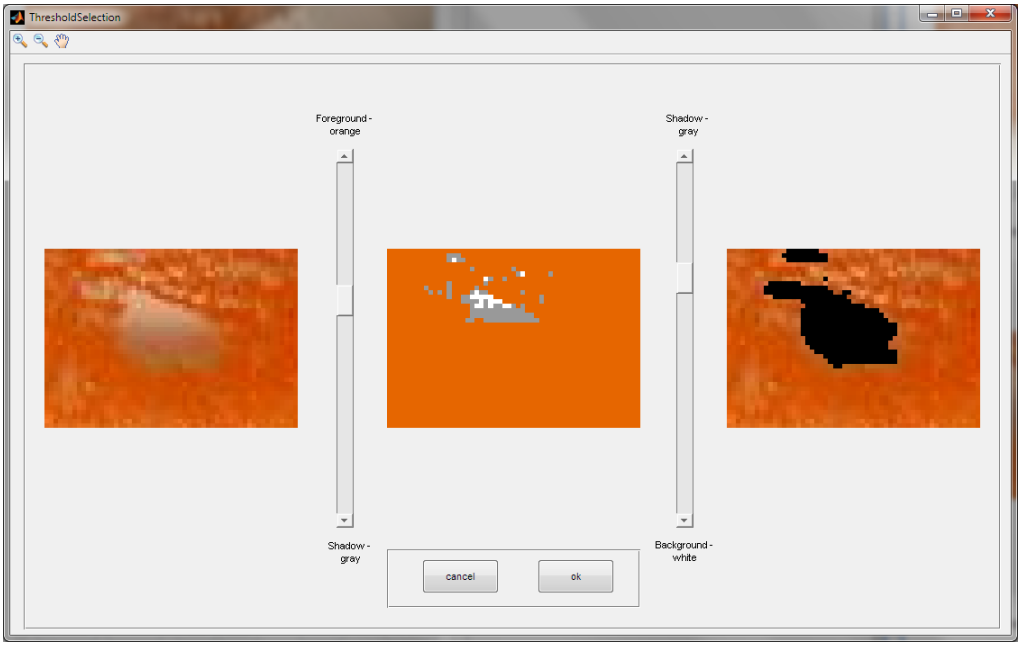


Fig. 7. Checking the automatic segmentation of a new accidental (chosen previously, Fig. 6).

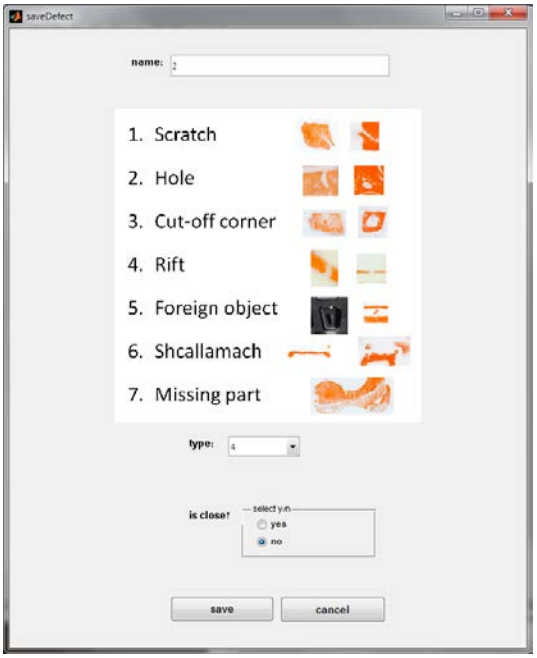


Fig. 8. Setting the different attributes (name, type, is close) of an accidental prior to saving.

B. CompareAccidentals

This tool is used to find the probability of having accidentals similar to the target accidentals by comparing the targets to a database of thousands of accidentals.

Many different functions can be performed using the CompareAccidentals tool as well. The main functions are: viewing a test impression image with its accidentals (Fig. 9), analyzing the statistics of an accidental (Fig. 10) and viewing the results (Fig. 11), displaying the most similar accidentals to a target one (Fig. 12) and calculating the probability of finding a configuration of accidentals similar to a target configuration (Fig. 13). The tool can also generate reports of the results.

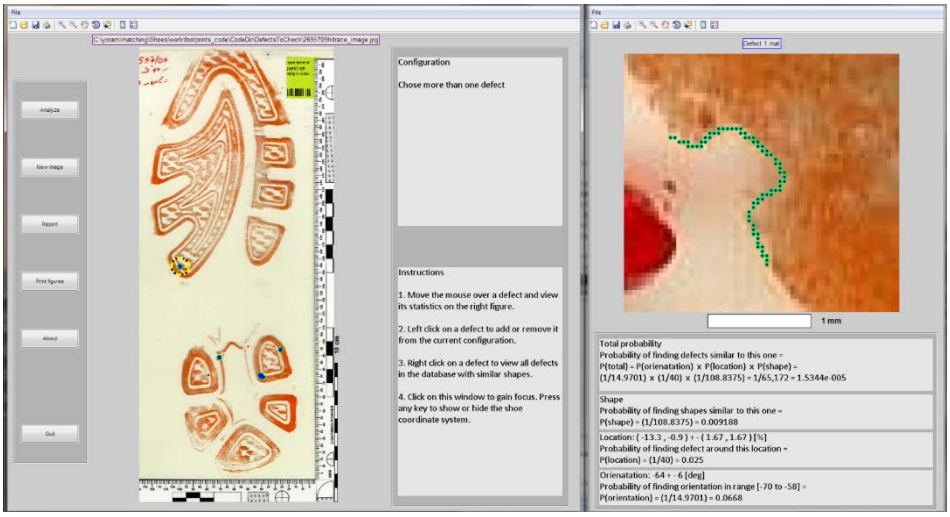


Fig. 9. A test impression image with its accidentals are loaded, and displayed using two figures.

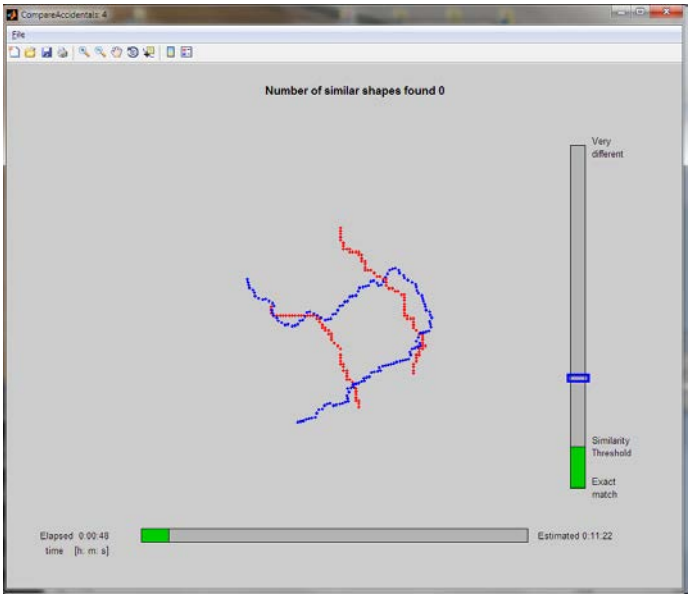


Fig. 10. Comparing a target accidental (red) to one of the other ~8900 accidentals (blue), during the process of analyzing its statistics.

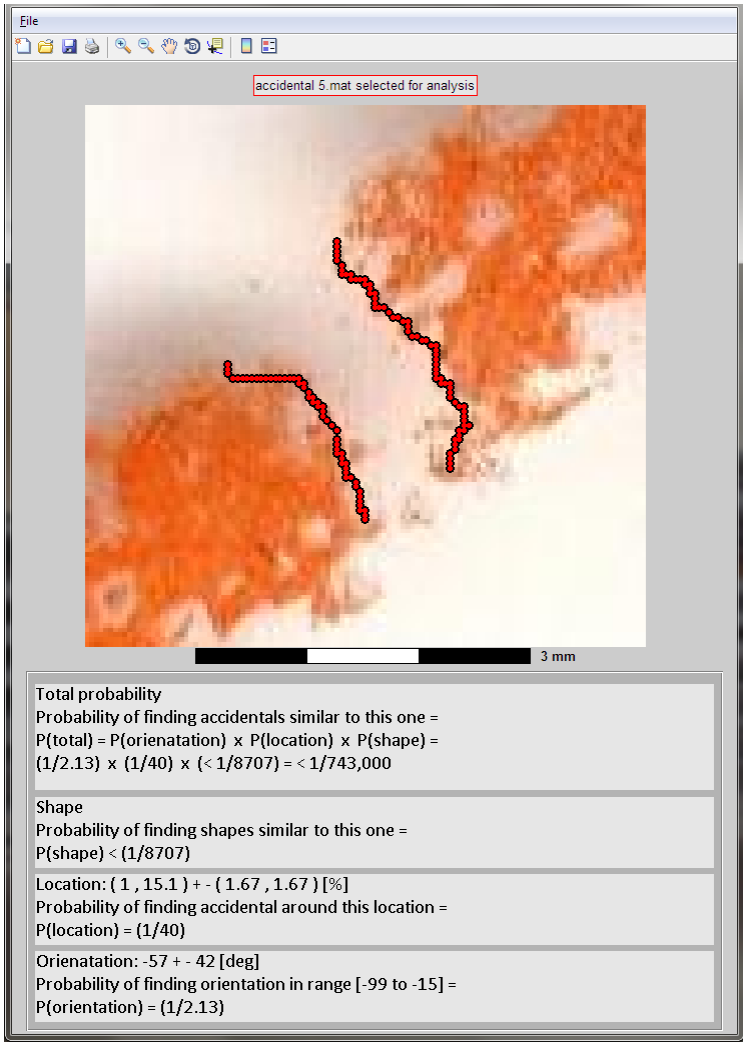


Fig. 11. The target accidental (from Fig. 10) and its statistics.

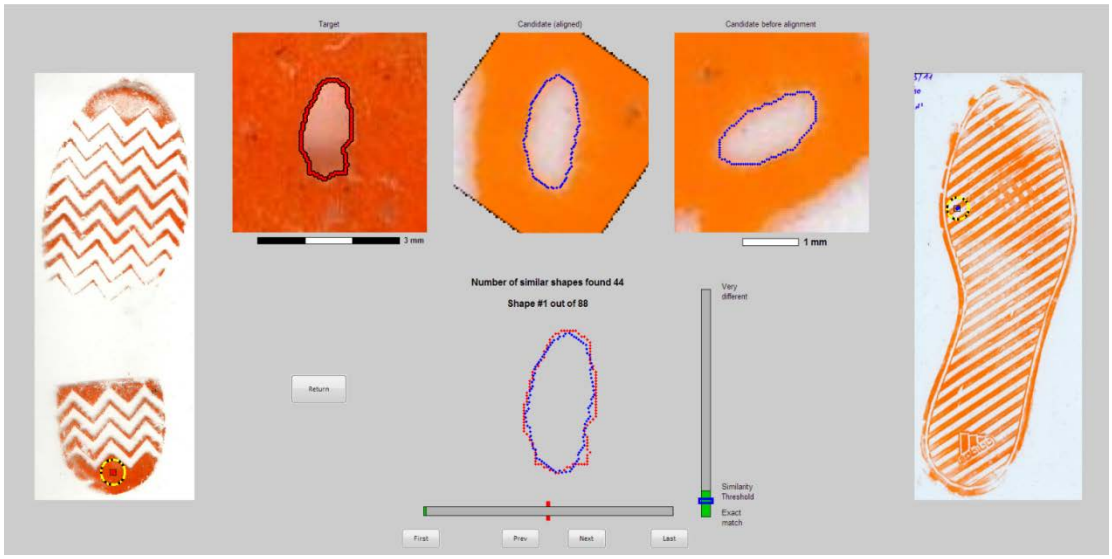


Fig. 12. Showing potential matching shapes. A target accidental (red), its test impression (left), the first candidate accidental (blue), and the test impression of the candidate (right). The middle bottom shows the two shapes superimposed. This target had 44 matches (out of ~8,900) that were similar enough to be considered originating from the same accidental (based on shape alone, and taking into account the error in shape estimation).

Fig. 13. The probability of finding a configuration of n accidentals is a product of the individual probabilities of the accidentals. A configuration of two (red marks on the test impression) and the estimated probability (top right).

2.3. Overview of the statistical model

One of our major goals in the project was to build a statistical model to estimate the rarity of a combination of accidentals. Our model is based on the assumption that the probabilities of finding accidentals are independent of each other.

Then, the probability of finding a combination of *n* accidentals is the product of the probabilities of finding each of the accidentals:

$$P(combination) = n! \cdot \prod_{i}^{n} P(accidental_i)$$

The *n!* factor comes from counting spatial configurations, as is explained below.

The probability of having each of the accidentals is itself a product of three terms:

$$P(accidental_i) = P(location_i) \cdot P(orientation_i) \cdot P(shape_i)$$

This equation is based on the assumption that each of the terms is independent of the others.

The statistical models and their dependencies are described in the Fig. 14.

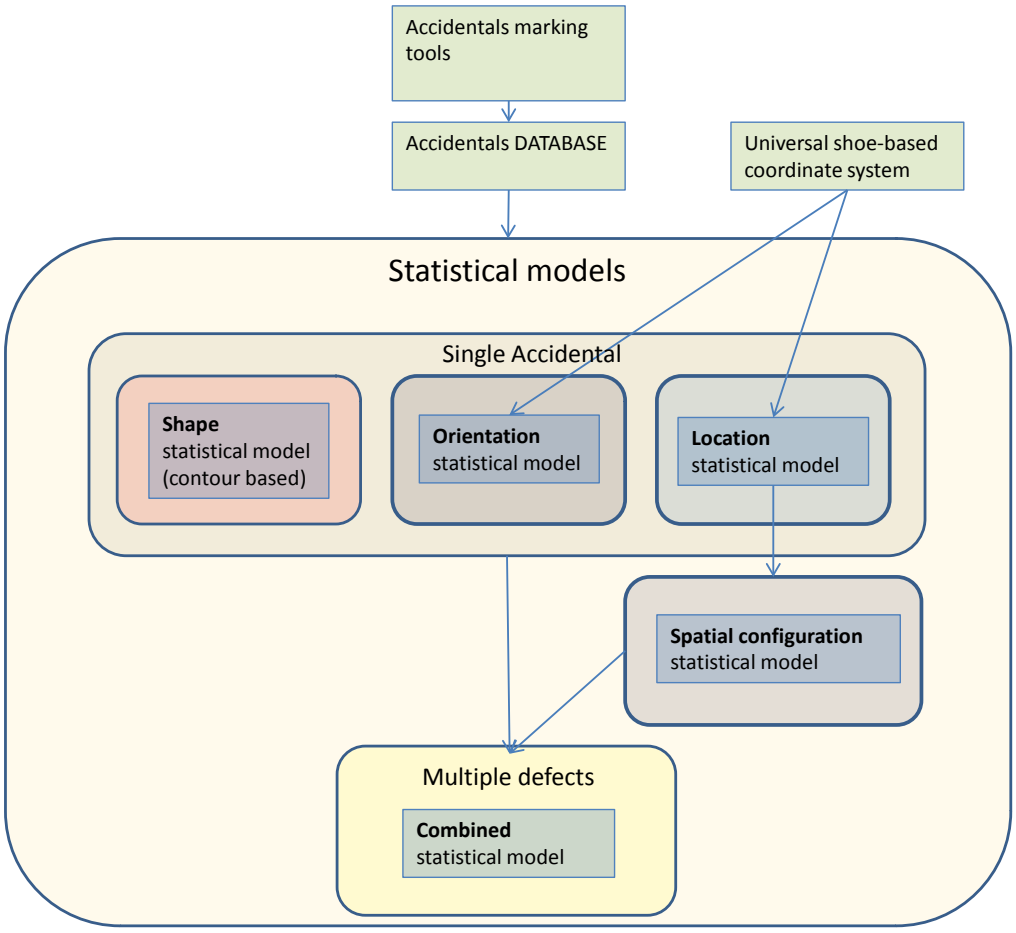


Fig. 14. The relations between the different statistical models in our system.

In this section we briefly describe each model. The detailed models are presented in section 2.4 to 2.6.

1. Spatial model – Location of accidentals

The probability of having an accidental in a specific location is explained using the grid metaphor. The process of estimating the probability can be viewed as dividing the shoe impression into cells and calculating the chance for an accidental to be in a specific cell. The first issue to determine is the size of a grid cell.

When comparing two test impressions of the same shoe, a very small grid cell, on the order of 1 mm^2 can be used. This is because we can achieve high quality alignment of the test impressions, and thus very accurately compare accidentals' locations on both prints. But when trying to collect the statistics of locations we need to compare numerous accidentals based on more than two test impressions. We need to align many test impressions, of many shoe types, models and sizes. This alignment process has inherently larger errors than the process of aligning two test impressions of the same shoe. Those larger errors must be estimated and taken into account. The size of the error in measuring the location of an accidental after aligning should be taken as the minimal size of grid possible.

Our error in estimating locations is on the order of 5 mm in each direction, leading to a grid cell size of (0.5)*(0.5)*cm^2 = 0.25 cm^2 (explained in detail in sections 2.4.4 - 2.4.5).

Since the area of a shoe sole is roughly 300 cm^2 the number of grid cells is 300/0.25 = 1,200.

Assuming a **uniform distribution** of locations, the probability of choosing a location at random is therefore:

1/1,200

Since we found an average of 30 accidentals per shoe, the final probability of having an accidental on a specific location is

$$P(location_i) = 30 \cdot \frac{1}{1{,}200} = \frac{1}{40}$$

The issue of uniformity of the distribution of locations is addressed in section 2.4.7, below.

2. Orientation

The probability of detecting an accidental with a specific orientation depends on two factors: the distribution of orientations in the population of accidentals and the error in estimating the orientation.

For demonstrative purposes only, let us assume that the distribution of orientations is uniform, and that the error in estimating orientation is 5° out of the 180° range. Then the probability of having a specific orientation is 5/180 = 1/36. $P(orientation_i) = \frac{1}{36}$

Estimating the distribution of orientations and its error are described in section 2.5, below.

3. Shape

The probability of having an accidental with a specific shape is a more complicated issue than the probabilities of location and orientation of the accidental. This is because shape is a high dimensional property, compared to the two dimensions of location and one dimension of orientation. It is impossible to collect enough data to fully estimate the enormous variety of shape. To overcome this problem we used the statistics of *shape comparisons* – the 1D (one dimensional) population of shape dissimilarity values.

So the question of the probability of a specific shape is answered by counting how many shapes in our database are similar **enough** to the target shape. The details are explained later. Now we will just give an example with numbers:

Let us assume that a specific shape is similar to 50 out of the 10,000 shapes in the database. The degree of similarity is such that all these shapes can be considered of the same shape. Therefore, the probability of having such a shape is

$$P(shape_i) = \frac{50}{10{,}000} = \frac{1}{200}$$

Of course, a more unique shape will have a lower probability.

The many details of estimating the shape probability are described in section 2.6, below.

4. Total probability of one accidental

The probability of a single accidental, using the numbers that were given above is

$$P(accidental_i) = P(location_i) \cdot P(orientation_i) \cdot P(shape_i) =$$

$$\frac{1}{40} \cdot \frac{1}{36} \cdot \frac{1}{200} = \frac{1}{288{,}000} \cong 3.5 * 10^{-6}$$

5. Combination of accidentals

What is the rarity of a specific combination of accidentals?

a. We assume that the orientations of the accidentals are independent:

$$P(orientation\ of\ combination\ of\ n \text{ accidentals}) = \prod_{i}^{n} P(orientation_i)$$

And for 3 accidentals, with varying probabilities (one of them given above) we have

$$P(orientation\ of\ combination \text{ of 3 accidentals}) =$$

$$P(orientation\ 1) \cdot P(orientation\ 2) \cdot P(orientation\ 3) =$$

$$\left(\frac{1}{18}\right) \cdot \left(\frac{1}{36}\right) \cdot \left(\frac{1}{50}\right) = \frac{1}{32{,}400}$$

b. We assume that the shapes of the accidentals are independent:

$$P(shapes\ of\ combination\ of\ n \text{ accidentals}) = \prod_{i}^{n} P(shape_i)$$

And for 3 accidentals, with varying probabilities (one of them given above) we have

$$P(shapes\ of\ combination\ of\ 3 \text{ accidentals}) =$$
$$P(shape\ 1) \cdot P(shape\ 2) \cdot P(shape\ 3) =$$
$$\left(\frac{1}{200}\right) \cdot \left(\frac{1}{30}\right) \cdot \left(\frac{1}{3{,}000}\right) = \frac{1}{18{,}000{,}000}$$

c. We also assume that their **locations** are independent, but every location taken, limits the number of possibilities to choose other locations.

Calculating the probability of having a specific configuration is done like this:

Since there are 30 accidentals per shoe (on average), and 1,200 possible locations per shoe, the first location has a probability of 30/1,200. This is like releasing 30 glass marbles into 1,200 holes, with equal probability for each hole.

The second location has a similar but not equal probability. It is 29 out of 1,199 locations (one marble was already released and one location has already been taken).

Calculating this for *n* accidentals, ($n < 30$) we have:

$$P(positions\ of\ combination\ of\ n \text{ accidentals}) = P(configuration)$$
$$= n! \cdot \prod_{i}^{n} \frac{(30 - i) + 1}{(1{,}200 - i) + 1}$$

The *n!* factor comes from the number of arrangements of the *n* accidentals. This simply comes from the fact that when we assign an index to an accidental, it is done arbitrarily since the order within the group of accidentals is not important. For example, if we have 2 locations that we term A and B, these can be arranged in 2 ways: AB and BA. When we look in our database for accidentals in these locations, we count both occurrences: the cases with the first accidental in location A and the second accidental in location B, *and* the cases where the first accidental is in location B and the second accidental is in location A. Here we must also count these two locations twice.

Thus, the probability of finding 3 accidentals in specific locations, out of 1,200 possible locations, when there are 30 accidentals on average per shoe is:

$$P(positions\ of\ combination\ of\ 3 \text{ accidentals})$$
$$= 3! \cdot \prod_{i}^{3} \frac{(30 - i) + 1}{(1{,}200 - i) + 1} = \frac{30}{1{,}200} \cdot \frac{29}{1{,}199} \cdot \frac{28}{1{,}198} = \frac{1}{11{,}793}$$

d. Now we can answer the original question: what is the rarity of a specific combination of accidentals?

$$P(combination) = P(configuration) \cdot \prod_{i}^{n} P(orientation_i) \cdot \prod_{i}^{n} P(shape_i) =$$

$$= n! \cdot \prod_{i}^{n} \frac{(30-i)+1}{(1{,}200-i)+1} \cdot \prod_{i}^{n} P(orientation_i) \cdot \prod_{i}^{n} P(shape_i)$$

And for our example with 3 accidentals with given probabilities:

$$P(combination\ of\ 3\ \text{accidentals}) = \frac{1}{11{,}793} \cdot \frac{1}{32{,}000} \cdot \frac{1}{18{,}000{,}000} \cong \frac{1}{6.8 \cdot 10^{15}}$$

$$= 1.47 \cdot 10^{-16}$$

Note that the result is not final. Each of the models (*location*, *orientation* and *shape*) must be refined, as will be explained next. The main issues are the uniform probability assumption and the estimated errors.

Now we discuss each of the models in more details.

2.4. Statistical model of location and configuration of accidentals

Location model

How do we estimate the probability of having an accidental as a function of location?

2.4.1. What is the definition of "location" of an accidental?

When first thinking about it, the location of an accidental seems a simple notion: it is the place where the accidental is situated. But does it mean the location of every part of the accidental, or maybe the location of one representative point for every accidental? Since we wanted a definition that would be as independent as possible of the other attributes (shape and orientation) we decided to represent the location of an accidental by one point - the centroid of the shape of the accidental. The centroid was either estimated by a human marker based on the image of the accidental, or found automatically from the shape of the accidental. The latter involved two steps: 1. Marking the shape of the accidental (its contour). 2. Programmatically finding the centroid of the contour.

We performed manual estimation of accidental locations by marking the centroids of ~20,000 accidentals – in the FAST data set.

We marked the shapes of ~9,000 accidentals and calculated their centroid locations – in the CONTOURS data set.

2.4.2. The need for universal coordinate system

The notion of location by itself is not well defined unless stated with respect to some coordinate system.

What coordinate system should be used? If we want to compare locations of accidentals situated on the same shoeprint, any coordinate system may suffice. However, this is not the task we intend to do. Instead, we would like to compare locations of accidentals across shoes, and to be able to answer the question: *What is the probability of finding an accidental on a specific location?*

This question has a meaning only if we can somehow relate locations on different shoes. To clarify this point, let us start with a simpler question: *does the front of a shoe sole have more accidentals than the back (ankle)?*

One can easily state an experimental method to answer the question: Take enough test impressions to construct a representative sample of the shoes population. Identify on each test impression the front part and the back part. Count how many accidentals there are per part, per shoe. Accumulate the numbers and calculate the statistics to get the answer.

An essential ingredient in this method is being able to identify the front and the back of each test impression.

Now let us ask a similar, yet more distinct question: *does the front right area of a shoe sole have more accidentals than the front left area?*

Answering this question would follow similar methods as the previous answer. Here, however, we must identify, for each shoe, the front right and front left areas.

To identify locations on an even finer grid (higher resolution), we would need a systematic way to define locations on a shoe. Furthermore, the definitions should work for every shoe in our database, and for future shoes as well.

This calls for a normalization technique that will align the shoes correctly. Aligning will enable treating the locations of accidentals from different shoes, as belonging to one sample, all with respect to the same coordinate system.

2.4.3. Marking the shoe alignment, universal coordinate system

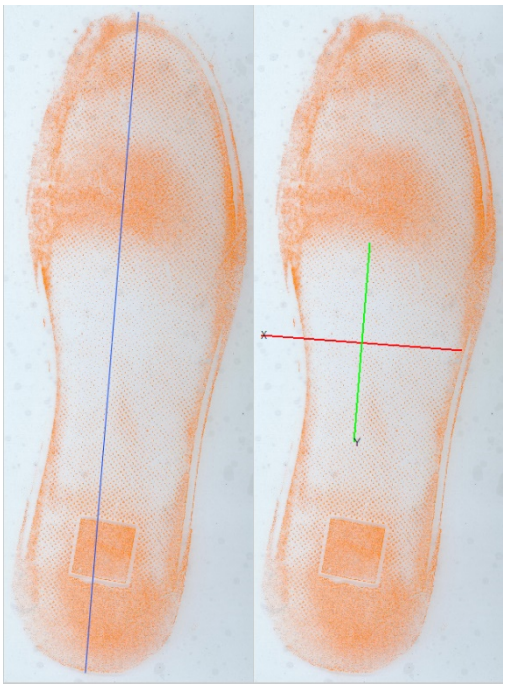
We developed a normalization technique that finds a standard (universal) coordinate system for shoes (Fig. 15).

Establishing the universal coordinate system on a test impression is done by manually marking the top and bottom of the shoe using a dedicated GUI[5] (that is part of the CompareAccidentals tool, Fig. 16). From the two extreme points we find:

1. The direction of the major axis of the shoeprint (and perpendicular to it, the direction of the minor axis).
2. The length of the shoeprint.
3. The origin of axes (at the middle of the marked line).

After marking the two points, the user is asked to mark another point on the outer side of the shoe, to establish the shoe's polarity, either a right or a left shoe.

Note that all measures used here are in pixels, and are relative to the coordinate system of the image.



A B

Fig. 15. Assigning a specific shoe coordinate system. *A*. The longitudinal axis of the shoe. *B*. The assigned coordinate system. The vertical axis (green) is aligned to the longitudinal axis of the shoe (blue), and the origin of the coordinate system is set at the middle of the longitudinal axis.

[5] GUI – graphical user interface. This software component allows interaction via mouse, keyboard and screen.

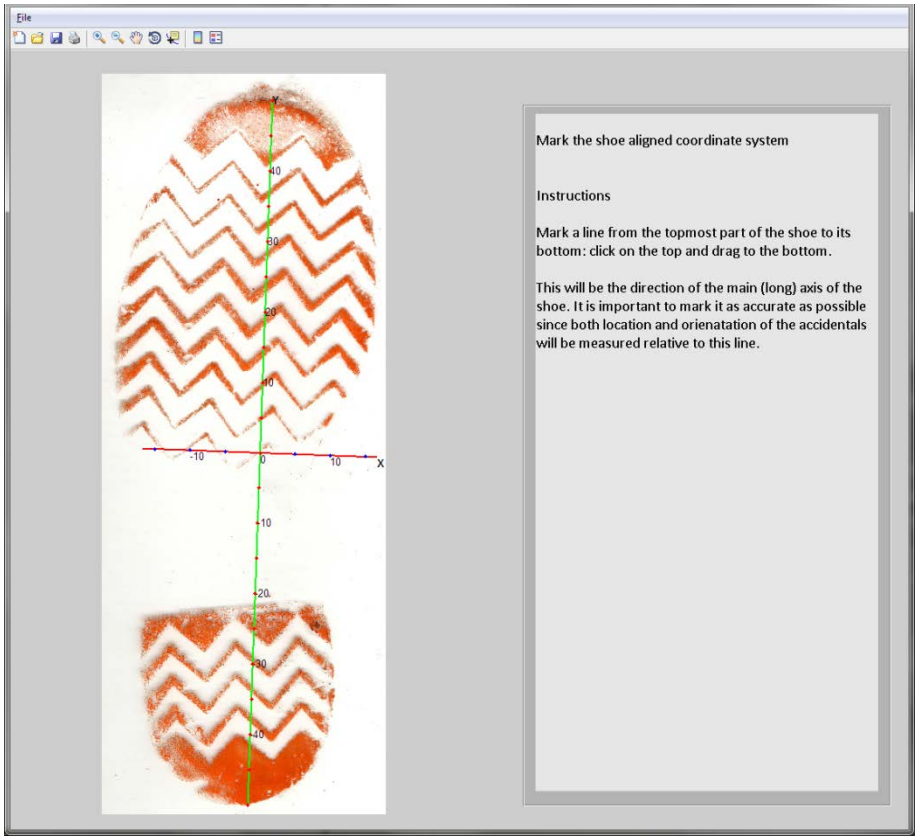


Fig. 16. The GUI to mark a shoe-specific coordinate system (part of CompareAccidentals tool).

Normalizing

The normalization is a transformation between coordinates systems, from image-based values (pixels) to normalized, shoe-based values. Each shoeprint is normalized using its shoe-centered coordinate system, following this procedure:

Ox,Oy = origin of the shoe aligned coordinate system [pixels].

R = rotation matrix. Its columns are unit vectors in the directions of the X and Y axes of the shoe aligned coordinate system.

L = the length of the shoe [pixels].

F = 1 or -1, for left or right shoe.

A. move (Ox,Oy) to (0,0) of the image. Now the origin of axes of the shoe aligned coordinate system is in (0,0).

B. Rotate using R. Now the axes of the shoe aligned coordinate system are in the directions of the main axes of the image (horizontal and vertical).

C. Scale by L. Now the length of the shoe is 1.

D. Multiply the horizontal axis by F. Now all shoes are turned to left shoes.

When the normalization is done for every accidental location, all accidentals are transformed to a single, universal coordinate system: (Fig. 23).

Aligning many test impressions: A similar normalization can also be performed for the test impressions themselves, not only for the accidentals (Fig. 24).

2.4.4. Estimating the error in location

There are two main sources of error in estimating the location of accidentals:

A. There is an inherent variability in estimating the center of accidentals. This happens in different ways for the two data sets we have. For our CONTOUR data set, i.e. accidentals whose contours we marked, we defined the location of the accidental as its center of gravity. This is approximated by the mean location of the contour pixels. Errors in estimation come from the errors (or variability) in the marking process. For our second data set (FAST), errors come from the variability in *estimati*ng the center of gravity of an accidental while marking (a cognitive task done by human markers), and from inaccuracies in the *marking* itself.

B. The errors described in A, in estimating the center of an accidental, are only one of the factors of the final error in the location. The second factor is due to errors in the process of assigning a shoe-specific coordinate system.

A more basic source of error influencing both A and B is caused by differences between the various test impressions. The shoe sole and the accidentals acquired on it are three dimensional. The test impressions are two dimensional, and the translation of the three

dimensional information from the shoe sole to a two dimensional representation is not identical every time. This creates variability in the exact appearance of the accidentals, due to differences in the stepping method and the amount of coloring medium (orange powder in this project) transferred from the shoe sole to the substrate.

To estimate the error in location we conducted two tests:

Test A. We made multiple test impressions of the same shoe. In each test impression we identified and marked 10 unique accidentals. This was done several times by 4 trained human markers. In total we have ~70 repetitions of each accidental.

We calculated the center of gravity for each accidental and considered this to be its location. Each of the test impressions was assigned a shoe-specific coordinate system, and the locations of accidentals were expressed in this system. Fig. 17 presents the 70 locations (aligned to a standard coordinate system) for each accidental. We also checked the distance between each pair of points, for all possible ~2400 pairs. This gave us a histogram of distances that approximates the distribution of error in location for each accidental (Fig. 18). All distances in Fig. 17 and 18 are expressed in pixels (with ratio of 23.25 pix/mm).

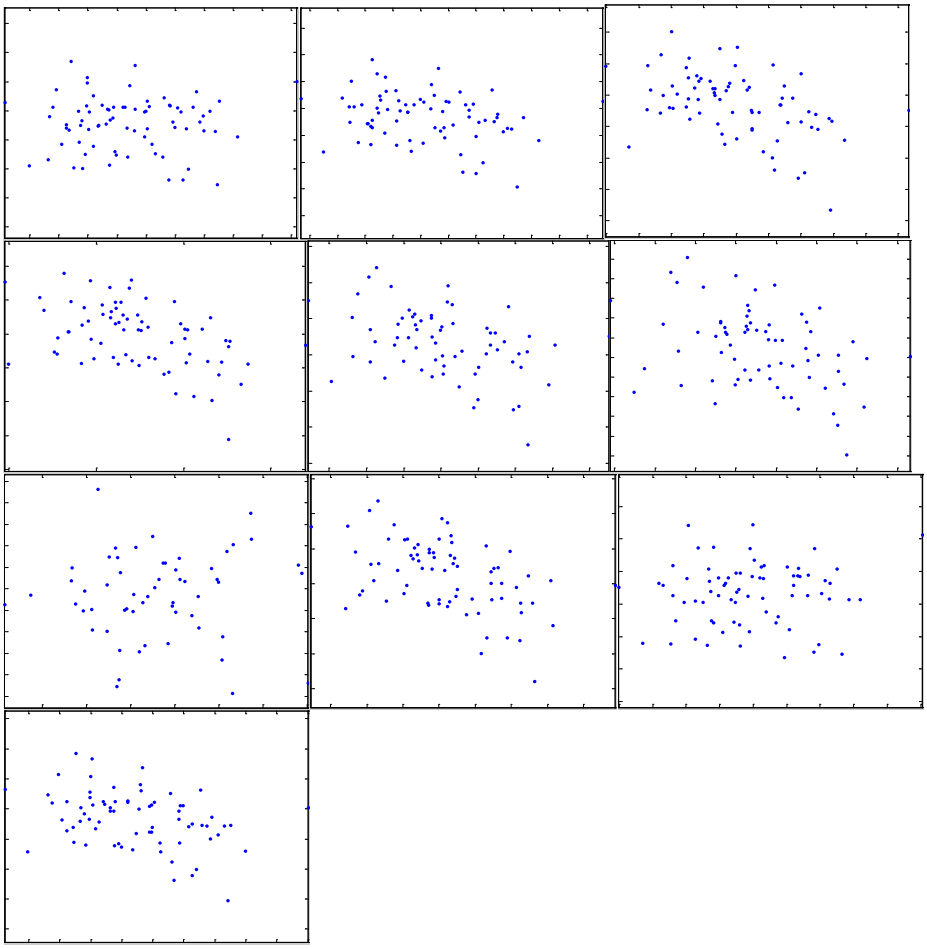


Fig. 17. Repeated marking: 10 accidentals were marked ~70 times each. Each plot shows the ~70 locations of a specific accidental. Plot sizes are 180x120 pixels = 7.74 x 5.16 mm.

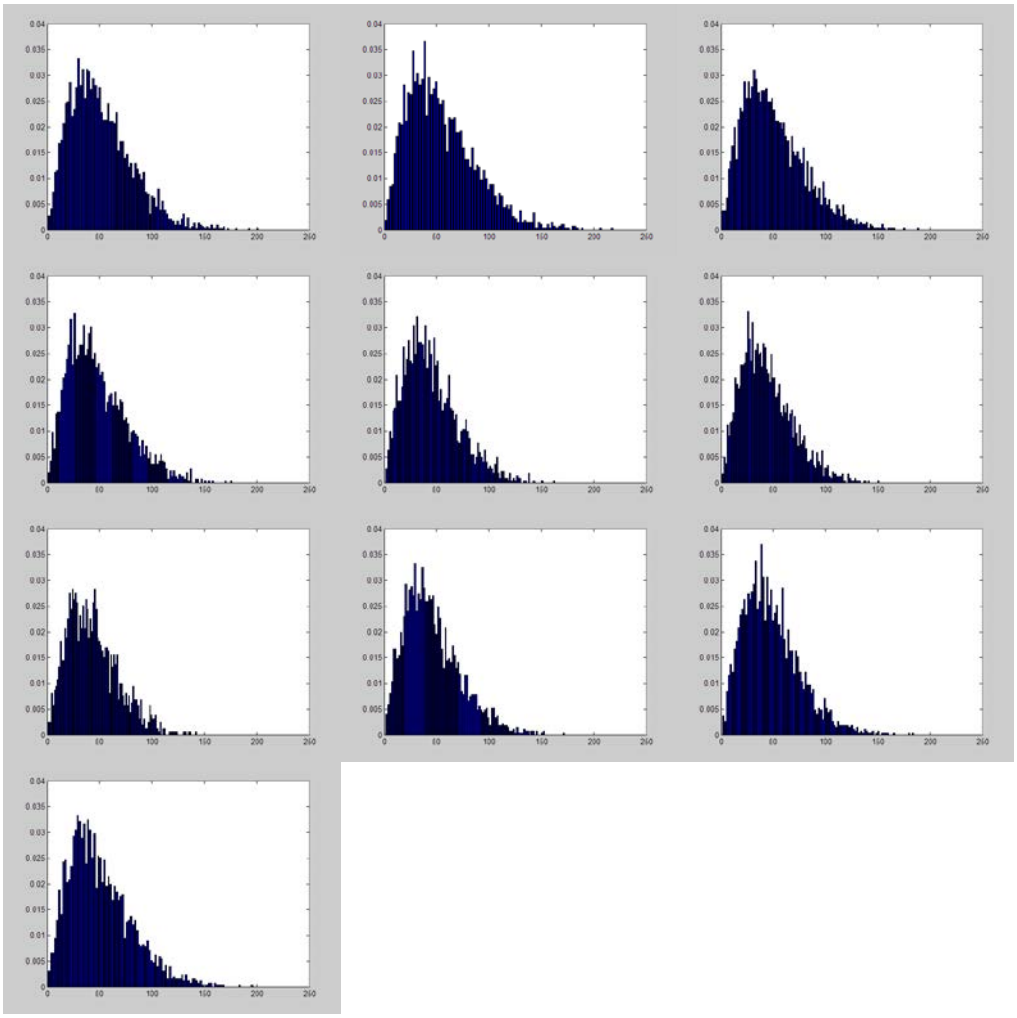


Fig. 18. Repeated marking: 10 accidentals were marked ~70 times each. Each plot shows histogram of distances between ~2,400 pairs of locations for each accidental. Distances are given in pixels with ratio of 23.25 pix/mm).

When collecting all 24,000 distances into one histogram (Fig. 19) the results are similar to those of each of the histograms for a single accidental.

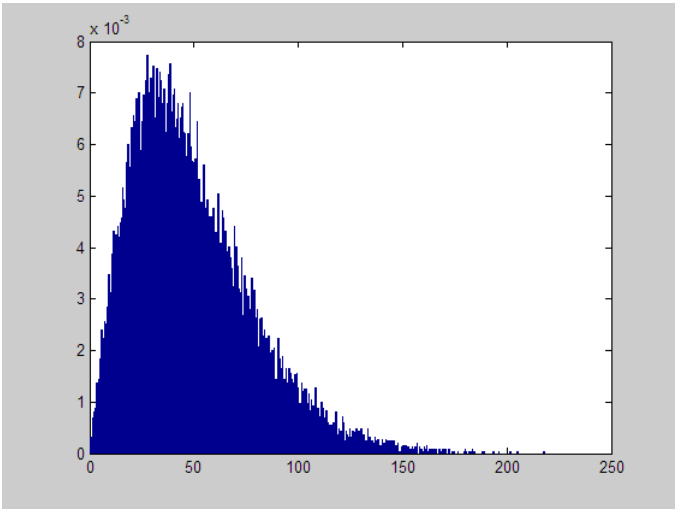


Fig. 19. Histogram of distances between ~24,000 pairs of locations.

Now let us estimate the error in location. Looking at Fig. 19, the maximal distance is around 200 pixels (8.6 mm), and the average is around 60 pixels (2.5 mm).

The distribution of the center of accidentals for the different accidentals looks similar. This hints that a large part of the variability in locations is due to the variability in assigning the coordinate system.

Test B. The second test was done using the same data set in an attempt to better isolate the second source - variability in assigning the coordinate system. The assumption was that marking the centers of known accidentals, which are easily identified, will keep the variability in the marking process low. We marked the center of the 10 accidentals manually (as is done for the FAST data set). This was done twice for ~20 test impressions and for 10 accidentals, yielding 38 repetitions for each of the 10 accidentals. A shoe-specific coordinate system was assigned to each of the test impressions. All locations are presented relative to an aligned coordinate system (Fig. 20).

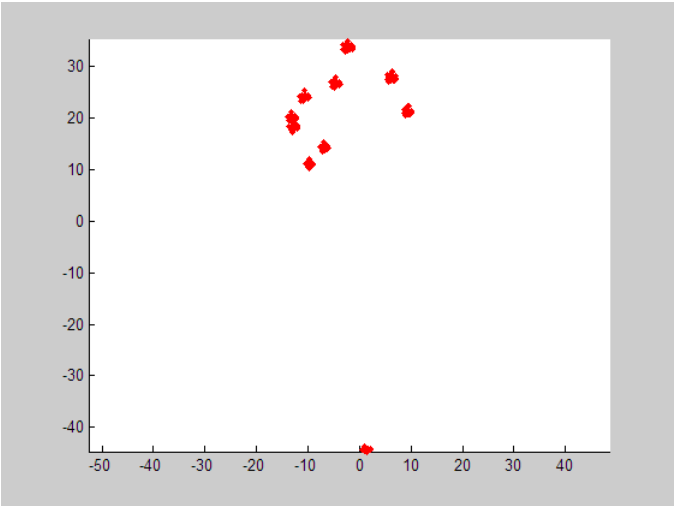


Fig. 20. Clusters of points for repeated marking of 10 accidentals, presented on one shoe-aligned coordinate system. Each accidental was marked 38 times. The distances are expressed in % of shoe length (100% = entire length from bottom to top of the shoe sole).

We then checked the distance between each pair of points, for all possible ~7,000 pairs. This gave a histogram of distances that approximates the distribution of errors (Fig. 21).

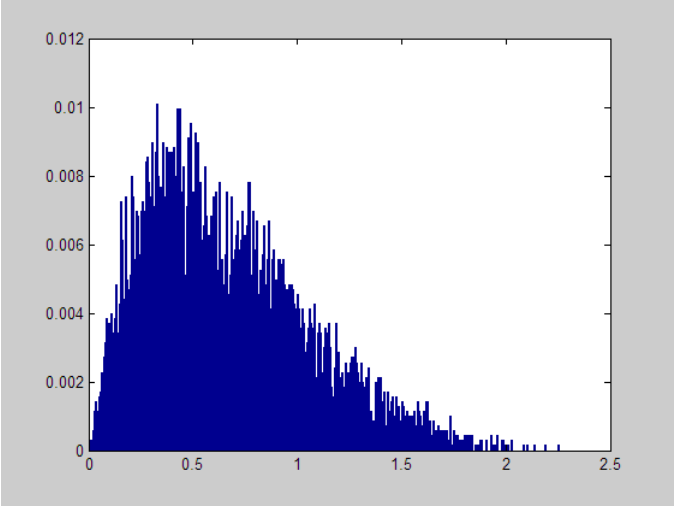


Fig. 21. Histogram of distances between 7,030 pairs of points from the same data that was used to create Fig. 20. The distances are expressed in % of shoe size.

Using the 7030 comparisons (Fig. 21), we can describe the statistics of location error (the distance between repetitions of the same accidentals):

The minimal error was 0.00980 (% of shoe size).

The maximal error was 2.25340 % → 6.75 mm for a 30 cm shoe.

The average error was 0.66788 % → 2.00 mm for a 30 cm shoe, with standard deviation of 0.39199.

The results of test B are similar to those reported above for test A, with the average error of 2 mm for a 30 cm shoe (~1.25 times smaller than the average error in test A).

Using the results of the test - the estimate of location error and its variability – we can turn to the question of grid size.

2.4.5. What should be the size of the grid?

The higher the error in estimating the location, the lower our ability to discriminate between two locations. This is expressed by the size of the grid: the higher the error, the lower the number of grid cells.

Our estimate of the average location error is ~2.5 mm (the results from tests A and B). This estimate is based on the operation of well-trained human markers. Taking into account that in practice larger errors might occur, and to be on the safe side, we double that number - 5 mm.

Assuming equal error in both X and Y directions, the size of a grid cell is (0.5 cm)x(0.5cm) = 0.25 cm^2.

Taking a shoe of size 30x10 cm, this leads to a grid of (300 / 0.25) = 1,200 cells.

Other researchers have placed various grids over the shoe sole. Stone [31] divided the shoe sole to one square millimeter units, and therefore a typical shoe sole contains approximately 16,000 possible locations. Petraco [4] suggested a different grid, specifically for the front of the shoe sole, of 324 cells that are not square and of varying sizes.

2.4.6. Configurations

Using the same data we also checked the statistics of two-accidental configurations. We defined the distance between configurations as the sum of distances between the individual locations. We compared pairs of configurations (all data used was from the repeated markings). We then compared all possible pairs, and for each, calculated the distance function:

$$dis\left(configuration(k,l)_i, configuration(k,l)_j\right)$$

Where k and l are two different accidentals, and i,j are two different repetitions of the accidental.

This yielded a histogram of distances between 196,850 pairs of configurations (Fig. 22 *left*). This histogram is almost exactly the same as the histogram for single locations (with 2X the values, as expected).

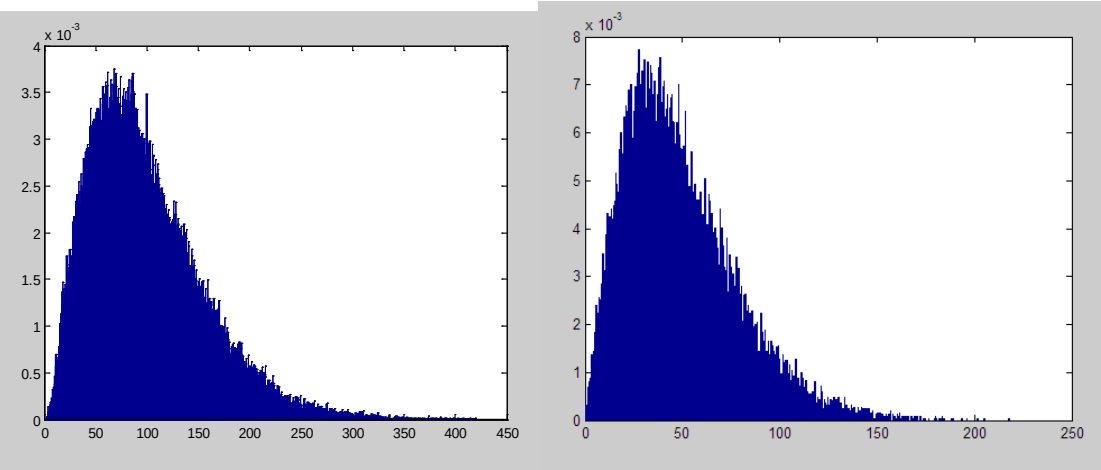


Fig. 22. *Left*: the histogram of distances between all 196,850 pairs of two-accidental configurations. *Right*: the histogram of ~24,000 pairs of single locations (the same as Fig. 19).

2.4.7. Analyzing the spatial distribution

After collecting the location of the accidentals we superimposed all onto one coordinate system (the CONTOURS data set, Fig. 23 *Left*). This of course raises the questions: are the accidentals uniformly distributed? Do they show specific patterns?

We used a 2D kernel estimation technique [32] to estimate the 2D Probability Density Function (PDF) of accidentals (Fig. 23 *Right*).

The result (Fig. 23 *right*) reveals a pattern that seems far from uniform. It also seems that the distribution of accidentals is highly influenced by the model of the shoe. How can one compensate for that influence?

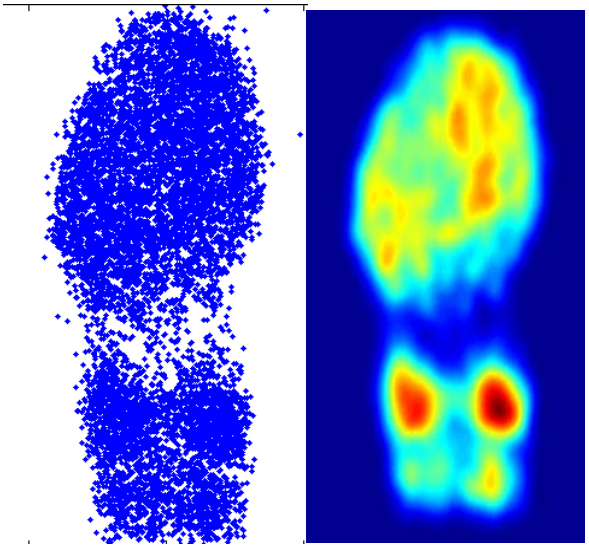


Fig. 23. *Left*: locations of ~8,900 accidentals of the CONTOURS data set. *Right*: estimated 2D pdf, color coded from blue (low probability) to red (high probability).

The discussions about this question led to the following understanding. Every shoe sole has specific contact areas with the surface upon which it is placed. These areas depend on the design of the shoe sole (shoe type, model, and size), and on the wear pattern of the specific shoe.

An accidental can leave its mark (using a test impression) *only* if it is situated on the contact areas of the shoe. We therefore decided to collect not only the location of the accidentals, but also the contact areas of all of the shoes in our database. Each test impression was segmented into two images: a foreground image (all pixels of its contact area) and a background image (all other pixels). All foreground images were added together giving an image that counts, for each pixel, how many times it was part of the contact area (Fig. 24).

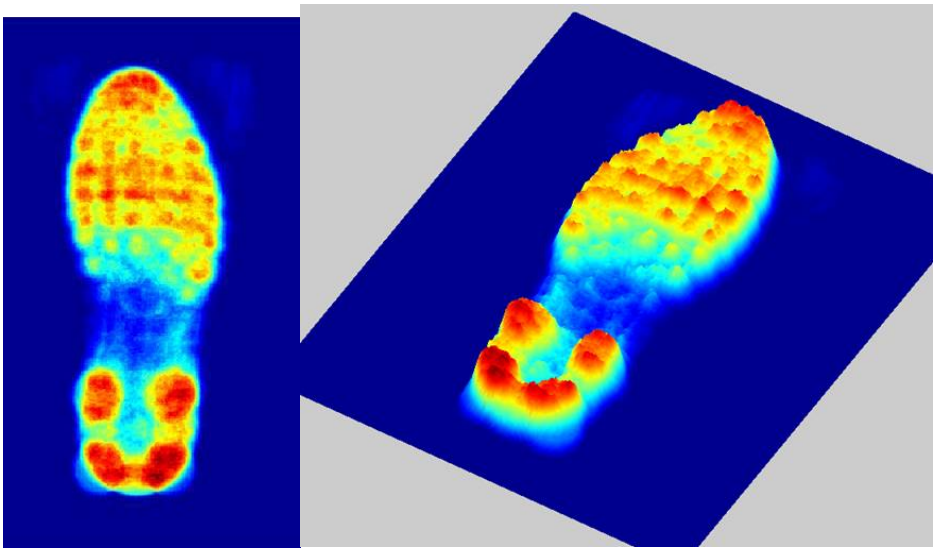


Fig. 24. Accumulated contact areas assembled from the 300 test impressions of the CONTOURS data set. *Left*: the contact area image. *Right*: a 3D view, showing the count in color as well as in height.

Next we aligned the two images – the 2D pdf and the accumulated contact areas image in order to estimate the probability of having an accidental in a specific location (Fig. 25).

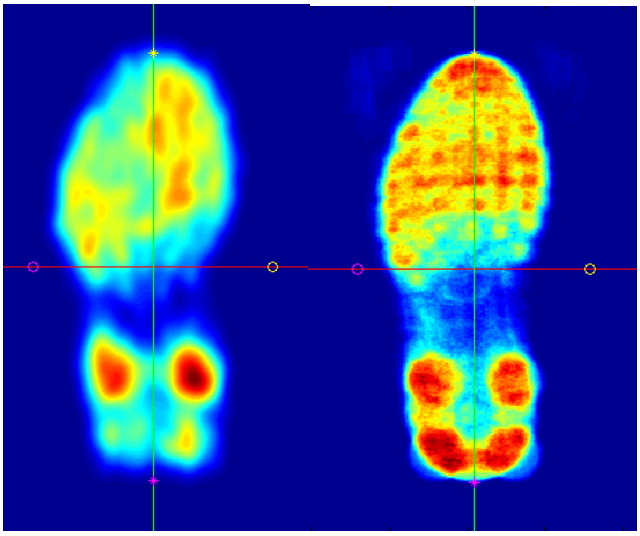


Fig. 25. Alignment of the estimated 2D pdf (*Left*) and the accumulated contact areas image (*Right*) from the 300 test impressions of the CONTOURS data set.

As described above, the spatial model for a single accidental should answer this question: "what is the chance for an accidental to appear in a specific location?" The model should take into account the error in defining a location, and the expected number of accidentals per shoe. It should also take into account the distribution of accidentals on the shoe sole. As described above, our initial assumption was that this distribution is uniform.

In order to estimate the distribution we attempted to normalize the 2D pdf (Fig. 25 *left*) by the accumulated contact area images (Fig. 25 *right*). The reasoning is this: when looking at the pdf only, we can see that some shoe models are more common than others among

the test impressions scanned into the database. We also do not compensate for the different wear patterns of the shoes in our database. Normalizing the pdf by the contact areas is a potential solution to perform the necessary compensations.

Basically the normalization should be a division of the raw 2D pdf (the value in each location) by the contact area (its value in the same location).

$$g(x,y) = f(x,y)/a(x,y)$$

Where $f(x,y)$ is the raw 2D pdf, $a(x,y)$ is the accumulated contact area function, $g(x,y)$ is the normalized pdf, and x and y are coordinates on a normalized shoe, belonging only to the area of the shoe.

A second normalization should follow: since we want to use the result as a 2D pdf, we need to normalize its integral to the sum of 1. This is done by dividing by the area under the pdf, i.e. the integral over the entire space:

$$\hat{g}(x,y) = \frac{g(x,y)}{\iint\limits_{x,y \in shoe} g(x,y)dxdy}$$

The first step in the normalization, i.e. the division by the accumulated contact area function ($f(x,y)/a(x,y)$), raises many technical difficulties.

The first and obvious problem with simple division is zero values of $a(x,y)$. Zero values of $a(x,y)$ indicates that in these specific locations no shoe sole was in contact with the surface. This should not be problematic since if no shoe sole contributed to those locations, we should not expect to have any accidentals in the same locations, i.e. the 2D pdf $f(x,y)$ should have zero values too in those locations. The solution then should be simple: for all *x,y* that $f(x,y)$ is zero, make $g(x,y)$ zero too, regardless of $a(x,y)$; anywhere else use the division as stated above. But this simple solution does not always work: $a(x,y)$ might have zero values in locations in which $f(x,y)$ is not zero. This is due to a basic difference between the ways $a(x,y)$ and $f(x,y)$ are constructed. $f(x,y)$ is a 2D pdf estimated from numerous normalized locations. It is smoother by nature than $a(x,y)$ (Fig. 25). This means that many of the zero valued locations in $f(x,y)$ will be smoothed out to have larger than zero values.

A possible ad-hoc solution to the division-by-zero problem is to add a small value to $a(x,y)$. This prevents the problem but with the price of an increase in the inaccuracy of the results; another way is to exclude altogether areas in $f(x,y)$ that have low values.

Thus, only areas of middle or high number of accidentals (that are expressed as higher values in $f(x,y)$) will be normalized. The excluded areas would be non-informative.

A second problem with simple division is that it relies on perfect alignment of $f(x,y)$ and $a(x,y)$. Indeed the alignment of the two functions is done exactly the same, and on the same test impression images, therefore this should be true. But any error or inaccuracy in the process of creating $f(x,y)$ (e.g. ill positioning of an accidental), or $a(x,y)$ (e.g. failing to accurately segment foreground from background due to some noise in the image) might cause local misalignment. A possible solution to that problem is to smooth both functions even more (especially $a(x,y)$). This will decrease the spatial resolution and accuracy of the result, but might increase its reliability.

A third problem with simple division is the boundaries of the shoes. Again, because of the way $f(x,y)$ is estimated, it smoothes edges. This may cause a drift of values from the boundary of the sole to locations outside the sole – where $a(x,y)$ is zero. A possible solution to that is to strictly crop $f(x,y)$ using a mask based on $a(x,y)$. Thus any position that is outside the sole would be excluded.

Another way of trying to overcome most of the problems was to increase the accuracy of our estimates of $f(x,y)$ (by increasing dramatically the number of accidentals, and using a simpler method to mark the centroid of each accidental), and of $a(x,y)$ (by using another set of test impressions with twice the number of images).

Therefore we used the same methods on the FAST data set as well, with ~20,000 new accidentals from 600 shoes (Fig. 26-28).

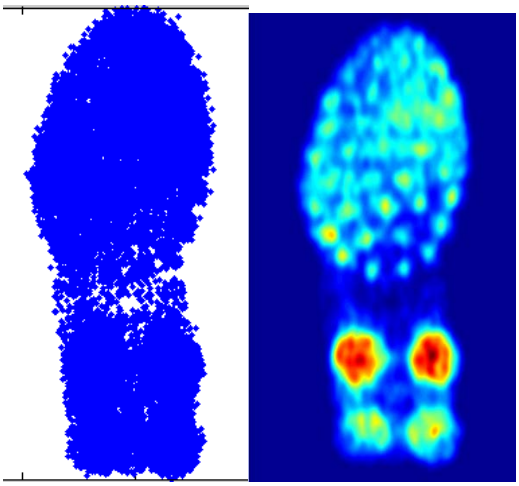


Fig. 26. *Left*: locations of ~20,000 accidentals of the FAST data set. *Right*: estimated 2D pdf, color coded fro blue (low probability) to red (high probability).

Note that for the FAST data set, the 2D pdf (Fig. 26 *right*) is more detailed and has higher resolution than for the CONTOURS data set (Fig. 23 *right*). This is due to two reasons. First, the number of accidentals marked is two times larger for the FAST dataset. Second, the locations of the CONTOURS dataset are less accurate – being derived from the bounding box that surrounds the accidentals and not the accidentals themselves. However, the general tendencies look very similar.

The accumulated contact areas images of the two data sets are also similar (Fig. 24 *left* and Fig. 27 *left*). This means that the 300 shoes of the CONTOURS data set and the 600 shoes of the FAST data set represent a similar sampling of the shoe soles space.

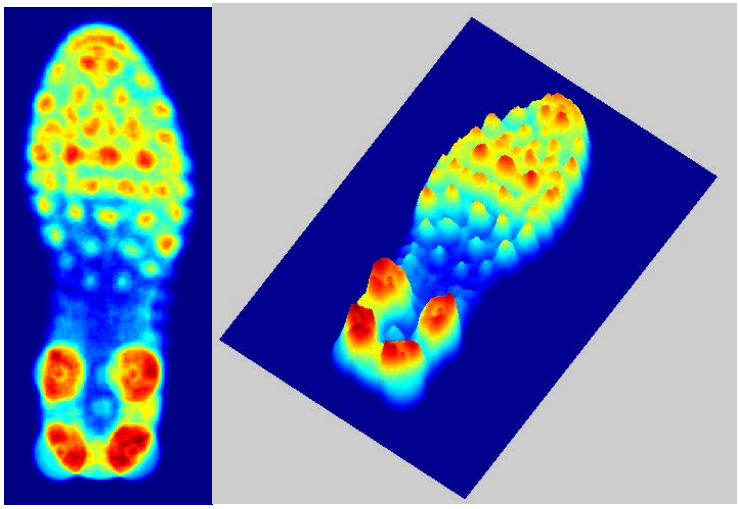


Fig. 27. Accumulated contact areas assembled from the 600 test impressions of the FAST data set. *Left*: the contact area image. *Right*: a 3D view, showing the count in color as well as in height.

The results are shown again with our attempted normalizations (Fig. 28).

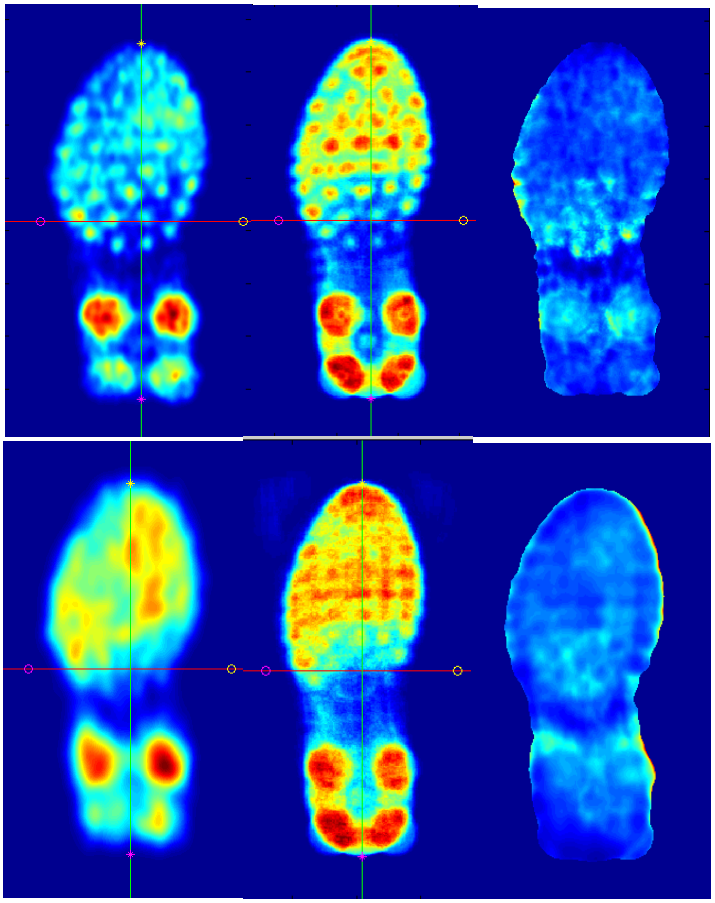


Fig. 28. Comparing the results of the two data sets. *Top*: the FAST data set. *Bottom*: the CONTOURS data set. *Left*: the 2D pdfs. *middle*: the accumulated contact areas images. *Right*: the normalized pdfs.

To summarize this subsection (on spatial distributions of accidentals)

1. We accumulated the locations of ~30,000 accidentals from ~900 shoes (using the CONTOURS and FAST datasets).
2. We estimated raw 2D pdfs from the data using kernel estimation techniques.
3. We gathered the accumulated contact areas of each dataset.
4. We only partially succeeded in normalizing the raw 2D pdfs by the accumulated contact areas. Our initial results suggest that the uniform distribution of locations is not bad as a first approximation. Since this is the best approximation we have so far, uniform distribution of locations is the model implemented in our deliverable software.
5. Using the results of this section we suggest a method of estimating the probability in future research.

2.4.8. How can the probability of having an accidental as a function of location be estimated?

Here is the (not fully completed) methodology we suggest using:

1. Collect many accidentals on many test impressions.

2. Align the shoes; e.g. by using the universal coordinate system we suggested.
3. Superimpose the normalized locations of the accidentals on one coordinate system.
4. Estimate $f(x,y)$ - the raw 2D pdf function from the superimposed locations.
5. Estimate $a(x,y)$- the accumulated contact areas.
6. Normalize the raw 2D pdf with respect to the accumulated contact areas by using something like $g(x,y) = f(x,y)/a(x,y)$. Note that a simple division is not suitable in practice and some technical details must be addressed, as described above.
7. Estimate the probability of finding an accidental at an area around a location using the estimated location error. Let (a,b) be the normalized location we are interested in, and ε the normalized error (in both direction). We are therefore looking for the probability of having an accidental in the range of $(a \pm \frac{\varepsilon}{2}, b \pm \frac{\varepsilon}{2})$.

   Defining $a_1, a_{2,}, b_1, b_2$ to be

   $$a_1 = a - \frac{\varepsilon}{2} \quad , \quad a_2 = a + \frac{\varepsilon}{2}$$
   $$b_1 = b - \frac{\varepsilon}{2} \quad , \quad b_2 = b + \frac{\varepsilon}{2}$$

   Then the probability is

   $$\Pr[a_1 \le X \le a_2, b_1 \le Y \le b_2] = \int_{a_1}^{a_2}\int_{b_1}^{b_2} g(x,y)dxdy$$

8. Multiply the result by the average number of accidentals per shoe (estimated from a large dataset).

2.5.Statistical model of orientation of accidentals

2.5.1. Introduction

Because of the hypothesis that shape and orientation of accidentals are independent, we decided to build a specific model for the orientation of accidentals and not to include it as part of the shape model. The best example of this independence is narrow scratches. The shape of such scratches is not rare, but they may appear in various orientations. Such independence could be explored to decompose a complex statistical model into a product

of two simpler models. Each such model is easier to construct, and more important, requires less data in its estimation.

This statistical model of orientations contains two parts:

A. Finding the distribution of orientations for numerous accidentals. This was done by analyzing the shape of all accidentals in the CONTOUR data set (n = ~8,900), and extracting their orientation relative to their shoe-specific coordinate system.
B. Finding the error in orientation. This was done using the multiple marking data sets. We found that the error in orientation depends largely on the shape of the accidental: longer shapes have lower errors (as is explained in the next section). This relation was then used to devise a fine tuned estimation of orientation error.

2.5.2. Extracting the orientation from the shape

To find the orientation of an accidental we used Principal Component Analysis (PCA) of the contour pixels [33]. PCA (when applied to a 2D data like the contour of the shape) finds the two orthogonal axes of the shape - the major and the minor axes. The major axis is the direction which explains most of the variability of the points. PCA also finds the variances along the two axes.

Finding the orientation of a shape amounts to finding its major (longer) axis, and finding the angle of that axis relative to the X axis of the shoe aligned coordinate system.

Orientation score

For elongated shapes, this procedure works well. However, for some of the more rounded shapes, the major and minor axes are not very distinct. We quantified this by defining a score for each shape: the ratio of variances along the two axes. A perfect circle would have a score of 1. An elongated shape would have a higher score (see Fig. 34 below).

2.5.3. Estimating the distribution of orientations

We calculated the orientation (in degrees, ranging from -90 to +90) for each of the ~8,900 accidentals of the CONTOURS data set. The results (Fig. 29) show a uniform distribution (with some noise).

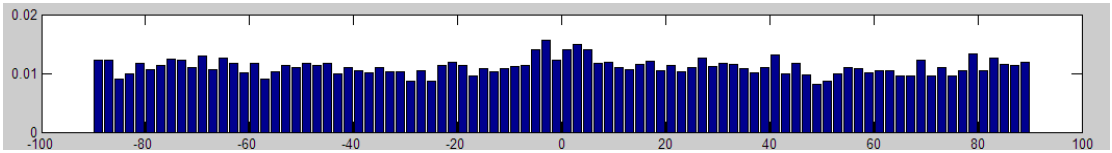


Fig. 29. Histogram (90 bins, each of 2 degrees) of orientations of the ~8,900 accidentals of the CONTOURS data set.

Even when the accidentals are divided into groups according to their shape (elongated vs. rounded), the distributions of orientations are largely uniform (section 2.5.6 below).

2.5.4. Estimating the error in orientation

The error in orientation has two main sources:

A. The error in estimating the orientation.

B. The error in the process of assigning specific coordinate system to a shoe.

Previously we concluded that the spatial error is on the order of 0.5 cm, and its main source is the variability in the coordinate systems. Orientation error is different. It is caused mainly by the problems in estimating the orientation, which are especially large for condensed, non-elongated shapes.

To estimate the error in the orientation extraction process we used the 5 datasets of repeated markings. For each of the accidentals we extracted the orientation angle (with respect to its shoe-specific coordinate system), and the orientation score.

We then calculated the difference in angle between every pair of repetitions of the same accidental. The total number of pairs was 82,595 (Fig. 30). These differences can be taken as the error in extracting the orientation of an accidental. Most of the orientation differences are below 10 degrees, but in some cases the error is as large as 80-90 degrees.

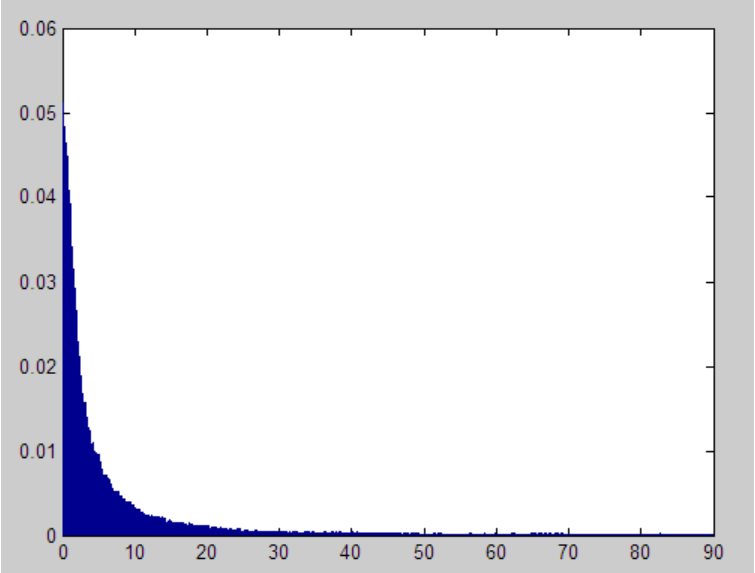


Fig. 30. Histogram of angle differences for 82,595 pairs of repetitions of the same accidental.

To investigate the source of such errors, let us look at the relation between orientation error and orientation score (Fig. 31). The results show a clear tendency: the error is reduced as the score is increased. Rounded accidentals with lower scores have very high orientation error, while elongated accidentals (high score) have low orientation error.

The orientation scores in Fig. 31 are accumulated into bins. Each bin in itself has numerous values, mostly of low error. To better describe this we calculated the average value and standard deviation of each bin. We then presented the average error + 1 standard deviation as a function of orientation score (Fig. 32). The data points were fitted by 3 exponentials in order to get a simple, usable functional description of the relations.

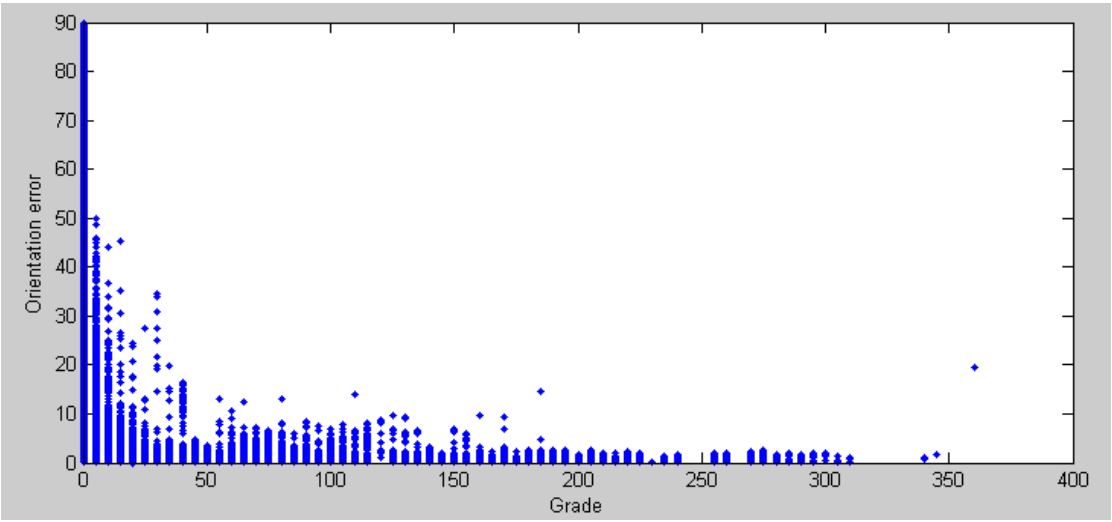


Fig. 31. Orientation error as a function of orientation score (or grade) using 82,595 pairs of repetitions of the same accidental. The orientation scores were accumulated into bins.

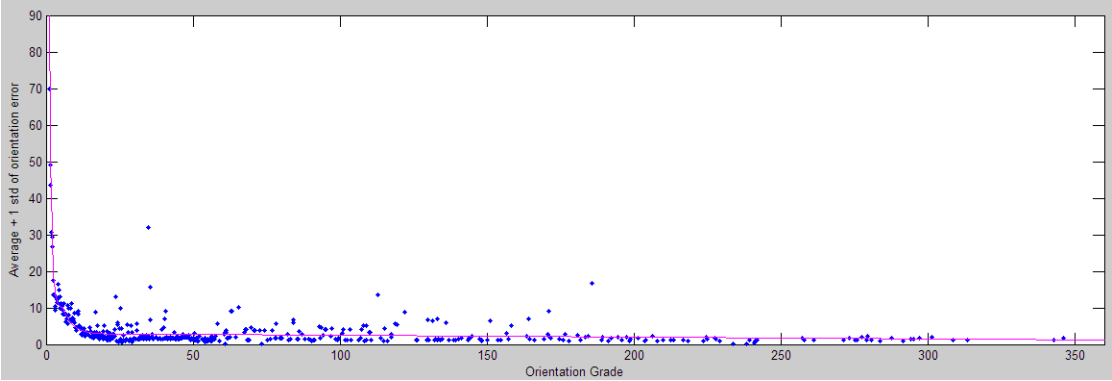


Fig. 32. Orientation error as a function of orientation score (or grade) using the average + 1 standard deviation of the bins of Fig. 31. The data points were fitted by 3 exponentials.

Next we describe how we use this estimate of orientation error in calculating the probability of finding an accidental with a specific orientation.

2.5.5. Relating the orientation error to probability

Given the error in estimating the orientation and the distribution of orientations, we can now estimate the probability of finding an accidental with a specific orientation.

Let $g(x)$ be the pdf of orientations, a be the orientation we are interested in, and ε the orientation error for a specific accidental. We are therefore looking for the probability of having an orientation in the range of $(a \pm \frac{\varepsilon}{2})$.

Defining $a_1, a_2,$ to be

$$a_1 = a - \frac{\varepsilon}{2} \quad , \quad a_2 = a + \frac{\varepsilon}{2}$$

Then the probability is

$$\Pr[a_1 \le X \le a_2] = \int_{a_1}^{a_2} g(x)dx$$

Assuming the orientations are distributed uniformly, the probability of having an orientation in the range of $(a \pm \frac{\varepsilon}{2})$ is

$$\Pr[a_1 \le X \le a_2] = \varepsilon / 180$$

Using the dependency of the error on the orientation score, we can calculate the probability of having an orientation in the range of $(a \pm \frac{\varepsilon}{2})$ as a function of orientation score (Fig. 33).

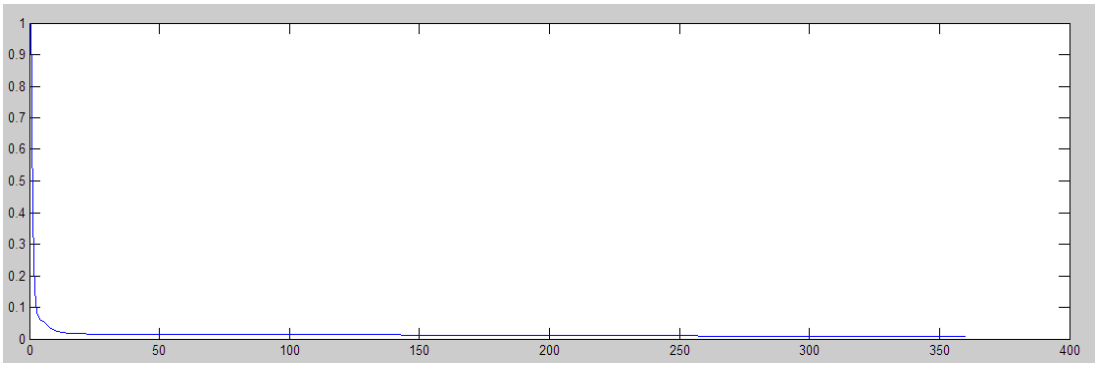
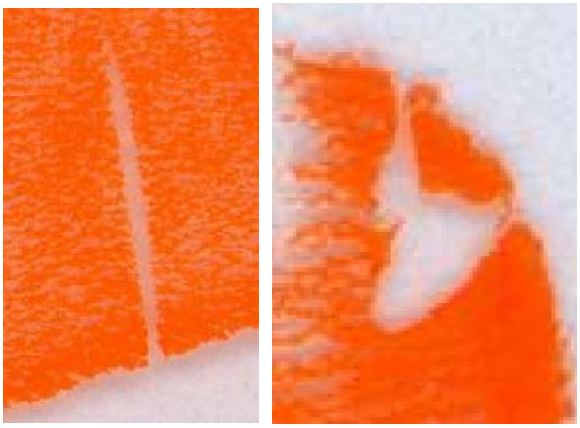


Fig. 33. Probability of orientation as a function of orientation score.

Let us finish this sub-section by an example of calculating the error and probabilities for 2 specific accidentals (Fig. 34).



Orientation score	**150**	**2.2**
Orientation error	**2.25°**	**26.67°**
Probability	**0.0125**	**0.148**

Fig. 34. Two accidentals with their orientation scores, orientation error, and probability.

The first accidental is very elongated, with a high orientation score of 150. Using the fit in Fig. 32 its orientation error is found to be only 2.25 degrees. Using the relation in Fig. 33 we find that the probability of finding such an accidental with a specific orientation within a range of 2.25 degrees is 0.0125 = 1/80. In other words, an elongated accidental's uniqueness is greatly increased if we know its orientation. On the other hand, the second accidental is much shorter and almost round. Its orientation score is 2.2 (remember that a circle has a score of 1), its orientation error is big: 26.67 degrees. The probability of finding an orientation within a range of 26.67° is almost 1/6.

2.5.6. Estimating the distribution of orientations for shape categories

In section 2.5.3 above, we described the uniform distribution of orientations in our dataset.

We also performed a finer estimation of the distribution by dividing the data into groups, according to their orientation score.

First we looked at the distribution of orientation score (Fig. 35). There are few very large values (very long and thin accidentals), but the majority of the values are from 1 to 10, i.e. from round shapes to moderately elongated.

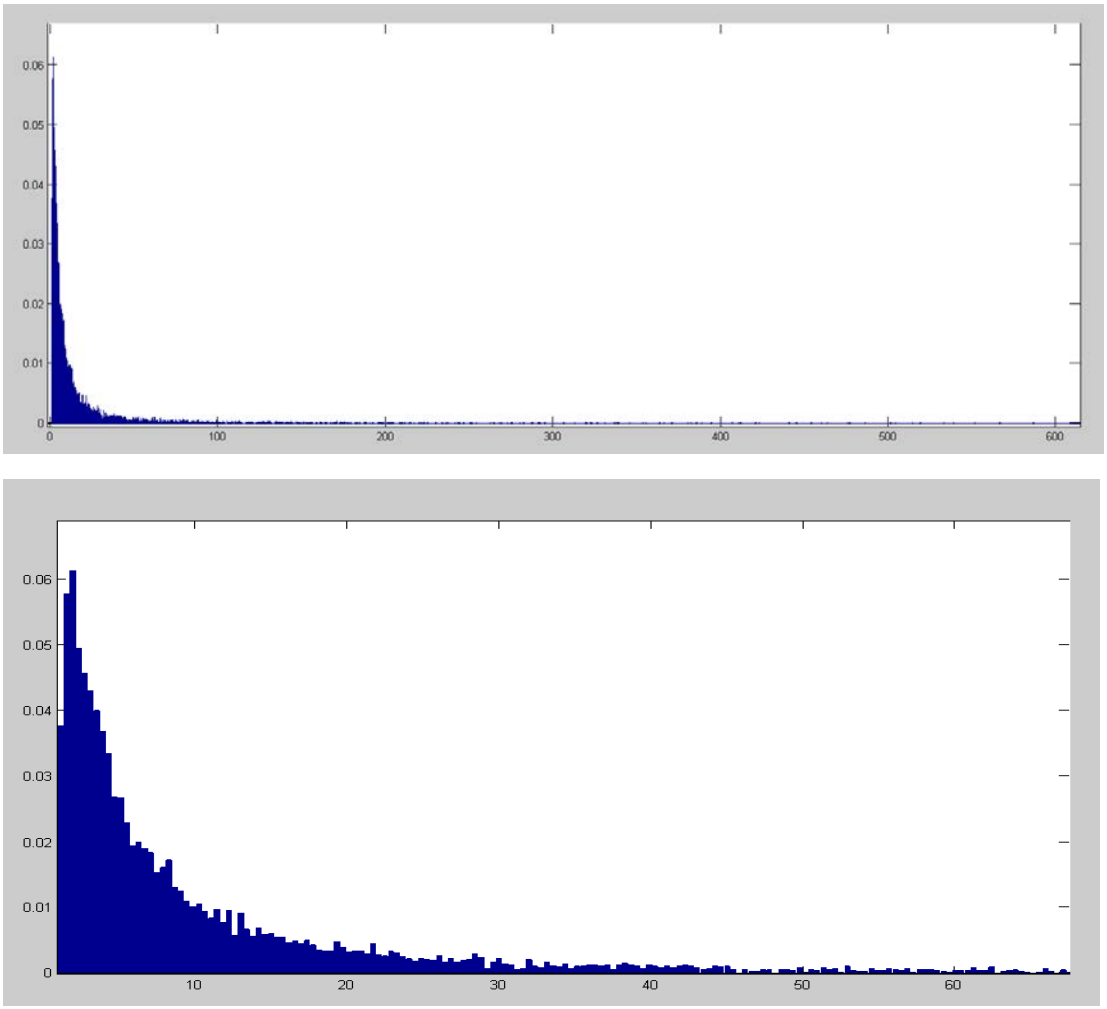


Fig. 35. The distribution of orientation score in the CONTOURS dataset (n = ~8,900). Top to bottom: Zooming in on the left part of the data.

Then we looked at the orientation vs. orientation score (Fig. 36). The results are that there is no obvious relation between the two.

Next we divided the accidentals into groups according to their orientation error. The whole population of ~8,900 accidentals was divided into groups of ~1,000 accidentals each. For each such group we calculated the mean orientation error. Using the errors we prepared histograms of orientations for each group. The results are presented in Fig. 37.

The results show that for almost all groups the distributions of orientations are nearly uniform (with noise), or at least do not have any clear tendency to any obvious orientation. The interesting exceptions are the last two groups of the most elongated shapes. There the distribution has distinctive peaks around orientation of 0 degrees (the long axis of the shoe), especially for the last group - the long scratches (with orientation score of 100 to 2,000). Note however that in the last group we have only 306 accidentals.

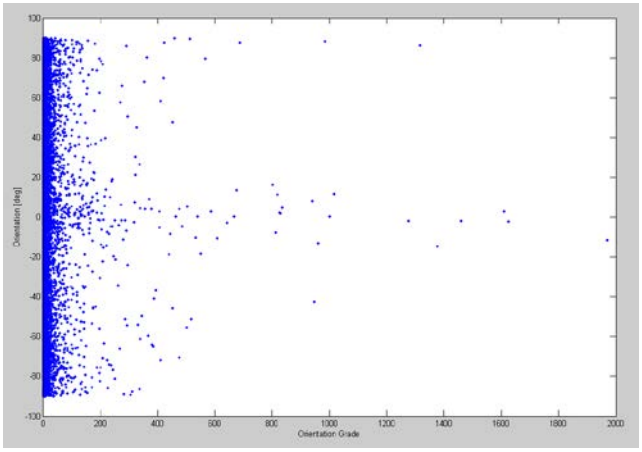
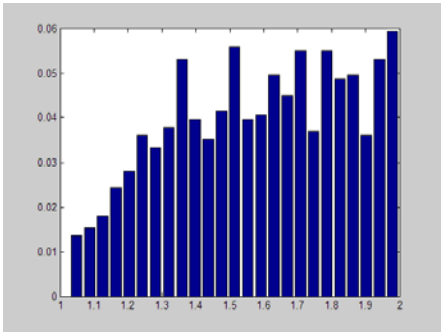
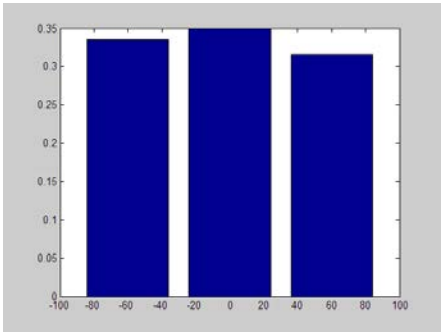


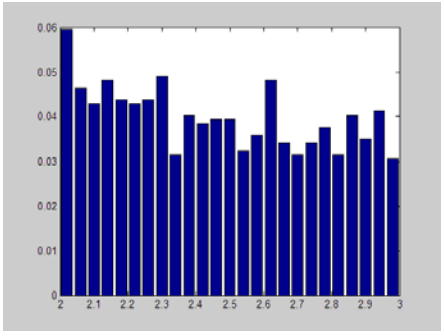
Fig. 36. Orientation [deg] vs. orientation score. Based on the accidentals from the CONTOURS dataset (n= ~8,900).

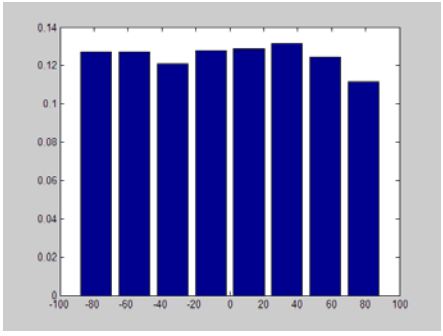
Distribution of the orientation score **Histogram of orientations**



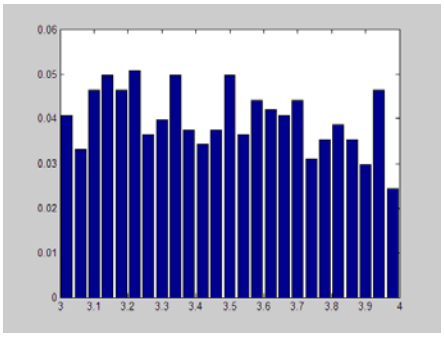
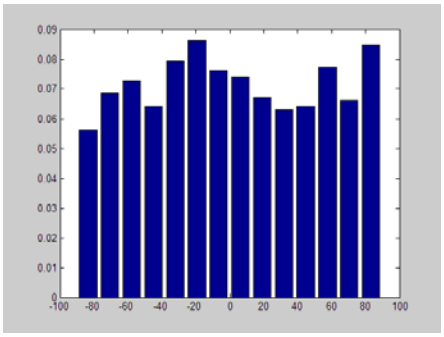


Score values 1 to 2. Mean error: 60° N = 1110.

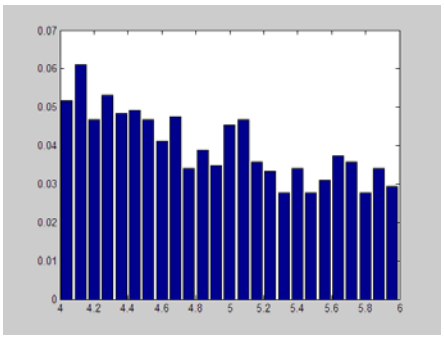
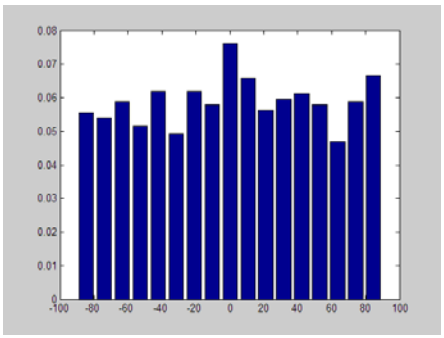




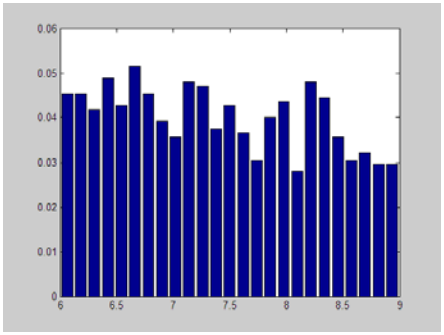
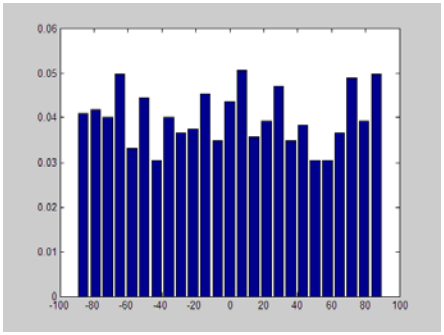
Score values 2 to 3. Mean error: 22°. N = 1140.

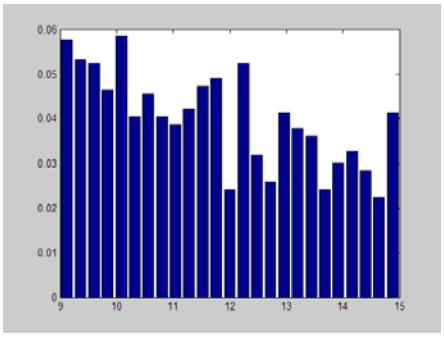
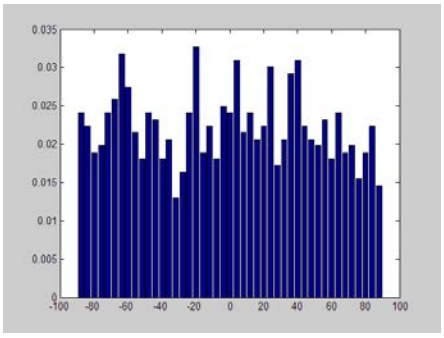
Score values 3 to 4. Mean error: 13° N = 906.

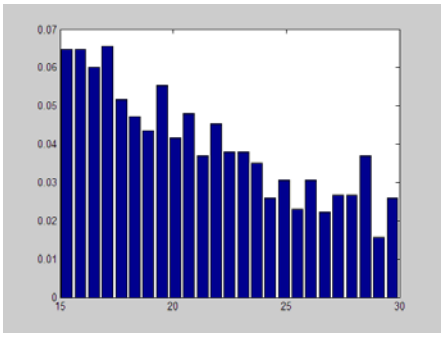
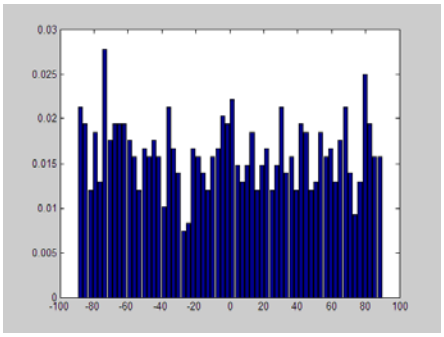
Score values 4 to 6. Mean error: 10° N = 1260.

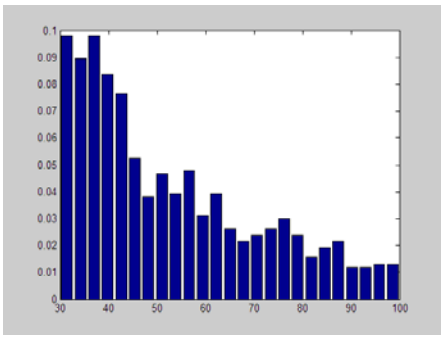
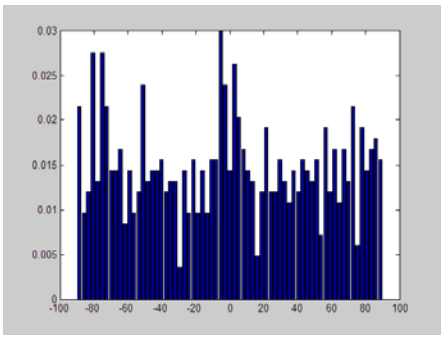
Score values 6 to 9. Mean error: 7° . N = 1147.

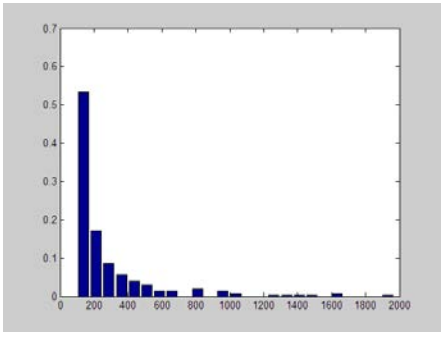
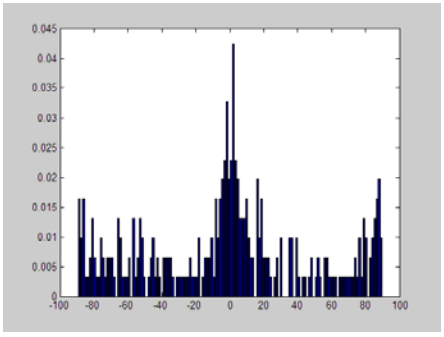
Score values 9 to 15. Mean error: 4° N = 1163.

Score values 15 to 30. Mean error: 3° N = 1082.

Score values 30 to 100. Mean error: 2.5° N = 836.

Score values 100 to 2000. Mean error: 1.25° N = 306.

Fig. 37. Distribution of orientations for different groups of accidentals, from the roundest shapes (first row) down to the most elongated shapes (last row). Each row shows the results for one group. Left: the distributions of orientation score for the group. Right: the histogram of orientations for the group. Note that the number of bins in each histogram is 180 / mean error. Additional information for each group (orientation score values, mean orientation error and number of samples) is presented below the plots.

The uniformity of the orientation of most accidentals shows that there isn't a preferred direction for accidentals on the shoe sole. The results of this project contradict the results obtained in a former work done as a M.Sc. in the University of Lausanne [34]. In this thesis the author found that more accidentals are formed on the long axis (north to south) than on the perpendicular axis. We found this to be true only for very long accidentals.

2.5.7. Summary of the statistical model of orientation

A. The distribution of orientations is uniform (for all but the longest shapes).
B. The shape of the accidental dictates its orientation error. The orientation of an elongated shape has low error and thus can be estimated accurately, but for a round, condensed shape the estimation of orientation is error prone and its orientation is poorly defined.
C. We used multiple repetitions of marking of the same accidentals to estimate the error in orientation. We found experimentally a functional relation between the shape (orientation score) and the orientation error.
D. Using the relation between orientation score and the orientation error, and the uniform distribution of orientations, we described the probability of finding an accidental with specific orientation (within the error range). The probability is simply *error value*/180.
E. The results (of the whole process) are that, while an elongated shape is very rare (i.e. has low probability of finding its specific orientation for a different accidental)–a rounded shape is not (since any orientation within the very large error range cannot be distinguished from any other).

Using the software tools delivered by this project, the process of finding the probability of orientation of a new accidental is:

1. Mark the contour of the accidental. (A semi-automatic operation in MarkAccidentals)
2. Mark the shoe aligned coordinate system. (A manual operation in CompareAccidentals)
3. Find the orientation of the accidental and its orientation score. (An automatic operation in CompareAccidentals)
4. Calculate the probability of finding an orientation within the range of error. (An automatic operation in CompareAccidentals)

2.6. Statistical model of accidental shapes

2.6.1. The statistical model of shapes - a conceptual framework

Errors in shape measurements: Like the location and orientation of an accidental, the shape has measurement errors too. These errors have to be quantified and taken into account when building statistical models.

The shape of an accidental is marked in a semi-automatic process using MarkAccidentals - a software tool developed in this project. The tool estimates the contour of an accidental using foreground from background segmentation, with the possible help of a human operator.

There are two main sources of variability in this process (hence errors in estimation of the shape). The first is the differences in actions of the human operators (the choice whether to intervene or not, the thresholds they choose, the contour parts they leave unchanged and those they delete, and the degree of manual marking they use to refine the results). The second source of shape variability is in the different appearance of an accidental in different test impressions of the same shoe.

Uniqueness of the shape attribute: There is a significant difference between the shape of an accidental and its location and orientation. Shape is a high dimensional property, compared to the two dimensions of location and one dimension of orientation. This makes the shape space (the set of all possible shapes) too big to be estimated directly. There is much literature on the different aspects of this problem. A common approach is to quantify a shape using a small number of descriptors, and then to use those descriptors in many shape related tasks such as shape recognition, shape compression, shape comparison, retrieval of shapes from databases, etc. [35]. However, estimating the statistics of shape descriptors can still require a lot of data, especially if the different descriptors are statistically dependent. Another approach (and the one we took) is to transform the problem from high dimensional to one dimensional using the statistics of shape *comparisons* – the 1D space of shape dissimilarity values. This approach has its drawbacks too, mainly losing information in the process of reducing the problem to a 1D (see [36] for a discussion of the drawbacks of this approach and the usage of probabilistic graphical models instead in the case of handwritten questioned document examination). In our case the reduction to 1D dissimilarity space was used successfully showing that the loss of information was not critical.

We now will briefly explain our shape dissimilarity measure, and then continue with the description of the statistical model of shape.

Shape dissimilarity: Our measure follows the measure of distance between points and curves used in the *Iterative closest point* algorithm [37]; for each point in one contour find the closest point in the second contour. The dissimilarity measure is then the normalized sum of distances of all points from one contour to the other, for both contours. This is a version of the shape distance #22 of the Modified Hausdorff Distance as appears in [38].

In practice we compute the dissimilarity using Matlab's distance transform (due to running time considerations), by the following steps (Fig. 38-40):

A. The two contours are registered by searching for the rotation and translation with minimal distance between contours.

B. Each contour is represented by its own matrix (the two contours are shown here on the same matrix).

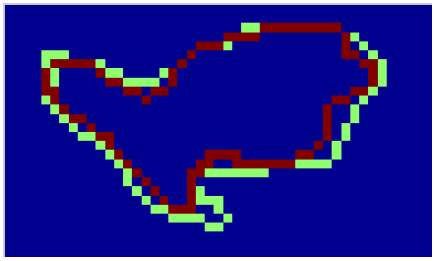


Fig. 38. Two aligned contours.

C. The ***distance transform matrix*** is computed for each contour.

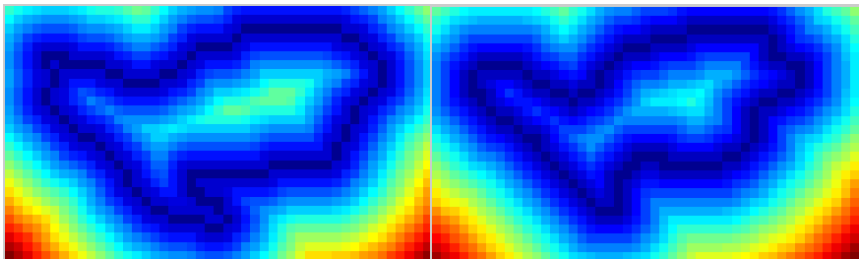


Fig. 39. The distance from each contour is color coded from blue (near) to red (far).

D. Each contour is superimposed onto the ***distance transform matrix*** of the other contour. A sum is calculated of the values of all matrix pixels that belong to the contour. This resulted in the values D12 and D21.

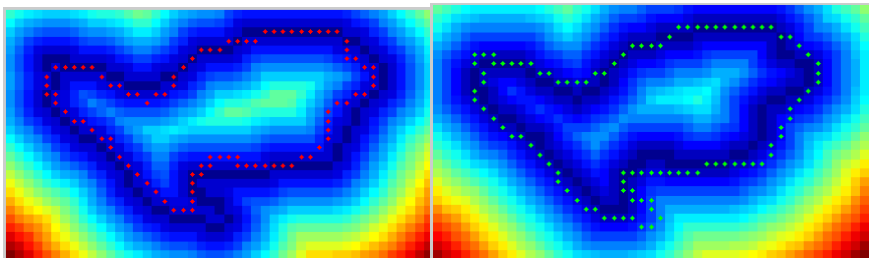


Fig. 40. Each contour is placed on the distance matrix of the other.

E. The two values D12 and D21 are averaged and divided by the average length of the contours.

Using this measure we quantified the variability in shape measurement of an accidental. This was done by marking the same accidental several times using the repeated markings datasets. But how does this help in finding what we want: how rare is a new shape, a shape that was marked only once?

Probability of a specific shape: The question of the probability of a specific shape is answered by counting how many shapes in our database are similar enough to the shape in question. Of course we need to define exactly what is 'similar enough'. This is done by comparing contours that originated from the same accidentals, and contours of different accidental, as is explained next.

Matches and non-matches: We define two populations. The first is the population of correct matches, i.e. pairs of contours that originated from the same accidental (and are different because of differences between the various test impressions and in the marking process). The second population is of the non-matches, i.e. pairs of contours that originated from a different accidental.

We expect that the matching error (degree of dissimilarity) between pairs in the 'correct matches' population will be lower than the matching error between pairs in 'non-matches' population. We expect the values to be drawn from two distributions that schematically look like this (Fig. 41):

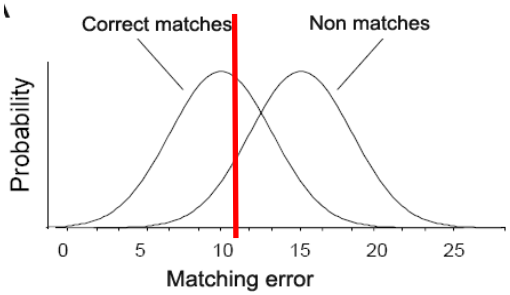


Fig. 41. A schematic representation of the expected matching error distributions of two populations of pairs of contours. The *correct matches* that originated from the same accidental, and the *non matches*, that originated from different accidentals. The red vertical line is the threshold between the two populations. The experimental distributions are shown in Fig. 44 below.

How can we use these distributions? First we can decide if two contours originated from the same source (accidental) or not. We compare the two contours and calculate their matching error. If it is to the *left* of the threshold (Fig. 41, red vertical line), the two contours are more likely to originate from the same accidental. If it is to the right of the threshold, they most likely originated from two different accidentals. Second, we can answer a different question 'What is the probability of an accidental's having a specific shape?' This is explained next.

The probability of a shape. The likely scenario for which our system is supposed to be used is this. We receive a new accidental and we want to find the probability of having this specific shape of an accidental in the database. The answer follows these steps:

1. Systematically comparing the shape against *all* the accidentals in our database.
2. For each comparison calculating its matching error.

3. Counting how many comparisons had a matching error value lower than the threshold. Such comparisons come from shapes that are too similar to be distinguished from our new shape. The number of these shape divided by the total number of the accidentals in our database is the result: the probability of having such a shape.

Can we do better? If we receive not only the image of the accidental to be checked, but also the shoe itself, we can create several test impressions, and mark the same accidental multiple times. This can be used to create its own 'matches' population in order to find a better value for the threshold.

2.6.2. Estimating the distributions of the matches and the non-matches populations

A. We checked several methods of matching error calculations and for normalizations, and chose the method that performs best (was most discriminative between the two populations).

B. We assembled histograms of both the *correct matches* population (Fig. 42, based on the multiple datasets, using more than 82,000 comparisons) and the *non-matches* population (Fig. 43, the CONTOURS dataset using more than 100,000 comparisons). Probability density functions (pdfs) were estimated from the histogram using Matlab's statistical toolbox (Fig. 44).

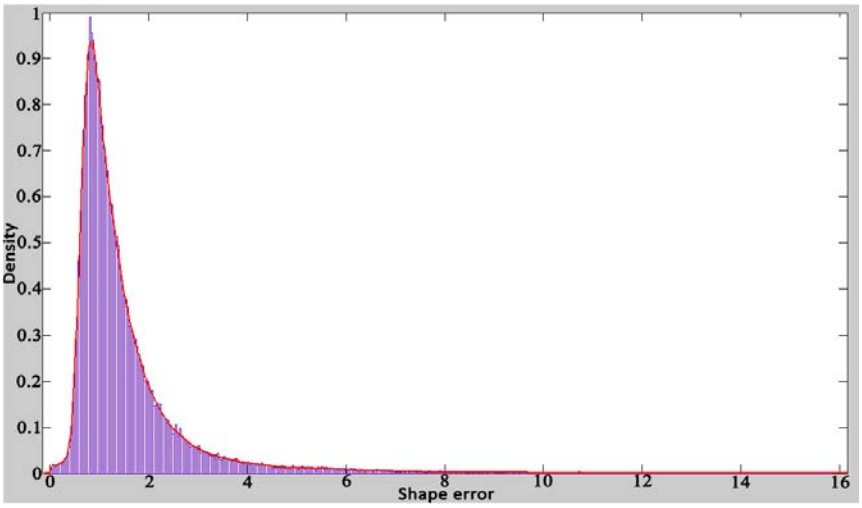


Fig. 42. Histogram of the correct matches population (blue) and a fit (red).

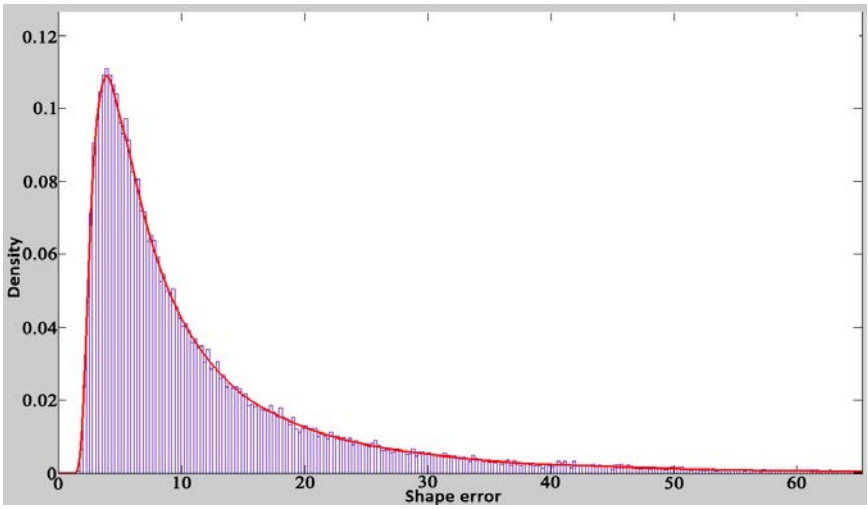


Fig. 43. Histogram of the non-matches population (blue) and a fit (red).

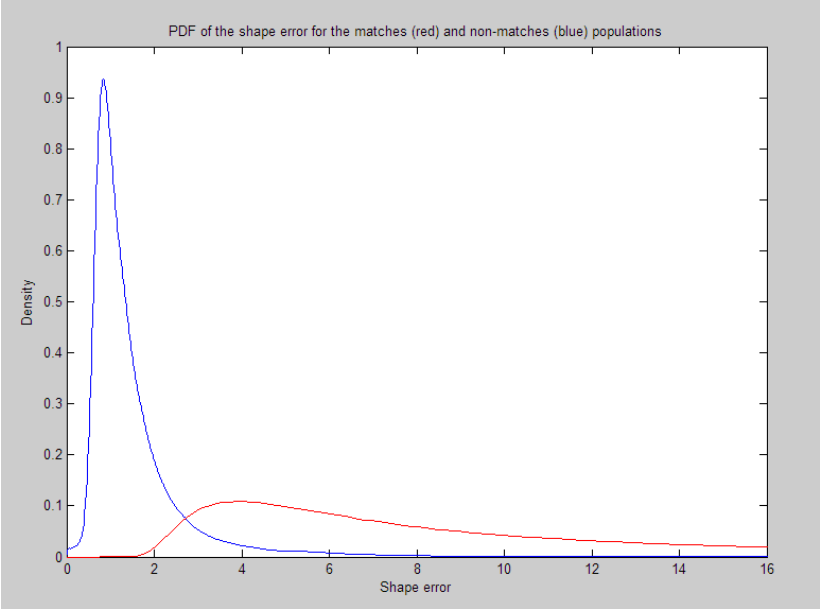


Fig. 44. The probability density functions (pdfs) of both populations. *Blue:* pdf of the correct matches population (n =82,592 comparisons). *Red*: pdf of the non-matches population (n = 110,700 comparisons).

2.6.3. Finding the optimal threshold

When deciding if a pair of shapes is similar enough to be considered a match, we should take into account several types of classification options and their implications. Classification of two shapes as a match may be either right (hit) or wrong (false alarm). Similarly, classifying a pair as a non-match, can be either right (correct reject) or wrong (miss). Each of these four types of classifications has a probability of occurrence that can be estimated using the distributions of two populations, the correct matches and the non-matches.

We shall use a schematic example to explain this type of analysis (Fig. 45). The left curve is the probability distribution function (pdf) of the matching error for the first group – the correct matches. The right curve is the pdf of the same measure for the group of non-matches. A specific value of the matching error is used as a threshold between the two groups (shown as a vertical line in Fig. 45B,C). Using the areas (light and dark gray) we can calculate the probabilities of hits, misses, false alarms and correct rejects.

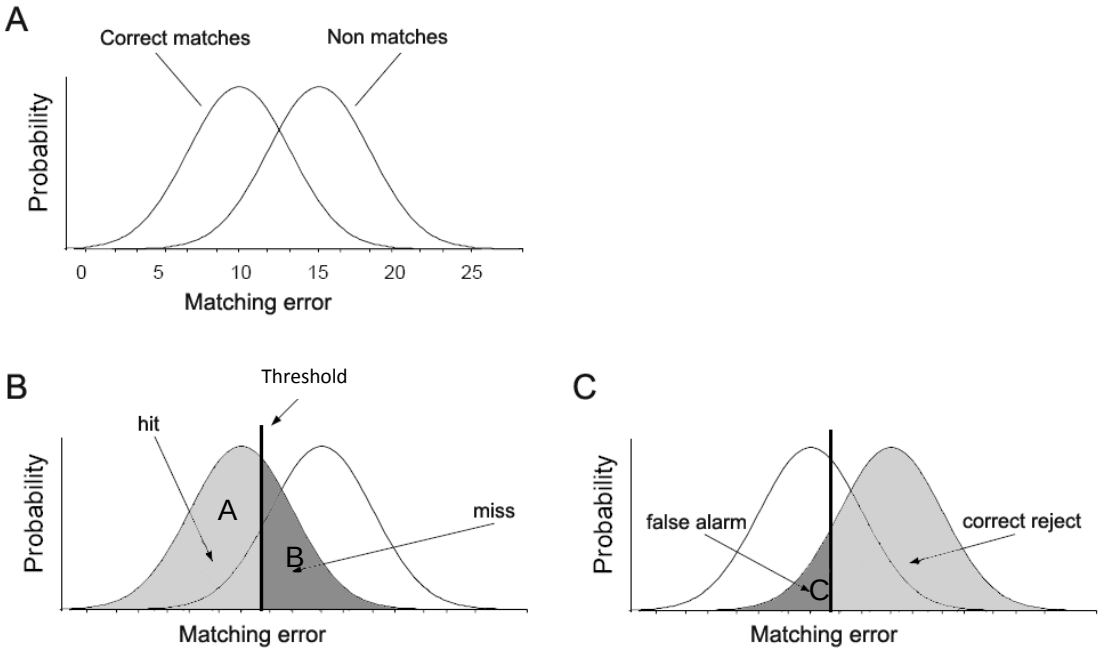


Fig. 45. Schematic illustration of the matching error probability density functions for correct matches and for non-matches. A. In this example the two pdfs have exactly the same shape with different means. B. All pairs of shapes with matching error values to the left of the chosen threshold will be classified as correct matches. Those with matching error values to the right of the threshold will be classified as non-matches. The light gray area (A) is the probability of a hit, i.e. a correct classification as a correct match. The dark gray area (B) is the probability of a miss, i.e. an erroneous classification as a non-match. C. The dark gray area (C) is the probability of a false alarm, i.e. an erroneous classification as a correct match. The light gray area is the probability of a correct reject, i.e. a correct classification as a non-match.

Deciding what threshold to use should be based on the relative significance we set for the different errors. If we can tolerate misses exactly the same as we tolerate false alarms, then the threshold will be positioned exactly in the middle of the two pdfs, being the optimal threshold for separation between the two groups that minimizes the sum of errors. If, on the other hand, we cannot tolerate false alarms, we should move the threshold to the left, until the false alarm area is minimal, unavoidably increasing the probability of misses. The effect of shifting the threshold is shown in Fig. 46.

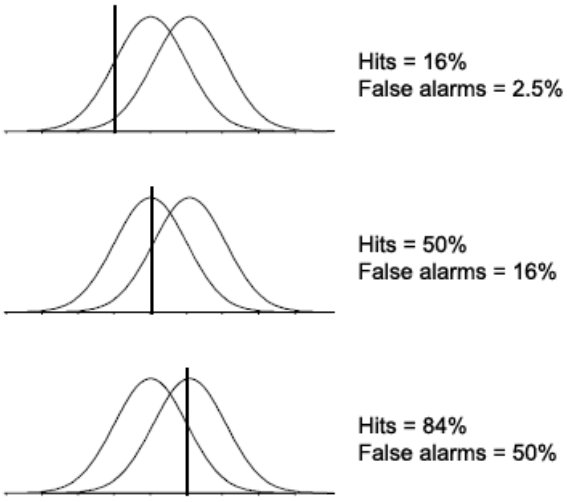


Fig. 46. Effect of shifting the threshold. In this example, the two pdfs are not well separated so reducing one type of error, unavoidably increases the other type. Moving the threshold to the left reduces the rate of false alarms to only 2.5% (top), but the hits rate is also low (16%) resulting in a high rate of misses (100-16% = 84%). Such poorly separated pdfs are expected for small length of the matched fracture lines, as shown below in our results.

Minimizing the errors: Looking at the correct matches pdf we define the two possible classifications:

1. A correct classification of a pair of shape as a match - its probability is the area **A** (left of the threshold).

2. An erroneous classification of a match as a non-match – its probability is the area **B** (right of the threshold).

Since the total area under a pdf curve is 1, we relate A and B:

B = 1-A

For the non-matches population, erroneous classification of a non-match as a match has probability **C** - the area left of the threshold.

Given no preference of one error over the other, the optimal threshold *T* will minimize the sum of two errors:

$$T = \arg\min_{matching\_error}(B + C) = \arg\min_{matching\_error}((1 - A) + C) =$$
$$= \arg\min_{matching\_error}(-A + C) = \arg\max_{matching\_error}(A - C)$$

The optimal threshold in our case was found to have a value of 2.71 (Fig. 47).

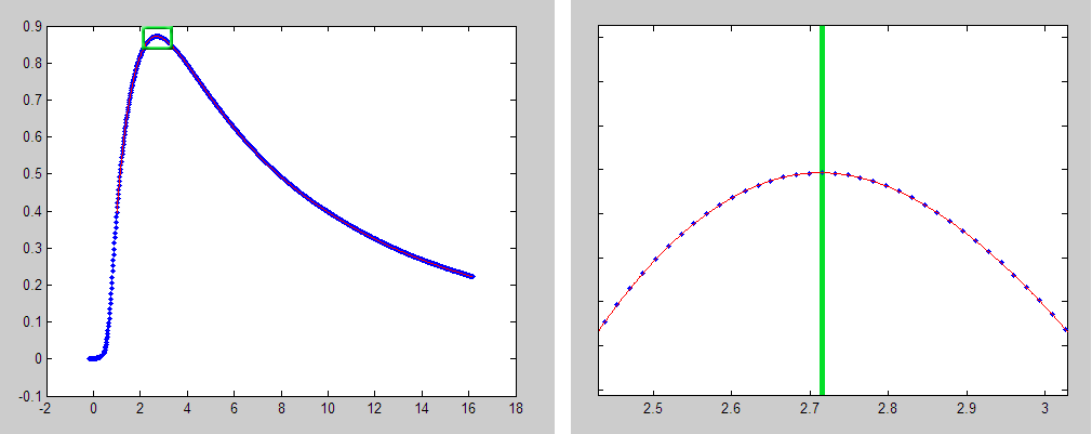


Fig. 47. Finding the optimal threshold. *Left*: the function (A-C) to be maximized. The area around the maximum is marked by a green rectangle. *Right*: zoomed in on the area of the maximum. Maximal value is found at threshold value of 2.71 (vertical green bar).

2.6.4. Finding the probability of a new shape - examples

The example (Fig. 48) shows a target accidental (in red) that was compared to the whole CONTOURS data set of ~8,900 accidentals. The 3 top matches are shown, before alignment (left; each accidental has a different orientation), and after alignment (right). The dissimilarity value (distance) is presented (title of right plots). Note that dissimilarity values lower than the threshold (of 2.7) mean that the two shapes are similar enough to be considered the same. For the target accidental we also present the dissimilarity values of all comparisons, in increasing order (Fig. 49). Another example is of a larger shape (Fig. 50) that turns out to be rarer.

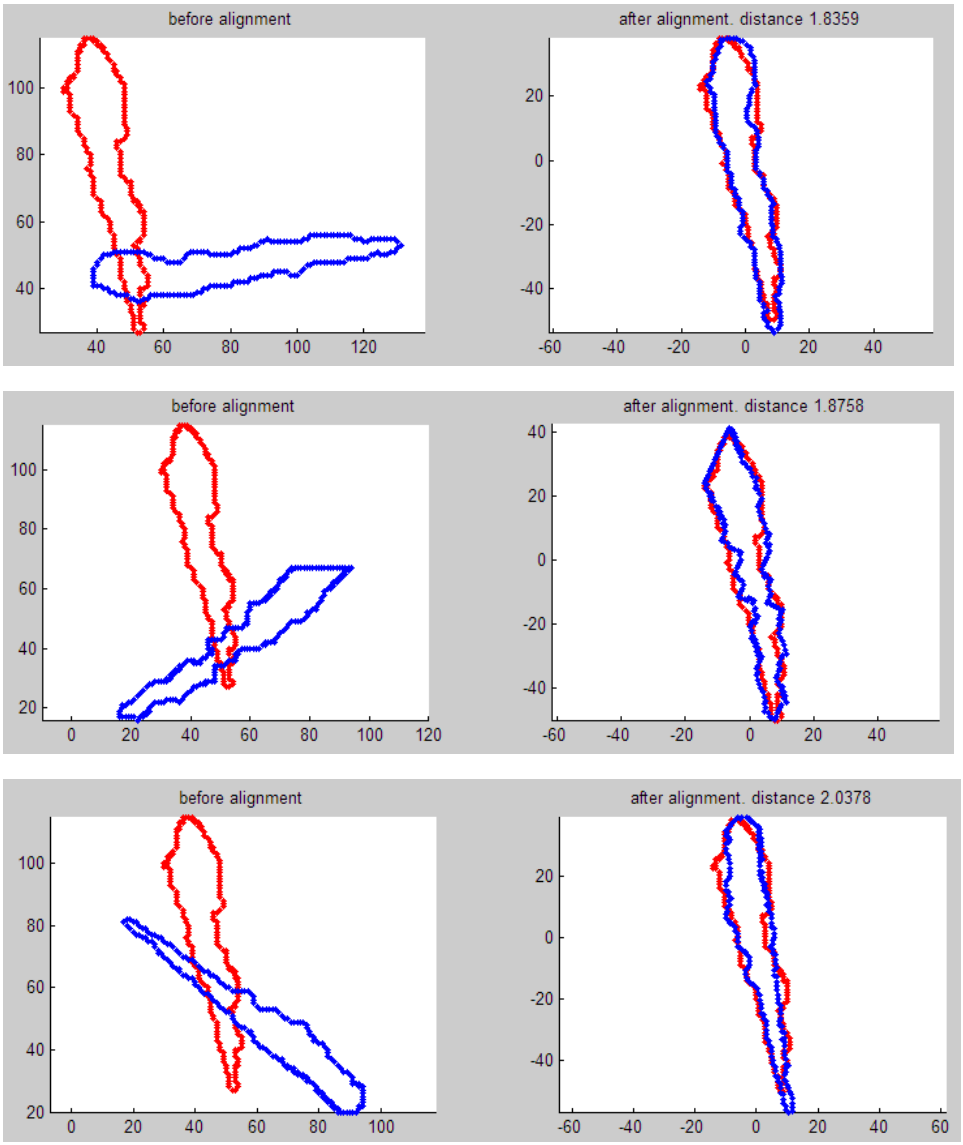


Fig. 48. Finding matches. Target shown in red, candidate match in blue. *Left*: the two shapes superimposed in their original orientation. *Right*: The two shapes after alignment.

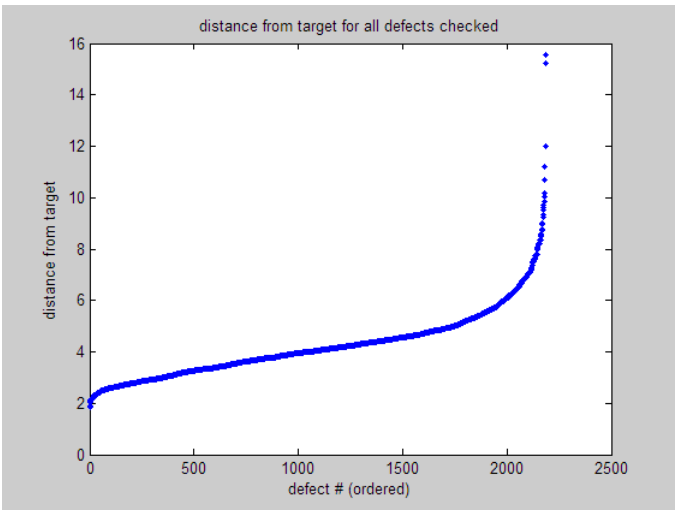


Fig. 49. Dissimilarity values of 2,200 comparisons (out of ~8,900) for the accidental in Fig. 47, in increasing order. There were 161 out of ~8,900 accidentals with distance below 2.7. The probability of finding a defect with this shape is therefore ~ 1/55.

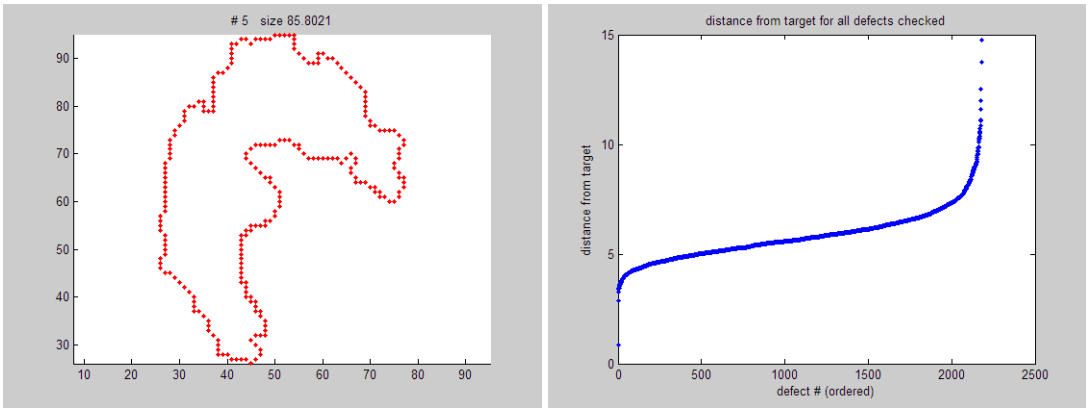


Fig. 50. A target accidental. Dissimilarity values of 2180 comparisons (out of ~8,900) for, in increasing order. There was only 1 out of ~8,900 accidentals with distance below 2.7. The probability of finding a defect with this shape is therefore ~ 1/8,900.

2.6.5. Summary of the shape model

A. The distribution of shapes is not estimated directly since shape is a high dimensional attribute. Instead, we use the one-dimensional results of shape comparisons to estimate the distributions of the correct matches (pairs of shapes originating from the same accidental) and of the non-matches (pairs of shapes originating from different accidentals).
B. We found an optimal threshold between the two distributions. A new pair of shapes can be compared and classified as originating from the same accidental (dissimilarity value < threshold) or from different accidentals (dissimilarity value > threshold).
C. The rarity of a new shape is found by matching it to the whole CONTOURS dataset (n = ~8,900) and counting how many comparisons yielded similarity values below the threshold.
D. The statistical shape model is implemented and used in the deliverable CompareAccidentals tool.

Using the software tools delivered by this project, the process of finding the probability of shape of a new accidental is:

1. Mark the contour of the accidental. (A semi-automatic operation in MarkAccidentals)
2. Mark the shoe aligned coordinate system. (A manual operation in CompareAccidentals)
3. Compare the shape of the accidental to the whole CONTOURS dataset. (An automatic operation in CompareAccidentals)
4. Calculate the probability of finding this shape. (An automatic operation in CompareAccidentals)

3. Conclusions

3.1. Discussion of findings

The problem of calculating the results of pattern comparison in a scientific, mathematically sound based manner has disturbed many forensic science practitioners and scientists. The 2009 NAS report focused on the lack of scientific procedure to evaluate the comparison results, or the "numerical" standard to move from one level of certainty to another.

This project solves, partially, the problem for shoeprints and could lead the way for other pattern comparison fields as well.

The methodology developed simultaneously with the advance of this project, leads today to the existence of a large database and algorithms to calculate the chance to get another accidental, similar (under restrictions) to the examined one.

The main findings and achievements

1. Development of software for semi-automatic marking of accidental contours. The initial contour marking is done automatically by the computer program, and the human operator further improves the marking.

2. Assembly of a large digital database of accidentals. The existing databases in the world were composed of actual photographs of accidentals and not their digital representation. Now, our database and software tools are accessible to the shoe expert community, allowing the computation of actual probabilities concerning the chance to find a similar accidental.

3. Development of a statistical model of the rarity of acquired accidentals and their combinations.

4. Development of software tools to assist the shoe experts in assessing the rarity of accidentals. The tools enable marking of new accidentals in a consistent way, and comparing them to a large database. The tools contain the following:

a) Developing a shoe aligned coordinate system that is universal – allows the accumulation of accidentals from many shoes onto one coordinate system.

b) Describing the distribution of locations of accidentals. Developing the notion of accumulated contact areas.

5. Presenting the problem of variability of the test impressions in a measurable way. One instance of an accidental is bound to be slightly different from another instance. We estimated the amount of dissimilarity expected to be found between multiple instances of the same accidental. This phenomenon, known for years, was not treated quantitatively before.

6. We discovered that the locations of the accidentals are, to a first approximation, evenly distributed on the sole of the shoe, when they are normalized and superimposed onto a universal shoe sole.

7. We discovered the even distribution of orientations of accidentals on the shoe sole. In contrast to several researches done on smaller populations the distribution is uniform for most shapes. A deviation from uniformity appeared only for the most elongated shapes which had higher probability of occurrence in the direction of the shoe's long axis.

8. We described a method for estimating shape distribution by using the notion of shape dissimilarity. First, a consistent way of measuring shape dissimilarity was developed. Second, we represented the distributions for two populations: shapes from the same source (correct matches) and shapes from different sources (non-correct matches). The overlap between the populations was used to calculate the error in measuring shape. This was used to estimate the probability of finding a similar shape to a target shape.

3.2. Implications of policy and practice

1. The development of practical tools for the shoeprints experts. The tools enable the experts to state their findings in a scientific based way, thus improving the quality, both from the scientific aspect and consistency of the expert opinion.

2. With the SESA project, experts around the world can start answering more scientifically in courts; Establishment of standards for moving towards quantitative methods in comparison of accidentals.

3. Today there is a potential for international cooperation in fortifying the databases of accidentals. The proficiency tests conducted from this day on could rely on the calculated value for each shoeprint (according its accidentals) in order to get to more uniform answer from all participants.

3.3. Implications for further research

1. The whole treatment in this project was on the level of test impressions. Dealing with accidentals on crime scene prints will require additional research and development. We identified the following critical items:
   a. The shapes of accidentals on crime scene's shoeprints are generally distorted to some degree and with much background noise of varied degree, which obscure or partially hide the print. Is it possible to evaluate the probability of an accidental (or its features) from real cases despite the complex and sometimes very significant noise?
   b. What data should be used when trying to model the statistics of real cases? The problem is to get enough data of cases that definitely came from the same origin, with both the crime scene and test impression accidentals in order to build the graph of correct matches between the accidentals as they appear on crime scene prints and on test impressions.
4. The statistical models that were developed and used in this project are based on assumptions that should be further checked. These include:
   a. The assumption that the accidental's features (location, orientation and shape) are independent. The joint distribution should be examined to find whether the features are independent, conditionally independent or dependent. Developing appropriate treatment should be derived according to the results.
   b. The underlying assumption that we can superimpose the locations of accidentals from several shoes together on one universal coordinate system must be further checked and verified.
   c. The underlying assumption that the statistics of accidentals do not depend (heavily) on the model of the shoe, its age or wear pattern, and can be aggregated together must be further checked and verified.
5. We presented a systematic approach for estimating the rarity of accidentals, but did not present an answer to the question of probability of identification - what is the probability that a specific accidental found on a crime scene shoeprint and another

accidental found on a test impression of a suspects' shoe comes from the same origin. We did not tackle this question for two reasons. First, dealing with crime scene evidence at that level was out of the scope of this project. Second, even for the comparison of test impressions accidentals, the way to determine the probability of identification is not obvious. The proper statistical framework should be chosen and used to develop a robust estimate of probability of identification.

4. Dissemination of research findings

We presented the results of this research project at several conferences:

An oral presentation at the ENFSI marks meeting in Lausanne (Switzerland), September 2011 [39,40]

An oral presentation at the 2012 IPES in Clearwater (FL), August 2012.

The ENFSI marks meeting in Bled (Slovenia), June 2013. The presentation included lectures presenting the developed methodology and the results of the project and a 3 hour workshop demonstrating how to work with the two computer programs. A DVD with the software was distributed to all the delegations [41, 42, 43].

A poster presentation at the 2015 IPTES in San Antonio (TX), August 2015.

An oral presentation at 2015 SAMSI forensic workshop in Durham – Rally (NC), September 2015.

5. Contributors

Students who contributed to this project (all from the Computer Science Department, Hadassah Academic College, Jerusalem):

M.Sc. project

Matthew Tovbin submitted his final project report titled "Shoeprint image feature extraction" in December 2011.

B.Sc. projects

Orit Shimon and Shir David submitted their final project report titled "*A system for marking shoeprint accidentals*" in September 2011.

Margarita Nikonova and Nura Ibrahem submitted their final project report titled "*A system for marking shoeprint accidentals – improvements and developments*" in August 2012.
The workers who scanned the images and marked the contours of the accidentals: Shir David, Shira Blitz, Tali Atzilov, Evelina Zaslavski and Osnat Cohen.

6. Acknowledgments

The authors would like to thank chief superintendent Eliot Springer (ret.), formerly the scientific deputy for the head of the Israel DIFS for his great efforts in getting the project on the move and assisting its progress.

7. References

1. Saks, M.J.; Koehler, J.J. The coming paradigm shift in forensic identification science. *Science*, **2005**, 309, 892-895.
2. Daubert vs. Merrel Dow Pharmaceuticals, INC. Supreme Court of the United States, **June 1993**.
3. National Research Council, Committee on Identifying the Needs of the Forensic Science Community, "Strengthening Forensic Science in the United States: A Path Forward", National Academies Press, Washington, DC, **2009**.
4. Petraco, D.K.N.; Gambino, C.; Kubic, A.T.; Olivio, D.; Petraco N. Statistical Discrimination of Footwear: A Method for the Comparison of Accidentals on Shoe Outsoles Inspired by Facial Recognition Techniques. *J Forensic Sci*. **2010,** 55(1), 34-41.
5. Hilderbrand, D.S. *Footwear, the missed evidence*, 2nd ed.; Staggs Publishing: Wildomar, CA, **2007.**
6. Keereweer, I. Guideline for Drawing Conclusions Regarding Shoeprint Examinations. Information Bulletin for Shoeprint/ Toolmark Examiners **2000**, 6, 47-62.
7. Keereweer, I.; van Beest, M.; van de Velde, J. M. *Guideline for Evaluating and Drawing Conclusions in Comparative Examination of Shoeprints*. Netherlands Forensic Institute: Netherlands, **2005**.
8. Keereweer, I. Footwear and foot impressions: comparison and identification. *encyclopedia of forensic science*, Wiley, **2009**.

9. ENFSI expert working group, Marks conclusion scale committee. Conclusion scale for shoeprint and toolmarks examination. *Journal of Forensic Identification,* **2006**, 56(2), 255-280.

10. Majamaa, H.; Ytti, A. Survey of the Conclusions Drawn of Similar Footwear Cases in Various Crime Laboratories. *For. Sci. Int*. **1996**, 82(1), 109-120.

11. Ytti, A.; Majamaa, H.; Virtanen, J. Survey of the Conclusions Drawn of Similar Shoeprint Cases, Part II. *Proceedings of the European Meeting for Shoeprint and Toolmark Examiners,* The Hague, Netherlands, **1997**, 157-169.

12. Shor, Y.; Weisner, S. A Survey on the Conclusions Drawn on the Same Footwear Marks Obtained in Actual Cases by Several Experts Throughout the World. *J. For. Sci.* **1999**, 44(2), 380-384.

13. Drews, F.; Zander, H. Toolmarks Comparison Identification with Methods of the Fuzzy Set Theory and the Image Processing. *Presentation at the 4th ENFSI SP/TM meeting,* Berlin, May **2001**.

14. Katterwe H.W.; Braune, M.; Ahlhorn, T.G. Image Processing Strategy and Automatic Comparison of Marks. *Presentation at the 4th ENFSI SP/TM meeting,* Berlin, May **2001**.

15. Lisounkin, A. Toolmarks Comparison and Identification with Pattern Recognition Methods based on 3D Surface Measurements. *Presentation at the 4th ENFSI SP/TM meeting,* Berlin, May **2001**.

16. Schreck, G. Computerized Comparison of Toolmarks by the PAMIR System. *Presentation at the 4th ENFSI SP/TM meeting,* Berlin, May **2001**.

17. Alexander, A.; Bouridane, A.; Crookes, D. Automatic classification and recognition of shoeprints. *Seventh International Conference on Image Processing and Its Applications* (Conf. Publ. No. 465) Vol. 2, **1999**, 638 – 641.

18. Bouridane, A.; Alexander, A.; Nibouche, M.; Crookes, D. Application of fractals to the detection and classification of shoeprints. *2000 International Conference on Image Processing,*Vol. 1, **2000**, 474 – 477.

19. Huynh, C.; de Chazal, P.; McErlean, D.; Reilly, R.B.; Hannigan, T.J.; Fleury, L.M. Automatic classification of shoeprints for use in forensic science based on the Fourier transform. *2003 International Conference on Image Processing*. Vol. 3, 14-17 Sept. **2003,** III - 569-572.

20. de Chazal, P.; Flynn, J.; Reilly, R.B. Automated processing of shoeprint images based on the Fourier transform for use in forensic science. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **2005**, 27(3), 341–350.

21. Pavlou, M. Allinson, N.M. Automatic Extraction and Classification of Footwear Patterns. *Lecture Notes in Computer Science, Intelligent Data Engineering and Automated Learning – IDEAL* **2006**.

22. Gueham, M.; Bouridane, A.; Crookes, D. Automatic Recognition of Partial Shoeprints Based on Phase-Only Correlation. *IEEE International Conference on Image Processing*, Sept. 16 2007-Oct. 19 **2007**, 4, IV - 441– 444.

23. Su, H., Crookes, D., Bouridane, A. Thresholding of noisy shoeprint images based on pixel context. *Pattern Recogn. Lett.* **2007**, 28(2), 301-307.

24. Su, H.; Crookes, D.; Bouridane, A.; Gueham, M. Shoeprint Image Retrieval Based on Local Image Features. *Third International Symposium on Information Assurance and Security*, 29-31 Aug. **2007,** 387 – 392.

25. Su, H.J.; Crookes, D.; Bouridane, A.; Gueham, M. Local Image Features for Shoeprint Image Retrieval. *British Machine Vision Conference* BMVC07, September **2007**, 10-13.

26. Tang Y.; Srihari S.N.; Kasiviswanathan H. Similarity and Clustering of Footwear Prints, San Jose, CA, August 15-16, **2010**.

27. Tang Y.; Srihari S.; Kasiviswanathan H.; Corso J. Footwear Print Retrieval System for Real Crime Scene Marks, Tokyo, Japan, Nov 11-12, **2010**.

28. Tang Y.; Kasiviswanathan H.; Srihari S.N. An efficient clustering-based retrieval framework for real crime scene footwear marks, *Int J Granular Computing Rough Sets and Intelligent Systems*, **2012**, 2(4), 327-360.

29. Srihari S.N. Analysis of Footwear Impression Evidence, *US DoJ Report*, March **2011**.

30. Bodziak, W.J. Footwear Impression Evidence, Detection Recovery and Examination, 2nd Ed.; CRC Press. **2000**, 335..

31. Stone, R.S. Footwear examination: mathematical probabilities of theoretical individual characteristics, *J Forensic Ident*. **2006**, 56(4), 577-599.

32. Botev, Z.I.; Grotowski, J.F.; Kroese, D.P. Kernel density estimation via diffusion. *The Annals of Statistics*, **2010**, 38(5), 2916-2957.

33. Jackson, J.E. A. User's Guide to Principal Components, *Wiley Series in probability and Statistics*, Wiley. **2005**.

34. Reymond J. Study of acquired characteristics on two general patterns of soles, *école des sciences criminelles Université de Lausanne travail de master en sciences forensiques, mention identification*, June **2010.**

35. Kunttu, I.; Lepisto, L. Shape-based retrieval of industrial surface defects using angular radius Fourier descriptor. *Image Processing, IET*, **2007**, 1(2), 231,236, doi: 10.1049/iet-ipr:20060113.

36. Srihari S.N. Computational Methods for Handwritten Questioned Document Examination, *US DoJ Report*, December **2010**.

37. Zhang, Z. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, **1994**, 13(12), 119–152. doi:10.1007/BF01427149.

38. Marie-Pierre Dubuisson and Anil K. Jain, A Modified Hausdorff Distance for Object Matching. *Proc. International Conference on Pattern Recognition*, Jerusalem, Israel, pp 566–568, **1994**.

39. Wiesner S, Shor Y, Tsach T and Yekutieli Y. (**2011**) Computerized system for aiding the expert in evaluating the degree of certainty in 2D shoeprints.*9th ENFSI SPTM 2011 meeting*. Lausanne, Switzerland.

40. Shor Y, Wiesner S, Tsach T and Yekutieli Y. (**2011**) The way to present degree of certainty. *9th ENFSI SPTM 2011 meeting*. Lausanne, Switzerland.

41. Wiesner S, Shor Y, Tsach T and Yekutieli Y. (**2013**) Statistic Evaluation of Shoeprint Accidentals (SESA) aids experts in evaluating the degree of certainty in 2D shoeprints.*10th ENFSI SPTM 2013 meeting*. Bled, Slovenia.

42. Shor Y, Wiesner S, Tsach T and Yekutieli Y. (**2013**) A methodological shift in the evaluation of shoeprints: present and future. *10th ENFSI SPTM 2013 meeting*. Bled, Slovenia.

43. Yekutieli Y, Wiesner S, Shor Y and TSach T. (**2013**) SESA Workshop: Why pay tomorrow for what you can get today for free? *10th ENFSI SPTM 2013 meeting*. Bled, Slovenia.