

Sexing Caucasian 2D footprints using convolutional neural networks

Marcin Budka^{1¶}, Matthew R. Bennet^{2¶*}, Sally Reynolds³ Shelby Barefoot² Sarah Reel^{4&}, Selina Reidy^{5&}, Jeremy Walker^{6&}

¹Department of Computing and Informatics, Bournemouth University, Poole, BH12 5BB, United Kingdom.

²Department of Environmental and Life Sciences, Bournemouth University, Poole, BH12 5BB, United Kingdom.

³Department of Archaeology and Anthropology, Bournemouth University, Poole, BH12 5BB, United Kingdom.

⁴Division of Podiatry and Clinical Sciences, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, United Kingdom.

⁵Identification Bureau, Yorkshire and the Humber Regional Scientific Support Services, Sir Alec Jeffreys Building, Peel Avenue, Calder Park, Wakefield, WF2 7UA, United Kingdom.

⁶Sheffield Teaching Hospital, NHS Trust, Podiatry Services, Woodhouse Clinic, 3 Skelton Lane, Sheffield, S13 7LY, United Kingdom.

* Corresponding author

Email: mbennett@bournemouth.ac.uk

¶ These authors contributed equally to this work.

& These authors contributed equally to this work.

Abstract

Footprints are left, or obtained, in a variety of scenarios from crime scenes to anthropological investigations. Determining the sex of a footprint can be useful in screening such impressions and attempts have been made to do so using single or multi landmark distances, shape analyses and via the density of friction ridges. Here we explore the relative importance of different components in sexing two-dimensional foot impressions namely, size, shape and texture. We use a machine learning approach and compare this to more traditional methods of discrimination. Two datasets are used, a pilot data set collected from students at Bournemouth University (N=196) and a larger data set collected by podiatrists at Sheffield NHS Teaching Hospital (N=2677). Our convolutional neural network can sex a footprint with accuracy of around 90% on a test set of N=267 footprint images using all image components, which is better than an expert can achieve. However, the quality of the impressions impacts on this success rate, but the results are promising and in time it may be possible to create an automated screening algorithm in which practitioners of whatever sort (medical or forensic) can obtain a first order sexing of a two-dimensional footprint.

Introduction

Within the forensic, podiatry and anthropological literature there have been several attempts to determine the sex of bare two-dimensional (2D) human footprints [1]. The aims and justification for these studies are varied and range from basic medical and anthropological description, via victim profiling as an aid in disaster identification, to use in criminal forensic casework [2]. In the latter context, 2D footprints are often found at crime scenes where bodily fluids have been tracked around a scene via bare feet or those encased in socks. Most forensic practitioners engaged in the preparation of judicial evidence, and/or its review, would currently refrain from determining footprint sex. This is due primarily to a lack of reliable tools for sex determination and the risk that false intelligence might be provided. If a tool could be shown to be reliable and universal, then this might change. The application of machine learning may provide such a tool and this forms the aim and focus of this paper [3]. This paper therefore focuses on an application of machine learning via the use of convoluted neural networks (CNN) rather than advancing the development and fine tuning of such algorithms and it is the application which is both novel and of interest here.

Our ability to sex a footprint is a function of the degree of sexual dimorphism in our species. Most species in the animal kingdom have larger females than males with the exception of mammals and birds [4]. Variation in dimorphism with body size follows the so-called Rensch's rule [5] in which taxa with larger males show greater dimorphism (hyperallometry). Human sexual dimorphism is best illustrated by stature within males who are typically 7% taller, and since stature and foot length correlate well, as indicated by the numerous empirical relationships [1], it should also be manifest in footprint length [6]. Dimorphism emerges primarily during postnatal growth, with male neonates only 1% longer than females at birth [7]. During childhood, dimorphism remains relatively minor until the onset of puberty when it becomes established. Sexing a footprint is therefore restricted to adult cases. In a forensic context this is complicated by variation between and within populations. Rodriguez et al. (2005) suggests that dimorphism is a function of phylogeny and selection pressures such as marriage systems [8][9][10][11], social stratification [10], sexual division of labour [11] and potentially nutritional standards [9,12]. Variation in stature with climate, as predicted by Bergmann's rule [13] has previously received some empirical support (e.g., [14,15]) and could potentially be a source of variance. The net result of this is that empirical relationships are likely to be subject to inter-population variance and when coupled with the modest (i.e., 7%) level of sexual dimorphism in *Homo sapiens* will limit the degree of discrimination that is possible.

The sexing of footprints has primarily proceeded by comparison of linear dimensions (e.g., [16–21]) and to a lesser extent via shape metrics (e.g., [22,23]). There is a known sex difference in the density of friction ridges on human fingers [24] which has been translated to feet in recent studies (e.g., [25–27]). Differences have also been observed in skeletal elements namely the tarsal and calcaneus bones (e.g., [3,28–34]). Here we develop a machine learning approach which looks at the entirety of a footprint in an attempt to determine its sex. We then partition discrimination between surface texture, inclusive of frictions ridges, and overall morphology.

Methods and materials

Data sources

A pilot study was first completed using data collected from student volunteers at Bournemouth University in the autumn of 2018. Volunteers left a standing footprint using an inkless pad system and were asked to confirm the sex they were registered as at birth. A total of 101 males and 132 female right footprints were obtained from adults between the ages of 18 and 25 years old and predominantly, although not exclusively white Caucasian. Difference in footprint size between traces left during walking (so-called dynamic traces) and those left in static tests are also well documented [35] and moreover one might expect a difference in the quality of friction ridges left between dynamic and static traces. In this initial study static traces were used for ease of execution, but clearly this something that should be factored into future research as discussed later. Footprint impressions were scanned and saved anonymously for analysis in accordance with ethical approval obtained from Bournemouth University (ID: 22317) and informed consent of the volunteers. A series of digital landmarks [1] identifying the main dimensional properties of the foot were placed digitally by an operator on each footprint image using DigTrace (www.digtrace.ac.uk) (Fig. 1a). Landmarks were exported as coordinates, and these were used to compute linear dimensions. A script was written to place 10 mm² sampling squares between selected landmarks and used to crop the image into a series of individual image fragments. The black and white pixel proportions were calculated for each of the squares and the ridge density was counted manually and a 10% sample verified for accuracy. The geometric morphometric analysis and discriminant analyses conducted on the landmark data were undertaken in the statistical software package PAST [36].

The author received additional data in the form of 2677 footprint images provided by Jeremy Walker of Sheffield NHS Trust, hereby referred to as the Walker-Data. This data was received in a fully anonymised state and subject's provided informed ethical consent for their anonymised data to be used in research. The original ethical approval was provided by the Sheffield NHS Trust. The data consisted of 2677 binary .tif images with 2240 x 3200 resolution at 200 ppi, of either a right or left print for 1483 females and 1194 males, aged between 16 and 81 with over 97.5% being Caucasians. Landmarks were placed digitally by an operator using the freeware DigTrace (www.DigTrace.co.uk) on a training

set of 200 randomly sampled from the 2677 population. This training set was used to train an automated landmark placement algorithm and applied to the remaining population. Sampling squares were obtained in the same way as for the pilot project however ridge density was not manually counted. A total of 20% of the images were selected at random and isolated for validation purposes. The images were cropped to a bounding box to eliminate background noise which has the potential to leak information. In order to reduce the computational requirements, the images were resized to 512x640. This aspect ratio is different from the original images as it better reflects the shape of the bounding boxes in the dataset; the aspect ratio of the footprints themselves was not altered. The sampling algorithm impacts on the information encoded within each image. The common default of BILINEAR interpolation results in loss of ridge detail but makes the texture appear softer and therefore more representative of a photographic image. From the other standard image down-sampling techniques, NEAREST neighbour preserves the binary, black and white, nature of the original images while HAMMING produces a sharper image than BILINEAR and does not have the dislocations at a local level that BOX does (Fig. 2). As a result of sensitivity testing and with the aim of maximising the information available, NEAREST neighbour, BILINEAR interpolation and HAMMING were the three channels selected of the same input image (Fig. 3a). An alternative texture-free version of the image as shown in Figure 3b, was created using erosion followed by dilation (fundamental morphological image processing operations) for 5 iterations each, with a 3 x 3 kernel. All the footprint images and associated metadata are available for the purposes of replication from Bournemouth University's Data Repository (<https://doi.org/10.18746/bmth.data.00000157>).

Machine learning algorithms

Over the past eight years there has been a rapid increase in applications of Convolution Neural Network (CNN) to a range of applications ranging from medical imaging [e.g., 37,38] to the analysis of forensic footwear impressions [39] and even the detection of building defects [40] to name just a few.

In order to build a predictive model for sex estimation (and age as a by-product), we experimented with several CNN architectures, namely ResNet34, ResNet50 and ResNet101 [41], pre-trained on the ImageNet2012 dataset. The input images have been resized to 512x640 pixels as discussed in the previous section. We have built a separate model for the following three tasks: (1) sex estimation; (2)

age estimation; and (3) combined sex and age estimation. For task (1) the network had a single output neuron producing an estimated probability of the print being female via the sigmoid function, and we have used the binary-cross entropy loss:

$$\mathcal{L}_c = -(y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c)) \quad (1)$$

where $y_c \in \{0, 1\}$ denotes the label and $\hat{y}_c \in (0, 1)$ is the network output.

For task (2) the network also had a single output neuron, this time producing an unconstrained age estimate, and was trained using the L1 loss:

$$\mathcal{L}_r = |y_r - \hat{y}_r| \quad (2)$$

where y_r is the actual age and \hat{y}_r is the network output or prediction.

For task (3) there were two outputs and we have used a weighted sum of the above two loss functions:

$$\mathcal{L} = \mathcal{L}_r + \lambda \cdot \mathcal{L}_c \quad (3)$$

where λ has been empirically set to 20.

For each task, we have replaced the final fully-connected layer from the pre-trained network with a custom head consisting of two linear layers with ReLU [42] non-linearity in-between, and used dropout [43] and batch normalisation [44]. We first trained the head only keeping the rest of the network fixed for 10 epochs using the Adam optimiser [45], and subsequently fine-tuned the whole network for another 10 epochs. One of the initial challenges of this model was excluding non-pertinent information. Of note, in an initial run of the model using the Bournemouth-Data it found a way of effectively cheating. The slip with the print code and sex was left attached initially with the whole image being used in its raw state. Figure 4 illustrates how the model quickly identified the significance of the tick placement denoting the sex of the participant.

In terms of evaluating the model there are a range of approaches. Cross-validation has been the gold standard in machine learning for several years [44,46]. This involves splitting the data into K approximately equal parts, training the model on K-1 parts and testing on the remaining part. This is then repeated K times always leaving a different part for testing. However, since the resurgence of deep neural networks [47] this has changed in favour of a hold-out method in which a single random split, usually 80-20, is used [41]. This is the approach we have used in the current work. One of the issues

cross-validation aimed to address was the problem of small datasets, where a hold-out set would be unlikely to be representative of the data distribution as a whole. With more than 2600 examples in the Walker-Data, with all being bare footprints collected in exactly the same way, this risk is reduced. In addition, the computational cost of training K models rather than 1 for each experimental setup is prohibitive for most modern CNN architectures. Finally, the vast majority of seminal neural network papers published in the last 7-8 years (e.g., [41,43,44]) do not use cross-validation for the similar reasons to those outlined above. We do acknowledge however that if our work were to be used in jurisprudence, more rigorous evaluation would be required.

Results

Conventional landmark analysis

Using both the Bournemouth and Walker data and the placed landmarks (Fig. 1) conventional geometric morphometric analyses were undertaken [1]. Figure 5 shows the distribution of landmarks in the dataset following a Generalised Procrustes Analysis (GPA) and the thin-plate spline for the male compared to female landmark means for the Bournemouth-Data. Less than 40% of the landmark variation is accounted for by the geometry of the longitudinal media arch, with arch being slightly less well-defined in the female footprints. Linear Discriminant Analysis (LDA) gives an initial sex discrimination of 76.92%, which when jack-knifed falls to 65.81% (Table 1). For the Walker-Data the results are similar 70.95% falling to 69.34% when jack-knifed (Table 2). Using the 153 possible inter-landmark distances without any GPA and therefore incorporating aspects of both shape and size the LDA gives an initial sex discrimination of 98.71%; reduced to 67.38% when jack-knifed (Table 1).

Friction ridges

Using the Bournemouth-Data friction ridges were used to examine their ability to sex the 2D footprints. Sample squares were initially cropped from the images automatically between equidistant selected landmarks (Fig 1a). The percentage of black versus white was computed automatically for each square using a high contrast version. In theory the greater the percentage of black the greater the ridge density should be. This resulted in a discrimination of only 59% falling to 53% when jack-knifed. Discrimination

values are reported in Table 1. The results are relatively poor which could be a function of the callus build in some areas of the foot, or the poor quality of some of the footprint images in areas of maximum plantar pressure. A second set of sample squares were extracted to avoid 'high contact' areas of the foot (Fig. 1b) and the analysis was repeated. The results obtained from the automated black-white percentage were better but still relatively poor. Ridge counting for each sampled image square followed the procedure of Krishan et al. [25] and [26] was then used as a further test. Using the data for 168 participants, 29 footprints were excluded due to being faint, an average total ridge count for females were obtained of 52 versus 47 for males across the three squares. Furthermore, the average ridge count of each individual square was higher for females than males. A discriminant analysis gave a 67% success rate (Table 1).

Machine learning

Table presents results for the five possible combinations of data input, namely: (1) Shape + Texture + Size; (2) Shape + Texture; (3) Shape + Size; (4) Shape only; and (5) Texture only. Inclusion of size is only possible in conjunction with shape and the texture only version is based on the sampling of squares from the footprint as shown in Figure 1. The aim here was to explore how important different elements of the impression were in making the classification. Each of these five inputs has been used for three different tasks: sex estimation, age estimation model, and combined sex and age estimation. This analysis was run on the Walker-Data being the larger of the data sets and using ResNet101 as it consequently outperformed the other variants tested. For all the experiments a random, 80/10/10 training/validation/test split of the dataset was used, and the metrics reported in the table apply to the test set.

The maximum discrimination that can be obtained is just under 90% with all elements (size, shape, and texture) used in the analysis; that is the entire image. Interestingly 83% is achieved for using texture alone but excluding it does not materially reduce the power of size or shape. Surprisingly, the best results are obtained when the model learns to estimate sex and age collectively. Although it might seem counterintuitive at first, it is possible that the richer feedback signal coming from two different sources instead of one helps the model converge on a better result. Therefore, improvement in one of the

subtasks does not necessarily compromise the performance in the other. The learnt intermediate representations of the patterns are probably also more semantically meaningful as the model is less likely to learn any idiosyncrasies associated with the individual subtasks. While we remain uncertain what the model is 'seeing' in the impression to estimate age, and therefore would treat this data with considerable caution, setting the model to estimate this as well helps improve the overall result.

The textured images within the Walker-Data footprints, scaled to preserve their relative size (Scenarios 1, 6 and 11, Table 2), were sexed correctly almost 90% of the time. Figure 6 shows the heat maps associated with the most successful classifications (i.e., the ones with the highest confidence). The model was also able to predict the age of the footprint-maker to a mean absolute error of 7.51 years. Repeating the analysis with the texture-free version of the footprints (Scenarios 2, 7 and 12, Table 2) reduces the classification accuracy only slightly to 89.13% and increases the age error to 8.79 years. This suggests that the majority of the discrimination is achieved via gross variation in size and shape of the foot images, and to a lesser extent by the texture associated with such things as ridge density. The proportion to which discrimination is partitioned between texture and gross morphology will vary with the resolution of the input images. For example, repeating the exercise using the Bournemouth-Data which has different image resolution gives a textured discrimination of 82.6% and a non-textured discrimination of 65%. The ridge detail is more prominent in the Bournemouth-Data and consequently appears to be doing a greater proportion of the work.

Comparison of the results for Scenarios 7 and 8 (i.e., size vs texture) reinforces this observation even further, replacing texture with size results in over 4% accuracy boost, although at the expense of the quality of age estimations. However, the Shape vs Texture comparison (Scenarios 9 and 10, Table 2) hints towards the latter having more discriminative power (83.89% vs 80.15%), once again at the expense of age estimation. The CNN outperforms conventional landmark-based analysis that involves shape only (Table 2, Scenario 16) but is equivalent to an inter-landmark analysis (shape + size; Table 2, Scenario 17). Manual landmark placement would make the CNN preferably due to the time saving, however it is worth noting that landmark placement was automated for the Walker-Data so in either case machine learning gives slightly better performance overall.

The heat maps provide an opportunity to examine which areas are most critical in determining the classification of a specific individual. Highlighted areas correspond to areas where a small change would

impact on the classification (Fig. 6). Essentially this allows us to determine which morphological areas are doing the most 'discriminative-work'. Two areas seem to be key one between the first and second toes and a second around the heel (Fig. 6). The latter makes sense since the width of the calcaneus/talus is an established measure of sexing the bones of the foot (e.g., [31,34,48,49]). The former is less easily interpreted, although features in the toe region (e.g., humps, phalange marks) have been identified as highly individual [21]. The shape in this region may also be influenced by gender based shoe choices[50].

Discussion and conclusion

This research demonstrates that accuracies of over 85% are possible when sexing footprints using a 'whole-foot' machine learning approach in which size, shape and texture are included in the model. This has been achieved with a relatively modest data library for training purposes, and improvement is no doubt possible with the availability of increased data volume. In the current model, the relative work done in achieving discrimination varies between the various image-components and is dependent on the image quality of the impressions used. The more texture (i.e., ridge detail) that the footprint images contain, the more that this aids sex discrimination. In a forensic application this is often obscured by socks and friction ridges may be obscured/clogged by blood or other fluid. Extraneous noise around the image may also be an issue. The current model should be considered as baseline only that can be expanded, and made more robust, by the addition of more training data across a range of potential applications. It shows potential however given more training data to improve its applicability.

Surprisingly, the model demonstrated some ability to age footprints as well, there is uncertainty however about what the model is actually 'seeing' unlike sexing, which seems to focus on the shape of the heel. Therefore, at this stage these results should be treated with considerable caution. However, setting the model to estimate age as well as sex helps improve the overall result. Feet stop growing following maturity therefore past this point, natural growth of the foot alone cannot be responsible for the machine's ability to age a footprint. However, changes to the feet may still occur in maturity due to systemic disease, musculoskeletal pathologies, and trauma. Furthermore, the skin can atrophy and become anhidrotic with age and pathology whilst plantar subcutaneous fat may also reduce. Whether the model is 'seeing' these changes remains uncertain. Irrespective of the validity of the age estimates

it is clear that improved sex discrimination is achieved by setting the machine learning model multiple tasks of which age is an obvious one.

In terms of applicability of this model and approach are varied provided that the model can be made robust enough to take a range of inputs. Forensic podiatry is an obvious future application. Currently forensic podiatry casework focuses on identifying similarity and dissimilarity between footprints left at a scene (questioned footprints) and sample footprints made by a suspect (reference footprints). This requires both questioned and reference footprints to be available for evidential comparison. In some circumstances, only the questioned footprint may be available, and its analysis may assist the investigation if accurate intelligence regarding the sex or age of potential suspects could be provided. In other scenarios however a first order assessment of sex may be appropriate for example in victim identification or rapid profiling to aid investigative direction.

As stated above the results look promising but a lot more training data of various types is needed to take the next step and produce something which can be deployed operationally. There are lots of issues to consider, of which the following are just a few. Most footprint examiners, at least in the UK, see partial marks so an operational model in the future will need to be able to give results on these. It should be possible in the short term to create 'partial prints' by cropping the existing training data and this is an obvious early win. However, the model has been trained with static prints, whereas most traces at crime scenes are dynamic prints (i.e., walking or running). Therefore 'real' forensic traces need to be obtained to train the model further. In addition, twisting and other forms of natural distortion all need to be considered. Recovered prints can also be presented as negatives and a future model will need to be able to invert these types of images for analysis. The current model works with either a right or a left foot and has yet to be applied to a trackway of prints. In theory sex determinations made on multiple tracks of the same person should increase the reliability of the conclusion. The final limitation of the current model is that both the data sets used are dominated by one racial group and an ethnically diverse training population would help. To amass the training data that is needed we need a community-based approach to this in which data is sourced from across the world in a variety of forms for training purposes. The sheer number of publications over the last decade which use two-dimensional print images is huge and illustrates the potential data that could be harnessed to develop a future app for

use by everyone. One final point worth stressing is that should the approach gather momentum with the addition of training data and before it is used in jurisprudence it would also be essential to explore different routes for validation, including perhaps the computationally intensive cross validation approached considered to be some as the gold standard.

1. Bennett MR, Morse SA. Human footprints: Fossilised locomotion? Human Footprints: Fossilised Locomotion? Springer Cham; 2014. doi:10.1007/978-3-319-08572-2
2. DiMaggio JA, Vernon W. Forensic Podiatry. Humana Press; 2011. doi:10.1007/978-1-61737-976-5_4
3. Navega D, Vicente R, Vieira DN, Ross AH, Cunha E. Sex estimation from the tarsal bones in a Portuguese sample: a machine learning approach. *Int J Legal Med.* 2015;129: 651–659. doi:10.1007/s00414-014-1070-5
4. Darwin C. *The Descent of Man, and Selection in Relation to Sex. The Descent of Man, and Selection in Relation to Sex.* Princeton University Press; 1871. doi:10.1038/011305a0
5. Rensch B. *Evolution Above the Species Level.* Columbia: Columbia University Press; 1959. doi:https://doi.org/10.7312/rens91062
6. Wells JCK. Sexual dimorphism of body composition. *Best Pract Res Clin Endocrinol Metab.* 2007;21: 415–430. doi:10.1016/j.beem.2007.04.007
7. Rodríguez G, Samper MP, Olivares JL, Ventura P, Moreno LA, Pérez-González JM. Skinfold measurements at birth: Sex and anthropometric influence. *Arch Dis Child Fetal Neonatal Ed.* 2005;90: 273–275. doi:10.1136/adc.2004.060723
8. Alexander, R.D., Hoogland, J.L., Howard, R.D., Noonan, K.M. and Sherman PW. Sexual dimorphisms and breeding systems in pinnipeds, ungulates, primates, and humans. *Evolutionary biology and human social behavior: An anthropological perspective.* 1979. pp. 402–435.
9. Gray JP, Wolfe LD. Height and sexual dimorphism of stature among human societies. *Am J Phys Anthropol.* 1980;53: 441–456. doi:https://doi.org/10.1002/ajpa.1330530314
10. Gaulin SJC, Boster JS. Human marriage systems and sexual dimorphism in stature. *Am J Phys Anthropol.* 1992;89: 467–475. doi:10.1002/ajpa.1330890408
11. Holden C, Mace R. Sexual dimorphism in stature and women's work: A phylogenetic cross-

- cultural analysis. *Am J Phys Anthropol.* 1999;110: 27–45.
doi:[https://doi.org/10.1002/\(SICI\)1096-8644\(199909\)110:1<27::AID-AJPA3>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1096-8644(199909)110:1<27::AID-AJPA3>3.0.CO;2-G)
12. Wolfe LD, Gray JP. A Cross-cultural Investigation Into the Sexual Dimorphism of Stature. Ember CR, editor. New Haven, Ct: Human Relations Area Files, Inc; 1982. Available: <https://hraf.yale.edu/ehc/documents/450>
 13. Bergmann KGLC. Über die Verhältnisse der wärmeökonomie der Thiere zu ihrer Grösse. *Göttinger Stud.* 1847;3: 595–708.
 14. Ruff CB. Morphological adaptation to climate in modern and fossil hominids. *Am J Phys Anthropol.* 1994;37: 65–107. doi:<https://doi.org/10.1002/ajpa.1330370605>
 15. Katzmarzyk PT, Leonard WR. Climatic influences on human body size and proportions: Ecological adaptations and secular trends. *Am J Phys Anthropol.* 1998;106: 483–503. doi:[https://doi.org/10.1002/\(SICI\)1096-8644\(199808\)106:4<483::AID-AJPA4>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1096-8644(199808)106:4<483::AID-AJPA4>3.0.CO;2-K)
 16. Zeybek G, Ergur I, Demiroglu Z. Stature and gender estimation using foot measurements. *Forensic Sci Int.* 2008;181: 54 e1–5. doi:10.1016/j.forsciint.2008.08.003
 17. Atamturk D. Estimation of Sex from the Dimensions of Foot, Footprints, and Shoe. *Anthropol Anzeiger.* 2010;68: 21–29. doi:10.2307/29543078
 18. Uhrová P, Beňuš R, Masnicová S, Neščáková E. Stature and sex estimate using foot dimensions. *Česká Antropol.* 2011;61: 32–35.
 19. Hemy N, Flavel A, Ishak NI, Franklin D. Sex estimation using anthropometry of feet and footprints in a Western Australian population. *Forensic Sci Int.* 2013;231: 402.e1-402.e6. doi:10.1016/j.forsciint.2013.05.029
 20. Jowaheer V, Agnihotri AK. Sex identification on the basis of hand and foot measurements in Indo-Mauritian population e A model based approach. 2011. doi:10.1016/j.jflm.2011.02.007
 21. Krishan K. Individualizing characteristics of footprints in Gujjars of North India--forensic aspects. *Forensic Sci Int.* 2007;169: 137–144. doi:10.1016/j.forsciint.2006.08.006
 22. Moudgil R, Kaur R, Menezes RG, Kanchan T, Garg RK. Foot index: Is it a tool for sex determination? *J Forensic Leg Med.* 2008;15: 223–226. doi:10.1016/j.jflm.2007.10.003
 23. Mukhra R, Krishan K, Kanchan T. Bare footprint metric analysis methods for comparison and identification in forensic examinations: A review of literature. 2018. doi:10.1016/j.jflm.2018.05.006

24. Champod C, Lennard CJ, Margot P, Stoilovic M. Fingerprints and other ridge skin impressions. Second Edi. CRC Press Inc; 2017.
25. Krishan K, Kanchan T, Pathania A, Sharma R, Dimaggio JA. Variability of footprint ridge density and its use in estimation of sex in forensic examinations. *Med Sci Law*. 2015;55: 248–290. doi:10.1177/0025802414557880
26. Kanchan T, Krishan K, Aparna KR, Shyamsunder S. Footprint ridge density: A new attribute for sexual dimorphism. *HOMO- J Comp Hum Biol*. 2012;63: 468–480. doi:10.1016/j.jchb.2012.09.004
27. Heathfield LJ, Prins A. M., Schall R. Comparison of footprint ridge density between two South African groups. *J Forensic Investig*. 2016;4: 4.
28. Harris SM, Case DT. Sexual Dimorphism in the Tarsal Bones: Implications for Sex Determination. *J Forensic Sci*. 2012;57: 295–305. doi:https://doi.org/10.1111/j.1556-4029.2011.02004.x
29. Riepert T, Drechsler T, Schild H, Nafe B, Mattern R. Estimation of sex on the basis of radiographs of the calcaneus. *Forensic Sci Int*. 1996;77: 133–140. doi:10.1016/0379-0738(95)01832-8
30. Steele DG. The estimation of sex on the basis of the talus and calcaneus. *Am J Phys Anthropol*. 1976;45: 581–588. doi:https://doi.org/10.1002/ajpa.1330450323
31. Zakaria MS, Mohammed AH, Habib SR, Fahiem AL. Calcaneus radiograph as a diagnostic tool for sexual dimorphism in Egyptians. *J Forensic Leg Med*. 2010;17: 378–382. doi:10.1016/j.jflm.2010.05.009
32. Bidmos MA, Dayal MR, Adegboye OA. Forensic Anthropology Population Data Measurements of the talus in the assessment of population affinity \$. *Forensic Sci Int J*. 2018;287: 221.e1-221.e7. doi:10.1016/j.forsciint.2018.03.016
33. Kim D-I, Kim Y-S, Lee U-Y, Han S-H. Forensic Anthropology Population Data Sex determination from calcaneus in Korean using discriminant analysis. *Forensic Sci Int*. 2013;228: 177.e1-177.e7. doi:10.1016/j.forsciint.2013.03.012
34. Nathana D, Michopoulou E, Kranioti EF. Forensic Anthropology Population Data Sexual dimorphism of the calcaneus in contemporary Cretans. *Forensic Sci Int*. 2017;277: 260.e1-260.e8. doi:10.1016/j.forsciint.2017.04.005

35. Reel S, Rouse S, Vernon W, Doherty P. Estimation of stature from static and dynamic footprints. *Forensic Sci Int.* 2012;219: 283.e1--5. doi:10.1016/j.forsciint.2011.11.018
36. Hammer Ø, Harper DA, Ryan PD. PAST: PALEONTOLOGICAL STATISTICS SOFTWARE PACKAGE FOR EDUCATION AND DATA ANALYSIS. *Palaeontol Electron.* 2001;4: p.9.
37. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. *RadioGraphics.* 2017;37: 505–515. doi:10.1148/rg.2017160130
38. Raza Ali A, Budka M. AN AUTOMATED APPROACH FOR TIMELY DIAGNOSIS AND PROGNOSIS OF CORONAVIRUS DISEASE.
39. Budka M, Wahid -UI-Ashraf A, Bennett M, Neville S, Mackrill A. DEEP MULTILABEL CNN FOR FORENSIC FOOTWEAR IMPRESSION DESCRIPTOR IDENTIFICATION (PREPRINT). 2021.
40. Perez H, Tah JHM, Mosavi A. Deep Learning for Detecting Building Defects Using Convolutional Neural Networks. doi:10.3390/s19163556
41. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016.
42. Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proc 27 th Int Conf Mach Learn Haifa, Isr.* 2010.
43. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res.* 2014;15: 1929–1958.
44. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Bach F, Blei D, editors. *PMLR*; 2015. pp. 448–456. Available: <http://proceedings.mlr.press/v37/ioffe15.pdf>
45. Kingma DP, Lei Ba J. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. *ICLR 2015 ADAM*: 2017.
46. Bishop, Christopher M. *Pattern recognition and machine learning.* Singapore: Springer; 2006.
47. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep learning.* Cambridge: MIT; 2016.
48. Bidmos MA, Asala SA. Sexual Dimorphism of the Calcaneus of South African Blacks. *J Forensic Sci.* 2004;49: JFS2003254-5. doi:10.1520/JFS2003254
49. Gualdi-Russo E. Sex determination from the talus and calcaneus measurements. 2006. doi:10.1016/j.forsciint.2006.10.014

50. Domjanic J, Fieder M, Seidler H, Mitteroecker P. Geometric morphometric footprint analysis of young women. *J Foot Ankle Res.* 2013;6: 27. doi:10.1186/1757-1146-6-27

Tables + Captions

Bournemouth Data	Protocol	% Discrimination	Jack-Knifed
N=233	Landmark coordinates (Fig. 1a)	76.92%	65.81%
N=233	Inter-landmark distances (Fig. 1a)	98.71%	67.38%
N=233	Binary % black, seven squares (Fig. 1a)	59.13%	53.48%
N=233	Binary % black, three squares (Fig. 1b)	50.64%	53.48%
N=99	Manual ridge counting three squares (Fig. 1b)	67%	43.35%
N=233	Machine Learning - Shape+Texture+Size	82.6%	-

Table 1: Sex discrimination on the Bournemouth data using various methods.

Scenario	Task	Shape	Texture	Size	Accuracy (Sex)	MAE (Age)
1	Sex estimation	Yes	Yes	Yes	88.76%	-
2		Yes	No	Yes	88.76%	-
3		Yes	Yes	No	79.40%	-
4		Yes	No	No	78.27%	-
5		No	Yes	No	80.15%	-
6	Sex & age estimation	Yes	Yes	Yes	89.89%	7.51
7		Yes	No	Yes	89.13%	8.79
8		Yes	Yes	No	85.01%	8.26
9		Yes	No	No	80.15%	8.98
10		No	Yes	No	83.89%	9.20
11	Age estimation	Yes	Yes	Yes	-	8.47
12		Yes	No	Yes	-	9.93
13		Yes	Yes	No	-	8.28
14		Yes	No	No	-	8.39
15		No	Yes	No	-	9.68
16	Sex estimation	Yes	No	No	69.64% (70.95%)	-
17	Sex estimation	Yes	No	Yes	89.03% (86.73%)	-

Table 2. Comparison of machine learning model accuracy (ResNet101) with different combinations of analysis using the Walker-Data. MAE is the mean age error, that is plus or minus x years. Scenario 16 is conventional landmark analysis based on coordinates subject to Generalised Procrustes Analysis (GPA) so including only elements of shape. The figure in parentheses is the jack-knifed result. Scenario 17 is conventional landmark analysis based on inter-landmark distances so includes elements of shape and size. The figure in parentheses is the jack-knifed result.

Figures



Figure 1: Landmarks and sampling areas. a. Initial sampling points mid-way between the placed landmarks. b. Revised and additional landmarks. c. Cropped sample areas.



Figure 2: The results of different resizing algorithms.



Figure 3: Image analysis. a. Image with channels composed using 3 resizing methods (R=NEAREST, G=BILINEAR, B=HAMMING). B. The same print with ridge detail removed using dilation operation followed by erosion operation.

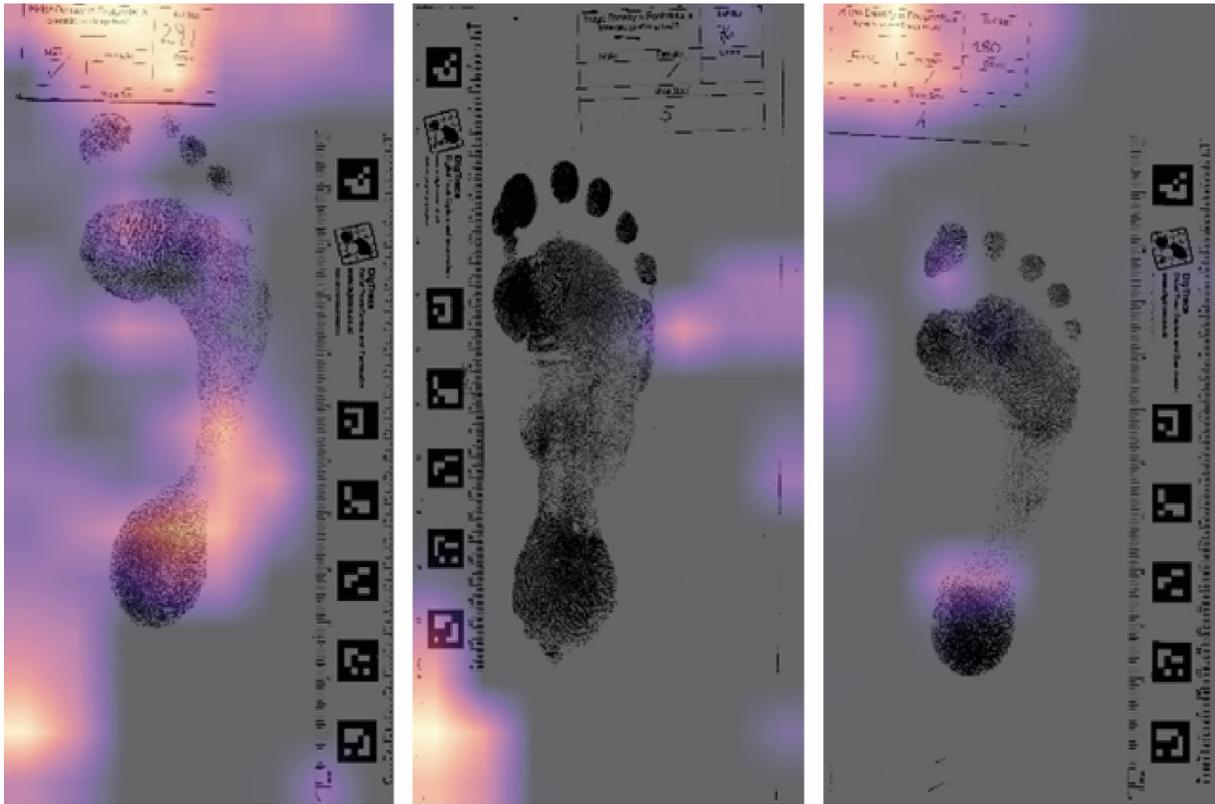


Figure 4: Example of the machine learning output where the focus for decision-making as indicated by the heat map is on the record card which has a sex indicator on it.

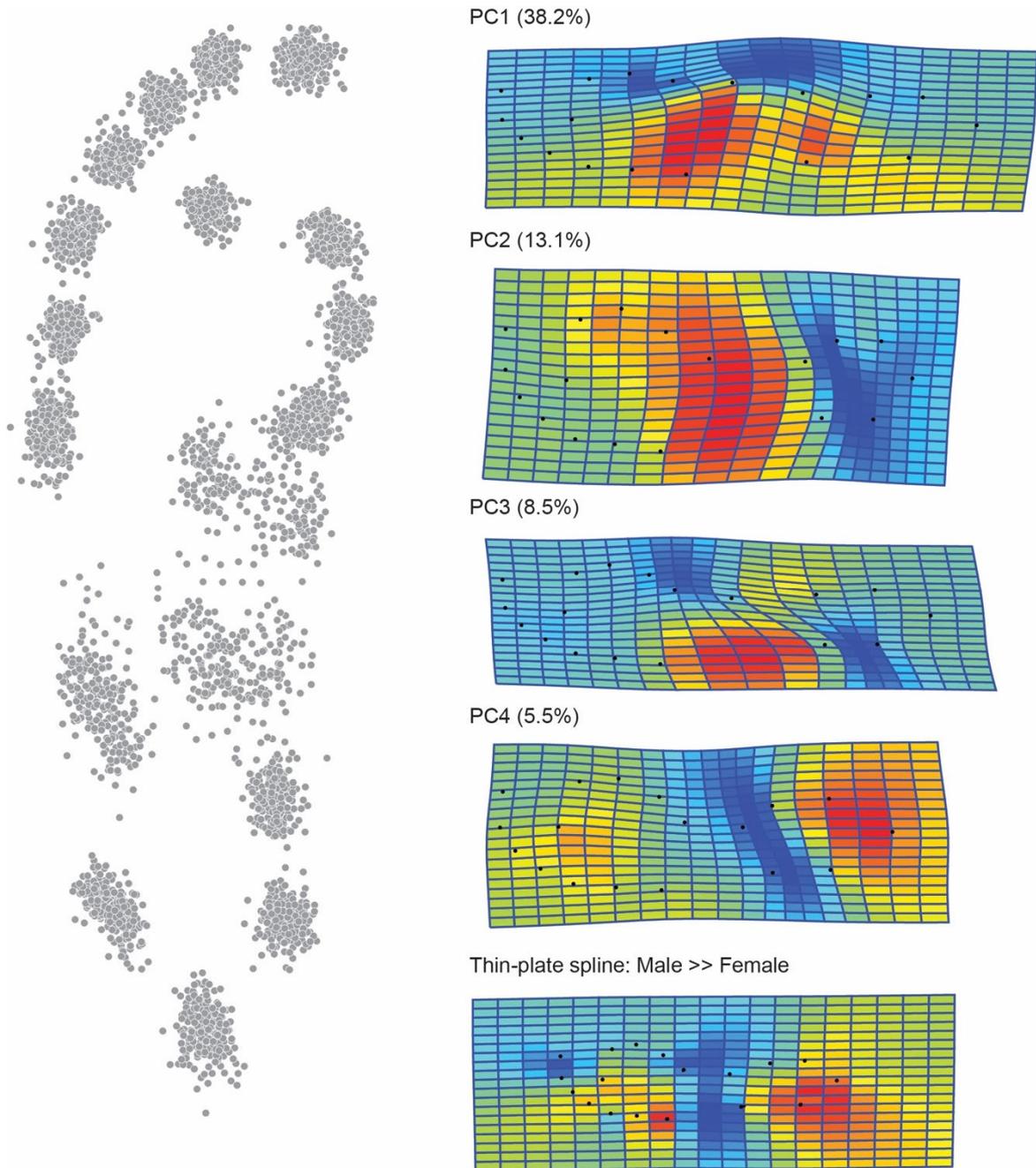


Figure 5: Geometric morphometrics on the Bournemouth-Data, showing the distribution of landmarks, the first four Principal Component Warps (65.3% of the variance) and thin-plate spline comparison of the male to female means. N= 132 males and 165 females.

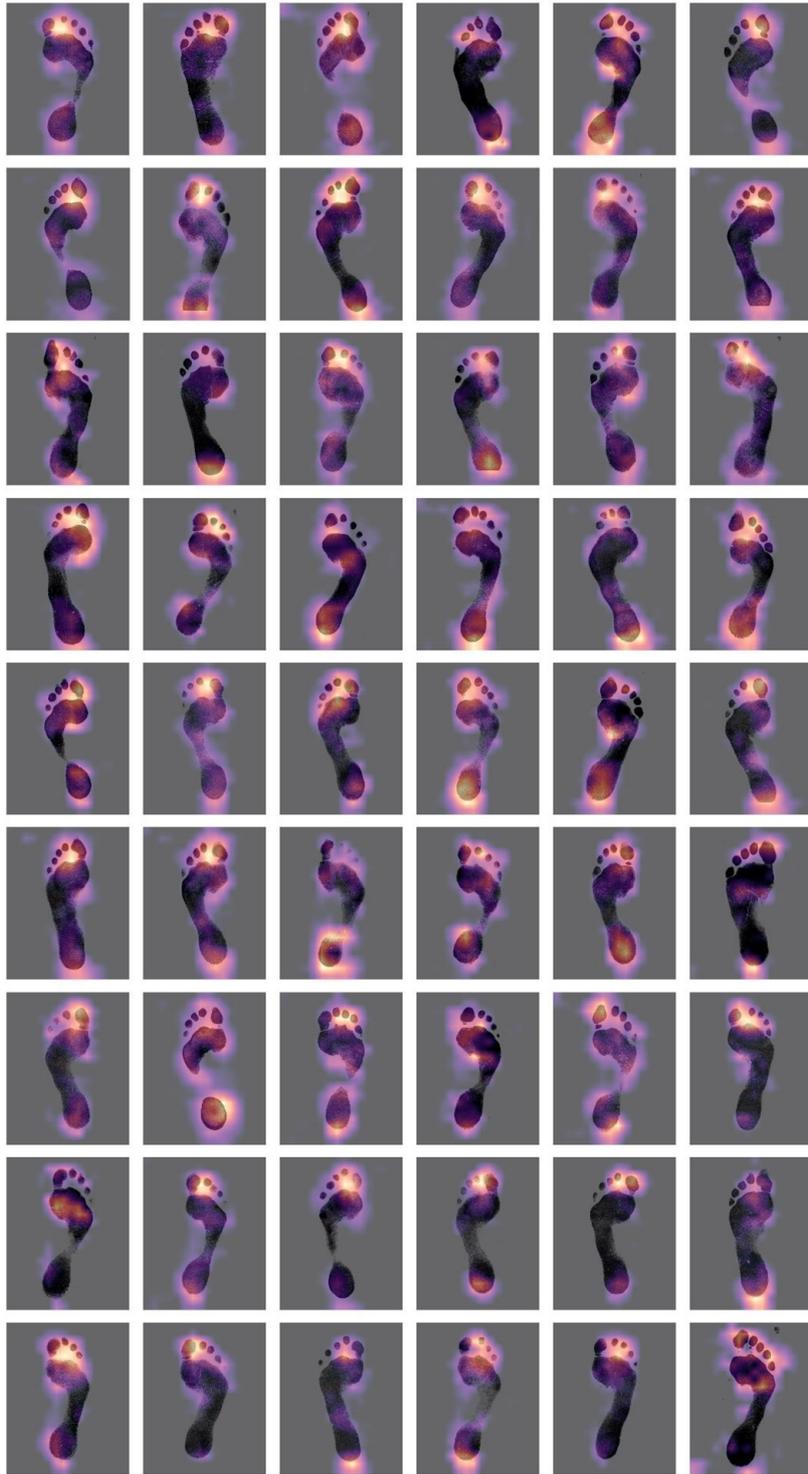


Figure 6: Grad-CAM heatmaps for textured inputs with the lowest entropy losses. The heat maps indicate areas which a key in the decision-making process. Note the concentration of bright areas around the heel and just behind the first and second toes.