Leveraging Key-Point Detection to Prevent Shoe Returns in Online Shopping

Andres PRADA GONZALEZ, Matthias BRENDEL
Footprint Technologies GmbH, Berlin, Germany

https://doi.org/10.15221/23.39

Abstract

More than half of all shoes ordered online are returned.1 Different shoe size systems, a lack of standards and deviating models lead to a high level of uncertainty on the part of consumers. As a result, retailers lose up to 20% of their revenues and the environment is unnecessarily polluted by packaging and shipping. Furthermore, users spend an average of 32 minutes of their time returning a shoe. We have developed a solution that allows consumers to measure their feet at home with their smartphone by processing two images of their feet on a DIN-A4 sheet. To accurately recommend a size, two sets of data are required: the feet measurements and the shoe dimensions. This is why, on the one hand, a foot measuring algorithm was developed using two computer vision techniques: semantic segmentation and key point detection. This paper compares both methodologies and evaluates which one performs better on a distance similarity metric. On the other hand, a database was built up where the Last data, i.e., the inner shoe dimensions of the shoes per manufacturer, model and size, are stored. When measuring the foot in 2D, there are factors such as the point of view of the images or the image resolution that can influence the perfect foot measurement. Hence, a guided capturing system was developed to minimize them. Finally, to match them to the perfect fitting size, we performed in-person tests with the manufacturer's shoes over the last year to optimize the way of recommending them.

Keywords: foot measurement, shoe size recommendation

1. Introduction

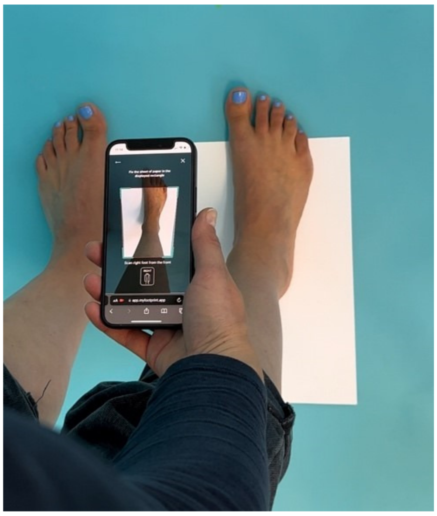
The lack of size standardization within the global shoe industry historically caused co-existing shoe-size systems. For example, many European shoe brands use a numerical system that is in the range of 18 to 50+, whereas in the UK it is common to see a range like 1 to 13. In America, shoe sizes are often accompanied by string labels, such as men's, women's, or kids. Today, the Paris stitch is considered the most common European shoe size system [1], where two consecutive shoe size labels (such as size 36 and 37) differ by 6.67 mm. Great Britain has developed its own shoe size system. This involved the introduction of its own unit of measurement for shoes, known as Barleycorn, which is based on inches. Nowadays, shoe sizes differ by half a barleycorn, which corresponds to 4.23 mm. The American length measurement system is based on the English one, with the scale starting earlier. Thus, the different shoe size systems mean that shoe sizes cannot be consistently converted. A further problem is manufacturing shoes on non-standardized Lasts, which, depending on the brand and the model, can deviate from the conventional measurement system in width and often length. In this respect, different shoe models from the same manufacturer, with the same size printed on sizing labels, may turn out to fit differently for the same person.

This problem causes more than half of all shoes ordered online to be returned because the size ordered does not fit, which causes a loss of sales for up to 20% for retailers and unnecessary pollution of the environment through packaging and return shipping. On average, users spend 32 minutes of their time returning a shoe, as well. The goal of the project is to enable the best possible fit recommendation of shoes for online shoppers, thereby avoiding size- and fit-related returns.

2. Technical solution

We have developed a solution that allows consumers to measure their feet at home with their smartphones. They are required to stand on a white, clean sheet of paper of size DIN A4 or on a US Letter and capture two photos (one from the front and one from the back) for each of their feet. These images are evaluated using a mixture of computer vision and machine learning techniques to deduce the relative size of the feet to the sheet of paper they stand on. This way, an accurate length and width of the feet can be determined. This technique allows other foot parameters to be computed, such as the bunions or metatarsal angle.

At the same time, a database was built in cooperation with shoe manufacturers and is growing continuously, in which the shoe's Last dimensions are stored for each model and size. This cooperation, in which the manufacturers release their Last data, is unique in the industry, even though it is indispensable for a correct fit recommendation, because shoe sizes do mostly not fully comply with any shoe size system, as described above. To ensure a high user adoption rate and low entry barriers, we adapted our foot measurement and size recommendation service to run fully on all common iOS and Android smartphones directly in a web browser without the need to install an additional app.



Picture of the Footprint Webapp being used to measure the right foot.

3. Approach for feet measurement

To achieve an optimal fit, it is crucial to compare the inner shoe length against the precise length of the foot plus an allowance for the toes. The difference in length between consecutive pairs of European shoe sizes amounts to 6.7 millimeters (mm), making a measurement accuracy goal of around 2 mm a reasonable aim, given the inherent process limitations.

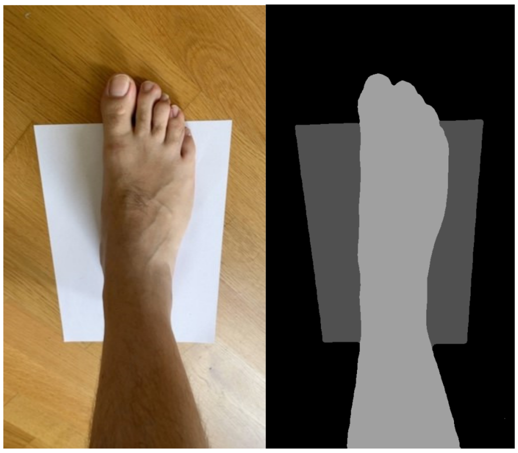
To meet this objective, we developed an algorithm that measures the distances between specific points of interest, referred to as "key point", In this context, for assessing length, these key points were selected to be the outermost toe and heel positions. We evaluated two machine learning approaches to detect these points of interest: semantic segmentation and key-point detection.

Considering the widely known accuracy and direct outcomes of Semantic Segmentation in different applications, as well as existing examples demonstrating its potential to achieve the desired precision, this method was chosen as the starting point.

3.1. Semantic Segmentation

The approach of semantic segmentation includes the generation of segmented masks for the paper and the foot, followed by the identification of relevant key points. The core of this approach was a semantic segmentation model based on the DeepLab v3+ [2] architecture, which was set to classify each pixel on the images into one of three categories: Background, Foot, and Paper.

The resulting pixel-based semantic mask is used to find the corner(s) of the paper on the image - mostly hidden by the leg - by extending the edges of the paper mask. The images are then cropped and corrected for distortion. The distance between the outermost points from the top and bottom of the contour of the foot represents the length of the foot in pixels, and the outmost right and left points would indicate the width of the foot.



Semantic segmentation mask obtained from a front frame where the black pixels correspond to the background, the dark gray to the sheet of paper and the light gray to the foot.

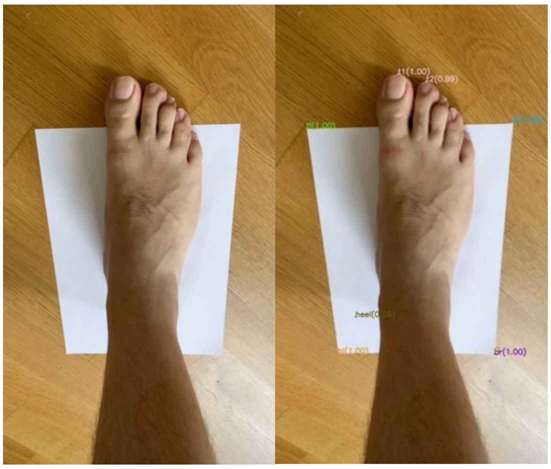
While the Semantic Segmentation method proved promising, it has its weaknesses due to an extremely complex and error-prone heuristic: the result of Semantic Segmentation is the pixel-based mask described above, i.e., the contour of the foot and the paper. To determine the exact key points on the contour to measure the foot's length and width, more advanced heuristics needed to be applied.

3.2. Key point Detection

Given the final task of the algorithm was to find the points of interest in RGB images, the second approach implied to train a model directly to perform it. Therefore, it offered the potential to significantly simplify and shorten the processing time. The processed data set could now be used to try out another approach and compare the results.

The key-point detection is a technique for predicting the position of points of interest on an RGB image. In our case, the key points of interest are, for example, a toe, the furthest out point of a ball of foot, or the corners of a sheet of paper.

The use case consisted of detecting the key point in an image containing a sheet of paper and a foot. Therefore, a bottom-up approach was implemented. The architecture chosen was HRNet [3] as the backbone network uses Associative Embedding to group key points. The use of HRNet provides a high-resolution representation of features, while the associative embedding approach ensures the correct mapping of key points to the corresponding objects in the image.



*Front frame processed through the key-point detection model.
The key points are drawn onto the image alongside their score [0-1].*

Once the points for the farthest toe and heel are found, the distances in pixels can be calculated, using the same procedure as for semantic segmentation. Using the DIN-A4 reference sheet, the output points of both meshes were scaled to the same scale.

4. Analysis and Evaluation

The 2D measurement algorithm has one task: measure the distance in pixels-, and then calculate the millimeters- between two given points. To compare the performance of both approaches to determine which performs better, a distance similarity metric was used. It was important to determine which of those could better predict the location of the key points and how reliable the final measurements were.

To achieve this, two key concepts were of interest: accuracy and precision. The first one refers to the maximum deviation of the measurement algorithm with respect to a given ground truth of the measurement. The second one aims to demonstrate the repeatability of the measurement under the same conditions.

4.1. Distance similarity metric

To compare and evaluate the performance of the key point detection model versus segmentation, a benchmark dataset was created. A spectrum of images ranging from the simplest to the most difficult cases were collected and labeled. This was done by manually defining all the key points of the image, i.e., the four corners of the sheet of paper, the longest point of each toe, the furthest outside point of the metatarsal outside and inside, and the heel point. The performance of the two algorithms was then examined and evaluated. This was done by finding the one that minimizes the distances (in pixels) between the predicted position (in the Cartesian coordinate system, i.e., x and y coordinates) of the point and its actual position.

It contained 85 labeled images none of the algorithms were trained on. The Object Keypoint Similarity (OKS) [4] metric was chosen as the benchmarking metric, which evaluates the calculated distances between key point and their ground truth.

$$\boldsymbol{OKS} = \boldsymbol{e}^{\frac{\left(-\left(\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}\right)^2\right)}{S}}$$

In the equation, (*x*,*y*) refers to the cartesian coordinates, *i* and *j* to the ground truth and the predicted key points respectively and *S* to the scaling factor. This Last parameter helps in smoothing out the outliers. For our experiments, it was set to 500.

4.2. Determine the real measurements of the foot

To determine the accuracy of a foot measurement in millimeters it is important to have a reliable ground truth. There are different ways to measure a foot's length and width. The one chosen to prove our algorithm was a 2D foot scanner. This device works as a normal DIN-A4 scanner, but it is encapsulated into a wooden box and has a thick glass to place your foot on. The output of the scanner was RGB images, and a post-processing step to obtain comparable one-to-one measurements with the measuring algorithm was required. Several things stood out in the process:

- Our measuring system always measured about 10 mm less than the scanner. This could be explained by the fact that the scanner measured at a heel height that the measurement algorithm cannot detect. This meant that a constant offset could be added to the algorithm to adjust for it and compare the values one-to-one.
- The results provided by the scanner varied considerably when the foot was not parallel in the same direction as the sheet of paper but was rotated. To study this effect further, several calibration tests were performed using a reference object with known dimensions on the scanner. This object - a standard credit card - was placed at different rotation angles and the measurements obtained for each of the edges were accurate to the millimeter. However, this did not happen when placing a foot on various positions. The millimeters measured were varying up to 12 mm depending on the rotation angle of the foot with respect of the edge of the scanner.

The conclusion from these experiments was that, although the scanner was providing consistent results when the foot was placed parallel to the edge of the scanner, it could not provide stable and repeatable results otherwise. Hence, it was discarded as a reliable approach to measure the length and the width of a foot.

4.3. Precision of the results

The only metric found to be representative of a good measurement system was the precision of the measurements. This metric can be understood as the maximum deviation between the largest and the smallest measurements of a certain distance – the length of the foot, for instance – in a series of consecutive measurements.

12 users tested the key-point measuring algorithm around 10 consecutive times, under the same conditions and stepping in and out of the DIN-A4 paper. This generated a total of 216 foot measurements. Each foot's length was computed, and once all the measurements were done, the mean per user was subtracted for each dimension. This step was used to normalize the measurements around the mean. We chose the standard deviation as the metric to represent the precision of the results. The mean standard deviation across all users for the key-point detection was 1.798 mm and 1.66 mm for the length and the width, respectively.

5. Development of recommendation system

The measuring algorithm represents the most critical part of the system, but it is not the only important part. To recommend accurately the shoe size for an individual two parts come into play. On the one hand, both feet must be measured accurately, and on the other hand, the dimensions of the shoe Lasts must be known.

In cooperation with the manufacturers, the exact inner shoe dimensions (shoe-Last data) of each model in each available size are stored in a database. When a manufacturer launches a new model, this must also be added.

Once the foot measurements were known, an algorithm was developed to match these measurements with the two-dimensional inner shoe dimensions. To check and improve the success of the matching algorithm, regular test actions are carried out with test people who first measure their feet with our measurement algorithm and then try on different sizes of a model and provide feedback. Further optimizations on the magnitude of toe-allowance were empirically carried out to ensure the accuracy of the size recommendations

6. Development of user interface and usability challenges

For the overall success of the product, obtaining an accurate technical measurement of the foot is crucial. To achieve this goal, it is imperative to acquire usable image material. This entails accurately positioning the feet on the sheet of paper during recording. Both feet necessitate two images: one captured from above to ascertain width and overall geometry, and another taken diagonally from behind to capture the heel and determine length. An essential determinant of the product's success is the high usability of the service. Beyond technical advancements, the development of a user-friendly application that intuitively guides users through the process is equally vital.

The technical application was transformed into a "user journey", outlining each step of the survey in detail. This process involved crafting individual screens that users would encounter, which were then organized into a "user flow." This provided a comprehensive view of the pathway and the choices accessible to users. To facilitate this, initial prototypes were designed and subjected to testing with designated users. These evaluations occurred periodically, involving up to 40 participants in a single day. The focus was on observing how participants interacted with their smartphones and identifying any challenging steps.

Notably, it was observed that some participants struggled with understanding the optimal way to hold their smartphones for capturing images from behind. To address this issue, various approaches—utilizing audio, video, and text—were developed to guide users through the recording process. This involved capturing images, creating explanatory videos, generating textual instructions, and recording an audio guide.

Subsequent testing was conducted to assess the effectiveness of these diverse strategies. The findings revealed that a combination of video and text proved to be the most effective approach.

7. Conclusions and future work

This paper has presented a shoe size recommendation system based on two technical approaches. Namely semantic segmentation and key-point detection. Initially, using a semantic segmentation approach to determine the contour of the foot and then finding the points of interest to measure different dimensions was detailed explained. However, given the goal of the problem finding points on the image, a key-point detection model served this purpose better by being more reliable in terms of better localizing the key points and thus, minimizing the deviance in a series of consecutive measurements of the same foot.

While grasping foot measurements is pivotal for accurate size recommendations, it only constitutes half of the system's functionality. A matching system was developed to correlate computed foot measurements with inner shoe Last dimensions. As each shoe delivers a unique fitting experience, consistent testing sessions are essential to ensure recommendation precision.

The system's upcoming research directions are based on two main ideas: making it even easier for users to use by giving better guidance and exploring how to create 3D models of feet. This goal will come with new challenges. For example, getting 2D details from shoemakers is already challenging in some cases. This includes not only convincing them that the data is important, but also dealing with technical problems, especially if many manufacturers do not have that kind of information.

References

[1] International Organization for Standardization. (2022). ISO 20685:2022 Footwear—Sizing—Conversion of sizing systems from Mondopoint system to the Paris point system. ISO.

[2] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. ArXiv. /abs/1802.02611

[3] Wang, Jingdong, et al. "Deep high-resolution representation learning for visual recognition." IEEE transactions on pattern analysis and machine intelligence 43.10 (2020): 3349-3364.

[4] Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. ArXiv. /abs/2204.06806