UNIVERSITY OF TARTU
FACULTY OF SCIENCE AND TECHNOLOGY
Institute of Computer Science
Software Engineering Curriculum

Gustav Amer

Foot Detection Method for Footwear Augmented Reality Applications

Master's Thesis (30 ECTS)

Supervisor: Amnir Hadachi, PhD

Tartu 2016

Foot Detection Method for Footwear Augmented Reality Applications

Abstract:

Augmented reality is gaining popularity as a technique for visualizing apparel usage. Ideally it allows users virtually to try out different clothes, shoes, and accessories, with only a camera and suitable application which encompasses different apparel choices.

Focusing on augmented reality for footwear, there is a multitude of different solutions on how to offer the reality augmentation experience to the end users. These solutions employ different methods to deliver the end result, such as using fixed camera and constant background or requiring markers on feet for detection. Among the variety of techniques used to approach the footwear reality augmentation, there is no single best, simplest, or fastest solution. The solutions' sources aren't usually even publicly available.

This thesis tries to come up with a solution for the footwear reality augmentation problem, which can be used as a base for any proceeding footwear augmented reality projects. This intentionally universal approach will be created by researching possible combinations of potential methods that can ensure a solutions regarding footwear reality augmentation.

In general, the idea behind this thesis work is to conduct a literature review about different techniques and come up with the best and robust algorithm or combination of methods that can be used for footwear augmented reality.

A researched, documented, implemented and publicized solution would allow any upcoming footwear augmented reality related project to start working from an established base, therefore reducing time waste on already solved issues and possibly improving the quality of the end result.

The solution presented in this thesis is developed with focus on augmented reality applications. The method is neither specific to any platform nor does it have heavy location requirements. The result is a foot detection algorithm, capable of working on commonly available hardware, which is beneficial for augmented reality application.

Keywords:

Image enhancement techniques, image processing methods, object detection algorithm, active contour detection algorithm

CERCS: P170

Jalatuvastus meetod jalatsite liitreaalsuse rakendustele

Lühikokkuvõte:

Liitreaalsus on populaarsust koguv platvorm rõivaste ning aksessuaaride kasutamise visualiseerimiseks. Ideaalis võimaldab see kasutajatel proovida erinevaid riideid, jalatseid ja aksessuaare, kasutades ainult üht kaamerat ning sobivat rakendust, mis võimaldab kuvada erinevaid valikuid.

Jalatsite liitreaalsuses on palju erinevaid lahendusi, et pakkuda kasutajatele liitreaalsuse kogemust. Need lahendused kasutavad erinevaid meetodeid, nagu fikseeritud kaamera, muutumatu taust ja markerid jalgadel tuvastuse hõlbsustamiseks. Nende meetodite hulgas pole ükski kindlalt parem, lihtsam või kiirem. Lisaks puudub tihtipeale avalikusel ligipääs arendatud rakendustele.

Käesolev magistritöö proovib leida universaalset lahendust, mis sobiks kasutamiseks kõigi tulevaste jalatsite liitreaalsuse rakendustega.

Võtmesõnad:

pildi parandamise tehnikad, pilditöötluse meetodid, objekti avastamise algoritm, aktiivne kontuuri leidmise algoritm

CERCS: P170

Table of Contents

Chapter 1: Introduction

1.1 General view and Background

Augmented reality is a view of the physical environment whose elements are supplemented by computer-generated input, for example, graphical elements. The aforementioned supplementation is conventionally done in real-time. With advanced technologies like object recognition and computer vision, augmented reality enables turning the information from surrounding real world interactive for the user. (1)

Augmented reality has a multitude of uses. It can be used for archaeological research, by supplementing the modern landscape with archaeological features and therefore enabling archaeologists to formulate conclusions regarding the site configuration and placement. Augmented reality can be helpful in the field of education. This is often done by implementing the education reading material with markers for an augmented reality device, which in turn can visualize the information and make the material easier to understand.

Possibly one of the biggest users of augmented reality is the commerce and advertising sector. With the use of similar techniques as in education, suppliers can offer customers a better overview of their product without having to provide a physical sample of the product. For example, embedding a marker in a magazine advertisement which, with a Smartphone and a corresponding application, customers can use to visualize a 3D model of the item on their Smartphone's display. Another way augmented reality is used by the commerce sector, is offering customers to virtually try out their products. This is called augmented reality dressing room, and it enables users to try out different apparel items without the need to manage them physically.

1.2 Objectives and Restrictions

The objective of this thesis is to create a method for foot detection in images. The intended method should be capable of functioning with minimal hardware and environmental requirements. In this case, the minimal hardware and environmental requirements mean that the method ought to be able to produce positive results in nearly any environment with a

Smartphone camera or a webcam. Later, in this document some existing solutions are described which rely heavily on controlling the environment and having specialized hardware. Therefore it is logical to make the assumption that the requirement restrictions posed in this thesis complicate creating the given method. The algorithm created during this thesis doesn't have heavy requirements for the hardware, but some environmental restrictions remain:

Background images - images of the same background in front of which the parseable image is taken. The background images are to be as close to identical as possible to the image where the foot is to be detected.

For the images to be near identical, the camera has to be static for all the images.

To keep the false positive rate down, the lighting should leave minimal visible shadows. For that the source(s) of lighting should be above the subject (person) or in front of the subject, near the camera.

1.3 Contributions and Relevance

While working on the thesis, multiple possible approaches to the stated problem were investigated, many more than just those which remain in the final solution. In addition to investigating and research, implementation prototyping was often used to determine any techniques effectiveness. For the final solution many pre-existing technologies were combined to produce a working method.

As previously stated, there is currently a distinct lack of functional foot detection algorithms. Therefore, this is an underdeveloped branch of image detection and computer vision in general. Therefore, any meaningful research connected to foot detection could be considered an important asset to the computer vision research field. Moreover, this dissertation can be used as a stepping stone for future advancements in foot detection algorithms.

1.4 Road Map

This document is divided into five chapters. The first and current chapter gives an overview of the purpose of this thesis, the general work that was done while creating the thesis, and the layout of the document.

The second chapter named State-of-the-art serves the purpose of describing a baseline of current existing solutions and technologies related to foot detection, object detection in general, etc. This chapter is divided into two main parts, related work and industrial solutions. The former includes analysis of articles Detection Of Moving Object Based On Background Subtraction and Background Subtraction for Object Detection Under Varying Environments. The latter covers Virtual Mirror and Augmented reality with markers.

The following chapter focuses on the author's contribution while working on the thesis as well as the methodology used. The third chapter is divided into five parts. In addition to introduction and conclusion, the parts described in this chapter include the problem statement, system design and architecture, and methodology.

The next chapter is about the results produced by the author's contribution detailed in the previous chapter. The fourth chapter also contains analysis of the given results. Main points of this chapter are the results from following techniques: picture enhancement and filtering, background extraction, active contour model integration, and foot orientation determination.

The fifth and final chapter is the conclusion, where it is presented a summarization of the thesis and possible perspectives of future work based on the contributions made within this thesis.

An alternative approach to foot detection researched while working on the dissertation, is presented in Annex.

Chapter 2: State-of-the-art

2.1 Introduction

This chapter of the thesis reviews existing work related to the topic of the thesis. The chapter starts by reviewing some related works, where exists some overlap of desired output or utilized techniques with this thesis. Afterwards, existing technologies which output is similar to the desired solution of this thesis, disregarding any requirements these technologies have, are surveyed.

2.2 Related work

2.2.1 Detection Of Moving Object Based On Background Subtraction

The authors of (2) employ background subtraction techniques to tackle challenges related to automated visual surveillance. Their goal is to create an algorithm capable of object detection in visual surveillance in real time with as little computational power as possible.

Their algorithm relies on utilizing a pre-existing background image, which is subtracted from every parseable frame. If the resulting difference exceeds a given threshold, work on the frame continues. For smoothing out noise and any random abnormalities the authors propose using a special 3x3 filter mask instead of the more conventional median filter. They believe the filter mask to be more efficient, as it doesn't require sorting the nearby values into order, and that edges are less likely to be smoothed out by its utilization.

2.2.2 Background Subtraction for Object Detection Under Varying Environments

The article (3) investigates object detection with the use of background subtraction. They make an explicit distinction of the environment, by dividing it into outdoor and indoor environments. This distinction enables them to better combat the undesired effects of shadows on background subtraction methods.

For outdoor environment images, they suggest making use of the fact, that background colour pallet is generally constant in RGB colour space. Their reasoning is that, for example sky areas have high blue intensity while ground has more green intensity. Therefore different pixels from the background but with similar green intensity on the ground area can be recognized as shadows, rather than new objects.

In case of indoor environment, the authors of the article suggested a different approach. They made the deduction, that because most indoor lighting is on walls or ceiling, the shadows from around the lower part of the object. By using data about detected objects shape and size from middle and upper part, the objects lower part can be approximated and the detected nearby shadows ignored.

2.3 Industrial solutions

2.3.1 Virtual Mirror

One of the best results in regards of footwear augmented reality is achieved by the Virtual Mirror. Specifically, Virtual Mirror developed for Adidas stores. (4) This mirror is a system that allows users to virtually try shoes and see the visualization in the mirror. The result is achieved by the use of augmented reality techniques which, in itself, include combining real video feed with 3D computer graphics model representation of virtual objects. The mirror environment removes the need for users to wear any special glasses. This Virtual Mirror system was first used in Adidas store in Champs Elysées, Paris, France, in October 2006. The application allows users, in addition of trying out existing models, also personalize the models by changing the design and colours of existing shoe models, as well as adding custom decorations and embroideries. The designed shoe can be seen worn via the Virtual Mirror. (4)

The virtual mirror utilizes a static camera that captures the video feed of the customer standing in front of the mirror. Instead of a real mirror is a screen, which displays the camera image that is horizontally flipped. The screen is mounted in such a way that the image displayed to the user would be similar to the reflection of a real mirror. The position and rotation of each foot is estimated with a 3D motion tracker. This approach is supposed to

be very robust and easily adaptable to new shoe models. After determining the position of the feet in 3D space, the shoe models, customized by the user, are rendered onto the video stream in a way that the user's real shoes are replaced with the virtual ones. To account for possible occlusion that can occur when the 3D virtual scene is converted into 2D video, the visibility of each shoe part is calculated for the given position. All the calculations take place in real time, so the user shouldn't experience any delay in the displayed image. (4)

In order to simplify the segmentation of the user's' feet from the background, the floor in the field of view of the camera is painted green and lighting is set under the camera to combat shadows generated by the general store lighting. The Virtual Mirror's algorithmic flow can be divided into 3 parts: segmentation and 2D image pre-processing, gradient-based 3D tracking, rendering and augmentation. (5)

Segmentation and 2D image pre-processing

The camera outputs a video feed with a 1024 x 768 resolution for processing. The camera is statically fixed, with all automatic control disabled and its shutter time is synchronized to the flickering of the room lighting. Also, to adjust for any change in the illumination, the camera gain is re-computed on every event when no user is detected in the camera's field of view. (4)

For speed efficiency, the image is downsampled by a factor of 2 until a 64 x 48 resolution is achieved. Then the segmentation algorithm starts with the smallest filtered image. It compares each pixel to a 3D RGB look up table consisting of elements, corresponding to the green background. Following the pixel classification, the image is modified to achieve a result where only 2 feet with shoes remain. The same operations are carried out on images of larger resolutions while some pixels are rejected for classification according to the lower resolution computation results. After segmentation of the largest (original) resolution image, if 2 shoes are considered completely visible, then the 3D motion tracking algorithm is started. (4)

Gradient-based 3D tracking

The tracking algorithm works by estimating the two body motion parameters, similar to those used in face tracking. Then the shoe models are rendered, according to the shape and orientation, onto the image. This enables getting the information regarding the shoes' dense depth and silhouettes from the graphics card z-buffer. A silhouette mask is generated from the retrieved information. (4)

In order to increase the efficiency of the tracking algorithm, the parameters used are generated by a gradient-based technique. The binary object borders are given constant gradient values, according to the pixel distance from the object – the smaller the distance, the higher the value. (4)

Rendering and augmentation

After calculating the motion parameters for both feet, the customized shoe models are rendered onto the image. The resulting displayed virtual model should cover its real world counterpart. All possible colours used in the personalized model have corresponding values stored in database for accurate visualization. (5)

When rendering the final image, an invisible transparent leg model is added. This allows correctly recovering all occlusions. (5)

Summary

The result achieved with the Virtual Mirror is very good. The output is created in with real time constraints and is quite lifelike. See figure 1 for an example.

Figure 1: Upper row: Scene captured with the Virtual Mirror camera. Lower row: Virtual Mirror output augmented with (5)

The negative aspects of this approach are its requirements. The algorithm depends on static camera, constant background and controlled lighting. All of which are unreasonable when dealing with Smartphone based system.

2.3.2 Augmented reality with markers

One way to get around the hardware requirements mentioned in the previous paragraph is to use markers to identify desired object for capture. This approach is described in (6). The article proposes using ARToolKits™ to calculate the camera orientation and position relative to physical markers. The usage of markers has its own limitations, for example it is usually required that the markers lie on planar object or a plane. As feet can be considered curved objects, therefore marker based detection methods are not suitable for foot detection and footwear augmented reality. The article continues suggesting specialized solutions, such as Microsoft Kinect™ for capturing depth information of 3D objects. By combining these two, markers and Microsoft Kinect™, methods, a two-stage tracking method is created. The method works by first estimating the foot location via markers. 6 markers are placed on specific locations on the foot, and their colour is to be distinctive from other objects in view. (6)

After the foot approximate location is estimated with markers, a 3D foot model is positioned onto the real foot recognized from the image. Then, by aligning the model to its physical counterpart based on depth data captured with Kinect™, the foot orientation can

be determined. Finally as the foot orientation and location is known, a shoe model can be rendered in the location with correct direction, onto the outgoing image. See figure 2 for an example.

Photorealistic rendering according to environmental lighting conditions is omitted, because it is considered too computation heavy for the real-time product customization application, the article focuses on. (6)

The approach described in (6) can be considered as step closer to Smartphone based footwear augmented reality, from the previous Virtual Mirror approach. Still, the use of markers and need for Kinect™ to gather depth data is not suitable for Smartphone based augmented reality system. (6)

There is also a negative aspect when comparing the marker & Kinect™ approach to the Virtual Mirror – as mentioned, the realistic lighting rendering is omitted and also occlusion detection has lower quality.

Figure 2: Virtual try-on of a customized shoe model (6)

2.4 Conclusion

In this chapter, we covered some researched articles that overlap with this thesis in some aspects. In addition, a few industrial solutions, which have similar goal to that of the current thesis, were inspected. Most notable of them was the Virtual Mirror, which has strict environment and hardware requirements, but delivers an excellent result. The next chapter focuses on contribution and methodology used while working on this thesis.

Chapter 3: Methodology and contribution

3.1 Introduction

In this chapter, we focus on the core of the thesis. We will cover the main problem statement, followed by descriptions of the system's design and architecture and finally cover the methodology used throughout the development process of this thesis.

3.2 Problem statement

The main objective of this thesis is detecting human foot in an image. The created algorithm should be usable with footwear augmented reality applications. Footwear focused augmented reality applications' one of the first steps in functioning is to detect foot or footwear in input video feed. The input feed would be processed as individual images, on each of which object detection, with focus on foot detection, would be used. While object detection is widely researched topic with many functioning solutions for variety of different object types, but no robust method for foot detection exists. The method created within this thesis utilizes image enhancement, background subtraction and active contour model with specific type of images to detect person's foot in the input image. The input for the algorithm is an image of a person's foot taken in an environment with controlled lighting in order to avoid more prominent shadows. During processing, the image is enhanced, its foreground is extracted, and active contour model is applied to produce the outline of the foot in the image.

3.3 System design and architecture

The following, figure 3, is a diagram describing action flow of foot detection algorithm created for this thesis. Items presented in an oval are artefacts of the algorithm, such as input and output images. The parallelograms contain actions or processes which use some artefacts as input and produce one or more new artefacts as output. The arrows display the flow of the diagram.

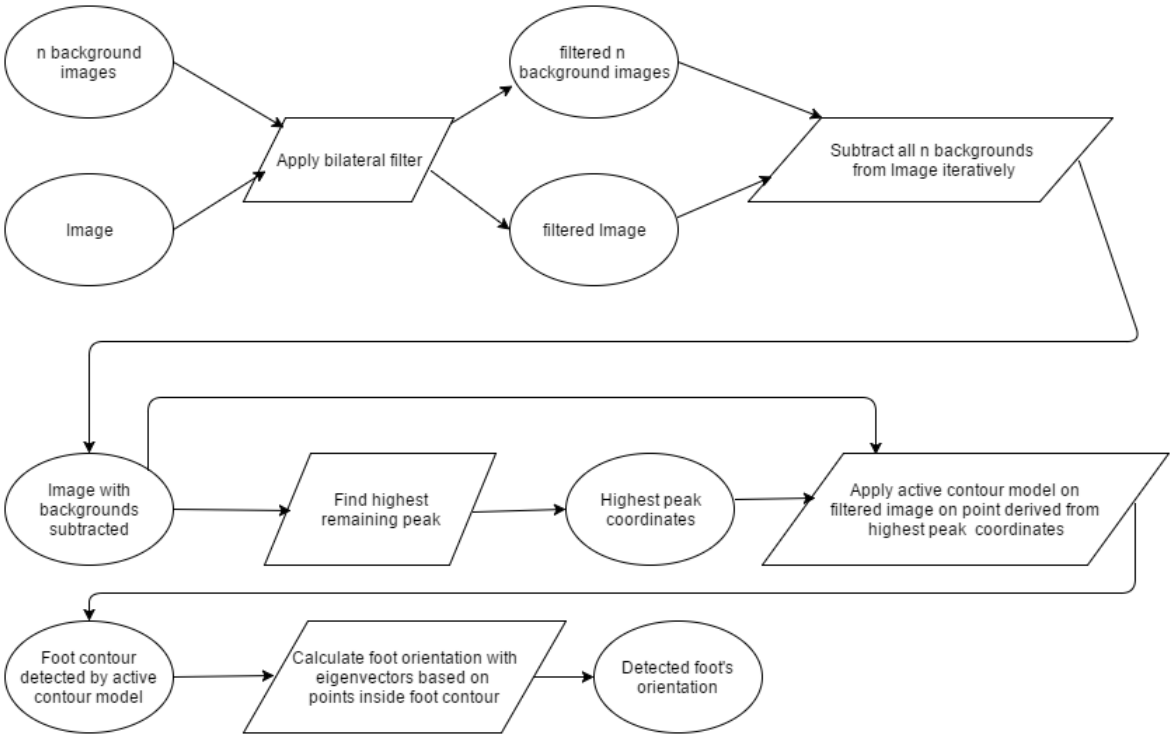

Figure 3: Algorithm flow diagram

3.4 Methodology

To begin with, the foot detection algorithm, in addition to the image with the foot to be detected, several background images are necessary. The background images and the main image must be taken with a fixed and static camera. For the benefit of the method, the photographing environmental lighting should be overhead or near camera pointing towards the subject.

3.4.1 Image enhancement

With all the described images present, the next step is enhancing the given images to improve background subtraction results.

Image enhancement is manipulating digitally stored in order to improve its quality, bring out specific features, or highlight certain characteristics. Image enhancement techniques can be divided into two main categories: spatial domain methods and frequency domain

methods. The former involves manipulating the pixels of the image directly, while the latter is about transforming the image via modifying the frequency. Frequency domain methods are useful for highlighting details and reducing the intensity variation across the image. Spatial domain methods are preferred choice for noise reduction, contrast and dynamic range modification, and edge enhancement and detection. Based on this, only spatial domain image enhancement methods are used during the processes of this algorithm.

For enhancing the images, histogram equalization, blur, Gaussian blur, median blur, and bilateral filtering methods were tested. The next following five subchapters focus on each of those filtering method explanation, result, and analysis.

3.4.1.1 Histogram equalization

Histogram equalization is useful technique for enhancing the contrast of an image.

An image can be represented by a matrix (m) sized rows (r) by columns (c) with matrix cell values the appropriate pixel values, usually ranging from 0 to 255. For the following sample, let's consider the highest possible value to be L, the normalized histogram of the image to be p and with a bin for every possible matrix cell value, and the histogram equalized image to be g. Therefore we get the following equation:

$$p_n = \frac{m_n}{m}$$

Where:

$N = 0, 1, \dots, L$

$m_n = number\ of\ cess\ with\ value\ n$

$m = total\ number\ of\ cells$

The image after applying the histogram equalization (g) can be defined by this equation:

$$g_{i,j} = floor(L \sum_{n=0}^{f_{i,j}} p_n)$$

Where the floor() is method for rounding down to the nearest integer.

The idea behind this transformation method comes from considering the image(f) and the equalized image(g) as continuous random variables X and Y on [0, L], where $Y = T(X)$

$$T(X) = L\int_0^X p_x(x)dx$$

The p_x is the probability density function of f.. T is the cumulative distribution function (CDF) of X multiplied by L. With the assumption that T is invertible and differentiable, it possible to show that Y defined by T(X) is distributed uniformly on [0,L]. Moreover, it can be shown that $p_Y(y) = \frac{1}{L}$

$$\int_0^y p_Y(z)dz = P(0 \leq Y \leq y) = P\big(0 \leq X \leq T^{-1}(y)\big) = \int_0^{T^{-1}} p_X(w)dw$$

$$\frac{d}{dy}\left(\int_0^y p_Y(z)dz\right) = p_y(y) = p_x(T^{-1}(y))\frac{d}{dy}(T^{-1}(y))$$

See figure 4 for an example of histogram equalization.

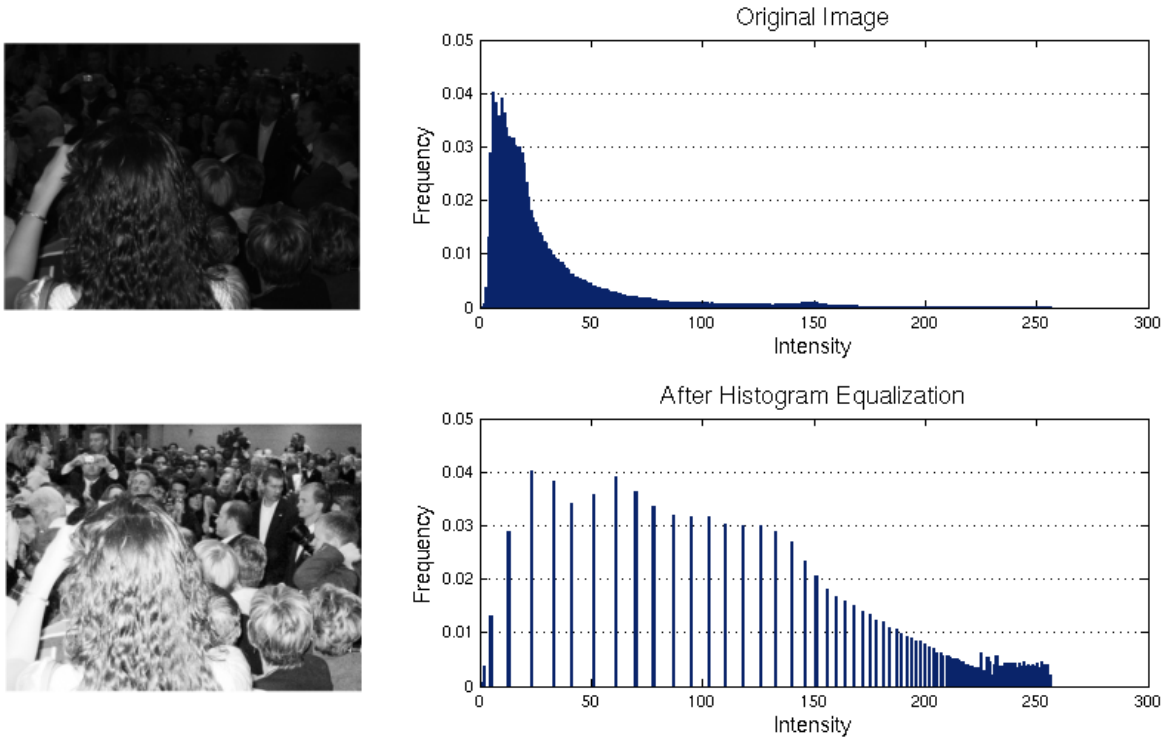

Figure 4: Histogram Equalization (7)

Enhancing images before background subtraction with histogram equalization resulted in worse background detection. The reason for it is that, with the addition of the new object in the image, the whole histogram is altered. Therefore applying the histogram for equalization produces significantly different results.

3.4.1.2 Average

Averaging is a quite simple image manipulation technique in image processing. The idea behind this method is to take the average value of all pixels under predetermined kernel and set the resulting value to the central cell. Essentially, the method is convolving the image with a normalized box filter.

A 5x5 normalized box filter looks like this:

$$K = \frac{1}{25}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Example of applying average filter with 3x3 kernel on 4x4 image:

1	40	10	90
20	35	20	10
25	80	75	15
5	65	50	0

The area under kernel is coloured green and in the sample it is used to recalculate the value in the cell currently filled with 35.

The sum of all values in area under the kernel: 1+40+10+20+35+20+25+80+75 = 306

The kernel is 3x3 therefore to find the average value, the sum 306 is to be divided with 3x3=9, so the new value for the cell, previously filled with 35 is 306/9=34

The image after applying the average filter on the one mentioned cell (pixel), the change is highlighted by yellow colour:

1	40	10	90
20	34	20	10
25	80	75	15
5	65	50	0

The following figure 5 is an example of image enhancement with average filter.

Figure 5: Example of average filtering - left is original, right is enhanced with average filter

Enhancing images with the averaging method, with 5x5 kernel, produced better result for background subtraction than unenhanced images or images enhanced with histogram equalization.

3.4.1.3 Gaussian blur

Gaussian blur is an image smoothing method, where the image is convolved with the Gaussian function.

One dimensional Gaussian function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where μ is the average of x.

When calculating average value, the centre point is the origin and μ = 0, therefore we get the following equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

Two dimensional Gaussian function:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

(8)

See figure 6 for an example of Gaussian blurring.

Figure 6: Example of Gaussian Blur (8)

Enhancing images with Gaussian blur with kernel size 5x5 produced slightly worse background subtraction than the averaging method.

3.4.1.4 Median blur

Median blur is an image manipulation technique that shares some similarities with average method above. For median blurring, a predetermined size kernel is moved over every pixel, like with the averaging method. Then, the values in under the kernel are sorted and the observed pixel value is substituted with the median value.

Example of median blur with 3x3 kernel on 4x4 image:

1	40	10	90
20	35	20	10
25	80	75	15
5	65	50	0

The area under kernel is marked by green and the values in those cells in sorted order is: 1, 10, 20, 20, 25, 35, 40, 75, 80 The values are sorted in order to retrieve the median value, which in this case is 25, Therefore we change the 35 in the original image with the median value 25.

The result after applying median filter on one pixel, the altered value is highlighted by yellow:

1	40	10	90
20	35	20	10
25	80	75	15
5	65	50	0

Following figure 7 is an example of median filtering.

Figure 7: Example of median filtering - left is original, right is enhanced with median filter

With median blurring, kernel equalling 5x5, the background subtraction results were somewhere between Averaging and Gaussian Blur results, quality wise.

3.4.1.5 Bilateral filtering

Bilateral filter is a non-linear, noise-reducing, and edge preserving smoothing filter. The bilateral filter combines domain and range filtering.

$$h(x) = k^{-1} \int_{\infty}^{\infty} \int_{\infty}^{\infty} f(\xi) c(f(\xi - x) s(f(\xi) - f(x)) d\xi$$

As the result, the bilateral filtering replaces the given pixel value (x in the equation) with the average of nearby similar pixel values.

(9)

See figure 8 for an example of bilateral filtering.

Figure 8: Example of bilateral filtering - left is original, right is enhanced with bilateral filter

Using bilateral filtering with background subtraction produced best results among the tested filtering methods.

3.4.2 Background subtraction

After processing the images with bilateral filtering, the background images can be subtracted from the main image. Background subtraction takes two images with identical dimensions and iterates over every pixel on both of the images simultaneously. If, on any location, the pixel values on both images are equal, the value in the same location in the output image is set to a pre destined value. Otherwise, the values in the output image are set from one of the original images, which is defined as foreground at start.

Example of background subtraction with 4x4 images:

Background:

25	25	25	25
25	25	25	25
25	25	25	25
25	25	25	25

Foreground:

25	25	25	25
25	80	85	25
25	25	55	25
25	25	25	25

Background - Foreground -> Subtraction (background replaced with 0)

25	25	25	25
25	25	25	25
25	25	25	25
25	25	25	25

25	25	25	25
25	80	85	25
25	25	55	25
25	25	25	25

0	0	0	0
0	80	85	0
0	0	55	0
0	0	0	0

For better quality result, many background images are subtracted, this helps to counter undesired effects caused, for example, by the flickering of lights. Subtracting multiple

background images is iterative process. Starting with the subject image and one background image, every pixel which is identical on both images, is removed, replaced with white, on the subject image. Then the process is continued with altered subject image and the next background image. By the end of this process, the resulting image should consist of completely white background, with only the subject - person or part of one, visible on the image. See figure 9 below.

Figure 9: Result of background subtraction

On the presented example, two types of defects can be seen. Firstly, the background isn't completely removed, patches of non-white areas exist. This is caused by the frequency of the lighting. In certain lighting conditions, on photographed images appear lighter and darker areas. See the following figure 10.

Figure 10: Effects of lighting on images.

Applying filtering reduces these effects, so does subtracting multiple background images. But in case the background images have similar effects, which all differ from those on the subject image, the background subtraction will produce imperfect results. Such is the case with the presented background subtraction example. These defects don't usually affect the foot detection algorithm, which will be explained in the next subchapter.

Before moving on to the foot location detection, there is a second type of defect noticeable on the example image. Background around the feet remains unremoved. This happens due to the shadow of the person altering the ground colouring, which then differs from the background images. To reduce the effect of shadows, which can decrease the quality of output of the algorithm, the lighting conditions should be selected so that it minimizes the amount of visible shadows on the image, especially in front of the subject.

3.4.3 Foot location detection

After the background of the image is removed, the next step is to find out the location of the foot on the image, so that we could move on to contouring it.

In the presented example (figure 9 Result of background subtraction) we can see one crucial aspect of the image, which helps to locate the person's foot. Even with the defects, the legs are the longest vertical non-white objects in the image. Therefore, by finding the location of the longest non-white vertical line, we can make the assumption that the searched foot is near the line's lowest point. The author has determined that the foot is at about 80% down from the top of the detected line.

3.4.4 Foot contouring with active contour model

With the foot's probable location determined, two steps of the algorithm remain: foot contouring and foot orientation determination. This subchapter focuses on contouring the foot.

To discover the foot contours, an outward expanding active contour model is applied.

3.4.4.1 Active contour model

The active contour model, also known as Snake, is an energy minimizing spline. It is influenced by image forces which pull it toward features like edges and lines. It is also guided by external constraint forces. Snake locks onto nearby edges, while localizing them accurately. A snake can be used for edge detection, line detection, detecting subjective contours, motion tracking and stereo matching.

In action, the internal spline force a piecewise smoothness constraint, while the image forces move the snake forward to prominent features of the image. The snake's energy functional can be described as the following, while the snake's position is represented parametrically:

$$v(s) = \big(c(s), y(s)\big)$$

Snake energy functional:

$$E^*_{snake} = \int_0^1 E_{snake}\big(v(s)\big)ds = \int_0^1 E_{int}\big(v(s)\big) + E_{image}\big(v(s)\big) + E_{con}\,(v(s))ds$$

In the equation, E_{int} is the spline's internal energy from bending, E_{image} is the image forces, and E_{con} is the external constraint forces.

The internal energy can be described thusly:

$$E_{int} = (\alpha(s)|v_s(s)|^2 + \beta(s)|v_{ss}(s)|^2)/2$$

The internal energy is composed by a first-order term and a second-order term. The $\alpha(s)$ is what controls the first-order term, and it makes the snake act similar to a membrane. The $\beta(s)$ affects the snake to behave as a thin plate. This is achieved due to the fact that $\beta(s)$ controls the second-order term.

The image forces can be described like this:

$$E_{image} = w_{line}\,E_{line} + w_{edge}\,E_{edge} + w_{term}\,E_{term}$$

(10)

For a depiction of snake advancement, see figure 11.

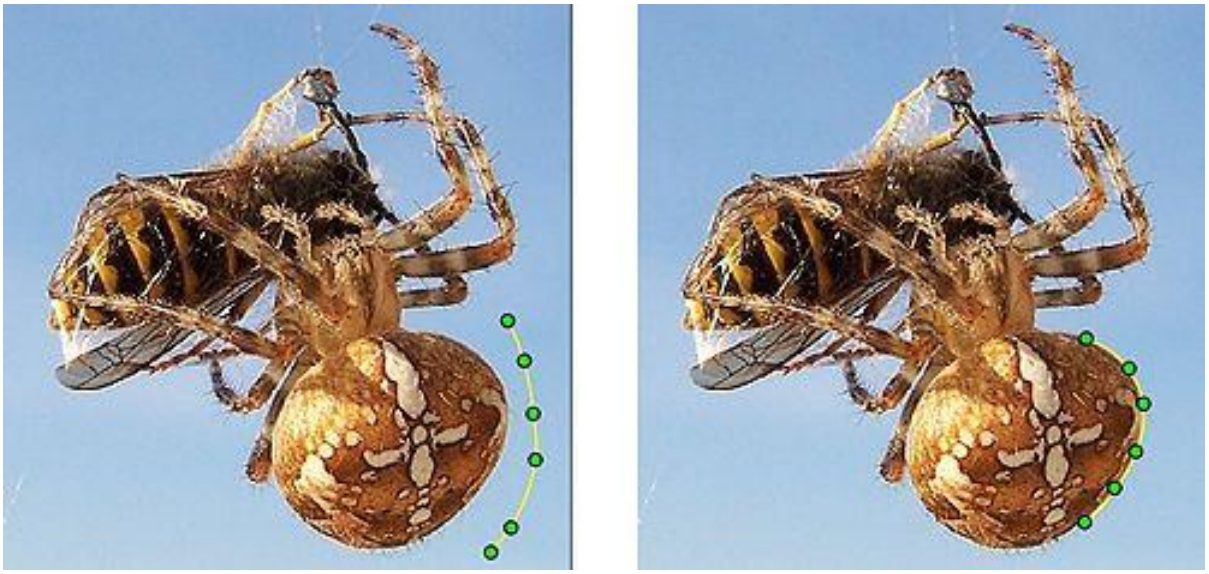

Figure 11: Example of snake advancement (11)

3.4.4.2 Applying active contour model - snake

As mentioned before, we use an outward progressing or expanding snake. The starting point is a small circle in the location deduced by the Foot location detection step. Then we let the snake grow for 200 iterations. After the iterations, the snake should have filled out the foot, stopping at its edges. Hence, outlining the foot.

3.4.5 Foot orientation

After applying the active contour model, we should have the data describing the edges of the foot. Then, the final step is in this algorithm is to determine the foot orientation based on the geographical data of the foot's edges. For calculating the foot's rotation, the pixels inside the foot outline are casted into data points. These data points can then be used to calculate eigenvectors, which enable us to determine the probable orientation of the foot.

3.5 Conclusion

In this chapter we covered the main problem statement of the thesis, the created algorithm design and architecture, and the methodology used. In the design and architecture sub-

chapter we depicted the whole process of the foot detection algorithm in a flow chart before focusing on every one of the five main steps in the process in the methodology part.

With the knowledge of the problem statement and methodology used to answer it, the next chapter covers the results from each of the main parts of the algorithm process described and analyses them with regard to the problem statement.

Chapter 4: Results and analysis

4.1 Introduction

This chapter is about the results produced by the method previously described, as well as analyses of the said results. The results are processed in the same order as the methods themselves were introduced in the previous chapter of this thesis. During the analyses part, the results are surveyed from the aspects of necessity in terms of the overall algorithm and the quality of the result including the effects varying quality may produce.

4.2 Picture enhancement and filtering

Necessity

Filtering the images with bilateral filter prepares the images for background subtraction. Enhancing the images smoothes out noise and anomalies created, for example, by the flickering of the environmental lighting. Smoothing should produce identical values for the background areas of the images. Mitigating the possibility an anomaly appearing in the image after smoothing or due to it, is the reason why multiple background images are used. So that the existence of an anomalous area in a background image after the filtering will not disrupt the algorithm.

Quality and its effect

Successfully filtered background and foreground images are (near) identical with the exception of the subject present in the foreground image. For that, any noise or anomalies present in any of the images must be smoothed and replaced with values equal to corresponding area values in other smoothed image.

Filtered images with identical background enable the background subtraction part of the algorithm to effectively subtract the background images and extract the person (subject) from the foreground image. See figure 12 for example of bilateral filtering applied to a foreground image.

Figure 12: Left: original, right: enhanced with bilateral filtering

Cases in which the filtering can be considered failure are when the smoothing enhances the anomalies in any of the images and therefore increases the differences between the foreground and background images. On these occasions, continuing with background subtraction, the anomalous areas could remain unremoved and be defined as parts of the foreground.

4.3 Background Subtraction

Necessity

Background subtraction extracts the subject from the foreground, which can then be used by the foot location detection part of the algorithm. Because the foot location detection works by finding the longest vertical line in the image with background subtracted, the background subtraction has a critical role in the algorithm.

Quality and its effect

A good quality background subtraction result would have only the subject remaining and all the background removed. Constantly perfect background subtraction results would improve the whole algorithm's result quality. But due to variety of possible photographing environments and lighting conditions, achieving flawless results every time, is near impossible. Regardless, decent results, where most of the background is removed and subject, with possibly some nearby shadows including, remain, are possible and imperative for the

functionality of this algorithm. See following figure 13 for a good background subtraction result.

Figure 13: Good background subtraction result

Outcome of the background subtraction method, where significant chunks of background remain after foreground extraction, could completely alter the outcome of the whole algorithm. Faulty outputs could be the result of failed image filtering, where noise or some other anomalies are increased rather than removed. Other possible cause of bad background subtraction is excessive subject's shadows present in the foreground image. These shadows can't be removed by smoothing and will be always left in by the subtraction. This can and must be countered by the selection of photographing environment and its lighting conditions. See figures 14 and 15 for low quality background subtraction results.

Figure 14: Bad background subtraction due to anomalies from lighting

Figure 15: Bad background subtraction result due to shadows

4.4 Foot location detection

Necessity

Foot location detection determines the location of the foot based on the longest vertical line remaining in the image of extracted foreground. The following step, the active contour model detection, requires a starting point, which must be inside the person's foot, to detect its contours. If the foot location detection fails and produces coordinates which aren't inside the foot, the active contour model will fail as well and as the whole algorithm. Therefore the foot location detection fills a critical role in the algorithm.

Quality and its effect

A successful foot location detection outputs coordinates which point to an area inside the foot. Correctly identified foot position enables the algorithm to move on to the next step

In contrast, a failed foot location detection would produce coordinates which don't point to the feet. This can be either due to an anomaly passed through image filtering and background subtraction or a shadow, which is then mistaken for the person. The other possible reason for incorrect foot location detection can be miss-calculating the foot height from the identified peak. In this case the output would point to location above the actual foot and the active contour model would extract a random blob from the person's leg instead of the foot.

4.5 Active contour model (Snake)

Necessity

The active contour model or Snake, as it is also known, starts from a given location and shape. The location is gained from the foot location detection and the shape is a simple circle with a small radius. During the Snake's method runtime, the starting circle expands until it outlines the foot.

Quality and its effect

Expected outcome of the Snake method is the extracted outline of the foot. Success of the method relies on two key aspects. Firstly, the location has to be inside the foot, this is handled by the previously explained foot location detection. Secondly, the active contour model utilized within this algorithm relies on a set number of iterations during which the shape (snake) moves and increases one step outward. For the snake to expand to the foot outline, the number of iterations must be sufficient, but not too much. See figures 16 and 17 for successful results of utilizing the active contour model.

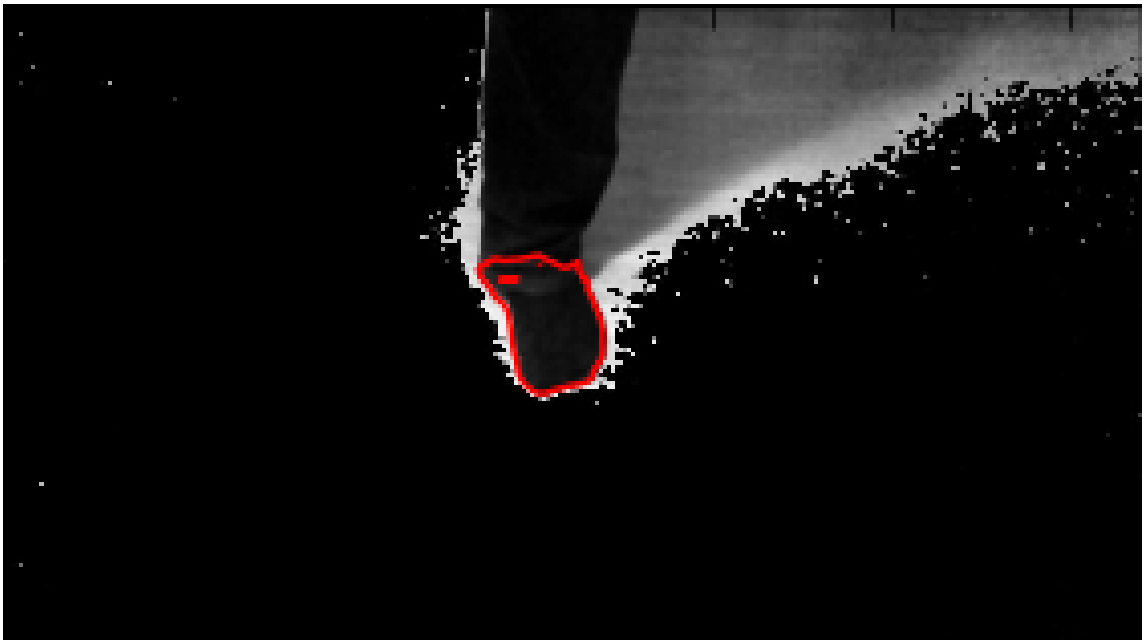

Figure 16: Successful snake utilization, subtracted background replaced with black

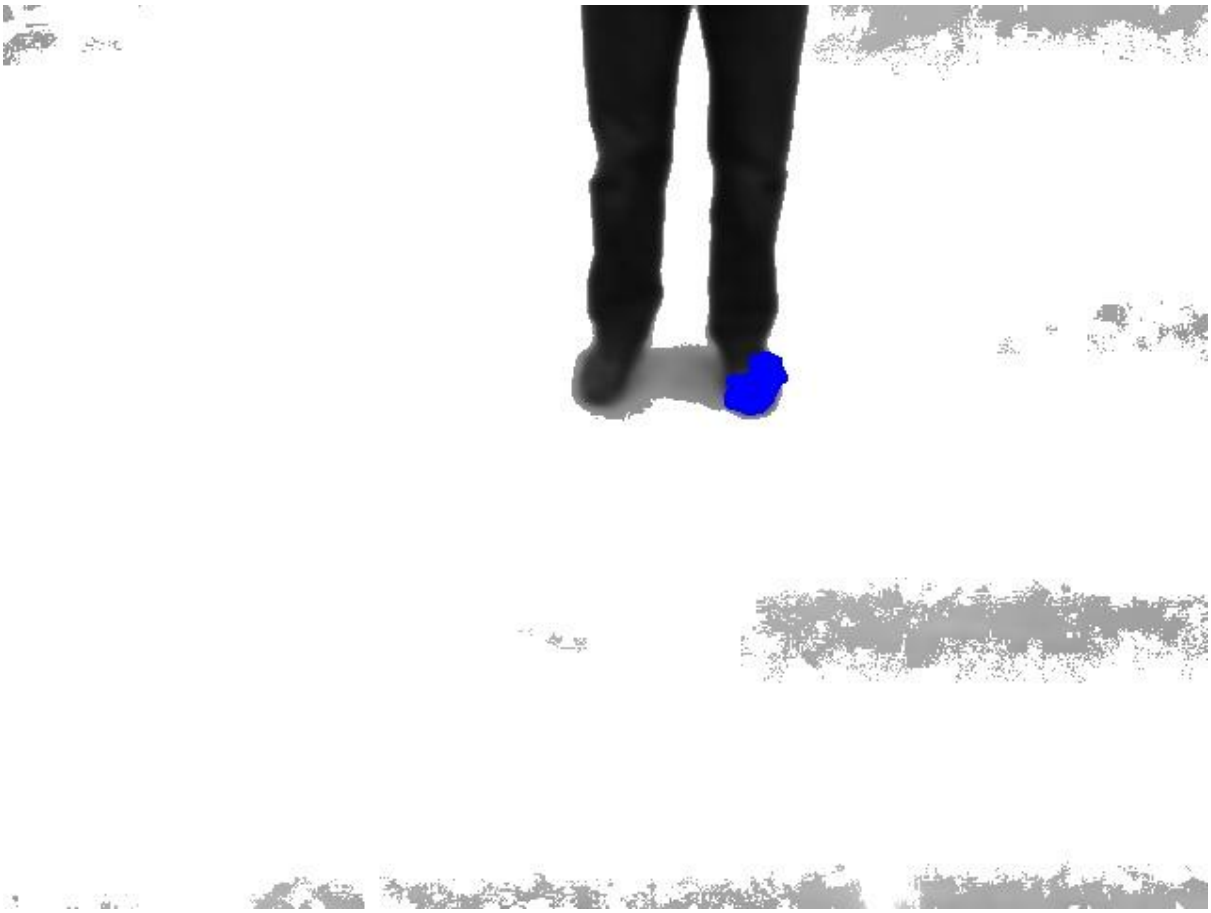

Figure 17: Detected foot (blue) projected onto original image

If the number of iterations is too large, the snake can expand past the contour of the foot. In case of too few iterations, the snake's expansion will not reach the foot contours. Both of those outcomes would be unsuitable for the foot detection algorithm. The active contour model could also fail to produce required results, if the starting point is not inside the foot.

4.6 Foot orientation

Necessity

Augmented reality applications add computer generated graphical elements and models onto real world items. In order for the projected model to look realistic, the applications require the orientation of the item. Without the orientation, the augmented reality applications wouldn't know the degree of rotation required to fit the models onto items. Therefore determining the orientation of the foot is an important step for an algorithm developed for augmented reality applications.

Quality and its effect

Determining the orientation of the foot is the last step in this algorithm. Therefore the output of this part is not used in any further processes in this algorithm. And because the foot's contours are received from the previous part, a low quality result would not cripple the outcome of the whole algorithm.

A falsely determined foot rotation degree is still undesired as this decreases the usability of this algorithm with augmented reality applications. Incorrect orientation can be caused by a bad quality foot contour detection. If the active contour model expands too far and includes part of the leg in addition to the foot, then the vertically oriented areas of the leg will alter the eigenvector calculations, which could result in incorrect foot orientation.

While a low quality result from foot orientation detection isn't a critical flaw for the algorithm from the point of view of foot detection, a positive rotation outcome significantly increases the usability of the algorithm with augmented reality applications. See figure 18 for an example of an output from the foot orientation determination step.

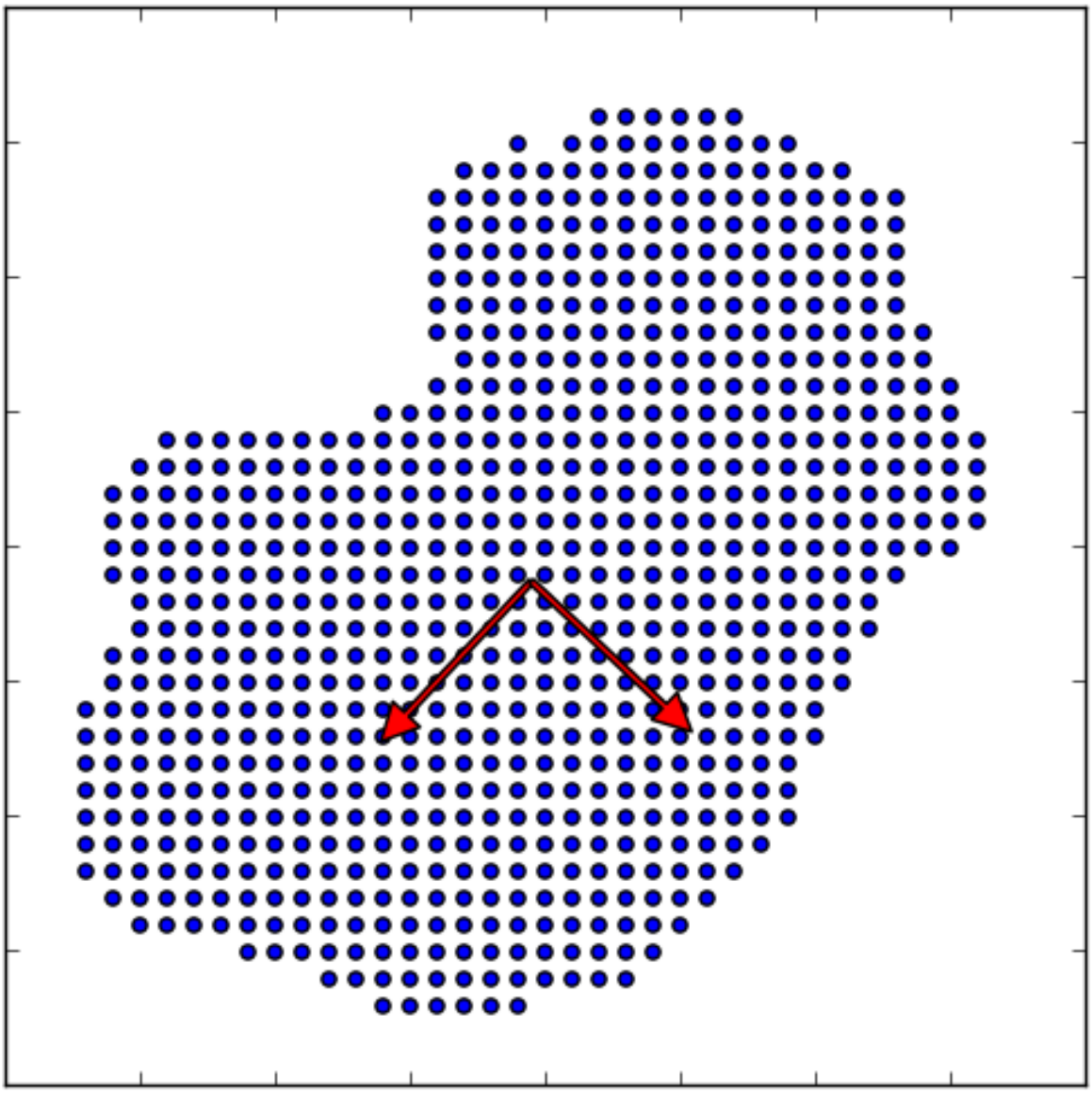

Figure 18: Detected foot plotted out via data points with arrows representing data growth directions

4.7 Conclusion

This chapter focused on the results of the different parts of the algorithm as well as their analysis. All the steps: picture enhancement, background subtraction, foot location detection, active contour model, and foot orientation determination, were covered individually.

Possible points of failure were discussed as well as the requirements for the outputs of each of the steps for successful algorithm functionality.

Next chapter is the conclusion of the whole thesis.

Chapter 5: Conclusion

5.1 Conclusion

This thesis presents an algorithm for foot detection dedicated to footwear augmented reality applications. The purpose of the foot detection method created within this thesis is to provide a reliable and robust baseline method which can be in conjunction with footwear augmented reality applications as well as bases for future foot detection methods.

The foot detection algorithm is divided into 5 steps: image enhancement, background subtraction, foot location detection, active contour model, and foot orientation.

Each of those steps is covered thoroughly in regard of necessity of the step, actions performed during the process in detail, and the analysis of the steps' output.

5.2 Future perspectives

Idea behind this thesis and the algorithm created during its creation is to provide a base method of foot detection usable with augmented reality footwear applications and with future work in computer vision field with focus on foot detection. Due to procedural realities discovered during the development process, the created algorithm has limitations, which hinder its intended universal usability. With this in mind, the future perspectives related to this thesis are probably research oriented. This dissertation gives a good overview of some practices usable for foot detection as well as their limitations and possible solutions for countering those.

Annex 1:

1.1 Pedestrian detection

Considerable advancements have been made in image processing and computer vision by automotive industry, specifically in regard of pedestrian detection. The article (12) describes an efficient real-time algorithm for pedestrian tracking and detection. The algorithm is based on using far infrared image. The reasoning behind this is that on far infrared image pedestrians' heads are the lightest objects and therefore a good starting place for image processing. (12)

The algorithm starts out by utilizing the SURF (Speeded up robust features) detector on the image and sorting the SURF points based on its Laplacian value. The Laplacian value is positive for dark areas and negative for light regions. Only points with negative Laplacian value are kept for further calculations, due to the previously mentioned characteristic that pedestrian heads are the lightest objects. (12)

Onward, the retained SURF points are clustered. To keep the time and space complexity down, reciprocal nearest neighbour search is used for clustering. The clustering is performed in the following manner – two clusters are merged if the Euclidean between the given cluster descriptors is below a threshold. To manage the thresholding, a tree structure is used. In the tree every level stands for a specific clustering threshold. (12)

After creating and evaluating the tree structure, a hierarchical codebook is utilized for matching. To accelerate the process, the matching is started from the top node and is recursively applied to lower levels only if the Euclidean distance to the top node centroid is smaller than the node radius. During this parsing, each extracted SURF point is evaluated with an activation value. (12)

After applying the hierarchical codebook, bounding boxes are generated based on matching features. The bounding boxes represent pedestrian full-body estimation and their height/width ratio is determined by training images. With the bounding boxes present, they are evaluated with SVM (Support Vector Machine). Bounding boxes for which the

evaluation determined to be non-pedestrian are removed, and highest valued bounding boxes are kept. (12)

The SVM evaluation is determined by the combination of hierarchical codebook based local features and SURF based global features. The hierarchical codebook based local features are essentially gathered by accumulating the already mentioned activation values. The SURF based global features are fast-computing global features, which characterize the texture and shape of the objects and therefore provide complementary information. (12)

Pedestrians are tracked with the use of temporal feature matching. This part can be divided into 2 steps: feature matching and hypotheses handling. For each of the pedestrians detected, the tracking algorithm is initialized. The tracking algorithm is based on nearest neighbour with the addition of the ambiguity rejection method. A number of hypothetical pedestrians are generated based on the image and a match is validated only if the ratio of distances is lower than a threshold derived from experiments. (12)

The algorithm described in (12) works well for detecting pedestrians in real-time, although the drawbacks in regards of using it for footwear augmented reality are quite obvious – the algorithm would have to be modified to detect feet instead of pedestrians and the implicit use of far infrared camera has to be countered.

1.2 Salient object detection

The algorithm in the previous chapter was based on the aspect that pedestrian heads are the lightest object in a far infrared image. By turning the desired object lighter than the rest of the image from a normal camera, the aforementioned algorithm could be used without a far infrared camera image.

This can be achieved with the use of FASA: Fast, Accurate, and Size-Aware Salient Object detection. (13)

The method described in (13) combines a global contrast map with a probability of saliency. In the beginning the image is quantized in order to reduce the number of colour pre-

sent in the image. Then the variances of the quantized colours and the spatial centre are calculated, to estimate the position and the size of the salient object. The probability of saliency is computed by evaluating an object model with the calculation results. Finally a single saliency map is generated by merging together the saliency probabilities of the contrast values and the colours. (13)

The FASA method is shown to be effective, capable of successfully detecting salient objects. Also it is reported to be reasonably fast, capable of processing image with a resolution of 400 x 400 pixels in 6 milliseconds. Taking all of the above into consideration, it might be possible to use the FASA method in real time image processing.

Bibliography

1. **Dictionaries, Oxford.** Augmented reality. *Oxford Dictionaries.* [Online] [Cited: 17 May 2016.] http://www.oxforddictionaries.com/definition/english/augmented-reality.

2. *Detection of Moving Object Based On Background Subtraction.* **Pawaskar, Mahesh C., Narkhede, N S and Athalye, Saurabh S.** 3, s.l. : International Journal of Emerging Trend & Technology in Computer Science, 2014, Vol. 3.

3. *Background Subtraction for Object Detection Under Varying Environments.* **Mashak, Saeed Vahabi, et al.** Johor, Malaysia : International Conference of Soft Computing and Pattern Recognition, 2010.

4. **Eisert, Peter, Fechteler, Philipp and Rurainsky, Jürgen.** *3-D Tracking of Shoes for Virtual Mirror Applications.* Berlin, Germany : s.n., June 2008.

5. **Eisert, Peter, Rurainsky, Jürgen and Fechteler, Philipp.** *VIRTUAL MIRROR: REAL-TIME TRACKING OF SHOES IN AUGMENTED REALITY.* Berlin, Germany : s.n., 2008.

6. *Augmented reality-based design customization of footwear.* **Yuan-Ping, Luh, et al.** 24 February 2012, Springer Science+Business.

7. **Fishbaugh, James.** Histogram Equalization. *University of Utah.* [Online] [Cited: 12 May 2016.] http://www.cs.utah.edu/~jfishbau/improc/project2/.

8. **阮一峰.** Gaussian Blur Algorithm. *PixelsTech.* [Online] [Cited: 12 May 2016.] http://www.pixelstech.net/article/1353768112-Gaussian-Blur-Algorithm.

9. Bilateral Filtering for Gray and Color Images. *University of Edinburgh School of Informatics.* [Online] [Cited: 12 May 2016.] http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MANDUCHI1/Bilateral_Filtering.html.

10. **M. Kass, A. Witkin, D. Terzopoulos.** *Snake: Active Contour Models.* Palo Alto : s.n., 1988.

11. Active Contour Model. *Wikipedia.* [Online] [Cited: 12 May 2016.] https://en.wikipedia.org/wiki/Active_contour_model.

12. *Pedestrian Detection in Far-Infrared Daytime Images Using a.* **Besbes, Bassem, et al.** 2015, Sensors.

13. *FASA: Fast, Accurate, and Size-Aware.* **Yildirim, Gökhan and Süsstrunk, Sabine.** Lausanne : Springer, 2015.

Appendix

License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Gustav Amer** (date of birth: 02.02.1991),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

of my thesis

Foot Detection Method for Footwear Augmented Reality Applications

supervised by Amnir Hadachi, PhD

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **13.05.2016**