Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics

Master's Thesis

Person Body Height Measurement by Using Surveillance Cameras

Bc. Jan Fabián

Supervisor: Ing. Jan Čech, Ph.D.

Study Programme: Open Informatics

Field of Study: Computer Vision and Image Processing

7. ledna 2015

Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Cybernetics

DIPLOMA THESIS ASSIGNMENT

Student: Bc. Jan F a b i á n

Study programme: Open Informatics

Specialisation: Computer Vision and Image Processing

Title of Diploma Thesis: Person Body Height Measurement by Using Surveillance Cameras

Guidelines:

The body height measurement can be used as a soft-biometric marker to identify a person. It may be important when the face recognition is not possible due to low resolution videos or when the subject is not facing the camera.

In the diploma thesis, do:
1. Propose a method for a monocular body height measurement from a single image and a short sequence. Assume the camera is fully calibrated and the ground plane location is known.
2. Design a ground-truth experiment.
3. Evaluate the proposed method with respect to the setup and the camera parameters.

Bibliography/Sources:

[1] A. Criminisi, I. Reid and A. Zisserman: Single View Metrology. International Journal of Computer Vision 40(2), 2000.
[2] Richard Hartley and Andrew Zisserman: Multiple View Geometry in Computer Vision Second Edition, Cambridge University Press, March 2004.
[3] Jan Cech, Vojtech Franc, Jiri Matas: A 3D Approach to Facial Landmarks: Detection, Refinement, and Tracking. In Proc. ICPR, 2014.

Diploma Thesis Supervisor: Ing. Jan Čech, Ph.D.

Valid until: the end of the winter semester of academic year 2015/2016

L.S.

doc. Dr. Ing. Jan Kybic
Head of Department

prof. Ing. Pavel Ripka, CSc.
Dean

Prague, September 17, 2014

České vysoké učení technické v Praze
Fakulta elektrotechnická

Katedra kybernetiky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Jan F a b i á n

Studijní program: Otevřená informatika (magisterský)

Obor: Počítačové vidění a digitální obraz

Název tématu: Měření výšky postavy z dohledových kamer

Pokyny pro vypracování:

Výška člověka může být jeden z atributů určení identity člověka. Především, pokud máme k dispozici záběr s nízkým rozlišením, kde není možná identifikace pomocí obličeje, nebo v záběrech, kde osoba stojí zády ke kameře.

V rámci diplomové práce:

1. Navrhněte automatickou metodu pro měření výšky postavy z jediné kamery, a to z jediného snímku i z krátké sekvence. Předpokládejte, že celková kalibrace kamery a roviny podlahy je známá.
2. Navrhněte experiment s referenčními daty.
3. Vyhodnoťte přesnost metody vzhledem k parametrům kamery a jejím umístění ve scéně.

Seznam odborné literatury:

[1] A. Criminisi, I. Reid and A. Zisserman: Single View Metrology. International Journal of Computer Vision 40(2), 2000.
[2] Richard Hartley and Andrew Zisserman: Multiple View Geometry in Computer Vision Second Edition, Cambridge University Press, March 2004.
[3] Jan Cech, Vojtech Franc, Jiri Matas: A 3D Approach to Facial Landmarks: Detection, Refinement, and Tracking. In Proc. ICPR, 2014.

Vedoucí diplomové práce: Ing. Jan Čech, Ph.D.

Platnost zadání: do konce zimního semestru 2015/2016

L.S.

doc. Dr. Ing. Jan Kybic
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.
Děkan

V Praze dne 17. 9. 2014

Aknowledgements

I would like to express my gratitude to my advisor Ing. Jan Čech, Ph.D., for the continuous support during the whole work on the diploma thesis. His guidance helped me in writing this thesis.
I am heartily thankful to my parents for supporting me throughout my life.

Prohlášení autora práce

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne 3.ledna 2015 .....

Abstract

This thesis deals with a person height measurement recorded by a single video cam. There are a few methods for the height measurement based on a reference length in an image. Such height estimation might be inaccurate. If we have a video sequence, it is possible to fuse height measurement from each frame to estimate an error and find the most likely height of the recorded person. In this thesis we formulate two new methods of the height estimation from a single image and two approaches how to fuse these sets of estimations into overall height estimation for the whole video sequence. Both methods require a calibrated camera. On top of that to use the first method we need to see the top and bottom point of a human figure and know a position of the ground floor. The second method requires visible and recognizable face of the person. These two methods complement each other to improve the overall height estimation. We did test measurements, which show, that the overall error of the estimation was under $< 4\%$.

Abstrakt

Tato práce se zabývá měřením výšky člověka zachyceného na kamerovém záznamu jedné kamery. Existuje několik metod umožňující změřit velikost člověka ze snímku pomocí referenční velikosti známého objektu ve scéně. Takto provedené měření může být nepřesné. Pokud ale máme k dispozici videozáznam, je možné spojit tato měření, odhadnout chybu a nejpravděpodobnější výšku pozorovaného objektu. V této práci navrhujeme dvě nové metody měření výšky z jednoho snímku a dva způsoby, jak spojit částečné odhady z jednotlivých snímků do celkového odhadu z videa. Obě metody odhadu výšky požadují kalibrovanou kameru. K použití první z metod je dále potřeba znalost umístění podlahy ve scéně a ze záznamu viditelnou hlavu a nohy pozorovaného člověka. Druhá metoda požaduje viditelný a rozpoznatelný obličej. Tyto dvě metody se navzájem doplňují a tím zlepšují celkovou přesnost v rámci celého záznamu. Provedli jsme měření na vlastních záznamech, které ukazálo, že chyba celkového odhadu se pomocí našeho způsobu pohybovala okolo $< 4\%$.

Contents

Chapter 1

Introduction

The aim of this thesis is to extract human height from a video sequence. An input to our algorithm is a video containing a record of a single moving person. An output is a height estimation of the person. The purpose of such estimation might vary.

Video camera surveillance systems are spread all around nonstop recording public places. They may record a crime or a law violation. The record is one of the evidence in the following investigation and contains information about the scene. When there is a person visible in the sequence, but the face can not be extracted, for instance the individual stands back to the camera or the image quality is poor, we search for another characteristics of the person. Height might be used as such identification attribute. Height can be also used as an additional support feature for human recognition.

One of the main problem of height estimation from a single monocular image is that we can not project an image point back to the scene as an 3D point. Hence we need some additional knowledge about the scene. The main approaches uses reference length of a known object in the scene. In this thesis we introduce two other approaches. One uses a position of the ground floor, the second use the facial landmark detector [4] and an average interpupillary distance to estimate the height of the person. These two methods complement each other. There are situations in which we can not use both of them, but one may still work and produce an estimation. Example of such situation is, when a human stands back to the camera, hence we can not detect his face, but his head and feet are in the camera range. On the other side if the person stands near to the camera and his feet are out of the camera range, we can detect his face but can not measure distance between upper and bottom boundary pixel in the image. Using both of them should minimize number of frames with no height estimation.

If we detect and estimate the height from a single image, there might be an error caused by uncertainties of the detecting. But we have whole video sequence, hence we have multiple height estimations, which can be merged together. Such overall estimation is more accurate. To find this overall estimation, we take the set of measurements as a statistical model containing probabilities about the height from each single measurement. We search for the most likely parameters of the model. These parameters are the height and interpupillary distance in our case.

The contributions of this thesis are:

1. We propose two methods for a human body height estimation from a single image, which complement each other.
2. We fuse multiple estimations from a video and We formulate an equation. These methods are based on maximum-likelihood and Bayes estimation strategy.
3. We perform multiple test of real video data and compare the proposed methods.

This thesis is organized as follows. In the next chapter we discuss present state of art in human body height estimation from a monocular image. In chapter 3 we present our methods of height estimation using a single image and fuse these approaches into one overall estimation. In chapter 4 we present experiment video data and the result of our methods applied on them. The final chapter 5 concludes this thesis.

Chapter 2

Background and Related Word

There are significant number of works dealing with the problem of height estimation. They differ in using calibrated or uncalibrated setup. In the uncalibrated case, a general approach for this problem is first computing the length ratio of the line segment with respect to another line segment in the scene of known length, called the reference length [5] [9, p. 221-222]. The main difficulty about this approach is that the reference length is not always available in the scene. Therefore other methods use a statistical model of human body [1] and [7], which calculates the height using vertical proportions.

In [5] various methods for estimating the height of an object using an uncalibrated camera are described. They describe how the affine 3D geometry of a scene can be measured by a single perspective image. Their approach is based on the estimation of the vanishing line of the ground plane and the vertical vanishing point. Nevertheless, errors appear in the estimated height of the object since these accrue from errors in estimating the ground plane and in estimating the 3D position of the object

In [11] authors expect knowledge of focal length, vanishing point and camera position, its height from the ground and pitch angle. Using ratio between the height of the camera in the image and the scene with the pixel height of the object, they obtain the physical height of the object. This approach is sensitive to correctly defined pitch angle.

In [9, p. 221-222] authors expect knowledge of a reference height of another object in the scene, which is parallel to the measured one.

The other approach is to use a statistical method in the Bayesian-like framework [1]. Such method does not need calibrated setup.

Chapter 3

Proposed Method

In this chapter we present a method to measure person height from a video sequence. Firstly we show two independent approaches to estimate the height from a single image. We derive an error of each approach and finally we aggregate the data from each frames and calculate the overall estimation from the whole video sequence. The chapter is organized as follows. In the section 3.1 we describe the method based on triangles similarity, then in the section 3.2 we propose an alternative method using the face detector [4]. In the last chapter we fuse the two methods and estimate the height of a person from the video sequence using maximum-likelihood and Bayesian estimation. We start by defining assumptions needed to determine the height by using either of the two methods.

3.0.1 Basic Assumption

- The camera is fully calibrated. We know extrinsic and intrinsic camera parameters.
- Position of the ground plane in the scene is known.
- Person is standing upright.

3.1 Triangles similarity method (TSM)

This part presents method for height measurement using triangles similarity (TSM). We derive the basic equation in the 3.1.1. We use detected points as input parameters for this equation, hence in section 3.1.3 we show how to find these points and in section 3.1.4 we use knowledge of the ground to refine the point detection.

This method is based on calculation On the top of the basic assumption this method requires the person to be fully visible in the frame, since we need image points of the top and bottom boundary of the person.

3.1.1 Overview

This method is based on calculation a ratio between the top and bottom scene point of the person, in other words between the head and the feet. Therefore on the top of the basic

assumption this method requires the person to be fully visible in the frame, since we need image points of the top and bottom boundary of the person.

The setup of the scene is illustrated in the fig. 3.1. Uppercase letters are the 3D scene points, while lowercase letters are their projection onto the image plane. A person is bounded by points $\mathbf{T}$ and $\mathbf{B}$ in the scene, the camera center is located at the point $\mathbf{C}$. The camera captures the scene under the pitch angle β. The camera vector rays $\mathbf{t}$ and $\mathbf{b}$ goes through $\mathbf{T}$ and $\mathbf{B}$ respectively. Points $\mathbf{T_G}$ and $\mathbf{B}$ show intersection of the rays with the ground plane. We observe that the two line segments in the scene, one created by the points $\mathbf{C}$ and $\mathbf{C_G}$ and the second created by the points $\mathbf{T}$ and $\mathbf{B}$, are both normal vectors of the ground plane.

We introduce length variables l_1 and l_2. l_1 holds the distance between $\mathbf{C_G}$ and $\mathbf{B}$, that is the distance between ground points of the camera and the person. l_2 is the distance between the intersection of the camera ray vector and the ground $\mathbf{T_G}$ and the camera ground point $\mathbf{C_G}$.

We find the person height h_o by using the equation

$$\frac{l_2}{h_c} = \frac{l_2 - l_1}{h_o} \tag{3.1}$$

which follows from similarity between triangles $\mathbf{CC_GT_G}$ and $\mathbf{TBT_G}$. h_c is the height of the camera and it is known from the calibration of the camera. In this equation we search for the h_o, hence we need to derive l_2 and l_1.

Firstly we start by deriving formula for camera ray vectors going from the center of the camera through the points on the image plane. Then we find the intersections of these camera ray vectors and the ground plane. We use these intersections to calculate the distances l_1 and l_2.

3.1.2 Height estimation

Camera ray vectors Assume we are given the image points $\mathbf{t} = \begin{bmatrix} u_t & v_t & 1 \end{bmatrix}^\intercal, \mathbf{b} = \begin{bmatrix} u_b & v_b & 1 \end{bmatrix}^\intercal$, which correspond to the top point and bottom point of the person detected in the image. These points are projection of the scene points $\mathbf{T}$ and $\mathbf{B}$ onto the image plane. Assuming we have projection matrix $\mathtt{P}$ of the camera, in the form $\mathtt{P} = \begin{bmatrix} \mathtt{Q} & \mathbf{p_4} \end{bmatrix}$, the camera rays are defined as

$$\begin{aligned} \mathbf{t} &= \mathtt{Q}^{-1}\mathbf{t} \\ \mathbf{b} &= \mathtt{Q}^{-1}\mathbf{b} \end{aligned} \tag{3.2}$$

Having these vectors in the 3D scene space, we now define the center of the camera in the scene $\mathbf{C} = -\mathtt{Q}^{-1}\mathbf{p_4}$. This gives us lines going through the center of the camera in the direction of the given camera ray vectors.

$$\begin{aligned} t &= \mathbf{C} + d_t\mathbf{t} \\ b &= \mathbf{C} + d_b\mathbf{b} \end{aligned} \tag{3.3}$$

Assuming we know the ground plane $\mathbf{G} : \mathbf{n}^\intercal(\mathbf{p} - \mathbf{p_0}) = 0$, where $\mathbf{p_0}$ is a scene point, that lies on the ground plane, and $\mathbf{n}$ is the normal vector of the ground plane.

Now we have to find such points on lines t and b, which lie also on the ground plane. This is done by finding appropriate line parameters d_t and d_b from equation (3.3). This gives us the points $\mathbf{B}$ and $\mathbf{T_G}$.

Intersection of rays with the ground plane To find the intersection of the lines t and b and the ground plane $\mathbf{G}$ we substitute the equations 3.3 into the plane equation $\mathbf{n}^\intercal(\mathbf{p}-\mathbf{p_0})=0$ and we get

$$\begin{aligned} \mathbf{n}^\intercal(\mathbf{C}+d_t\mathbf{t}-\mathbf{p_0}) &= 0 \\ \mathbf{n}^\intercal(\mathbf{C}+d_b\mathbf{b}-\mathbf{p_0}) &= 0 \end{aligned} \tag{3.4}$$

after rearranging the variables, we solve the equations for the d_t and d_b

$$\begin{aligned} d_t &= \frac{\mathbf{n}^\intercal(\mathbf{p_0}-\mathbf{C})}{\mathbf{n}^\intercal\mathbf{t}} \\ d_b &= \frac{\mathbf{n}^\intercal(\mathbf{p_0}-\mathbf{C})}{\mathbf{n}^\intercal\mathbf{b}} \end{aligned} \tag{3.5}$$

Having the parameters d_t and d_b we define the scene ground points using the equations 3.3 and we get

$$\begin{aligned} \mathbf{T_G} &= \mathbf{C}+\frac{\mathbf{n}^\intercal(\mathbf{p_0}-\mathbf{C})}{\mathbf{n}^\intercal\mathbf{t}}\mathbf{t} \\ \mathbf{B} &= \mathbf{C}+\frac{\mathbf{n}^\intercal(\mathbf{p_0}-\mathbf{C})}{\mathbf{n}^\intercal\mathbf{b}}\mathbf{b} \end{aligned} \tag{3.6}$$

We use the scene ground points $\mathbf{T_G}$ and $\mathbf{B}$ to define the length variable l_1 and l_2. In the figure 3.1 we see, that l_1 is the distance between the scene ground points $\mathbf{C_G}$ and $\mathbf{B}$ and l_2 is the distance between the scene ground points $\mathbf{C_G}$ and $\mathbf{T_G}$, hence we derive the point $\mathbf{C_G}$ using the same approach as in the 3.6.

Point $\mathbf{C_G}$ is the closest point to the camera center $\mathbf{C}$ on the ground plane $\mathbf{G}$. We search for the intersection of the line perpendicular to the ground plane with the ground plane itself. This line goes through the scene point $\mathbf{C}$ and it is determined by the equation $c=\mathbf{C}+d_c\mathbf{n}$. We substitute the equation for the line into the equation for the plane as in the 3.4 and get the equation for the parameter $d_c=\frac{\mathbf{n}^\intercal(\mathbf{p_0}-\mathbf{C})}{\mathbf{n}^\intercal\mathbf{n}}$. Then for the scene ground point $\mathbf{C_G}$, which is the closest to the camera center $\mathbf{C}$, holds

$$\mathbf{C_G}=\mathbf{C}+\frac{\mathbf{n}^\intercal(\mathbf{p_0}-\mathbf{C})}{\mathbf{n}^\intercal\mathbf{n}}\mathbf{n} \tag{3.7}$$

We evaluate the length l_1 and l_2 as follows

$$\begin{aligned} l_1 &= \|\mathbf{C_G}-\mathbf{B}\|_2 \\ l_2 &= \|\mathbf{C_G},\mathbf{T_G}\|_2 \end{aligned} \tag{3.8}$$

Having the lengths l_1 and l_2 we recall equation (3.1) and solve it for h_0

$$h_o=\frac{(l_2-l_1)h_c}{l_2} \tag{3.9}$$

and by substitution the Euclidean distance function for the length we obtain the height of the person in centimeters

$$\boxed{h_o=\frac{(\|\mathbf{C_G}-\mathbf{T_G}\|_2-\|\mathbf{C_G}-\mathbf{B})\|_2 h_c}{\|\mathbf{C_G}-\mathbf{T_G}\|_2}} \tag{3.10}$$

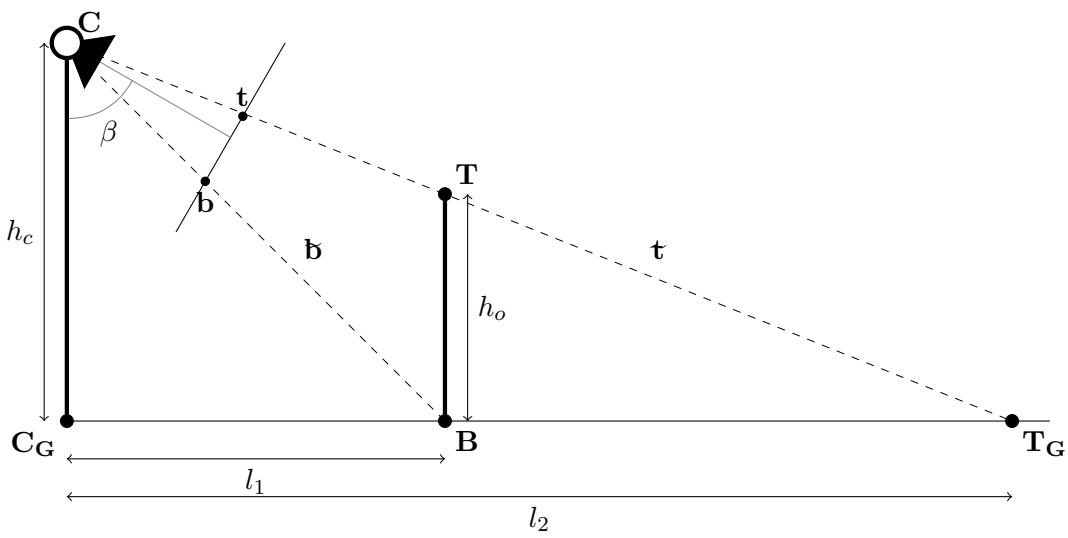


Figure 3.1: Overview of scene, which is captured by camera C. Object is defined by points B and T.

We defined the equation for the height estimation from the two image points **t** and **b**, which are the projection of the head and feet of a person. In the next subsection 3.1.3 we show how to detect these image points from a video sequence automatically.

3.1.3 Automatic detection

The method in section 3.1.2 needs image points **t** and **b** to be given. We will present procedure for for automatic detection of these points by using the background subtraction. As we use the scene points **T** and **B** for the top and bottom boundaries of the person, in this section we inspect obtaining their image projections. We define the general case, where is no relation between the scene points **T** and **B**. But as long as a human body stands upright, we assume, that the head and feet are located on the normal vector of the ground plane. Hence in the in the subsection 3.1.4 we introduce a constraint on the ground plane to estimate the bottom scene point. To obtain the top and bottom image point of a person from a video, we detect moving object in the frame. The camera setup is static, therefore we use one sequence frame as a background and calculate the absolute value of difference between the background $\mathbf{I}_{bg}$ and the current frame $\mathbf{I}_i$.

$$\mathbf{I}_d = |\mathbf{I}_{bg} - \mathbf{I}_i| \tag{3.11}$$

There are sophisticated methods for background subtraction. In [8] authors create a background model and continuously upgrade it. But since we have good lightning condition in our setup, a straightforward method worked well. The images are of dimensions $(h \times w \times 3)$, since we use a color cam, therefore one pixel is a vector of the length of 3. To find the highest difference we choose the maximum over the 3 color channels.

$$\mathbf{I}_{dm} = \max_{k=\{1,2,3\}} (\mathbf{I}_{d_{i,j,k}}) \tag{3.12}$$

(a) Background image of the sequence $\mathbf{I}_{bg}$

(b) Iterated image $\mathbf{I}_i$

(c) Maximum of the absolute difference of the $\mathbf{I}_{bg}$ and $\mathbf{I}_i$

(d) The result of the Canny edge segmentation.

Figure 3.2: The process of splitting the background away from the person.

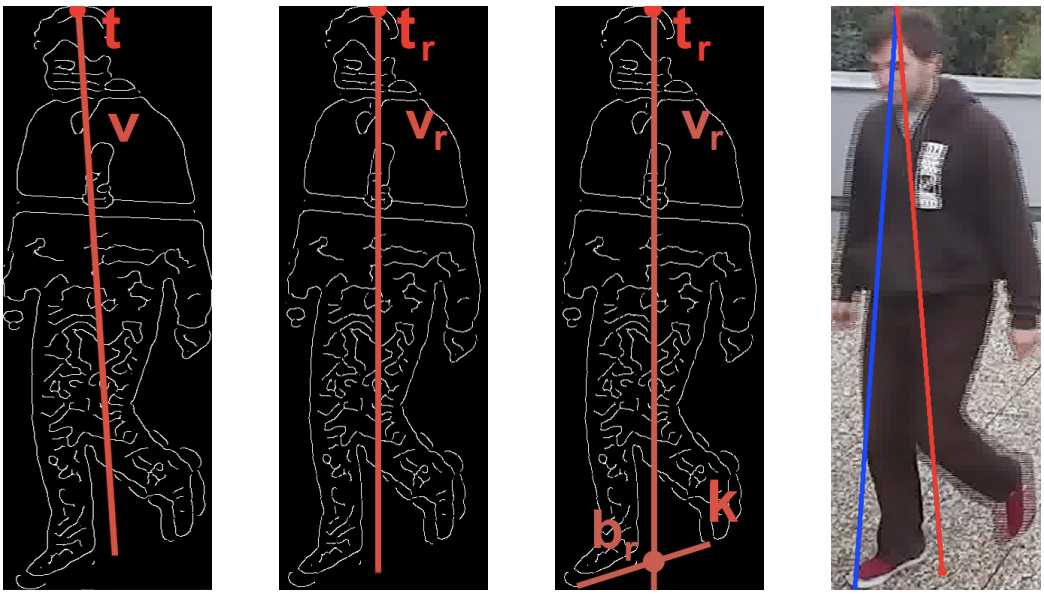
The image $\mathbf{I}_{dm}$ contains only one color channel and shows the maximum difference between the background image and the current image. As the video sequence is affected by a noise, there is a noise also in the difference image $\mathbf{I}_{dm}$. We can remove the noise pixels by thresholding. But in our setup we lost information about the person by using thresholding, some parts of the body were often marked as the background and the height estimation was inaccurate. We use edge detection instead, because it removes the noise by blurring the image firstly, but it preserves the shape of the body.

Canny edge detector To detect edges of the person we use the method developed by John F. Canny [3]. This algorithm is tuned by two parameters, sigma and threshold. The very first step is to blur the input image with Gaussian filter. There comes the sigma as the scale of the blur kernel, the size of the kernel window is derived from the sigma, usually as three times larger. We apply Sobel-operator on the blurred image in the next step. We get the horizontal and vertical approximation of the gradient of the image. The blurred edges are sharpened by non-maximum suppression. Threshold, as the next parameter, is used to distinguish real edges from false positive. The result of such method is in 3.2d.

With the result of the Canny edge detection we now have bounding box containing image of the human body. We can easily detect the top point of the head. Detection of the bottom image point is complicated since a posture of the person might vary. Feet can be together or apart, we can not simply use the lowest foreground pixel. There comes handy the knowledge of the ground plane.

3.1.4 Constraint on the ground plane

With reference to figure 3.2d, by knowing the top and bottom image points of the human, the top point $\mathbf{t}$ can be detected, as the first point in the vertical direction from the top of the image. The naive approach to find the bottom point of the body $\mathbf{b}$ by searching the image space from the bottom of the image. The result is displayed in the figure 3.3d with



(a) Original cropped image with the line created by the direction vector $\mathbf{v}$

(b) Transformed image by $\mathtt{A}$. $\mathbf{v}_r$ corresponds to $\mathbf{v}$ in the original image and it is parallel to the *y-axis*.

(c) Line $\mathbf{l}$ connecting the lowest edge points on both halves of the image. Point $\mathbf{b}_r$ is the intersection of the $\mathbf{l}$ and $\mathbf{v}_r$.

(d) Different approaches of finding the bottom body point; the lowest edge point is displayed with blue color, the red line is constructed with the knowledge of normal of the ground floor.

Figure 3.3

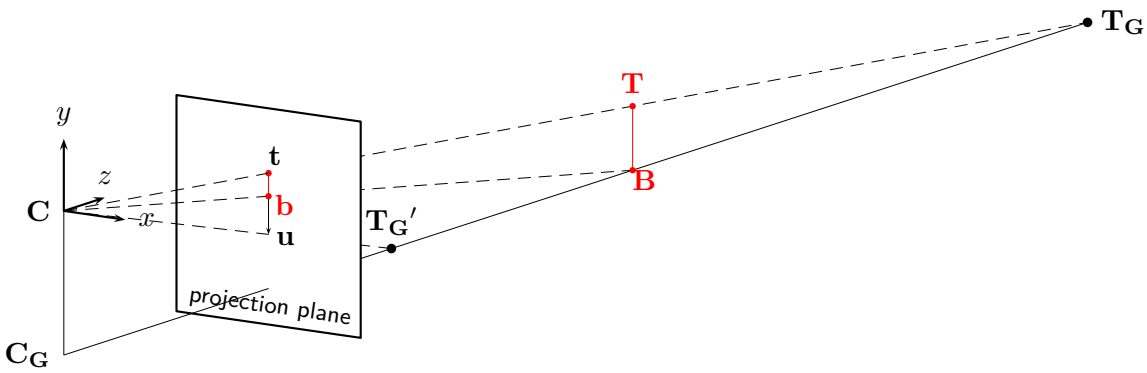


Figure 3.4: Model of the scene. Scene points $\mathbf{T_G}$ and $\mathbf{T_G}'$ are intersection of the rays and the ground plane. $\mathbf{t}$ and $\mathbf{u}$ their projection to the image plane. Red colored $\mathbf{T}$, $\mathbf{B}$ and $\mathbf{b}$ are unknown variables.

blue color. We will use the knowledge of the ground plane position and orientation. Such attitude is show in the figure 3.3d as the red line.

At this step of the algorithm we need to project the normal of the ground into the image in such way, that the line will go through point $\mathbf{t}$. To achieve that we choose any scene point on the line formed by the points $\mathbf{T_G}$ and $\mathbf{C_G}$, we call it $\mathbf{T_G}'$, it is displayed in the figure 3.4. If we project $\mathbf{T_G}'$ back to the image and connect this new image point $\mathbf{u}$ with $\mathbf{t}$, we get line corresponding to the normal of the ground plane. This line divides the bounding box of the person into two halves. If we calculate direction vector of the line as $\mathbf{v} = \mathbf{t} - \mathbf{u}$, we rotate the image in such way, that this line is parallel to the vertical axis. For $\mathbf{v} = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$

$$\omega = \arccos\left(\frac{v_2}{\sqrt{v_1^2 + v_2^2}}\right) \tag{3.13}$$

gives us angle between the vector $\mathbf{v}$ and the vertical axis. Knowing ω we rectify the image by

$$\mathtt{A} = \begin{bmatrix} \cos(\omega) & -\sin(\omega) & 0 \\ \sin(\omega) & \cos(\omega) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.14}$$

The vector $\mathbf{v}$ is parallel to the vertical axis in the rectified image and divides the image into two rectangles. We search each rectangle for the edge pixel with the lowest vertical value. By connecting these two points we get line $\mathbf{k}$. Intersection of lines $\mathbf{k}$ and $\mathbf{v}$ results in the point $\mathbf{b}_r$.

Knowing the image projection of the normal vector of the ground plane $\mathbf{v}$, we reformulate the task of the automatic detection. Instead of two separate image points, the top point $\mathbf{t}$ and the bottom point $\mathbf{b}$, we search for the top point $\mathbf{t}$ and the length l. The bottom image point is then expressed as

$$\mathbf{b} = l\frac{\mathbf{v}}{\|\mathbf{v}\|} + \mathbf{t} \tag{3.15}$$

. The length l is the distance between points $\mathbf{t}_r$ and $\mathbf{b}_r$, the two image points taken from the rectified image.

Now we have the height estimation equation 3.10 and the two needed image points. There is an uncertainty about the correct position of the image point $\mathbf{t}$ and the length l.

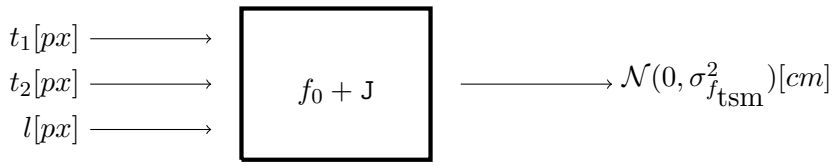


Figure 3.5: Overview of the function of the TSM method. Inputs are variances of the top image point position t_1 and t_2 and the distance to the bottom point l. The error is normal distribution with the variance $\sigma^2_{f_{\text{tsm}}}$.

3.1.5 Error analysis

In this section we inspect how an pixel error made during the estimation of the $\mathbf{t}$ and l affects the height estimation h_o in centimeters. We use this information in the final estimation for the whole video sequence. We do not inspect the uncertainty in the projection matrix $\mathtt{P}$, since the major uncertainty arises from the automatic detection and the camera is static.

The image point $\mathbf{b}$ is expressed as $\mathbf{b} = l\frac{\mathbf{v}}{\|\mathbf{v}\|} + \mathbf{t}$. We substitute this equation in the 3.6 and see that our TSM method is based on the variables $\mathbf{t} = \begin{bmatrix} t_1 & t_2 \end{bmatrix}$ and l. As long as the camera setup is static all the other variables are constants. We define function $f_{\text{tsm}}(t_1, t_2, l)$ which calculates the estimation using equation (3.10).

Variables t_1, t_2 and l are certainly non-linear combination, hence we approximate the function f_tsm to a first order Taylor series expansion $f_{\text{tsm}} \approx f_0 + \mathtt{J}\begin{bmatrix} t_1 & t_2 & l \end{bmatrix}^\intercal$ with Jacobian matrix $\mathtt{J} = \begin{bmatrix} \frac{\partial f_{\text{tsm}}}{\partial t_1} & \frac{\partial f_{\text{tsm}}}{\partial t_2} & \frac{\partial f_{\text{tsm}}}{\partial l} \end{bmatrix}$. f_0 is a constant, therefore we do not inspect its affect on the error. Let the covariance matrix $\mathtt{C}$ is a diagonal matrix representing the error in t_1, t_2 and l in pixels.

$$\mathtt{C} = \begin{bmatrix} \sigma^2_{t_1} & 0 & 0 \\ 0 & \sigma^2_{t_2} & 0 \\ 0 & 0 & \sigma^2_l \end{bmatrix}$$

the variance σ^2_i of the height estimation function f_{tsm} in the i-th frame of the video.

$$\sigma^2_i = \mathtt{J}^{(i)}\mathtt{C}\mathtt{J}^{(i)\intercal} \tag{3.16}$$

Aside from the height estimation from each frame of the sequence we obtain also the variance of the normal distribution of the error.

We emphasize that such error estimation is based on the local linearization, hence it is valid for reasonably small σ_{t_1}, σ_{t_2} and σ_l.

3.2 Face detecting Method (FDM)

3.2.1 Overview

This alternative method is based on a face detection. The face is not always visible if the person is not facing the camera or the resolution is to low for the detector. Nevertheless, if

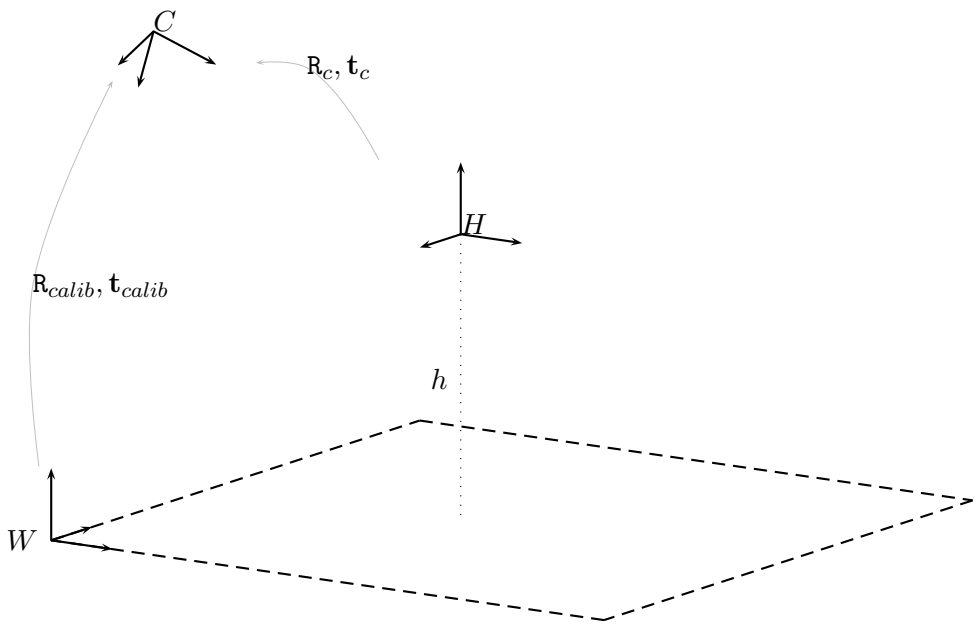


Figure 3.6: Overview of coordinate systems used in the scene. W stands for the global world coordinate system, C is the system of the camera and H reflects position of the head and its axes.

there are frames in the sequence, when the face can be spotted, we use the information as another support parameter. This method uses the algorithm [4].

The algorithm provides 6 degrees of freedom head pose from a single image. This is achieved by localization of the facial landmarks in the image. The algorithm uses then a generic face model, which is a set of 3D points of a generic face. When the landmarks are found, the algorithm projects the generic face model to the image and minimize the distance between these projected image points and the facial landmarks. The output is the head rotation and translation with with respect to the camera coordinate system. Since we know the camera pose from the calibration, we estimate the head pose in the world coordinate system and from that the height.

The scene is displayed in figure 3.6. There are three coordinates systems for world coordinates W, camera coordinates C and head coordinates H. Rotation matrix and translation vector are used to translate the point from one coordination system to another. From the camera calibration part we obtain $\mathtt{R}_{calib}$ and $\mathbf{t}_{calib}$, the landmark detector gives $\mathtt{R}_c$ and $\mathbf{t}_c$, a head pose in the camera coordinate system.

3.2.2 Height estimation

A point X_C, that is a point expressed in camera basis, can be evaluate as

$$\mathbf{X}_C = \mathtt{R}_{calib}\mathbf{X}_W + \mathbf{t}_{calib} \tag{3.17}$$

To express height, we want to project the point from the camera system to the world system.

$$\mathbf{X}_W = \mathtt{R}_{calib}^{-1}\mathbf{X}_C + \mathtt{R}_{calib}^{-1}\mathbf{t}_{calib} \tag{3.18}$$

We introduce new variables $\mathtt{R}_W = \mathtt{R}_{calib}^{-1}$ and $\mathtt{t}_W = \mathtt{R}_{calib}^{-1}\mathbf{t}_{calib}$ and get

$$\mathbf{X}_W = \mathtt{R}_W \mathbf{X}_C + \mathtt{t}_W \tag{3.19}$$

The landmark detector algorithm uses a generic normalized face model. It means, that there is unit distance between pupils. The translation vector of [4] is also given up to scale. If we multiply the translation vector with the known mean interpupillary distance, we get the translation vector in world coordinates. According to [6] these values are listed in table 3.1.

Gender	Mean s_0	Variance σ_{ipd}^2
Male	6.47	0.37
Female	6.23	0.36

Table 3.1: Interpupillary distance (cm)

We introduce a new translation $\mathbf{t}_{cm} = \mathbf{t}_c(s_o + \delta)$, where s_0 comes from the table above 3.1 and δ is an error which follows the normal distribution $\mathcal{N}(0, \sigma_{ipd}^2)$. We use it to derive a formula for $\mathbf{X}_C$

$$\mathbf{X}_C = \mathtt{R}_c \mathbf{X}_H + \mathbf{t}_{cm} \tag{3.20}$$

As long as we are not interested in orientation of the head, just its position in the scene, we shifted the face model to have the $\mathbf{0}$ at the top of the head. Then $\mathbf{X}_H = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^\mathsf{T}$ and we simplify equation (3.20)

$$\mathbf{X}_C = \mathbf{t}_{cm} \tag{3.21}$$

Now we substitute into equation (3.19)

$$\mathbf{X}_W = \mathtt{R}_W \mathbf{t}_{cm} + \mathbf{t}_W \tag{3.22}$$

$$\mathbf{X}_W = \mathtt{R}_W \mathbf{t}_c(s_o + \delta) + \mathbf{t}_W = \mathtt{R}_W \mathbf{t}_c s_o + \mathbf{t}_W + \mathtt{R}_W \mathbf{t}_c \delta \tag{3.23}$$

Since we need just the height, which in our world system is the third coordinate, we simplify the equation

$$\mathtt{R}_W = \begin{bmatrix} \mathtt{r}_{W1}^\mathsf{T} \\ \mathtt{r}_{W2}^\mathsf{T} \\ \mathtt{r}_{W3}^\mathsf{T} \end{bmatrix}, \mathtt{t}_W = \begin{bmatrix} t_{W1} \\ t_{W2} \\ t_{W3} \end{bmatrix}$$

$$\boxed{h = \underbrace{(\mathtt{r}_{W3}^\mathsf{T} \mathbf{t}_c s_o + t_{W3})}_{h_0} + \underbrace{\mathtt{r}_{W3}^\mathsf{T} \mathbf{t}_c \delta}_{\mathrm{J}\delta}} \tag{3.24}$$

We have the equation for the height estimation using the face detector. We have seen that there is uncertainty in the estimation of the interpupillary distance (*ipd*). Hence we analyze how an input error in the *ipd* affects an estimation of the height.

3.2.3 Error analysis

The closed-form equation 3.24 can be rewritten as

$$h = h_0 + J\delta \tag{3.25}$$

where $h_0 = \mathtt{r}_{W3}^{\mathsf{T}}\mathbf{t}_c s_o + t_{W3}$, $J = \mathtt{r}_{W3}^{\mathsf{T}}\mathbf{t}_c$.

Assuming that estimation of the j-th frame in the video follows the normal distribution $h^{(j)} \sim \mathcal{N}(h_j - J_j\delta,\ \sigma_j^2)$, we measure the uncertainty in the translation vector $\mathbf{t}_c$ given by the face detector script [4]. The same package includes another function, which takes as parameters calibration matrix $\mathtt{K}$, rotation matrix $\mathtt{R}_c$ and translation of the camera $\mathbf{t}_c$, points of the mean face model $\mathbf{X}$ and parameter ϵ representing pixel error in percentage of the face. It returns covariance matrix $\mathtt{C}_\mathbf{t}$ of the translation. Taking matrix $\mathtt{C}_\mathbf{t}$, we evaluate covariance σ_i^2 as follows

$$\sigma_j^2 = s_0\mathtt{r}_{W3}^{\mathsf{T}}\mathtt{C}_\mathbf{t}^{(j)}s_0\mathtt{r}_{W3} = s_0^2\mathtt{r}_{W3}^{\mathsf{T}}\mathtt{C}_\mathbf{t}^{(j)}\mathtt{r}_{W3} \tag{3.26}$$

$\sqrt{\sigma_j^2}$ then represents a standard deviation in the height measurement for the j-th frame using the face detector algorithm.

3.3 Fusion of methods applied on a video

So far we have inspected methods of the height estimation for a single picture. For each frame of the video we have measured the height and an error. Now we fuse these estimations and estimate the height of the person according to the measurement from the whole video sequence.

A video contains k frames, we track the person n times using the TSM method and m times using the FDM method, where $n \leq k$ and $m \leq k$.

$$x_1, x_2, \ldots x_n \text{ - estimation using the TSM method, where } x_i \sim \mathcal{N}(h,\ \sigma_i^2)$$

$$y_1, y_2, \ldots y_m \text{ - estimation using the FDM method, where } y_j \sim \mathcal{N}(h - J_j\delta,\ \sigma_j^2)$$

We have statistical model containing two parameters height h and uncertainty of the interpupillary distance δ. We calculate these parameters using maximum-likelihood estimation (MLE) and Bayesian estimation.

3.3.1 Maximum-Likelihood estimation

We have the set of data, which are the independent observations $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_m$. Now we define the *likelihood function*. MLE maximize this function by adjusting the parameters, which are h and δ. We start with conditional probability

$$\begin{aligned} p(x_1, x_2, \ldots x_n, y_1, y_2, \ldots y_m|h, \delta) &= \textstyle\prod_{i=1}^n p(x_i|h, \delta) \prod_{j=1}^n p(y_j|h, \delta) \\ p(x_1, x_2, \ldots x_n, y_1, y_2, \ldots y_m|h, \delta) &= \textstyle\prod_{i=1}^n c_i e^{-\frac{(h-x_i)^2}{2\sigma_i^2}} \prod_{j=1}^m c_j e^{-\frac{(h-J_j\delta-y_j)^2}{2\sigma_j^2}} \end{aligned} \tag{3.27}$$

By applying the logarithm our likelihood function is

$$L = \log_e(p(x_1, x_2, \dots x_n, y_1, y_2, \dots y_m|h, \delta)) = \sum_{i=1}^{n}\left(K_i - \frac{(h-x_i)^2}{2\sigma_i^2}\right) + \sum_{j=1}^{m}\left(K_j - \frac{(h-J_j\delta-y_j)^2}{2\sigma_j^2}\right) \tag{3.28}$$

To find the maximum of the function L the derivatives of the function L with respect of h and δ equal zero, which leads to a system of two equations.

$$\begin{aligned} \frac{\partial L}{\partial h} = \sum_{i=1}^{n} -\frac{(h-x_i)}{\sigma_i^2} - \sum_{j=1}^{m}\frac{(h-J_j\delta-y_j)}{\sigma_j^2} = 0 \\ \frac{\partial L}{\partial \delta} = \sum_{j=1}^{m}\frac{(h-J_j\delta-y_j)J_j}{\sigma_j^2} = 0 \end{aligned} \tag{3.29}$$

We expand each of the two equations

$$\begin{aligned} -h\sum_{i=1}^{n}\frac{1}{\sigma_i^2} + \sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} - h\sum_{j=1}^{m}\frac{1}{\sigma_j^2} + \delta\sum_{j=1}^{m}\frac{J_j}{\sigma_j^2} + \sum_{j=1}^{m}\frac{y_j}{\sigma_j^2} = 0 \\ h\sum_{j=1}^{m}\frac{J_j}{\sigma_j^2} - \delta\sum_{j=1}^{m}\frac{J_j^2}{\sigma_j^2} - \sum_{j=1}^{m}\frac{J_jy_j}{\sigma_j^2} = 0 \end{aligned} \tag{3.30}$$

and express h and δ as they are the two unknowns in the system.

$$\begin{aligned} (-\textstyle\sum_{i=1}^{n}\frac{1}{\sigma_i^2} - \sum_{j=1}^{m}\frac{1}{\sigma_j^2})h + (\sum_{j=1}^{m}\frac{J_j}{\sigma_j^2})\delta &= -\textstyle\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} - \sum_{j=1}^{m}\frac{y_j}{\sigma_j^2} \\ (\textstyle\sum_{j=1}^{m}\frac{J_j}{\sigma_j^2})h - (\sum_{j=1}^{m}\frac{J_j^2}{\sigma_j^2})\delta &= \textstyle\sum_{j=1}^{m}\frac{J_jy_j}{\sigma_j^2} \end{aligned} \tag{3.31}$$

The linear system is solve for certain height and interpupillary distance, we call them $\hat{h}$ and $\hat{\delta}$, which maximize the likelihood function, hence these are the estimates for the given sequence.

3.3.2 Bayesian estimation

This is the other approach of how to use the set of data $x_1, x_2, \dots x_n$ and $y_1, y_2, \dots y_m$. Bayesian estimate allows us incorporate prior of the unknown parameters. In our case the parameters h and δ are height and offset of the interpupillary distance. As we stated in equation (3.27), our estimations are independent conditional probabilities of the normal distribution with known mean value and variance. The height of people and their interpupillary distance are random variables, which follow a distribution with some probability density function.

For simplicity we assume $p(h, \delta) = p(h)p(\delta)$. Probability of human height h and variation of the interpupillary distance δ can be found in [10] and [6]. In our statistical model we use Gaussian mixture of two components - one for male, the other for female.

$$\begin{aligned} p(h) &= \tfrac{1}{2}c_M^h e^{-\frac{(h-\mu_M^h)^2}{2\sigma_M^{h^2}}} + \tfrac{1}{2}c_W^h e^{-\frac{(h-\mu_W^h)^2}{2\sigma_W^{h^2}}} \\ p(\delta) &= \tfrac{1}{2}c_M^{ipd} e^{-\frac{(h-\mu_M^{ipd})^2}{2\sigma_M^{ipd^2}}} + \tfrac{1}{2}c_W^{ipd} e^{-\frac{(h-\mu_W^{ipd})^2}{2\sigma_W^{ipd^2}}} \end{aligned} \tag{3.32}$$

Using likelihood (3.27) and the prior probability (3.32) we define posterior probability as

$$p(h,\delta|x_1,x_2,\dots x_n,y_1,y_2,\dots y_m) = \frac{1}{z}\underbrace{p(x_1,x_2,\dots x_n,y_1,y_2,\dots y_m|h,\delta)p(h,\delta)}_{\tilde{p}(h,\delta|x_1,x_2,\dots x_n,y_1,y_2,\dots y_m)} \tag{3.33}$$

The Bayesian height estimate $\hat{h}$ over the whole video sequence is calculated as

$$\hat{h} = \frac{1}{z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h\tilde{p}(h,\delta|x_1,x_2,\dots x_n,y_1,y_2,\dots y_m)\mathrm{d}h\mathrm{d}\delta \tag{3.34}$$

and the offset of interpupillary distance estimate

$$\hat{\delta} = \frac{1}{z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta\tilde{p}(h,\delta|x_1,x_2,\dots x_n,y_1,y_2,\dots y_m)\mathrm{d}h\mathrm{d}\delta \tag{3.35}$$

The constant z reflects the fact that the probability sums to one

$$z = \int_{\infty}^{\infty}\int_{\infty}^{\infty} \tilde{p}(h,\delta|x_1,x_2,\dots x_n,y_1,y_2,\dots y_m)\mathrm{d}h\mathrm{d}\delta$$

.

The integrals are computed numerically. In 2D-case, the space of the integrals is very large. In our model the non-zero contributions are well localized. We inspect space of height h in interval between 80 cm and 240 cm, the offset of a interpupillary distance space between -1.5 cm and 1.5 cm. We use trapezoidal rule in our implementation. Otherwise such integral can be computed using Monte Carlo sampling method [2].

We presented methods to obtain height estimation from a set of estimations based on the maximum-likelihood and Bayesian estimation.

Chapter 4

Experiments

In this chapter we describe experiments we made to test the proposed methods of height estimation. We use our own video sequences, we record multiple persons and measure their real height and interpupillary distance. Then we compare the results of proposed methods with the measured values.

Firstly we describe the videos and scene setup and our approach to get the camera calibration. We present result of frame-by-frame height estimation and, then we fuse the methods and evaluate the height estimation for each video sequence.

4.1 Video sequences

We recorded 8 video sequences with different subjects listed in the table 4.1. The footage resolution is full HD 1920x1080 pixels. We record the scene using camera **Panasonic HC-V130** standing on a tripod. Our camera is static. To eliminate changes of light conditions we record all the footage during cloudy weather. Such condition is also ideal to avoid shadows. The recorded videos were significantly interlaced,

The people captured in the videos walked in front of the camera approximately 20 seconds. During this time period they faced the cam in different angles.

4.2 Setup description

We inspect finding a projection matrix of the camera setup. Since the camera is static, we need to estimate the projection matrix of the camera just once for each video. In the following subsection we describe approach we use to find the projection matrix.

4.2.1 Camera resectioning

One of the assumption of the both height estimation methods is to know extrinsic and intrinsic camera parameters. Since we know the scene, we can measure the distances between significant points in the scene and manually choose these points in the picture taken from a video. We use six points algorithm to estimate projection camera matrix P [9, p. 181]. The

Preview	Id	Height	Interpupillary distance
	1	167.2 cm	6.2cm
	2	173.2 cm	6.1cm
	3	185.2 cm	6.4cm
	4	110 cm	5.8cm
	8	171 cm	6.3cm
	9	180 cm	6.3cm
	9	185 cm	6.4cm

Table 4.1: Overview of the taken videos. We measured each person real height and interpupillary distance.

task is to solve homogeneous equation, which gives as all the twelve numbers of the matrix P. To create the system of equations we need at least six 2D $\leftrightarrow$ 3D correspondences. The more correspondences we have, the more combination of them we inspect to find more accurate projection matrix. As listed in Fig. 4.1 we choose and measure eleven scene points $\mathbf{X}_i$, their

Figure 4.1: Measured scene points and their image projections. The **0** point is located on the ground floor. Three values show x, y, z coordinates of each scene point.

image correspondences $\mathbf{u}_i$, $\mathbf{X}_i \leftrightarrow \mathbf{u}_i$. It gives us $\binom{11}{6}$ combinations of choosing six points. According to [9] solving one possible combination returns twelve possible matrices P. For $12 \cdot \binom{11}{6}$ matrices P_j we project the eleven scene points back to the image $\mathbf{x}_i$ and measure the projection error $e_{pj} = \sum_{i=1}^{11} \|\mathbf{x}_i - \mathbf{u}_i\|$. Matrix P with the minimal projection error $e_{pmin} = \min_j(e_{pj})$ is used for the height estimation methods. In Fig. 4.2 we show the e_{pmin} for each video. Inaccuracies in clicking of the image correspondences cause the differences of the errors.

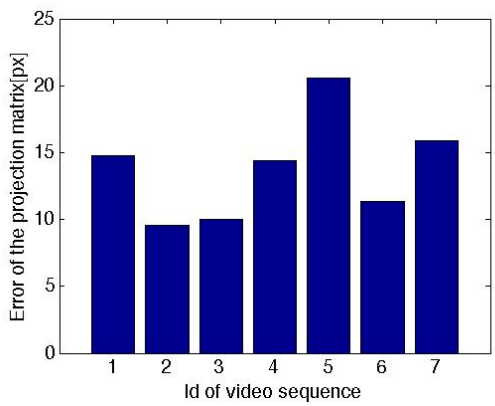


Figure 4.2: Sum of projection errors of chosen matrix P for each video sequence e_{pmin}.

4.3 Evaluation of the proposed methods

Recorded videos are evaluated frame by frame and height is measured by methods:

1. **triangle-sim**. Proposed method based on triangle similarity of the scene points in section 3.1.

2. **face-detect**. Method using the facial landmark detector [4] referred in section 3.2.

After this step we have two sets of independent measurements with known error probability distribution.

We fuse the two sets using

1. **Maximum-likelihood**. We define a likelihood function and search for its maximum.

2. **Bayesian estimate** We calculate posterior probability of the likelihood defined by our height estimation of each frame and prior probability of height and interpupillary distance.

In triangle-sim method we use the Canny edge detector. In our setup we choose parameters $\sigma = 0.13$ as the standard deviation of the Gaussian filter and $thr = 10$ specifies sensitivity thresholds for the Canny method.

The face detector method needs mean values of interpupillary distance μ_{ipd} and the difference from the mean value δ_{ipd}. We use $\mu_{\text{ipd}} = 6.4cm$ and $\delta_{\text{ipd}} = 0.4$ for male subjects, $\mu_{\text{ipd}} = 6.2cm$ and $\delta_{\text{ipd}} = 0.3$ for female subjects and $\mu_{\text{ipd}} = 5.5cm$ and $\delta_{\text{ipd}} = 0.2$ for children.

Bayesian estimate uses prior probability to calculate the posterior. The prior probability in our case is defined by two Gaussian mixtures, one for height distribution among the population and the second for the interpupillary distance. These values differs for people around the world, we used values listed in [6] and [10] for white race, since our setup is located in the Czech republic and all the subjects are of white race. We use the deviation of the height distribution for male $\sigma_{mbox[M]} = 9.6561cm$, female $\sigma_{mbox[F]} = 7.4654cm$, mean values for male $\mu_{mbox[M]} = 180.31cm$ and $\mu_{mbox[M]} = 167.22cm$ for female.

4.4 Videos

In this section we present results of the experiment. Each experiment shows three sample screenshots of the height estimate. We include graph overview showing single image height estimates. For better readability we display every tenth estimate. The graph includes error bars representing deviation along a curve. There parts of the graph missing the error bars, it means that during these frames the estimator does not estimate any height.

Then we present statistics of the estimates. There are two tables in each experiment section. The first one represents statistics of the single image height estimators. We show how many estimations have been measured, what is their mean value and variance. The second table shows fuse methods over the whole sequence. We listed the aggregated results and the difference between the estimate and real height.

(a) $h = 165.9$ cm, $e = 3.4$ cm (b) $h = 168.9$ cm, $e = 4.3$ cm (c) $h = 167.1$ cm, $e = 3.4$ cm

Figure 4.3: Results of the TSM method for sample frames with the measured height h and error calculated as $3 \cdot \sqrt{\sigma^2}$

4.4.1 Experiment video sequence id. 1

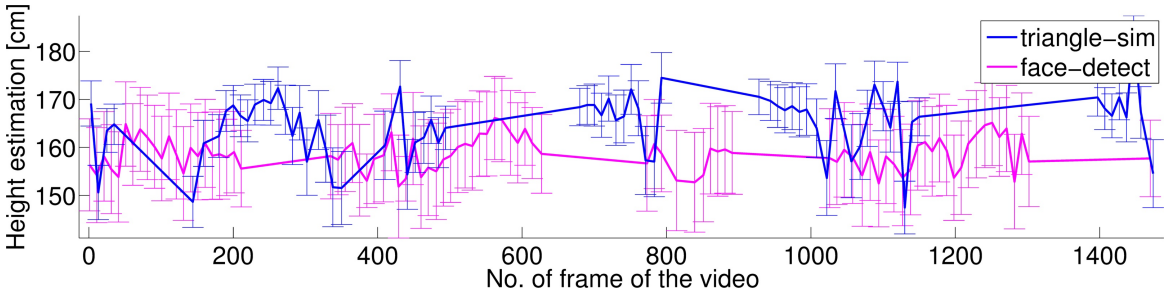


Figure 4.4: Results of the TSM method for sample frames with the measured height h and error calculated as $3 \cdot \sqrt{\sigma^2}$

Table 4.2: Single image height estimation methods

Method	Number of estimations	Mean	Variance
triangle-sim	703	165.3909 cm	5.2274 cm
face-detect	871	159.0094 cm	9.0963 cm

Table 4.3: Methods for estimation height over the whole video

Method	Estimated height h	Error e_h
max-likelihood	171.8313 cm	4.6313 cm
Bayes est.	168.6459 cm	**1.4459** cm

The first video captures a women of height 165.3 cm. The footage lasts for 60 s and during that time both estimators capture comparable amount of the single image estimations. In Fig. 4.4 it can be seen, that the estimations are in range of 150 - 170 cm.

(a) $h = 176.7$ cm, $e = 1.8$ cm (b) $h = 179$ cm, $e = 2.1$ cm (c) $h = 178.6$ cm, $e = 2.0$ cm

Figure 4.5: Results of the TSM method for sample frames with the measured height h and error calculated as $3 \cdot \sqrt{\sigma^2}$

4.4.2 Experiment video sequence id. 2

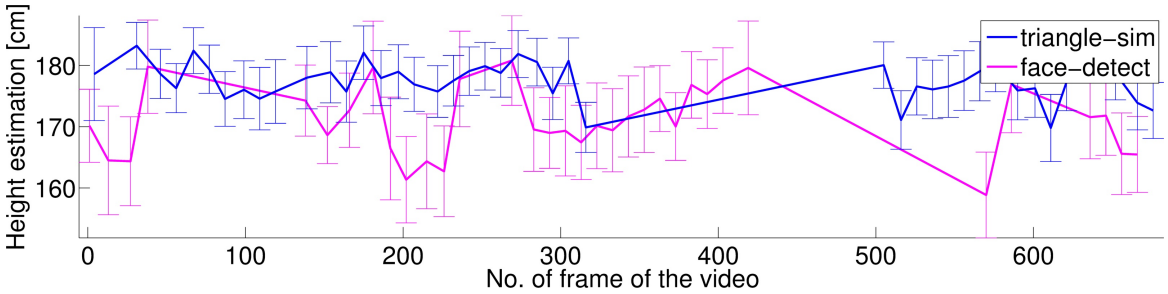


Figure 4.6: Results of the TSM method for sample frames with the measured height h and error calculated as $3 \cdot \sqrt{\sigma^2}$

Table 4.4: Single image height estimation methods

Method	Number of estimations	Mean	Variance
triangle-sim	411	177.2541 cm	4.4835 cm
face-detect	334	170.4134 cm	6.8701 cm

Table 4.5: Methods for estimation height over the whole video

Method	Estimated height h	Error e_h
max-likelihood	178.4592 cm	5.2592 cm
Bayes est.	175.8615 cm	**2.6615** cm

The second video captures a women of height 173.3 cm. The footage lasts for 27 s and during that time both estimators capture comparable amount of the single image estimations. In Fig. 4.4 it can be seen, that the estimations are in range of 150 - 170 cm.

(a) h = 187.84 cm, e = 1.7 cm (b) h = 190.75 cm, e = 1.9 cm (c) h = 187.6 cm, e = 2.5 cm

Figure 4.7: Results of the TSM method for sample frames with the measured height h and error calculated as $3 \cdot \sqrt{\sigma^2}$

4.4.3 Experiment video sequence id. 3

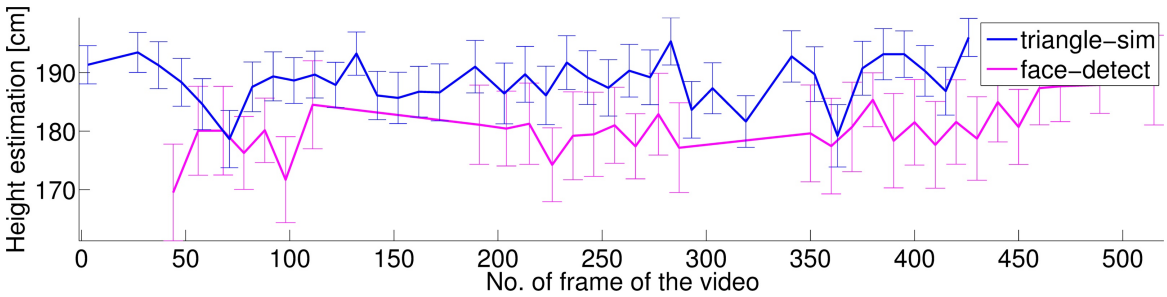


Figure 4.8: Sum of projection errors of chosen matrix P for each video sequence e_{pmin}.

Table 4.6: Single image height estimation methods

Method	Number of estimations	Mean	Variance
triangle-sim	374	188.3329 cm	4.3981 cm
face-detect	307	179.8799 cm	6.9114 cm

Table 4.7: Methods for estimation height over the whole video

Method	Estimated height h	Error e_h
max-likelihood	182.1835 cm	3.0165 cm
Bayes est.	186.2275 cm	**1.0275** cm

The video sequence with id 3 shows a men of height 185.2 cm. The face-detect method

The first video captures a women of height 165.3 cm. The footage lasts for 60 s and during that time both estimators capture comparable amount of the single image estimations. In Fig. 4.4 it can be seen, that the estimations are in range of 150 - 170 cm.

(a) $h = 134.1$ cm, $e = 10$ cm (b) $h = 128.1$ cm, $e = 10.5$ cm (c) $h = 134.5$ cm, $e = 9.5$ cm

Figure 4.9: Results of the TSM method for sample frames with the measured height h and error e.

4.4.4 Experiment video sequence id. 4

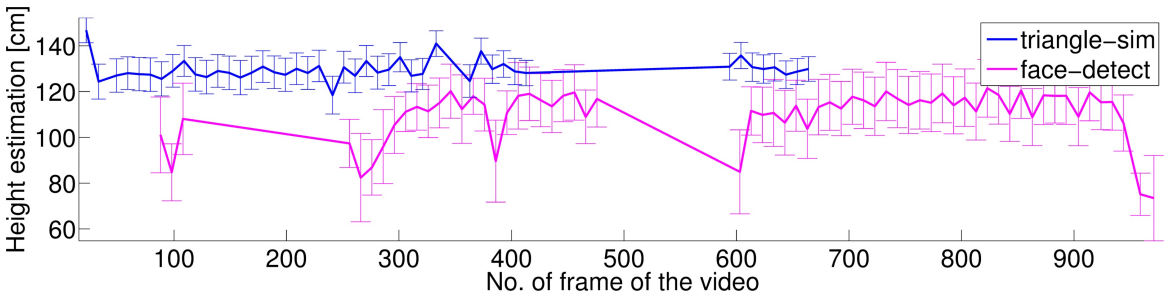


Figure 4.10: Estimated height per each frame of the video sequence. Error bars show approximated error of each height measurement $3 \cdot \sqrt{\sigma^2}$.

Table 4.8: Single image height estimation methods

Method	Number of estimations	Mean	Variance
triangle-sim	448	129.1271 cm	6.7522 cm
face-detect	628	108.6169 cm	12.6409 cm

Table 4.9: Methods for estimation height over the whole video

Method	Estimated height h	Error e_h
max-likelihood	123.1653 cm	13.1653 cm
Bayes est.	120.6833 cm	**10.6833** cm

(a) $h = 172.4$ cm, $e = 7.8$ cm (b) $h = 169.6$ cm, $e = 7.4$ cm (c) $h = 168.7$ cm, $e = 7.8$ cm

Figure 4.11: Results of the TSM method for sample frames with the measured height h and error e.

4.4.5 Experiment video sequence id. 5

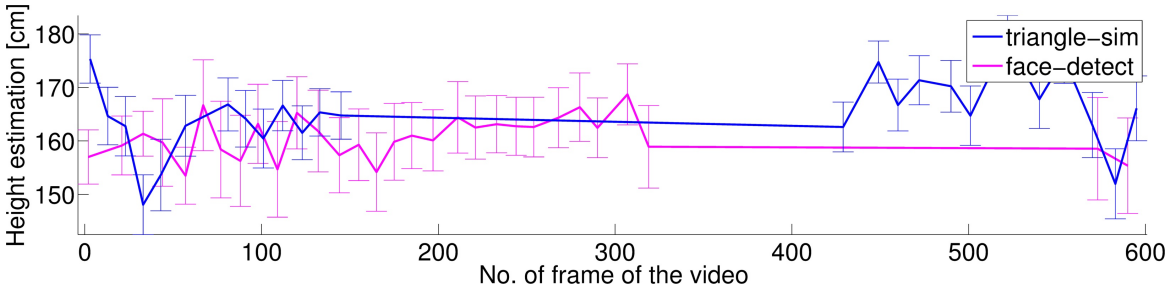


Figure 4.12: Estimated height per each frame of the video sequence. Error bars show approximated error of each height measurement $3 \cdot \sqrt{\sigma^2}$.

Table 4.10: Single image height estimation methods

Method	Number of estimations	Mean	Variance
triangle-sim	246	166.3539 cm	5.1195 cm
face-detect	302	161.6232 cm	7.0069 cm

Table 4.11: Methods for estimation height over the whole video

Method	Estimated height h	Error e_h
max-likelihood	175.7464 cm	**4.7464** cm
Bayes est.	165.5158 cm	5.4842 cm

(a) $h = 183.585$ cm, $e = 6.8$ cm (b) $h = 179.0$ cm, $e = 6.8$ cm (c) $h = 179.5$ cm, $e = 6.8$ cm

Figure 4.13: Results of the TSM method for sample frames with the measured height h and error e.

4.4.6 Experiment video sequence id. 6

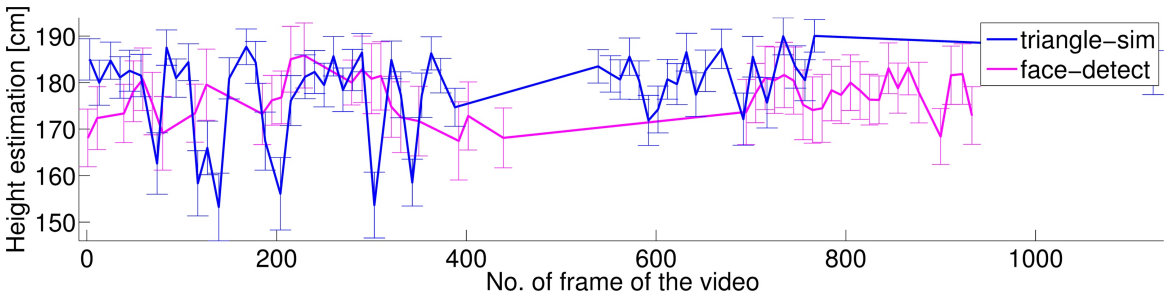


Figure 4.14: Estimated height per each frame of the video sequence. Error bars show approximated error of each height measurement $3 \cdot \sqrt{\sigma^2}$.

Table 4.12: Single image height estimation methods

Method	Number of estimations	Mean	Variance
triangle-sim	553	177.7367 cm	4.8331 cm
face-detect	470	177.9759 cm	6.565 cm

Table 4.13: Methods for estimation height over the whole video

Method	Estimated height h	Error e_h
max-likelihood	180.4829 cm	**0.4829** cm
Bayes est.	175.8621 cm	4.1379 cm

(a) $h = 187.0$ cm, $e = 4.9$ cm (b) $h = 184.5$ cm, $e = 4.9$ cm (c) $h = 188.9$ cm, $e = 4.9$ cm

Figure 4.15: Results of the TSM method for sample frames with the measured height h and error e.

4.4.7 Experiment video sequence id. 7

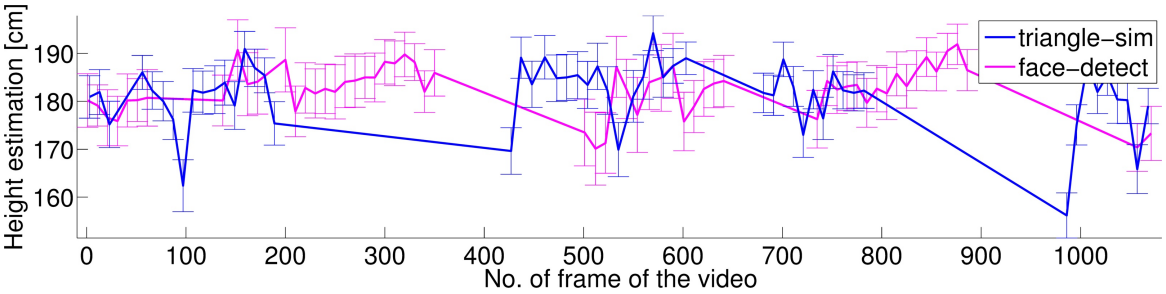


Figure 4.16: Estimated height per each frame of the video sequence. Error bars show approximated error of each height measurement $3 \cdot \sqrt{\sigma^2}$.

Table 4.14: Single image height estimation methods

Method	Number of estimations	Mean	Variance
triangle-sim	553	177.7367 cm	4.8331 cm
face-detect	470	177.9759 cm	6.565 cm

Table 4.15: Methods for estimation height over the whole video

Method	Estimated height h	Error e_h
max-likelihood	183.4966 cm	**1.5034** cm
Bayes est.	182.7586 cm	2.2414 cm

Chapter 5

Conclusion

We proposed two independent methods of body height estimation in this thesis. The first one is based on the position of the ground floor and it is called Triangle similarity method and the second one is based on the landmark detector [4] and is called Face detector method. We created a sensitivity analysis of each of the method so in every frame we have a body height estimate and its error. The Face detector method gives also estimate of the interpupillary distance. We suggested method for fuse single image estimates and aggregation over the whole sequence. This is done as Maximum Likelihood estimation and Bayesian estimation.

We presented an experiment test including multiple videos. These experiments show, that even though single measurements fluctuate, the aggregated body height measurement is estimated precisely. The overall best estimation differs in 0.5 cm. The average error is about 3 cm on the test data.

5.1 Future work

We proposed a very simple automatic detection of person moving in front of a static camera. There are better and sophisticated methods of background subtraction.

Knowing the shape of a person more precisely would help us with recognizing the bottom boundary pixel of the human body. As we stated, the feet position vary, a person can have feet apart or together. If the thresholding works reasonably accurate, we can use learning algorithm. It crawls the neighborhood of the line defined by the ground floor normal image projection and searches for the most probable position of the feet.

Experiments show, that recording using a surveillance camera placed outside can bring other uncertainty. People can wear hat or women can have heels, which affect the overall estimation.

The solution is to detect abnormality and increase expected error in Triangle Similarity Method. Since Face Detector method is not affected by these situations, its impact on the overall video sequence height estimation would rise.

Next step is plugging into a broader tracking system, which would support a re-detection of lost trackings.

Bibliography

[1] C. BenAbdelkader and Y. Yacoob. Statistical body height estimation from a single image. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–7, Sept 2008.

[2] R. E. Caflisch. Monte carlo and quasi-monte carlo methods. *Acta Numerica vol. 7*, 1998.

[3] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6):679–698, Nov 1986.

[4] J. Cech, V. Franc, and J. Matas. A 3d approach to facial landmarks: Detection, refinement, and tracking. *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 2173 – 2178, 2014.

[5] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 434–441 vol.1, 1999.

[6] N. A. Dodgson. Variation and extrema of human interpupillary distance.

[7] Y.-P. Guan. Unsupervised human height estimation from a single image. *Journal of Biomedical Science and Engineering*, 1998.

[8] I. Haritaoglu, D. Harwood, and L. Davis. W4: real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):809–830, Aug 2000.

[9] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[10] M. MA, F. CD, H. R, and O. CL. Anthropometric reference data for children and adults: U.s. population, 1999–2002. *Advance data from vital and health statistics; no 361. Hyattsville, MD: National Center for Health Statistics*, 2005.

[11] M. Momeni-K., S. C. Diamantas, F. Ruggiero, and B. Siciliano. Height estimation from a single camera view. In *VISAPP (1)*, pages 358–364, 2012.

Appendix A

CD content

The thesis comes with a compact disk which contains source codes of the experiments presented in the thesis and electronic version of the thesis. The CD has the following structure:

doc This folder contains electronic version of the thesis.

toolbox This folder contains helper Matlab scripts

experiments This folder contains Matlab scripts implementing the experiments

- **bayesian** Fuse script using Bayesian estimation
- **ml** Fuse script using Maximum-likelihood estimation
- **test-fdm** Script for frame by frame height estimation using Face detector method
- **test-tsm** Script for frame by frame height estimation using Triangle similarity method