MVA2015 IAPR International Conference on Machine Vision Applications, May 18-22, 2015, Tokyo, JAPAN

Annotation driven MAP Search Space Estimation for Sliding-Window based Person Detection

Stefan Becker
stefan.becker@iosb.fraunhofer.de

Wolfgang Hübner
wolfgang.huebner@iosb.fraunhofer.de

Michael Arens
michael.arens@iosb.fraunhofer.de
Fraunhofer IOSB
Gutleuthausstr. 1, 76275 Ettlingen, Germany

Abstract

A common method for performing multi-scale person detection is a sliding window classification. For every window location and scale a binary classification is done. Many state-of-the-art person detectors follow this sliding window paradigm. Not only this exhaustive search space strategy is computationally expensive, it usually produces large number of false positives.

In order to estimate an optimal reduced search space, we derive a maximum a posteriori probability (MAP) solution given only the person annotations of a dataset. The proposed MAP solution considers the naturally height distribution of persons, deviations from a flat world assumption, and annotation uncertainty. The effectiveness compared to the traditional uniform sliding window selection strategy is shown on different realistic monocular pedestrian detection datasets. Moreover the MAP search space estimation provides design parameters for modeling the tradeoff between detection performance and runtime constraints.

1 Introduction

The detection of pedestrians is an essential component for advance driver assistance systems (ADAS). In order to localize the pedestrians in the image, the first step is to select image regions likely to contain persons. After this candidate generation process, the selected regions are classified as pedestrian or non-pedestrian. The most common combination is the sliding window classification. It consists in exhaustively scanning of the image and a binary classification for every window location and scale. The final object location is determined by non-maximum suppression. Many state-of-the-art pedestrian detectors follow this sliding window paradigm like the approach of Felzenszwalb et al. [1] and the Fastest Pedestrian Detector in the West [2].

Despite the fact that such approaches can differ in the way of scaling the image or the sliding window, their shared drawback is the computationally inefficiency of an exhaustive scanning and the large number of produced false positives in inappropriate regions like the sky or windows inconsistent with perspective constraints [3].

The integration of prior knowledge about the scene can be used to restrict the search space. Under a flat world assumption the space of corresponding valid pedestrian location can be reduced with the information about the initial camera position with respect to the ground or a fixed horizontal line. By just sampling the candidate windows projected on the ground the number of candidates is drastically reduced. For pedestrian detection this concept is for example used in the approaches of Bertozzi et al. [4] and Gavrila et al. [5]. In order to overcome the limitations of a fixed camera ground position, some methods adaptively estimate the horizon line or the ground plane. Such approaches often rely on 3D information and use a stereo camera setup. A good overview on 3D approaches for an effective candidate generation can be found in Geronimo et al. [3] or Llorca et al. [6].

Nevertheless, the road position is adaptively estimated, this information can be used to filter the standard sliding window algorithm by geometric constraints. For example, Sudowe et al. [7] derive an analytical solution for valid sliding window locations which can be directly integrated in the detection process.

An alternative approach to reduce the number of candidates is to perform a coarse-to-fine search. As example, the approach of Pedersoli et al. [8] is mentioned. There, the search space is pruned based on the confidence of the classifier on a particular scale. The classifier output can also be used for a branch and bound technique. In the work of Lampert et al. [9], the output of a support vector machine as classifier is bounded to find a globally optimal solution at sublinear time.

In recent years there are more detection algorithms which shun sliding window in favor of a segmentation pre-processing step for candidate generation (see [10]). Nevertheless there exist numerous alternative methods for candidate generation, we focus on approaches which localize pedestrians based on a sliding window classification. As described above, many state-of-the-art pedestrian detectors follow this concept (see [11]).

In this work we propose a MAP solution for estimating an optimal search space based on ground truth annotations of the enclosing bounding boxes of pedestrians. Thereby, geometric constraints for valid window locations based on a naturally height distribution of persons for a fixed camera to ground distance are indirectly included. But in contrast to approaches like Sudowe et al. [7] or Hoiem et al. [12], an active ground plane or horizon line estimation is not needed. Here, deviations of a flat world assumption and errors due to manual labeling are offline determined and considered for estimating valid window locations only from annotations.

This paper is structured as follows: The maximum likelihood solution for reducing the search space based on annotations is derived in section 2. The proposed MAP solution is than presented in section 3. In section 4 the effectiveness compared to the traditional uniform sliding window selection strategy is shown in a qualitatively evaluation. Section 5 provides a conclusion.

2 ML Search Space Estimation

The maximum likelihood (ML) solution for reducing the search space corresponds to find the most likely scanning window. This means, that for a particular scale only the most likely image region with the corresponding fixed classifier window is scanned. In order to estimate the ML solution from the pedestrian annotations, we assume a normally distributed pedestrian height and a normally distributed deviation from the flat world assumption. According to this assumption the optimal search space for a sliding window detector corresponds to an average pedestrian on the ground plane.

A shared concept of detectors which use a rigid representation model is a fixed feature constellation inside the window. Therefore, training a classifier on a rigid representation model using training samples with strong position variations is not effective and can lead to a loss of dominant property and meaningful characteristic [13]. Hence, the sample height is normed to a fix height, and the training data variation is mainly focused on capturing variation of pose, texture etc. and not of different person heights. For training a rigid representation model for a particular scale the samples are resized to a common size including a padding of pixels on top and at the bottom. Accordingly the classifier is optimized for a normed person height and the assignment between the training height and the optimal search space for an average person is logical. Here, we define the search space for only capturing the person height and not the padding pixels. Therefore the foot point of a person standing on the ground can directly be linked to a y coordinate and for classification the chosen padding pixels have to be added. A common choice for training a pedestrian classifier is the freely available INRIA dataset [14]. There the standard window (scale 1) complies with a common size of 64 × 128 pixels including a padding of approximately 16 pixels on top and at the bottom. So the search space for a classifier trained on INRIA has a height of 96 pixels and the bottom coordinate of the annotations correspond to the ground plane.

As basis for estimating optimal parameter settings we first analyze the dependence between the pedestrian height in pixels and the y coordinate of the associated bounding box. Figure 1 shows the scatter plots for the annotations of the Daimler Mono Pedestrian dataset [15] and the Caltech Pedestrian dataset [11]. In the case of the Daimler dataset only annotation from the first 10895 frames from the driving sequence are considered. For Caltech the sets 01-05 from the training sequences are used. Based on our assumption, the least squares model fit to the line $Y = \beta_0 + \beta_1 X$ through the data pairs (X_i, Y_i) with Y_i as the bottom coordinate of the search space and X_i the person height obtains the line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ where $\hat{\beta}_1 = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\sum XY = \sum_{i=1}^{n} X_i Y_i$, $\sum X^2 = \sum_{i=1}^{n} X_i^2$. Considering a probabilistic model for linear regression $Y_i = \beta_0 + \beta_1 X + \epsilon_i$ and assuming the ϵ_i are i.i.d. with $\mathcal{N}(0, \sigma^2)$. The least square estimate for $\mu_i = \beta_0 + \beta_1 x_i$ is exactly the ML estimate. For the model $Y_i = \mu_i + \epsilon_i$ the ML estimates of β_0 and β_1 are the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. For guessing a particular search space at a particular scale respectively height, say X, the regression model guesses $\hat{\beta}_0 + \hat{\beta}_1 X$. The ML search spaces for the Daimler and the Caltech dataset are depicted in figure 1. The average gradient magnitude image over the positive training examples of the INRIA dataset are overlaid, so that the person foot point corresponds to the ML solution and the person top corresponds to the upper y coordinate of the search space.

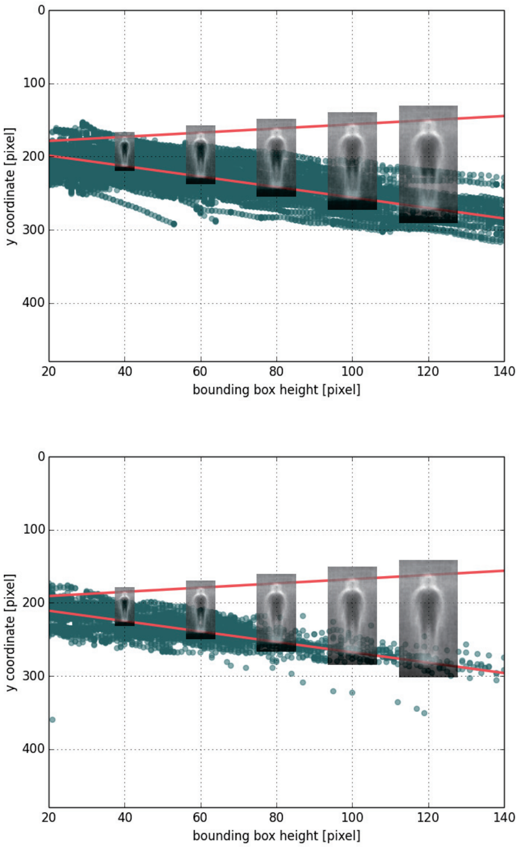

Figure 1. Distribution of the bounding box height with respect to y-bottom coordinate and resulting ML search space. On top for the Caltech dataset and on bottom for the Daimler dataset.

For evaluating how well the assumptions of an average flat world fits, we compare the horizon line estimated with our ML solution to the horizon line estimated with the given camera parameters from the Daimler dataset. If the camera tilt is small, estimation of the horizon line y_0 can be simplified by using the known camera height h_c, the object world height and object image height [12]. By assigning the average person (we set the average height $\bar{h}$ to 171.5cm and the standard deviation σ_h to 12cm; see [16]) to the scale 1 search space the resulting horizon position is given by $y_0 = 96\frac{h_c}{\bar{h}} + \hat{Y}_{96}$. Figure 2 illustrates the good agreement between both results, where the ML solution is on the right and the solution estimated with the external and internal camera parameters is shown on the left.

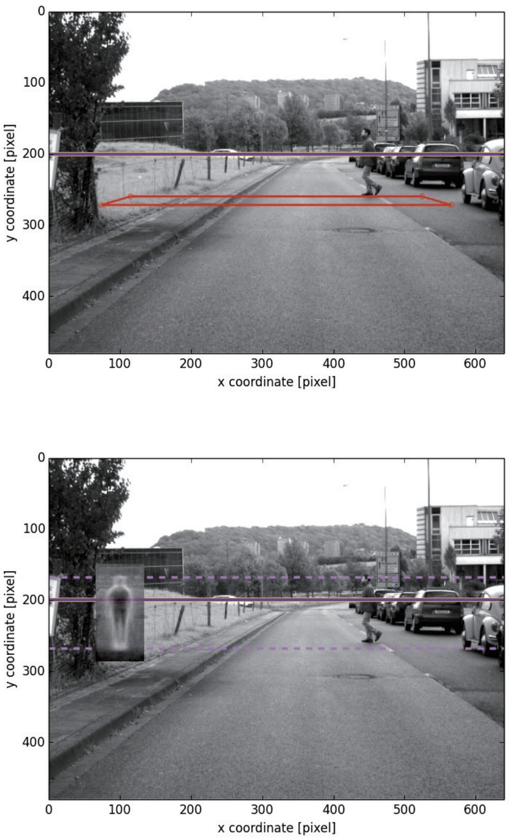

Figure 2. Horizon line estimation with the extrinsic and intrinsic camera parameter (top) and derived from the ML fit based on annotations and known height above ground (bottom).

3 MAP Search Space Estimation

The ML estimation of the reduced search space $\hat{\theta}_{ML}$ (the bottom point of the regression fit plus the chosen height rather scale) is optimal for a fixed camera to ground distance and for an average person standing on the flat ground. In many scenarios, there are some deviations to this ideal situation. The main errors are caused by the height variation of persons, deviation of a flat world, and annotation errors. For example, under a small camera tilt, which is valid for a ADAS setup, the foot point of a smaller person is related to a lower y coordinate (see [7]). Further, the same deviation to the ML search space can also be caused by an average height person standing on lower ground. These deviations are independent from each other and additive.

In order to find the optimal reduced search space, S is considered to be a random variable given the annotations of the person height H in pixels. The MAP estimate is defined as $\hat{\theta}_{MAP} = argmax_{s \in S} P(S|H)$ and can be written in terms of $\hat{\theta}_{ML}$ (see [17]) leading to $\hat{\theta}_{MAP} = \hat{\theta}_{ML} P(H)$. The definition of the prior $P(H)$ is based on the chosen probabilistic model for linear regression $Y_i = \beta_0 + \beta_1 X + \epsilon_i$. The difference between the observed y coordinates from the annotations Y_i and predicted outcome $\hat{Y}_i$ is the residual $e_i = Y_i - \hat{Y}_i$, which is the vertical distance to the regression line. Least squares (ML) estimate complies with minimizing the sum of squared errors (SSE) $\sum_{i=1}^{n} e_i^2$ and e_i can be seen as an estimate of the ϵ_i. The variation of the person height is proven to be normally distributed. In addition the deviation of the flat world is also assumed to be normally distrusted and independent from height variations. The annotation errors can be interpreted as additive Gaussian noise. Overall, one receives an error distribution, which is the superposition of the three error sources. In our experiments, the assumptions of a Gaussian distribution residual error fits well for the Caltech and Daimler dataset annotations.

In addition, to just modeling the prior based on the residual error distribution, the prediction interval of the regression fit is used. A prediction interval is an estimate of an interval in which future observations will fall with a certain probability, given what has already been observed. This captures the variability in the data. Here the upper and the lower bound of the prediction interval extends the ML search space for a given quantile and allows to reduce the search space at a chosen scale. These and the following derivation are based on classical regression theory (see for example Draper and Smith [18] for further details). For identically distributed errors the $\frac{\hat{\theta}_{ML} - \theta_j}{\hat{\sigma}}$ statistics follows a t-distribution with $(n-2)$ degrees of freedom and a normal distribution for a large number of annotations n. The ML estimate of that variance is given by $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$ so that $E(\hat{\sigma}^2) = \sigma^2$. This is also called the mean square error (MSE) of the regression. If the true model is known and hence $\hat{\beta_0}$ and $\hat{\beta_1}$ are the true parameters then the computed coefficients β_0 and β_1 are estimates of $\hat{\beta_0}$ and $\hat{\beta_1}$, respectively. The standard errors of the parameter can be neglected if the number of annotations is high or the regression parameters β_0 and β_1 are known. Contrary, the prediction interval would still have width and is a good choice to extend the search space for the detector. The standard deviation of a future observation at point x_p can be calculated as follows:

$$\hat{\sigma}_{y_p} = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x^2 - n\bar{x}^2}}. \tag{1}$$

The corresponding $(1-\alpha)$ prediction interval can be constructed using a t-quantile read at $n-2$ degrees of freedom:

$$y_p \in [\hat{y}_p - \hat{\sigma}_{y_p} t_{1-\alpha/2;n-2}, \hat{y}_p + \hat{\sigma}_{y_p} t_{1-\alpha/2;n-2}]. \tag{2}$$

The results for the annotation data for the selected sequences for the Caltech and Daimler dataset are shown in figure 3 and the estimated parameter for the search space reduction are summarized in table 1.

The limits of the prediction interval are used to estimate the MAP search space. By choosing a t-quantile and thereby the percentage of pedestrians inside the search space, a tradeoff between detection performance and runtime constraints can be achieved.

4 Results

In this section, we evaluate the effectiveness of the proposed approach and discuss several effects, which

Table 1. Parameters for the optimal search space estimation.

dataset	n	$\hat{\beta}_1$	$\hat{\beta}_0$	R^2	$\hat{\sigma}$
Daimler	8615	0.711	196.505	0.827	13.048
Caltech	52117	0.716	184.316	0.848	15.403

can lead to a non-optimal result. First, the MAP solution based on the annotations is compared to a MAP solution that takes only the persons height into account. On this account, the camera to ground distance is assumed to be fixed. Hence, all deviations from the ML search space bottom y coordinate are caused by the height variation. For modeling the prior, specific statistics about the person height distribution can be used (e.g. we set $\bar{h}$ to 171.5cm and the standard deviation σ_h to 12cm). With the chosen height distribution, one can estimate the new limits for every scale by using equation 2. The results with $t_{1-\frac{\alpha}{2}} = 1.96$ are shown in figure 3. The figures clearly show how the proposed model captures deviations from a fixed camera ground distance or violations from a flat world assumption. Too big or small annotation boxes can easily be considered as an additional Gaussian noise term. Normally one would expects that the variance of the residual error increases strongly for larger scales. Firstly, due to height distribution of occurring annotation heights in the data, the larger scales are under-represented (see [11]). Further, the larger scales are closer to the camera and in typical street scenarios there is less variation for short range distances to the car. Hence, errors compared due to height variations seem to be more common for close ranges. For small scales there is a gradual shift towards variation caused by changes in the camera ground distance. A constant variation of all error sources would lead to a heteroskedasticity of the error distribution. Classical linear regression models rely on the fact that there is no heteroscedasticity. Heteroscedasticity does not cause ordinary least squares coefficient estimates to be biased, but it can cause standard errors obtained from data analysis to be above or below the true variance. In the used annotation data, we could not determine this effect. Moreover, by dividing the pedestrian height into different scale ranges and individually estimating the error variance a too strong bias can be avoided. Motivated by the distribution of annotation heights in the data set, only the scales between 20 and 140 pixels are considered, which captures according to Dollár et al. [11] pedestrian from near, middle and far scales. Because the higher rather near scales are under-represented, a further division into the corresponding scales ranges is not applied.

In order to determine the goodness of the regression fit, the coefficient of determination R^2 is calculated. R^2 is the percentage of the variation that is explained by the model. The values of R^2 for the Caltech and Daimler dataset are shown in table 1.

Inside the chosen scale range the error has almost a constant variance. By using this error distribution to model the prior of the MAP search space estimate, we can compare the resulting reduced search space with an exhaustive search. When performing a sliding window classification, the image is traversed from the top-left corner with a certain stride on both axes. Figure 3 also includes the comparison of an exhaustive search with a minimum stride of 1 pixel to a MAP estimate with $\alpha = 0.025$ and $t_{1-\frac{\alpha}{2}} = 1.96$ for the scales between 20 and 140. Areas which are excluded from scanning, lie outside of the prediction interval. By adapting the confidence value for the prediction interval and the stride of the sliding window, one is able to gradually choose between fixed time constrains and the probability of missing persons outside the MAP search space.

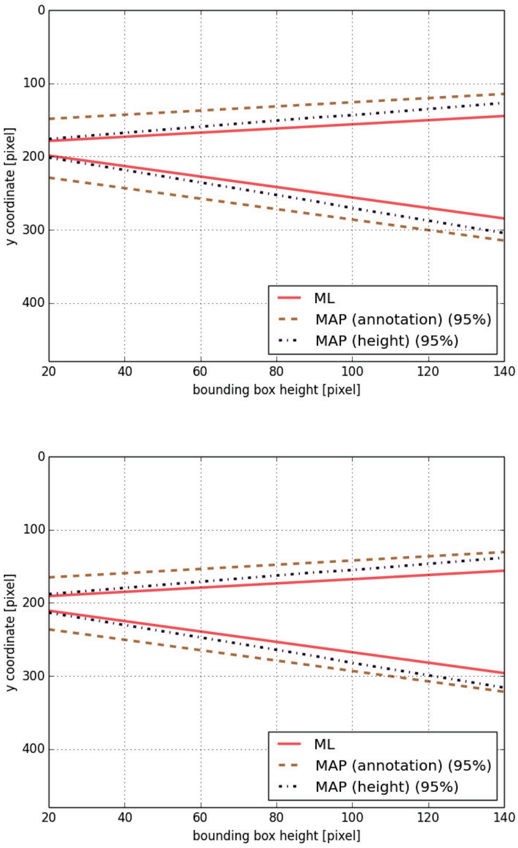

Figure 3. The MAP search space estimate for a prediction interval ($\alpha = 0.025$ and $t_{1-\frac{\alpha}{2}} = 1.96$) with a annotation based prior model is compared to the MAP solution with a prior model only from a person height statistics. On top for the Caltech dataset and on bottom for the Daimler dataset.

For a person height of 96 pixels complying with a window size of 64 × 128 pixels (scale 1) the MAP search space is visualized over the heatmap for all occurring annotation bounding boxes for heights smaller than 96 (see figure 4). By choosing a prediction level of 95% and as reference detector the approach of Dollár et al. [2], one is able to reduce the false positive rate in case of a persistent miss rate on both datasets. Accordingly, the performance is slightly improved. This performance boost in addition to the reduced runtime vanishes for smaller prediction levels. Leading to a tradeoff between detection performance and runtime constraints.

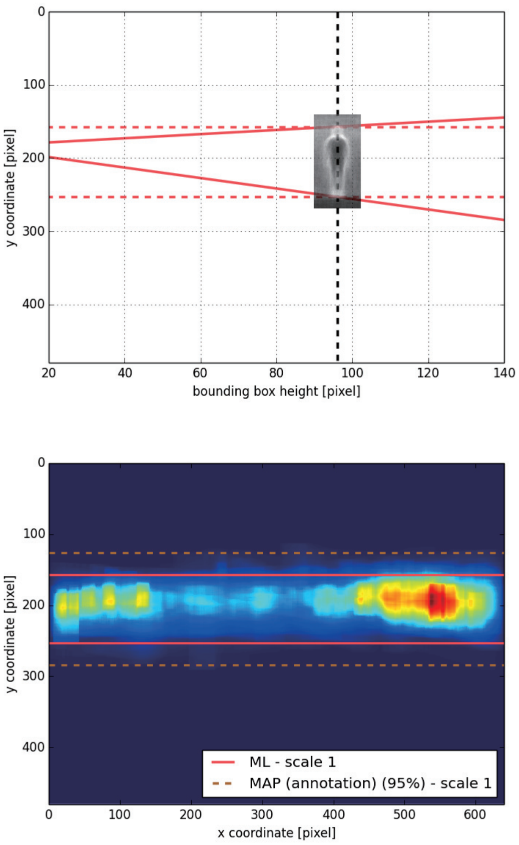

Figure 4. ML search space for the Caltech dataset for scale 1 (top; horizontal dotted lines). On bottom the heat map of the annotation bounding boxes with the MAP search space for 95% prediction limit (scale 1).

5 Conclusion

In this paper, we presented an MAP solution for reducing the search space of a sliding window pedestrian recognition scheme. In addition to a ML solution, the deviation of a flat world assumption and the real world person height distribution is considered in the model. All calculations are only based on annotation data from realistic monocular pedestrian datasets and also take manual annotation errors into account. The methods provides design parameters for the probability of missing persons outside of a detection corridor while taking maximum advantage of the sliding window scheme. The effectiveness of the proposed approach and several effects, which can lead to non-optimal output are discussed on the Daimler and on the Caltech datasets.

References

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan: "Object Detection with Discriminatively Trained Part-Based Models," *Transactions on Pattern Analysis and Machine Intelligence* , 2010

[2] P. Dollár, S. Belongie, and P. Perona: "The Fastest Pedestrian Detector in the West," *British Machine Vision Conference*, 2010

[3] D. Gerónimo and A.M. López: "Vision-based Pedestrian Protection Systems for Intelligent Vehicles," *Springer Briefs in Computer Science*, 2014

[4] M. Bertozzi, A. Broggi, R. Chapuis, F. Chausse, A. Fascioli, and A. Tibaldi: "Shape-based pedestrian detection and localization," *International Conference on Intelligent Transportation Systems*, 2003

[5] D.M. Gavrila, J. Giebel, and S. Munder: "Vision-based pedestrian detection: the PROTECTOR system," *Intelligent Vehicles Symposium*, 2004

[6] D.F. Llorca, M.A. Sotelo, A.M. Hellín, A. Orellana, M. Gavilán, I.G. Daza, and A.G. Lorente: "Stereo regions-of-interest selection for pedestrian protection: A survey," *Transportation Research Part C: Emerging Technologies* , 2012

[7] P. Sudowe and B. Leibe: "Efficient Use of Geometric Constraints for Sliding-Window Object Detection in Video," *International Conference on Computer Vision Systems*, 2011

[8] M. Pedersoli, A. Vedaldi, and J. Gonzàlez:, "A coarse-to-fine approach for fast deformable object detection," *Conference on Computer Vision and Pattern Recognition*, 2011

[9] C.H. Lampert, M.B. Blaschko, and T. Hofmann: "Beyond Sliding Windows: Object Localization by Efficient Subwindow Search," *Conference on Computer Vision and Pattern Recognition*, 2008

[10] J.H. Hosang, R. Benenson, and B. Schiele:, "How good are detection proposals, really?," *British Machine Vision Conference*, 2014

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona: "Pedestrian Detection: An Evaluation of the State of the Art," *Transactions on Pattern Analysis and Machine Intelligence*, 2012

[12] D. Hoiem, A. Efros, and M. Hebert: "Putting Objects in Perspective;" *International Journal of Computer Vision*, 2008

[13] S. Becker, A. Voelcker, H. Kieritz, W. Hübner, and M. Arens: "Automated Generation of High-Quality Training Data for Appearance-based Object Models," *Unmanned / Unattended Sensors and Sensor Networks*, 2013

[14] N. Dalal and B. Triggs: "Histograms of oriented gradients for human detection," *Conference on Computer Vision and Pattern Recognition*, 2005

[15] M. Enzweiler and D.M. Gavrila: "Monocular Pedestrian Detection: Survey and Experiments," *Transactions on Pattern Analysis and Machine Intelligence*, 2008

[16] Federal Statistical Office of Germany: "Height, weight and body mass index of the population by sex and age-groups; Results of Microcensus,"2009

[17] J.O. Berger: "Statistical decision theory and Bayesian analysis," *Springer Series in Statistics*, 1985

[18] N.R. Draper, H. Smith: "Applied regression analysis (2. ed.)," *Wiley series in probability and mathematical statistics*, 1981