19th Computer Vision Winter Workshop
Zuzana Kúkelová and Jan Heller (eds.)
Křtiny, Czech Republic, February 3–5, 2014

Pedestrian Recognition and Localisation in Image Sequences as Bayesian Inference

Tobias Klinger, Franz Rottensteiner, and Christian Heipke

Institute of Photogrammetry and GeoInformation,
Leibniz Universität Hannover, Germany
{klinger, rottensteiner, heipke}@ipi.uni-hannover.de

Abstract *An exhaustive-search people detector such as the HOG/SVM has at least two drawbacks w.r.t. the accuracy of recognition and geometric precision: first, it achieves high recall rates only among several false positive detections and second, the geometric precision of the positioning of the underlying object is poor due to the non-rigid body shape of people and background structures. However, the fact that the HOG/SVM does potentially provide high recall rates makes it a fair basis for hypothesise-and-validate-frameworks. We build upon the outcome of the HOG-detector and improve the recognition performance and geometric precision of the same using Bayesian Networks and apply statistical knowledge that we learn from training data for the definition of the probability functions. The approach is evaluated on two real image sequences and achieves results that can compare with the state of the art.*

1 Introduction

Automatic object detection is a key discipline in photogrammetry and computer vision. The term detection involves the recognition, i.e. the decision that an object of a specific object class is present and at least a coarse localisation of the object. Most state-of-the-art pedestrian detectors like the HOG/SVM [23] or the AdaBoost based detector [7] scan the entire image at different scales with a sliding window and classify its content as either person or background. Though these approaches give a solution to the recognition and localisation problem at the same time, the results are not particularly reliable and precise. In a comparative study of 16 different people detection systems [8] the authors point out that acceptable recognition rates are only achieved, if many false positive detections are accepted as well. Such systems, if applied permissively, have a high chance for false positive detections, because they usually rely on a single type of low-level features, which is typically not discriminative enough against similar object classes. It is hence reasonable to classify also against other similar object classes like in [16], or to evaluate additional information like foreground information [12], [24] or shape [13], [17] prior to further processing. Sequential processing has the drawback that false decisions taken at a single step cannot be recovered later. Other approaches involve context information, e.g. [22], which constrains detections to plausible regions in the image. [9], [11], [19] constrain the detections only to the ground plane, requiring a holistic understanding of the scene, which is a formidable task in itself on one hand, and which is very restrictive, because they disregard all objects that do not stand on the ground plane on the other.

However, there has been considerable success in the improvement of people recognition in [2], [11], [19], all of which use Bayesian Networks for the inference about the presence or absence of people and their positions. Bayesian Networks are directed graphical models in which observations and hidden parameters are treated as random variables in a generative Bayesian manner. The random variables are represented by nodes and the conditional independence properties of their joint distribution are represented by directed edges, see, e.g. [3] for details.

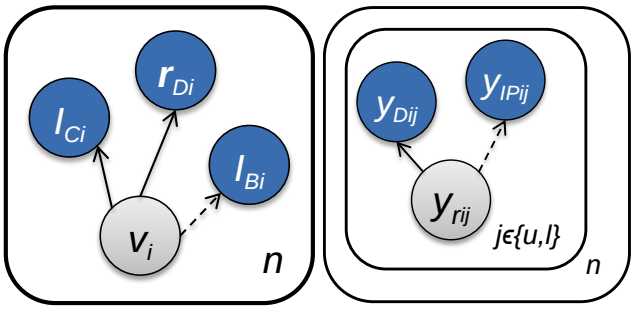
Though there is a lot of work related to the recognition of people, only few papers address the geometric accuracy of the detections. The positions of the detected persons are usually broken down to the location of the classification window, which does not always align well to the actual extents of people in the image and thus only gives an approximate position. For the evaluation of automatic detection results with reference data from manual annotations, the PASCAL VOC challenge, for instance, requires that the ratio between intersection and union area of the two rectangles is larger than 50% [10]. This criterion should highlight the object recognition performance and does not address the geometric accuracy of the detections. For many realistic applications like visual odometry with landmarks, collision avoidance in driver assistance systems or the analysis of motion and interactions of people in sport science or video surveillance, the geometric accuracy is crucial. In such applications, the 2D position of an object, usually its highest or lowest point in the image, is projected into 3D space. In [15] the importance of a correct segmentation of objects in the image for the geometric accuracy in 3D is pointed out. Comaniciou's Mean-Shift tracker [5], for instance, progressively finds the best fitting position of a tracked target by iteratively moving it to the region that best coincides with the colour histogram of the target, but only estimates the centroids of the objects, which makes the actual positioning in 3D difficult. In [19] the authors use stereo-image pairs as input and jointly estimate the object position on and the parameters of the ground plane in the scene. The locations of the pedestrians, given by a HOG-detector, are then optimised in 3D by joint prob-

abilistic modeling with the ground plane parameters. Here, it is indirectly assumed that the detector already delivers the correct bounding boxes in the image. In [6] pixel-wise segmentation of approximately positioned objects is conducted by integrating edge and colour cues. As the location estimated by the detector is only used as initialisation, the segmentation is prone to drift away from the underlying object.

In this work, we stick to the Bayesian probability theory and evaluate various sources of information about the presence or absence and the positions of people in a probabilistic model. In contrast to the related work, we do not require a holistic scene model and we restrict detections only to parts of the scene that are accessed by people during a training sequence. We aim to achieve state-of-the-art recognition results in challenging indoor and outdoor scenarios and to improve the geometric accuracy of the localisation of the detected objects at the same time. The validity of a detection hypothesis and the location in the image are treated as hidden parameters in two different Bayesian Networks. Like [6], we break with the assumption of unbiased results of the detector, but go further and learn the uncertainty of positioning people from image sequences. The prior and conditional probabilities for the Bayesian Networks are all learnt from training data. The recognition performance and the geometric accuracy are evaluated on two common benchmark datasets.

2 Method

The central building block for our work on people recognition is the HOG/SVM-detector, which is capable of achieving relatively high recall rates, but is also prone to false positive detections. Convenient detection results can only be achieved when the false positives are distinguished from the true positives. In this paper people recognition is stated as a binary classification problem in which the results of the HOG-detector as well as additional information are regarded as input for a joint probabilistic model. In order to achieve the highest possible recall rate, no thresholding is applied in the HOG/SVM framework. In the remainder of this section, we introduce two different Bayesian Networks, one for the solution of the recognition problem (Sec. 2.1) and one for the improvement of the localisation accuracy of people (Sec. 2.2). For the positioning we consider the bounding box which results from the application of the HOG-detector only as an approximate position and observe a second source of information about the object position that we derive from the analysis of optical flow points. The positions of people in the image are represented by minimal spanning rectangles around the visible parts of the persons. We model the highest and the lowest row coordinate of the persons as random variables and estimate the posterior position by applying Bayesian inference. The position estimated by the second graphical model is used as observed variable in the first graphical model. Using a refined position of a detection hypothesis has the advantage that a detection candidate will only be discarded, if even at the refined position the classifier does not strike. This gives rise to the possibility that misplaced detection windows - which is often a problem, e.g. when people are observed from the side - are corrected before they are further evaluated.



(a) (b)

Figure 1: Graphical models used (a) for people recognition and (b) for people localisation. Observable variables are shown in blue, the hidden parameters in gray. The dashed edges denote the possibility of the associated observations to be omitted, see text.

2.1 Model for People Recognition

As mentioned above, people recognition is regarded as a binary classification problem and hence we determine a binary label v_i that indicates whether the ith of n detection candidates observed in each frame by the HOG/SVM-detector really corresponds to a person or not. For simplicity of notation, we omit the indexes i in the remainder of this paper.

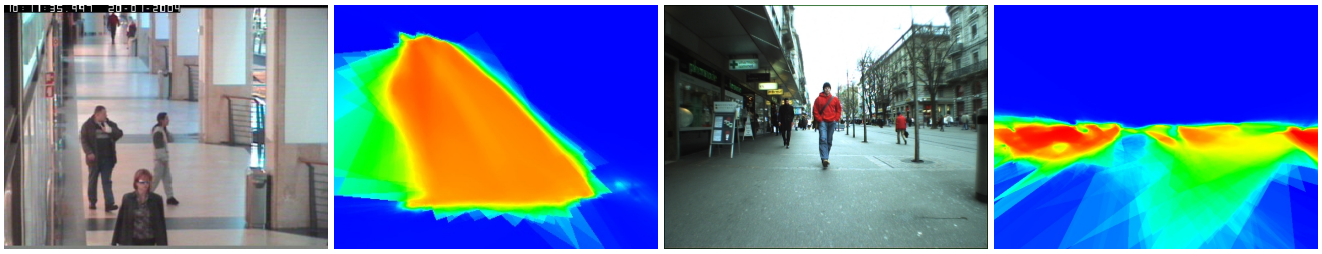
For the determination of v we apply Bayesian inference in the context of a directed graphical model, depicted in Fig. 1a. The observed variables, depicted as blue nodes in the model, are

- The surrounding rectangle $\mathbf{r}_D = [x_{Dl}, y_{Du}, x_{Dr}, y_{Dl}]^T$ around a person given by the HOG-detector, defined by its upper left (x_{Dl}, y_{Du}) and its lower right point (x_{Dr}, y_{Dl}) with their row (y) and column (x) coordinates
- The confidence value I_C that is proportional to the certainty about the binary classification (person vs. not person) of the SVM classifier used in the HOG framework
- An observation obtained from background subtraction, i.e. the fraction of foreground pixels I_B inside $\mathbf{r}_D$.

Following the standard notation for graphical models (see, e.g. [3]), each directed edge represents a conditional probability function for the child node given the parent node. The joint probability density of the involved variables can be written in accordance with the network design in Fig. 1a as Eq. (1):

$$\begin{aligned} P(v|I_B, I_C, \mathbf{r}_D) &\propto P(v, I_B, I_C, \mathbf{r}_D) \\ &= P(v)P(I_B|v)P(I_C|v)P(\mathbf{r}_D|v) \end{aligned} \tag{1}$$

For each edge in Fig. 1a we train an individual Random Forest (RF) classifier [4] using the according observation I_C, $\mathbf{r}_D$ or I_B as features. For training, we apply the HOG-detector on image sequences with available annotations of the bounding rectangles around the visible people in the scene. The detections are then divided into sets of positive



(a) ThreePastShop1 image sample (b) $P(v_i = 1|\mathbf{r}_{Di})$ (c) ETHZ-Bahnhof image sample (d) $P(v_i = 1|\mathbf{r}_{Di})$

Figure 2: Posterior probability of a detection to be correct given the center of the rectangle around the detected object in the image. For processing, the upper left and lower right points of the rectangle is used. Red: $P(v_i = true|\mathbf{r}_{Di})$ is high, blue: $P(v_i = true|\mathbf{r}_{Di})$ is low.

training samples (for v =true) and negative training samples (for v =false) by validation with reference data, using the bounding box intersection-over-union score [10]. The priors $P(v)$ are defined as the ratios of true positive and false positive detections in the training data.

The edge related to the conditional probability density function (pdf) $P(\mathbf{r}_D|v)$ represents the probability, that $\mathbf{r}_D$ is the surrounding rectangle if a person is present or absent. Given the posterior probability $P(v|\mathbf{r}_D)$ by classification with the RF and the priors $P(v)$, the likelihood $P(\mathbf{r}_D|v)$ as required for the factorisation of the joint pdf (Eq. (1)) can be written as

$$P(\mathbf{r}_D|v) \propto \frac{P(v|\mathbf{r}_D)}{P(v)}.$$

In Fig. 2 example images for both datasets used in this work are shown together with the posterior probabilities that are generated by a classifier (the RF used for visualisation is only trained with the 2D center point of the rectangle as feature) for each possible position of $\mathbf{r}_D$ in the image. The incorporation of the pdf $P(\mathbf{r}_D|v)$ is beneficial for two reasons; first, we achieve a very fine differentiation of likely vs. unlikely regions for detections and second, we do not require to interpret the scene geometry prior to the detection.

The pdf $P(I_C|v)$ is the probability density for a confidence value being observed given the presence or absence of a person. Related to posterior probability and the priors, the likelihood $P(I_C|v)$ can be written as

$$P(I_C|v) \propto \frac{P(v|I_C)}{P(v)}.$$

The pdf $P(I_B|v)$ related to the results of background subtractions is integrated into our model due to the assumption that people differ from the background because of their motion, hence the observations I_B is a strong indicator for the presence or absence of a person. The likelihood can be written as

$$P(I_B|v) \propto \frac{P(v|I_B)}{P(v)}.$$

We derive I_B only from images captured by a camera with constant exterior orientation (w.r.t. 6 degrees of freedom), i.e. where an algorithm for background subtraction such as [21] can be applied without adjustments. Therefore, the edge connected with I_B is drawn as a dashed line in Fig. 1a. For image sequences from moving camera platforms we do not apply background subtraction and the variable I_B and the according likelihood $P(I_B|v)$ is excluded from Eq. (1). For image sequences captured by a static camera, we apply the algorithm of [21] for background subtraction.

The unknown parameter v is determined to be the label that achieves the maximum a posteriori (MAP) probability among the two possible states (true and false) given the observations. The decision rule is formalised in Eq. (2).

$$v = \begin{cases} true, & \text{if} \frac{P(v=true,I_B,I_C,\mathbf{r}_D)}{P(v=false,I_B,I_C,\mathbf{r}_D)} > 1, \\ false, & \text{otherwise.} \end{cases} \quad (2)$$

2.2 Model for People Localisation

In the model described in Sec. 2.1 the minimal spanning rectangle around a person is considered observable. In fact, the location that is given by the HOG/SVM-detector is only an approximation to the true position. In this section, we define a second graphical model, see Fig. 1b, which considers the row coordinates of the highest (y_{riu}) and lowest points (y_{ril}) of a person related to the ith detection as hidden parameters. Again, we omit the indexes i for the sake of simplicity. As observed variables of this model we consider the upper and lower row coordinates y_{Du} and y_{Dl} of the HOG-detection window as well as an additional pair of observations of the row coordinates that we derive from the analysis of optical flow, i.e. y_{IPu} and y_{IPl}.

We define y_{IPj} as the row coordinates of the highest and lowest interest points on the visible parts of a person, that are tracked by an optical flow algorithm. We apply the algorithm of [20] for the selection of interest points and track them by the algorithm of [14]. For each hypothesis about the presence of a person given by the HOG-detection we apply the following strategy for the measurement of y_{IPj} (see Figs. 3 and 4 for an illustration):

1. We establish a search space for the person by expansion of the rectangle given by the HOG-detector (visualised as red rectangles in Fig. 3b-e) in vertical direction by a third of its size, in order to assure that the person is within the search space. We consider the upper 25% of this area as search space for the head point (upper yellow rectangles in Fig. 3b-e) and the lower 25% as search space for the foot point (lower yellow rectangles).

(a) (b) (c) (d) (e)

Figure 3: Examples from the ETHZ-sequence for observations used for the localisation of people. In (a) a schematic representation of the pdfs used for the inference of the posterior locations is given. The red curves represent the pdfs based on observations by the HOG-detector, the blue curves those for the observations based on optical flow. The black rectangle symbolises the true position. In (b)-(e), the yellow rectangles indicate the raw detection as result of the HOG-detector. The blue rectangles indicate the search spaces for the head and foot position, respectively. The blue horizontal lines inside the rectangles are the measured positions by the analysis of optical flow, the green horizontal lines the inferred posterior position. In (e) the position of the person in the background is not estimated by our approach correctly due to partial occlusion by the person in the foreground.

2. In both regions, we take all optical flow vectors ending in the upper and lower search space, respectively, of the current image (indicated by the short lines in Fig. 3b-e).

3. We generate a histogram of magnitudes of optical flow vectors, using 10 histogram bins and consider only flow vectors with a magnitude between 0 and 30 pixels. If the histogram has two or more local maxima, we suppose that the flow vectors related to the maximum with the smallest magnitudes originate from the background and discard the flow vectors from further consideration. Of the remaining flow vectors we also remove those that do not have more than a minimum number of neighbours in a predefined radius (we set the minimum number of neighbours to three and the radius to 20 pixels; the discarded flow vectors are visualised by red, the remaining flow vectors by green lines in Fig. 3b-e). We set the image row coordinates of the highest and lowest interest points that remain as observations of the head point (y_{IPu}) and food point (y_{IPl}), respectively (indicated by the blue horizontal lines in Fig. 3). If the histogram only has one local maximum, we do not evaluate the observation x_{IP} for the according person in the current image, because the interest points cannot be separated into points originating from the foreground and the background by our approach.

We model the likelihoods for the measurements y_{IPj} and y_{Dj} to be observed at a distance Δy from the true position y_{rj}, i.e. $\Delta y_{IPj} = y_{IPj} - y_{rj}$ and $\Delta y_{Dj} = y_{Dj} - y_{rj}$, respectively, by normal distributions:

$$P(y_{IPj}|y_{rj}) \propto e^{-\frac{1}{2}(\frac{\Delta y_{IPj} - \mu_{\Delta yIPj}}{\sigma_{\Delta yIPj}})^2} \tag{3}$$

and

$$P(y_{Dj}|y_{rj}) \propto e^{-\frac{1}{2}(\frac{\Delta y_{Dj} - \mu_{\Delta yDj}}{\sigma_{\Delta yDj}})^2} \tag{4}$$

with mean $\mu_{\Delta yDj}$ and $\mu_{\Delta yIPj}$, respectively, and standard deviation $\sigma_{\Delta yDj}$ and $\sigma_{\Delta yIPj}$, respectively. The parameters of the pdf in Eq. (3) are determined from the distribution of deviations of the measured positions y_{IPj} from the (true) positions given by reference data and those of Eq. (4) from the deviations of y_{Dj} from the reference data. A visualisation of two exemplary distributions together with the fitted Gaussians is given in Fig. 4.

Given the observations y_{Dj} and y_{IPj} measured for each detection in each consecutive frame in the evaluation phase and the parameters $\mu_{\Delta yDj}$, $\sigma_{\Delta yDj}$, $\mu_{\Delta yIPj}$ and $\sigma_{\Delta yIPj}$ of the pdfs (3) and (4) learnt from training data, the posterior position y_{rj} can be inferred for each detection candidate as the expected value

$$E(y_{rj}) = \mu_{rj}$$
$$= \frac{\sigma_{rj}^2}{\sigma_{\Delta yDj}^2}(y_{Dj} - \mu_{\Delta yDj}) + \frac{\sigma_{rj}^2}{\sigma_{\Delta yIPj}^2}(y_{IPj} - \mu_{\Delta yIPj}) \tag{5}$$

with

$$\sigma_{rj}^2 = (\frac{1}{\sigma_{\Delta yDj}^2} + \frac{1}{\sigma_{\Delta yIPj}^2})^{-1} \tag{6}$$

The first term of Eq. (5) considers the influence of the observed position by the HOG-detector, weighted by the variance of the measurements in the training sequence. The second term refers to the position measured by the analysis of the optical flow vectors, also weighted by the variance of the measurements. The observation with the lower variance hence has the stronger influence on the posterior position

μ_{rj}. The subtrahends in the brackets incorporate the mean deviations of the measurements from the reference data as corrections. If the head or the feet point cannot be measured by the analysis of the optical flow vectors, the second terms of Eq. (5) and (6) are set to zero and only the first term, related to the HOG-detection, influences the posterior.

3 Experiments and Results

Experiments are conducted on two publicly available datasets, one from an indoor sequence with constant exterior orientation of the camera, the CAVIAR dataset [1] and the other from the ETHZ dataset [9] captured from a moving platform in an outdoor scenario. For the experiments involving our method from Sec. 2.2, the posterior row coordinates are used as the row coordinates of the observed rectangle in the graphical model from Sec. 2.1, maintaining the column coordinates of the HOG-detections.

3.1 Datasets

In the CAVIAR scenario, the sequence ThreePastShop1, consisting of 1650 images, is taken for training and the sequence ThreePastShop2 with 1521 images for testing. From the ETHZ dataset we take the Bahnhof-sequence of 1000 images, split the data in two halves and apply cross-validation. The training and test sequences hence follow on from one another. As the camera is mounted on a moving platform in the ETHZ-sequence, the observation I_B is excluded from the graphical model in this case. Though in the ETHZ-sequence the position of the camera changes over time, the tilt angle relative to the ground does not change significantly. We hence assume that the probabilities related to the position in the image are transferable from training to test sequences within an acceptable range of validity. The HOG/SVM-detector is configured without internal threshold, so that the results are as complete as possible. Only people with a minimum height of 48 pixels[1] are considered for processing. The bounding rectangles are shrinked in order to compress the systematic margin around people in the training data.

3.2 Detector Recognition Accuracy

The accuracy of the people detector is evaluated in terms of its recall capability and the number of false positive detections per image (fppi). Experiments with the Bayesian Network (Fig. 1a) are conducted with and without the refinement of the position by inference on the graphical model in Fig. 1b. The results are compared with results achieved by the classification of all observations in a single feature vector $[x_{Dl}, y_{Du}, x_{Dr}, y_{Dl}, I_C, I_B]^T$ (with I_B only evaluated for the static camera) with a Naive Bayes model [3], a Gaussian Mixture model (GMM) [18] and Random Forests. The results are plotted in Fig. 5.

We conclude from the plots, that among the three classification techniques Naive Bayes, GMM and RF, the Random Forest features the highest recall rates, though also the false-positive (FP) rates are higher than those of the rest. The Naive Bayes and the GMM classifier deliver similar recall rates with more false positives per image but also a higher recall rate in case of the CAVIAR-sequence on the side of the former one. Using the Bayesian Network from Fig. 1a, the results for the fppi-rates in the ETHZ-sequence are a few percent higher than those of the RF classifier, but the recall rate could be increased slightly as well. The application of the Bayesian Networks on the CAVIAR-sequence delivers less false positives and similar recall values than the RF classier. The extension of the model with the observation of the optical flow points (Fig. 1b) increases the recall rate in the upper part of the curve only in the ETHZ-sequence, while also the FP-rate increases slightly. W.r.t. the quality ($Q = \frac{TP}{TP+FN+FP}$, according to the measure of true positives (TP), false negatives (FN) and FPs) of object detection, in the CAVIAR-sequence the best results are achieved by the Bayesian Networks, both of which achieve the same score ($Q = 64\%$) superior to the quality achieved with the RF (63%). In the ETHZ-sequence, the quality does not differ considerably between the Random Forests and the Bayesian Networks (all around 59%). Compared to the results one achieves by varying the internal threshold of the HOG/SVM, plotted as red dashed lines, where the maximal recall score is achieved at a fppi rate of 6.6 in the CAVIAR-sequence and 6.3 in the ETHZ-sequence, respectively, all of the applied classifiers reduce the fppi at least by a factor of four, while the recall drops, in the best case, only about two percent in the CAVIAR test case and about six percent in the ETHZ test case, respectively. The achieved recognition accuracy is comparable to that of the single frame but stereo based detector in [19] and outperforms [9], see Fig. 5b.

3.3 Detector Position Accuracy

In this section we evaluate the average error of the measured and inferred positions by comparison with reference data. For each true positive detection, the deviations of the row coordinates of the head and the feet points from the reference data are recorded in a histogram. From the 95%-quantile of this histogram a Gaussian $N(\mu_{0.95}, \sigma^2_{0.95})$ with mean $\mu_{0.95}$ and standard deviation $\sigma_{0.95}$ is fitted. For the evaluation of the position accuracy the mean value of the Gaussian is tested for accordance with reference data using a statistical test of differences between two mean values. For the reference data a labeling uncertainty of one pixel is assumed so that the reference values follow a standard normal distribution $N(\mu_{ref}, \sigma^2_{ref}) = N(0, 1)$. As metric for similarity between the mean of the measurements and the mean of the reference data we apply

$$y_d = \frac{\mu_{0.95} - \mu_{ref}}{\sigma_d} \tag{7}$$

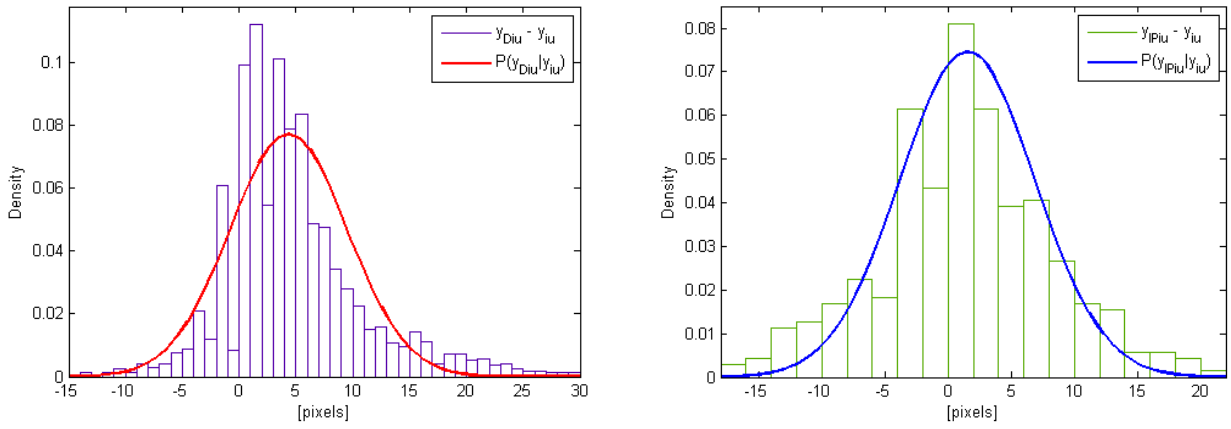
with

$$\sigma_d^2 = \sigma^2_{0.95} + \sigma^2_{ref} \tag{8}$$

resulting from variance propagation of uncorrelated observations.

The average deviation of the observations y_{Dj} and y_{IPj} from the reference data has already been considered as corrections in Sec. 2.2, see also Fig. 4. We calculate the metric

[1] We apply the HOG/SVM-detector of OpenCV, trained with the INRIA person dataset (http://pascal.inrialpes.fr/data/human/) with a height of 96 pixels for the people. We scale the input images by the factor 2 and achieve detections of people appearing with a minimal height of 48 pixels.



(a) (b)

Figure 4: Example histograms of differences in the row coordinates (a) between HOG-detection result and reference and (b) between IP-based locations and reference data. The differences are normalised according to a height of 96 pixel. The distributions are approximated by normal distributions.

	$\mu_{0.95}[\%]$	$\mu_{0.95}[px]$	$\sigma_{0.95}[px]$	y_d
HOG-Head	0.8	0.8	1.6	0.42
HOG-Feet	2.6	2.5	3.6	0.67
IP-Head	5.9	-5.7	5.1	-1.10
IP-Feet	2.0	1.9	4.3	0.43
Inferred Head	1.4	-1.3	2.2	-0.54
Inferred Feet	0.5	0.5	2.7	0.17

Table 1: CAVIAR-sequence: Difference between measured and inferred row coordinates from reference values in percent of the object height and in pixels.

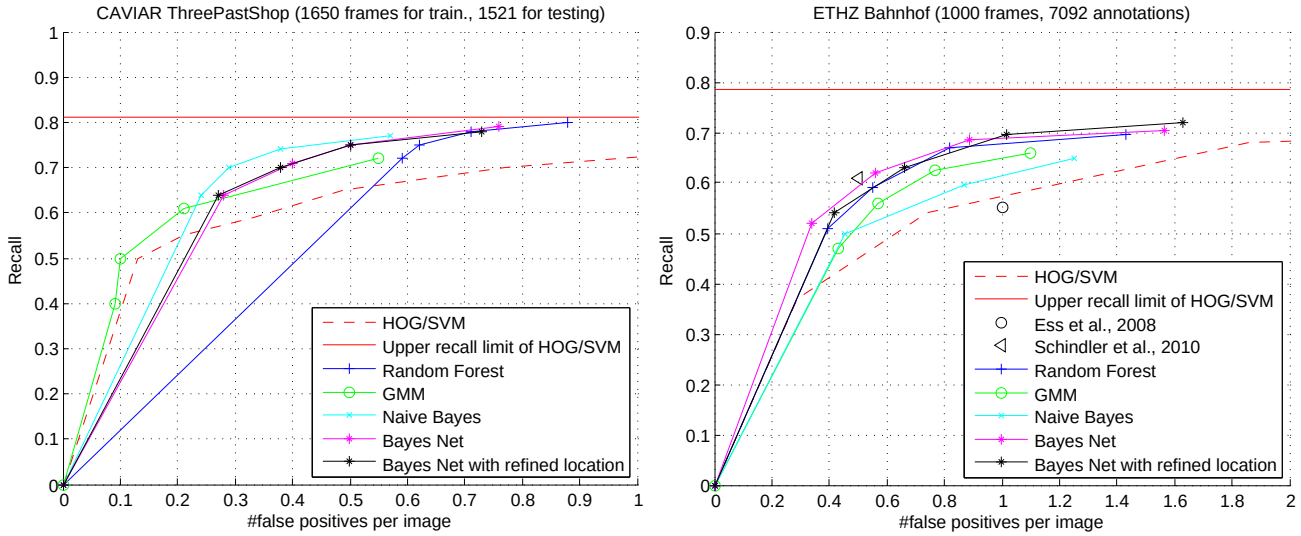
	$\mu_{0.95}[\%]$	$\mu_{0.95}[px]$	$\sigma_{0.95}[px]$	y_d
HOG-Head	5.2	5.0	5.4	0.91
HOG-Feet	0.4	0.4	3.3	0.12
IP-Head	0.9	0.9	6.4	0.14
IP-Feet	1.0	-1.0	5.3	-0.19
Inferred Head	0.6	-0.6	4.9	-0.12
Inferred Feet	0.4	-0.4	2.7	-0.14

Table 2: ETHZ-sequence: Difference between measured and inferred row coordinates from reference values in percent of the object height and in pixels.

y_d for the observed positions and the inferred positions as measure for accordance with the reference data and compare the results in Tab. 1 and 2. The deviations from the reference data are normalised w.r.t. a standard height of 96 pixels. From Eq. (5) it can be concluded, that among the mean values of the positions given by the HOG-detector (referred to as HOG-head and HOG-feet in the tables) and the ones given by the analysis of optical flow (IP-head and -feet) the one with the lower standard deviation has the stronger influence on the posterior, which in any of the test cases (Inferred Head and Feet in the CAVIAR- and ETHZ-sequence) is the position given by the detector. From the tables we conclude that the applied approach does not improve the position accuracy, if the position given by the detector already lies in the sub-pixel domain. In turn, when the error lies in the magnitude of some pixels, the inference of the posterior does enhance the alignment of the posterior positions with the reference. In either case, where the mean deviation of the position given by the detector lies around two and five pixels (see HOG-Feet in Tab. 1 and HOG-Head in Tab. 2), the posterior positions coincide much better with the reference data, which is reflected in the smaller values for y_d.

4 Conclusions

We conclude from this work that by the joint evaluation of the available information in the image that is linked to hidden parameters by a graphical model, the detection performance can be improved, even without the understanding of the 3D scene geometry and with a single camera as measuring device. Our approach leads to recognition results that are comparable with the state-of-the-art and at the same time improves the geometric accuracy of the results. The proposed method is a good starting point for tracking-by-detection systems, because the number of false positive detections from the underlying people detector that would lead to spurious trajectories are reduced significantly. The average geometric accuracy of the estimated location of pedestrians in the image is in any case around one pixel. We estimated the unknown parameters, i.e. the validity flag of a detection and its refined position, in two different graphical models. We plan to combine these models and estimate the parameters in a joint probabilistic model in future work, which opens the possibility for that position to be assigned to the detection, which most supports the joint probability.



(a) Recall and fppi values achieved in the CAVIAR-sequence

(b) Recall and fppi values achieved in the ETHZ-sequence

Figure 5: Recall and fppi values of the investigated classifiers and our proposed methods plotted in (a) for the CAVIAR-sequence and in (b) for the ETHZ-sequence. Also the HOG/SVM applied with internal thresholding (red dashed line) is drawn as baseline. The red horizontal lines indicate the maximum recall values reached by the HOG/SVM, towards which the red dashed line converges at fppi=6.6 in (a) and at fppi=6.3 in (b), respectively. In (b) also two samples from the results of related work are depicted.

References

[1] Caviar test case scenarios. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/, 2004.

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[3] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564 – 577, 2003.

[6] Q. Dai and D. Hoiem. Learning to localize detected objects. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3322–3329. IEEE, 2012.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Carlo Tomasi Cordelia Schmid, Stefano Soatto, editor, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886 –893 vol. 1. IEEE Computer Society, Los Alamitos, june 2005.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (4).*, pages 743–761, 2011.

[9] A. Ess, B. Leibe, K. Schindler, , and L. van Gool. A mobile vision system for robust multi-person tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8. IEEE Press, June 2008.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[11] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[12] S. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *European Conference on Computer Vision, 2006*, volume 3954 of *LNCS*, pages 133–146, 2006.

[13] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 878–885. IEEE, 2005.

[14] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, volume 81, pages 674–679, 1981.

[15] M. Menze, T. Klinger, D. Muhle, J. Metzler, and C. Heipke. A stereoscopic approach for the association of people tracks in video surveillance systems. *PFG Photogrammetrie, Fernerkundung, Geoinformation*, 2013(2):83–92, 05 2013.

[16] B. Ommer, T. Mader, and J. M. Buhmann. Seeing the objects behind the dots: Recognition in videos from a moving camera. *International Journal of Computer Vision*, 83(1):57–71, 2009.

[17] D. Ramanan. Using segmentation to verify object hypotheses. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[18] D. A. Reynolds. Gaussian mixture models. in: Encyclopedia of biometrics, pp. 659-663. springer us,

2009.

[19] K. Schindler, A. Ess, B. Leibe, and L. Van Gool. Automatic detection and tracking of pedestrians from a moving stereo rig. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):523–537, 2010.

[20] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994.

[21] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[22] A. Torralba and P. Sinha. Statistical context priming for object detection. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 763–770. IEEE, 2001.

[23] P. Viola, M.J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[24] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1208–1221, 2004.