Pedestrian Detection by Probabilistic Component Assembly

Martin Rapus[1,2], Stefan Munder[1], Gregory Baratoff[1], and Joachim Denzler[2]

[1] Continental AG, ADC Automotive Distance Control Systems GmbH
Kemptener Str. 99, 88131 Lindau, Germany
{martin.rapus,stefan.munder,gregory.baratoff}@continental-corporation.com
[2] Chair for Computer Vision, Friedrich Schiller University of Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany
joachim.denzler@uni-jena.de

Abstract. We present a novel pedestrian detection system based on probabilistic component assembly. A part-based model is proposed which uses three parts consisting of head-shoulder, torso and legs of a pedestrian. Components are detected using histograms of oriented gradients and Support Vector Machines (SVM). Optimal features are selected from a large feature pool by boosting techniques, in order to calculate a compact representation suitable for SVM. A Bayesian approach is used for the component grouping, consisting of an appearance model and a spatial model. The probabilistic grouping integrates the results, scale and position of the components. To distinguish both classes, pedestrian and non-pedestrian, a spatial model is trained for each class. Below miss rates of 8% our approach outperforms state of the art detectors. Above, performance is similar.

1 Introduction

Pedestrian recognition is one of the main research topics in computer vision with applications ranging from security problems, where e.g. humans are observed or counted, to automotive safety area, for vulnerable road user protection. The varying challenges are given by appearance of pedestrians, due to clothing and posture, and occlusions, for example pedestrians walking in groups or behind car hoods. For automotive safety applications the real-time performance needs to be combined with high accuracy and low false positive rate.

Earlier approaches employed full-body classification. Most popular: Papageorgiou et al. [12] applies Haar-wavelets with SVM [15]. Instead of SVM, a cascade based on AdaBoost [6] is used by Viola and Jones [16], to achieve real-time performance. An extensive experimental evaluation of histograms of oriented gradients (HOG) for pedestrian recognition is made by Dalal and Triggs [2]. In place of the constant histogram selection [2], Zhu et al. [19] use a variable selection made by an AdaBoost cascade, which achieves better results. Gavrila and Munder [7] recognize pedestrians with local receptive fields and several neural networks.

J. Denzler, G. Notni, and H. Süße (Eds.): DAGM 2009, LNCS 5748, pp. 91–100, 2009.

The achieved performance with full-body classification is still not good enough to handle the big variability in human posture. To achieve better performance, part-based approaches are used. These approaches are more robust against partial occlusions. Part-based approaches often consist of two steps, the first one detects components, mostly by classification approaches, while the second step groups them to pedestrians. One possible way to group components is to use classification techniques. Mohan et al. [11] use the approach proposed in [12] for the component detection. The best results per component are classified by a SVM into pedestrian and non-pedestrian. In Dalal's thesis [3], the HOG-approach [2] is used for the component detectors. A spatial histogram for each component weighted by the results is classified by a SVM. Felzenszwalb et al. [5] determine the component model parameters (size and position) in the training process. For the pedestrian classification the HOG component feature vectors and geometrical parameters (scale and position) are used as input for a linear SVM.

The fixed ROI configuration used in these approaches puts a limit on the variability of part configurations they can handle. To overcome this limitation, spatial models that explicitly describe the arrangement of components were introduced. In general, these approaches incorporate an appearance model and a spatial model. One of the first approaches is from Mikolajczyk et al. [10]. The components are detected by SIFT-like features and AdaBoost. An iterative process with thresholding is used to generate the global result via a probabilistic assembly of the components, using the geometric relations: distance vector and scale ratio between two parts, modeled by a Gaussian. Wu and Nevatia [18] use a component hierarchy with 12 parts and the full-body as root-component. The component detection is done by edgelet features [17] and Boosting [14]. For the probabilistic grouping the position, scale and a visibility value is incorporated. Only the inter-occlusion of pedestrians is considered. The Maximum-A-Posteriori (MAP) configuration is computed by the Hungarian algorithm. All results above a threshold are regarded as pedestrian. Bergtholdt et al. [1] use all possible relations between 13 components. For the component detection SIFT and color features are classified through randomized classification trees. The MAP configuration is computed with A*-search. A great number of parts is used by the last two approaches for robustness against partial occlusions. The computation time for the probabilistic grouping grows non-linearly with the number of components used, and with the number of component detection results. As a consequence these probabilistic based methods have no real-time performance on an actual desktop PC.

Our approach is part-based. For real-time purpose our pedestrian detector is divided into the three parts, head-shoulder, torso, legs and for better classification performance we distinguish between frontal/rear and side view. HOGs [2] are used as component features. We make use of a variable histogram selection by AdaBoost. The selected histograms are classified through a linear SVM. Because similar histograms are selected with weighted fisher discriminant analysis (wFDA) [9] in comparison to a linear SVM, but in less training time, we apply wFDA as weak classifier. A Bayesian-based approach is used for component

grouping. To reduce the number of component detections thresholding is applied, keeping 99% true positive component detection rate. Our probabilistic grouping approach consists of an image matching and a spatial matching of the components. To use the component results for the image matching they are converted into probabilistic values. Invariance against scale and translation is achieved by using the distance vector, normalized through scale, and the scale ratio between two components. In comparison to existing approaches the spatial distributions are not approximated, instead the distribution histograms are used directly. We also differentiate component arrangements by class. Below miss rates of 8% our approach outperforms state of the art detectors. Above, performance is similar.

The paper is organized as follows. Sect. 2 describes the component detection step, followed by the component grouping through a probabilistic model in Sect. 3. The results for the component detection and grouping step are discussed in Sect. 4. The conclusion forms Sect. 5 and the paper ends with an outlook in Sect. 6.

2 Component Detection

HOG features were proven best in [2], and thus adopted here for the component detection. The averaged gradient magnitude and HOG images for our components, derived through the INRIA Person dataset [2], are visualized in Fig. 1 and Fig. 2. Instead of the histograms the corresponding edges with weighted edge length are shown. Pedestrian contours are well preserved in the average edge images, while irrelevant edges are suppressed. A (slight) difference can be seen in the head component. In the frontal view, the whole contour is preserved and in the side view it is only the head contour, while the shoulder contour is blurred. Two different methods for the histogram selection are examined. One is a constant selection [2]: the image is divided into non-overlapping histograms, followed by an extraction of normalized blocks neighboring histograms. The other approach is similar to [19] and uses variable selection. The best histogram blocks (varying size and position) are selected using AdaBoost. We use the weighted Fisher discriminant analysis [9] as weak classifier. The classification of the generated feature vector is done by a linear SVM.

3 Probabilistic Component Assembly

This step builds the global pedestrian detections out of the detected components $V = \{v_{HS}, v_T, v_L\}$, where the superscripts HS, T and L stand for head-shoulder, torso and legs respectively, by applying the appearance and the spatial relationship.

The probability $P(L|I)$ to find a pedestrian, consisting of the mentioned components, with configuration $L = \{\mathbf{l_{HS}}, \mathbf{l_T}, \mathbf{l_L}\}$ in the actual image I, with $\mathbf{l_i}$ as position and scale for the ith component, is given by Bayes rule:

$$P(L|I) \propto P(I|L) \cdot P(L) \ . \tag{1}$$

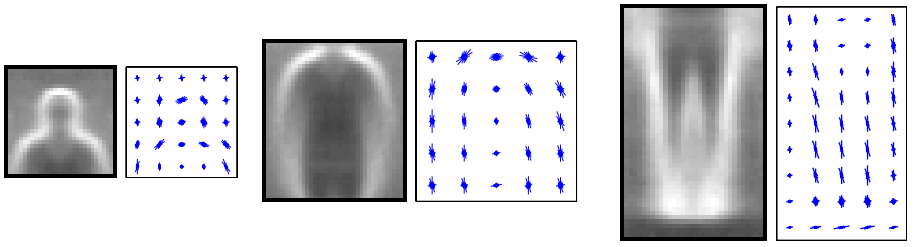


Fig. 1. Average gradient magnitudes and average HOGs for the frontal/rear view components (head, torso and legs) - INRIA Person dataset

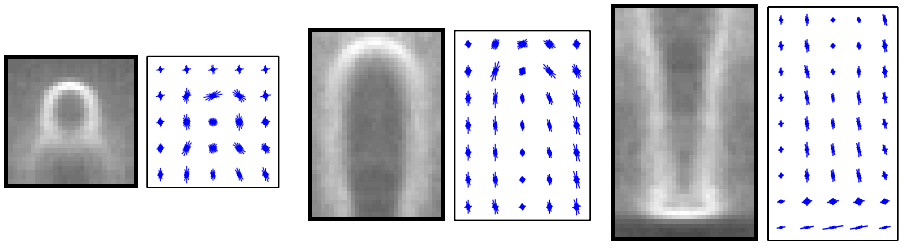


Fig. 2. Average gradient magnitudes and average HOGs for the side view components (head, torso and legs) - INRIA Person dataset

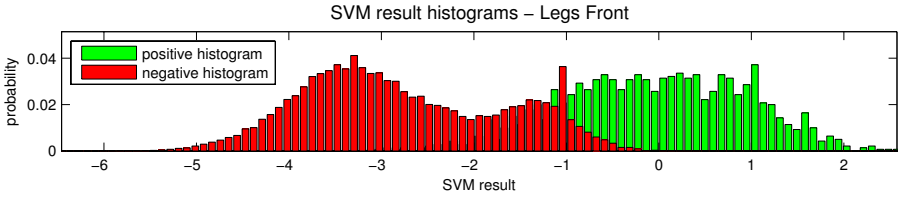
The first factor $P(I|L)$ is the detection probability of the components, at the position and scale given by L. The second factor $P(L)$ represents the prior probability of a positive pedestrian component arrangement. Every Head-Shoulder detection is used as start point to find the corresponding MAP configuration by greedy search. In the following sections we will go further into detail.

3.1 Probabilistic Appearance Model

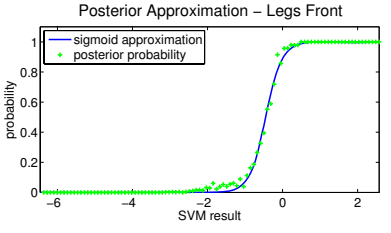
To compute $P(I|L)$ the component results of the detection step are used. For this purpose the SVM results $f(\mathbf{x})$ are converted into probabilistic values. From the many choices available, we preferred an approximation of the a posteriori curve $P(y = 1|f(\mathbf{x}))$, that for a specific SVM result $f(\mathbf{x})$ a pedestrian component $y = 1$ is given, because the best fit was achieved by this model. By using Bayes rule with the priors $P(y = -1)$ and $P(y = 1)$, and class-conditional densities $p(f(\mathbf{x})|y = -1)$ and $p(f(\mathbf{x})|y = 1)$, we get:

$$P(y = 1|f(\mathbf{x})) = \frac{p(f(\mathbf{x})|y = 1)P(y = 1)}{\sum\limits_{i=-1,1} p(f(\mathbf{x})|y = i)P(y = i)} \quad . \tag{2}$$

The resulting a posteriori values for the frontal legs training set are shown in Fig. 3(b), derived with the class-conditional densities, which can be seen in Fig. 3(a). A sigmoid function $s(z = f(\mathbf{x})) = \frac{1}{1+exp(Az+B)}$ is used to approximate



(a)



(b)

Fig. 3. (a) Distribution histograms and (b) the approximated a posterior curve by a sigmoid function for the frontal legs

the posterior. The parameters for $s(z)$ are determined by the Maximum Likelihood method proposed by Platt [13], using the Levenberg-Marquardt method. To compute the sigmoid parameters, training sets for each component and view are used. Fig. 3(b) shows the approximated curve for the frontal legs.

By assuming independence between the detectors for each component v_i, $P(I|L)$ is given by:

$$P(I|L) = \prod_{v_i \in V} P(y = 1|f_i(\mathbf{x_i})) \tag{3}$$

with $\mathbf{x_i}$ as the extracted feature vector and f_i as the result of the ith component.

3.2 Probabilistic Geometric Model

Besides the appearance likelihood value $P(I|L)$, for component configuration L, the probability for the spatial arrangement $P(L)$ has to be computed. Invariance against scale and translation is achieved by using the relative distance vector $\mathbf{d_{ij}}$ and the scale ratio $\Delta s_{ij} = \frac{s_i}{s_j}$ between two components i and j:

$$\mathbf{d_{ij}} = \begin{pmatrix} dx_{ij} \\ dy_{ij} \end{pmatrix} = \frac{1}{s_i} \cdot \begin{pmatrix} x_j - x_i \\ y_j - y_i \end{pmatrix} \quad . \tag{4}$$

As in common literature [4] the model is expressed as a graph $G = (V, E)$, with the components v_i as vertices and the possible relations as edges e_{ij} between component i and j. Our model regard all possible component relations, except those between the same component in different views. Every edge e_{ij} gets a weight $w_{ij} \in [0, 1]$, to account that component pairs of the same view appear

more likely, than component pairs of different views. The weights are generated from the component training sets.

With the priors $P(l_i, l_j) = P(\mathbf{d_{ij}}, \Delta s_{ij})$ the probability of the component arrangement L is given as:

$$P(L) = \prod_{e_{ij} \in E} w_{ij} P(l_i, l_j) = \prod_{e_{ij} \in E} w_{ij} P(\mathbf{d_{ij}}, \Delta s_{ij}) \ . \tag{5}$$

The generated distribution histograms for the geometrical parameters $\mathbf{d_{ij}}$ and Δs_{ij} are used for the priors $P(l_i, l_j)$. To distinguish between a pedestrian-like and non-pedestrian-like component arrangement, two spatial distributions are generated, one for the positive $P_p(L)$ and one for the negative class $P_n(L)$. Distribution histograms are also used for the negative class. The distributions are computed as follows: First the positive spatial distribution histograms are computed from training data. Afterwards, the spatial distributions for the negative class are generated, using only the hard ones, i.e. those lying in the distribution histogram range for the positive class. As final spatial result the difference between the positive and negative spatial result is used.

4 Experiments

The INRIA Person dataset [2] is used for our experiments. This dataset contains a training set with 2416 pedestrian labels and 1218 images without any pedestrians and a test set with 1132 pedestrian images and 453 images not containing any pedestrians. Both sets have only global labels. For the component evaluation, part labels are needed, so in a first step we applied our component labels: head-shoulder, torso and legs, in front/rear and side view. In a second step the average label sizes were determined, see Table 1. Smaller labels were resized to the size given in Table 1. The number of positive training samples and test samples, for every component and view, are listed in Table 1. Some images have no component training labels because of occlusions.

In a first experiment the component detection was evaluated, followed by testing the proposed probabilistic model from Sect. 3. Finally, the probabilistic method is compared to state of the art detectors. Receiver Operating Characteristic (ROC) curves in loglog scale are used for the experimental evaluation of the miss rate $\left(\frac{FalseNeg}{TruePos+FalseNeg}\right)$ against the false-positive rate. Matching criteria is 75% overlap between detection and corresponding label.

4.1 Component Detection

The proposed component detection in Sect. 2 is evaluated. "Unsigned" gradients, 9 orientation bins and a block size of 2x2 histogram cells are used as parameters for the HOG features. In this test the constant histogram selection is compared against a variable selection, as described in Sect. 2. The block sizes for the constant selection are: 16x16 pixels for the frontal torso and 12x12 pixels for

Table 1. Component sizes and the number of positive training/test samples

Part	View	Width	Height	# pos. Training-Samples	# pos. Test-Samples
head	front	32	32	1726	870
	side	32	32	678	262
torso	front	40	45	1668	846
	side	32	45	646	286
legs	front	34	55	1400	756
	side	34	55	668	376

the remaining components/views. For the variable selection, block size range is 8x8 to maximum, not limited to a specific scale. The negative training set was created by using the bootstrapping method given in [2]. The generation of regions of interest (ROI) is done by a sliding window approach. ROI's are generated in different scales. The factor 1.2 is used between two scales. In all scales the step size is 4 pixel in both directions. For the SVM classifier training we use SVMlight [8].

The ROC-curves for the component detection are shown in Fig. 4 and Fig. 5, divided into frontal/rear and side views. It confirms that variable selection (solid lines) yields better results than constant selection (dotted lines), except for the frontal head-component. The results for the frontal/rear head with constant selection are slightly better as those with variable selection. An interesting observation is the obvious difference between the head and leg results, which is stronger in the frontal/rear view than the side view. The leg component produces at 10% miss rate three times fewer false positives than the head. In the frontal view, similar results are recieved by head and torso. The ROC-curves of the side torso and side legs intersect at 10% miss rate. Below 10% miss rate, fewer false positives are produced by the torso and above 10% miss rate the legs generate less false positives.

The computation time per component ROI is in average 0.025 ms, on a 1.8 GHz dual core PC, using only one core. At a resolution of 320x240 pixels, 20000 search windows are generated in average per component and view. The component detection at this resolution with full search takes about 3.1 seconds.

4.2 Probabilistic Component Assembly

The proposed Bayesian approach to component assembly from Sect. 3 is evaluated here. In a first step the probabilistic approach is tested with and without the use of spatial distribution histograms for the negative class, and afterwards compared against state of the art detectors. These detectors are the one from Dalal [2] and the cascade from Viola and Jones [16]. Again the INRIA Person dataset is used as test set.

First the probabilistic approach is evaluated. The results are given in Fig. 6. By using spatial distribution histograms from both classes we achieve better results. The difference between both curves is greater at higher false positive

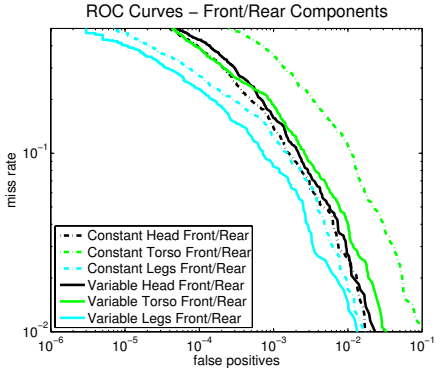


Fig. 4. Front/Rear component results

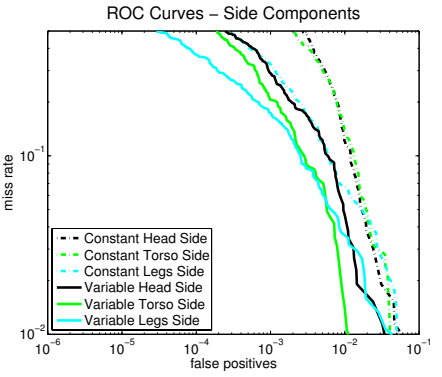


Fig. 5. Side component results

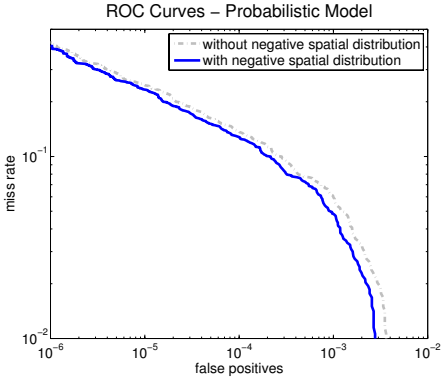


Fig. 6. Probabilistic grouping results

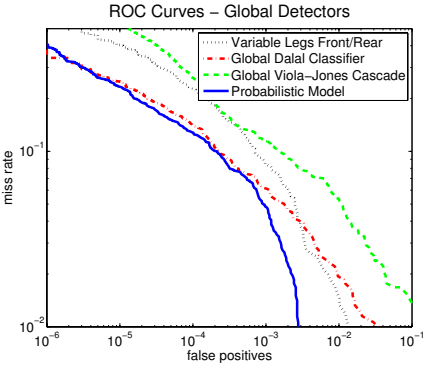


Fig. 7. State of the art detectors in comparison to our approach (blue line)

rates. At low miss rates the extra usage of spatial distributions for the negative class reduce the number of false positives compared to the common approach. In the following experiment the probabilistic approach is compared against state of the art detectors. Fig. 7 shows the best probabilistic detector in comparison to the mentioned standard detectors and the best component result (frontal/rear legs). The results of our part-based approach are slightly better as the best state of the art detector. Below 8% miss rate our probabilistic method outperforms the state of the art detectors. Note that Dalal's detector takes a larger margin around a person, so in comparison to our approach more contextual information is incorporated. Fig. 8 shows some typical results of our approach.

At a resolution of 320x240 pixels, after applying thresholding to the component detection results, we get on average about 400 detections per component and view. For this resolution, our probabilistic grouping approach takes 190 milliseconds in average on a 1.8 GHz PC.

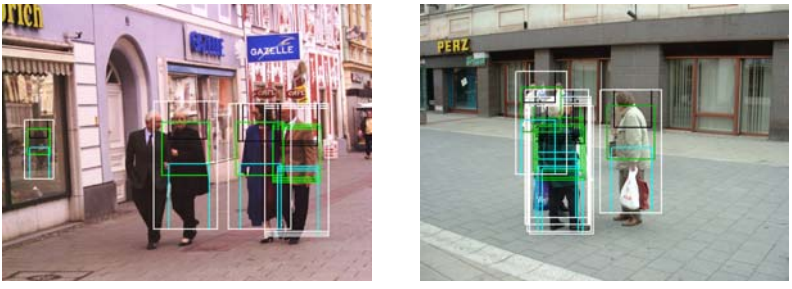


Fig. 8. Some detection results (white - full body, black - head, green - torso, cyan - legs). No post-processing was applied to the images.

5 Conclusion

In this paper a Bayesian component-based approach for pedestrian recognition in single frames was proposed. Our pedestrian detector is composed of the head-shoulder, torso and legs, divided into front/rear and side view for better recognition. For the component detection a variable selection of histograms of oriented gradients and SVM classification is applied. In the next step, the components are grouped by a Bayes-based approach. To shrink the number of candidates for the probabilistic grouping, thresholding is applied to all component results, so that 99% true positive component detection rate remains. Invariance against scale and translation is achieved by using the relative distance vector and scale ratio between the components. To make a better separation into positive and negative spatial component arrangements, distributions for both classes are generated. Instead of approximating these distributions, for example by a Gaussian, the computed distribution histograms are used directly. The results confirm the positive benefit of using distributions for both classes and not only for one. Below miss rates of 8% our approach outperforms state of the art detectors. Above, performance is similar.

6 Future Work

One main drawback of our approach is computation time, mainly of the component detection. Using a cascaded classifier would make the component detection faster. To improve the performance of our approach the narrow field of a pedestrian can be included as contextual information. First experiments show promising results. The performance of the front/rear views is much better than for the side views. To overcome this, left and right side views could be separated.

References

1. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A Study of Parts-Based Object Class Detection Using Complete Graphs. In: IJCV (in press, 2009)
2. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR, vol. 1, pp. 886–893 (2005)

3. Dalal, N.: Finding People in Images and Videos, PhD thesis, Institut National Polytechnique de Grenoble (July 2006)
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial Structures for Object Recognition. IJCV 61(1), 55–79 (2005)
5. Felzenszwalb, P., Mcallester, D., Ramanan, D.: A Discriminatively Trained, Multiscale, Deformable Part Model. In: CVPR, Anchorage, Alaska, June 2008, pp. 1–8 (2008)
6. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
7. Gavrila, D.M., Munder, S.: Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. IJCV 73, 41–59 (2007)
8. Joachims, T.: Making large-Scale SVM Learning Practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge (1999)
9. Laptev, I.: Improvements of Object Detection Using Boosted Histograms. In: British Machine Vision Conference, September 2006, vol. 3, pp. 949–958 (2006)
10. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 69–82. Springer, Heidelberg (2004)
11. Mohan, A., Papageorgiou, C., Poggio, T.: Example-Based Object Detection in Images by Components. PAMI 23(4), 349–361 (2001)
12. Papageorgiou, C., Evgeniou, T., Poggio, T.: A Trainable Pedestrian Detection System. In: IVS, pp. 241–246 (1998)
13. Platt, J.: Probabilities for SV Machines. In: Press, M. (ed.) Advances in Large Margin Classifiers, pp. 61–74 (1999)
14. Schapire, R.E.: The Strength of Weak Learnability. Machine Learning 5(2), 197–227 (1990)
15. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc, New York (1995)
16. Viola, P., Jones, M.: Robust Real-time Object Detection. IJCV 57(2), 137–154 (2004)
17. Wu, B., Nevatia, R.: Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors. In: ICCV, vol. 1, pp. 90–97 (2005)
18. Wu, B., Nevatia, R.: Detection and Segmentation of Multiple, Partially Occluded Objects by Grouping, Merging, Assigning Part Detection Responses. IJCV 82(2), 185–204 (2009)
19. Zhu, Q., Yeh, M.-C., Cheng, K.-T., Avidan, S.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: CVPR, pp. 1491–1498 (2006)