High-Level Fusion of Depth and Intensity for Pedestrian Classification

Marcus Rohrbach[1,3], Markus Enzweiler[2] and Dariu M. Gavrila[1,4]

[1]Environment Perception, Group Research, Daimler AG, Ulm, Germany
[2]Image & Pattern Analysis Group, Dept. of Math.
and Computer Science, Univ. of Heidelberg, Germany
[3]Dept. of Computer Science, TU Darmstadt, Germany
[4]Intelligent Systems Lab, Fac. of Science, Univ. of Amsterdam, The Netherlands

marcus.rohrbach@online.de {uni-heidelberg.enzweiler,dariu.gavrila}@daimler.com

Abstract. This paper presents a novel approach to pedestrian classification which involves a high-level fusion of depth and intensity cues. Instead of utilizing depth information only in a pre-processing step, we propose to extract discriminative spatial features (gradient orientation histograms and local receptive fields) directly from (dense) depth and intensity images. Both modalities are represented in terms of individual feature spaces, in each of which a discriminative model is learned to distinguish between pedestrians and non-pedestrians. We refrain from the construction of a joint feature space, but instead employ a high-level fusion of depth and intensity at classifier-level.
Our experiments on a large real-world dataset demonstrate a significant performance improvement of the combined intensity-depth representation over depth-only and intensity-only models (factor four reduction in false positives at comparable detection rates). Moreover, high-level fusion outperforms low-level fusion using a joint feature space approach.

1 Introduction

Pedestrian recognition is an important problem in domains such as intelligent vehicles or surveillance. It is particularly difficult, as pedestrians tend to occupy only a small part of the image (low resolution), have different poses (shape) and clothing (appearance), varying background, or might be partially occluded. Most state-of-the-art systems derive feature sets from intensity images, i.e. gray-scale (or colour) images, and apply learning-based approaches to detect people [1, 3, 9, 22, 23].

Besides image intensity, depth information can provide additional cues for pedestrian recognition. Up to now, the use of depth information has been limited to recovering high-level scene geometry [5, 11] and focus-of-attention mechanisms

Marcus Rohrbach and Markus Enzweiler acknowledge the support of the Studienstiftung des deutschen Volkes (German National Academic Foundation).

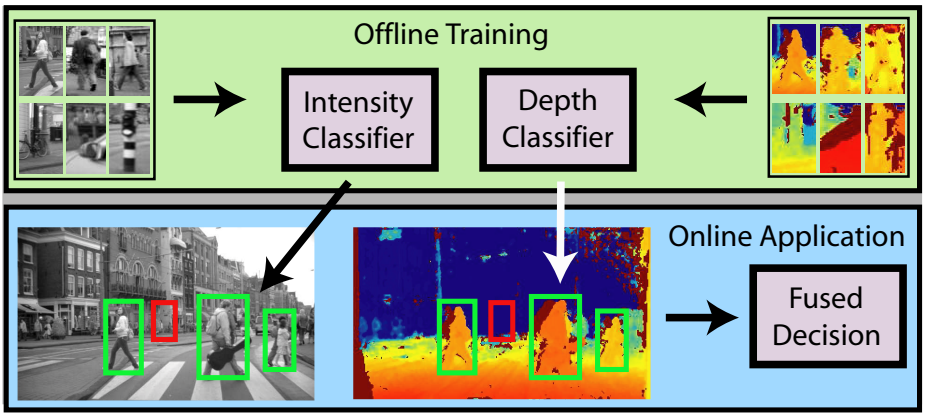


Fig. 1. Framework overview. Individual classifiers are trained offline on intensity and corresponding depth images. Online, both classifiers are fused to a combined decision. For depth images, warmer colors represent closer distances from the camera.

[8]. Given the availability of real-time high-resolution dense stereo algorithms [6, 20], we propose to enrich an intensity-based feature space for pedestrian classification with features operating on dense depth images (Sect. 3). Depth information is computed from a calibrated stereo camera rig using semi-global matching [6]. Individual classifiers are trained offline on features derived from intensity and depth images depicting pedestrian and non-pedestrian samples. Online, the outputs of both classifiers are fused to a combined decision (Sect. 4). See Fig. 1.

2 Related Work

A large amount of literature covers image-based classification of pedestrians. See [3] for a recent survey and a challenging benchmark dataset. Classification typically involves a combination of feature extraction and a discriminative model (classifier), which learns to separate object classes by estimating discriminative functions within an underlying feature space.

Most proposed feature sets are based on image intensity. Such features can be categorized into texture-based and gradient-based. Non-adaptive Haar wavelet features have been popularized by [15] and adapted by many others [14, 22], with manual [14, 15] and automatic feature selection [22]. Adaptive feature sets were proposed, e.g. local receptive fields [23], where the spatial structure is able to adapt to the data. Another class of texture-based features involves codebook patches which are extracted around salient points in the image [11, 18].

Gradient-based features have focused on discontinuities in image brightness. Local gradient orientation histograms were applied in both sparse (SIFT) [12] and dense representations (HOG) [1, 7, 25, 26]. Covariance descriptors involving a model of spatial variation and correlation of local gradients were also used [19].

(a) Pedestrian (b) Non-Pedestrian

Fig. 2. Intensity and depth images for pedestrian (a) and non-pedestrian samples (b). From left to right: intensity image, gradient magnitude of intensity, depth image, gradient magnitude of depth

Yet others proposed local shape filters exploiting characteristic patterns in the spatial configuration of salient edges [13, 24].

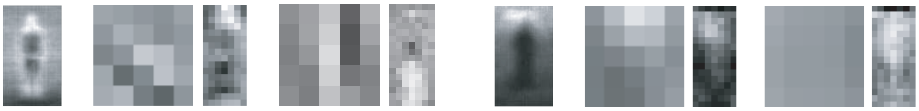
In terms of discriminative models, support vector machines (SVM) [21] are widely used in both linear [1, 25, 26] and non-linear variants [14, 15]. Other popular classifiers include neural networks [9, 10, 23] and AdaBoost cascades [13, 19, 22, 24–26]. Some approaches additionally applied a component-based representation of pedestrians as an ensemble of body parts [13, 14, 24].

Others combined features from different modalities, e.g. intensity, motion, depth, etc. Multi-cue combination can be performed at different levels: On *module-level*, depth [5, 9, 11] or motion [4] can be used in a pre-processing step to provide knowledge of the scene geometry and focus-of-attention for a subsequent (intensity-based) classification module. Other approaches have fused information from different modalities on *feature-level* by establishing a joint feature space (low-level fusion): [1, 22] combined gray-level intensity with motion. In [17], intensity and depth features derived from a 3D camera with very low resolution (pedestrian heights between 4 and 8 pixels) were utilized. Finally, fusion can occur on *classifier-level* [1, 2]. Here, individual classifiers are trained within each feature space and their outputs are combined (high-level fusion).

We consider the main contribution of our paper to be the use of spatial depth features based on dense stereo images for pedestrian classification at medium resolution (pedestrian heights up to 80 pixels). A secondary contribution concerns fusion techniques of depth and intensity. We follow a high-level fusion strategy which allows to tune features specifically to each modality and base the final decision on a combined vote of the individual classifiers. As opposed to low-level fusion approaches [17, 22], this strategy does not suffer from the increased dimensionality of a joint feature space.

3 Spatial Depth and Intensity Features

Dense stereo provides information for most image areas, apart from regions which are visible only by one camera (stereo shadow). See the dark red areas to the left of the pedestrian torso in Fig. 2(a). Spatial features can be based on either depth Z (in meters) or disparity d (in pixels). Both are inverse proportional given the



(a) Intensity features (b) Depth features

Fig. 3. Visualization of gradient magnitude (related to HOG) and LRF features on (a) intensity and (b) depth images. From left to right: Average gradient magnitude of pedestrian training samples, two exemplary 5×5-pixel local receptive field features and their activation maps, highlighting spatial regions of the training samples where the corresponding LRFs are most discriminative with regard to the pedestrian and non-pedestrian classes.

camera geometry with focal length f and the distance between the two cameras B:

$$Z(x, y) = \frac{fB}{d(x, y)} \text{ at pixel } (x, y) \tag{1}$$

Objects in the scene have similar foreground/background gradients in depth space, irrespective of their location relative to the camera. In disparity space however, such gradients are larger, the closer the object is to the camera. To remove this variability, we derive spatial features from depth instead of disparity. We refer to an image with depth values $Z(x, y)$ at each pixel (x, y) as *depth image*.

A visual inspection of the depth image vs. the intensity image in Fig. 2 reveals that pedestrians have a distinct depth contour and texture which is different from the intensity domain. In intensity images, lower body features (shape and appearance of legs) are the most significant feature of a pedestrian (see results of part-based approaches, e.g. [14]). In contrast, the upper body area has dominant foreground/background gradients and is particularly characteristic for a pedestrian in the depth image. Additionally, the stereo shadow is clearly visible in this area (to the left of the pedestrian torso) and represents a significant local depth discontinuity. This might not be a disadvantage but rather a distinctive feature. The various salient regions in depth and intensity images motivate our use of fusion approaches between both modalities to benefit from the individual strengths, see Sect. 4.

To instantiate feature spaces involving depth and intensity, we utilize well-known state-of-the-art features, which focus on local discontinuities: Non-adaptive histogram of oriented gradients with a linear SVM (HOG/linSVM) [1] and a neural network using adaptive local receptive fields (NN/LRF) [23]. For classifier training, the feature vectors are normalized to $[-1; +1]$ per dimension.

To get an insight into HOG and LRF features, Fig. 3 depicts the average gradient magnitude of all pedestrian training samples (related to HOG), as well as exemplary local receptive field features and their activation maps (LRF), for both intensity and depth. We observe that gradient magnitude is particularly high around the upper body contour for the depth image, while being more evenly distributed for the intensity image. Further, almost no depth gradients

are present on areas corresponding to the pedestrian body. During training, the local receptive field features have developed to detect very fine grained structures in the image intensity domain. The two features depicted in Fig. 3(a) can be regarded as specialized "head-shoulder" and "leg" detectors and are especially activated in the corresponding areas. For depth images, LRF features respond to larger structures in the image, see Fig. 3(b). Here, characteristic features focus on the coarse depth contrast between the upper-body head/torso area. The mostly uniform depth texture on the pedestrian body is a prominent feature as well.

4 Fusion on Classifier-Level

A popular strategy to improve classification is to split-up a classification problem into more manageable sub-parts on data-level, e.g. using mixture-of-experts or component-based approaches [3]. A similar strategy can be pursued on classifier-level. Here, multiple classifiers are learned on the full dataset and their outputs combined to a single decision. Particularly, when the classifiers involve uncorrelated features, benefits can be expected. We follow a *Parallel Combination* strategy [2], where multiple feature sets (i.e. based on depth and intensity, see Sect. 3) are extracted from the same underlying data. Each feature set is then used as input to a single classifier and their outputs combined (high-level fusion).

For classifier fusion, we utilize a set of fusion rules which are explained below. An important prerequisite is that the individual classifier outputs are normalized, so that they can be combined homogeneously. The outputs of many state-of-the-art classifiers can be converted to an estimate of posterior probabilities [10, 16]. We use this sigmoidal mapping in our experiments.

Let $\mathbf{x}_k, k = 1, \ldots, n$, denote a (vectorized) sample. The posterior for the k-th sample with respect to the j-th object class (e.g. pedestrian, non-pedestrian), estimated by the i-th classifier, $i = 1, \ldots, m$, is given by: $p_{ij}(\mathbf{x}_k)$. Posterior probabilities are normalized across object classes for each sample, so that:

$$\sum_j \left(p_{ij}\left(\mathbf{x}_k\right)\right) = 1 \tag{2}$$

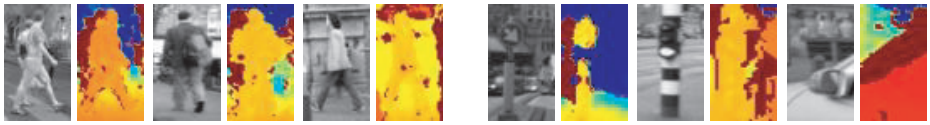
Classifier-level fusion involves the derivation of a new set of class-specific confidence values for each data point, $q_j(\mathbf{x}_k)$, out of the posteriors of the individual classifiers, $p_{ij}(\mathbf{x}_k)$. The final classification decision $\omega(\mathbf{x}_k)$ results from selecting the object class with the highest confidence:

$$\omega(\mathbf{x}_k) = \arg\max_j \left(q_j\left(\mathbf{x}_k\right)\right) \tag{3}$$

We consider the following fusion rules to determine the confidence $q_j\left(\mathbf{x}_k\right)$ of the k-th sample with respect to the j-th object class:

Maximum Rule The maximum rule bases the final confidence value on the classifier with the highest estimated posterior probability:

$$q_j\left(\mathbf{x}_k\right) = \max_i \left(p_{ij}\left(\mathbf{x}_k\right)\right) \tag{4}$$



(a) Pedestrian samples (b) Non-Pedestrian samples

Fig. 4. Overview of (a) pedestrian and (b) non-pedestrian samples (intensity and corresponding depth images).

Product Rule Individual posterior probabilities are multiplied to derive the combined confidence:

$$q_j\left(\mathbf{x}_k\right) = \prod_i \left(p_{ij}\left(\mathbf{x}_k\right)\right) \tag{5}$$

Sum Rule The combined confidence is computed as the average of individual posteriors, with m denoting the number of individual classifiers:

$$q_j\left(\mathbf{x}_k\right) = \frac{1}{m}\sum_i \left(p_{ij}\left(\mathbf{x}_k\right)\right) \tag{6}$$

SVM Rule A support vector machine is trained as a fusion classifier to discriminate between object classes in the space of posterior probabilities of the individual classifiers:

Let $\mathbf{p}_{jk} = \left(p_{1j}\left(\mathbf{x}_k\right), \ldots, p_{mj}\left(\mathbf{x}_k\right)\right)$ denote the m-dimensional vector of individual posteriors for sample $\mathbf{x}_k$ with respect to the j-th object class. The corresponding hyperplane is defined by:

$$f_j\left(\mathbf{p}_{jk}\right) = \sum_l y_l \alpha_l \cdot K\left(\mathbf{p}_{jk}, \mathbf{p}_{jl}\right) + b \tag{7}$$

Here, $\mathbf{p}_{jl}$ denotes the set of support vectors with labels y_l and Lagrange multipliers α_l. $K(\cdot,\cdot)$ represents the SVM Kernel function. We use a non-linear RBF kernel in our experiments. The SVM decision value $f_j\left(\mathbf{p}_{jk}\right)$ (distance to the hyperplane) is used as confidence value:

$$q_j\left(\mathbf{x}_k\right) = f_j\left(\mathbf{p}_{jk}\right) \tag{8}$$

5 Experiments

5.1 Experimental Setup

The presented feature/classifier combinations and fusion strategies, see Sects. 3 and 4, were evaluated in experiments on pedestrian classification. Training and test samples comprise non-occluded pedestrian and non-pedestrian cut-outs from intensity and corresponding depth images, captured from a moving vehicle in an urban environment. See Table 1 and Fig. 4 for an overview of the dataset.

	Pedestrians (labelled)	Pedestrians (jittered)	Non-Pedestrians
Training Set (2 parts)	10998	43992	43046
Test Set (1 part)	5499	21996	21523
Total	16497	65988	64569

Table 1. Dataset statistics. The same numbers apply to samples from depth and intensity images.

All samples are scaled to 48 × 96 pixels (HOG/linSVM) and 18 × 36 pixels (NN/LRF) with an eight-pixel (HOG/linSVM) and two-pixel border (NN/LRF) to retain contour information. For each manually labelled pedestrian bounding box we randomly created four samples by mirroring and geometric jittering. Non-pedestrian samples resulted from a pedestrian shape detection step with relaxed threshold setting, i.e. containing a bias towards more difficult patterns.

HOG features were extracted from those samples using 8×8 pixel cells, accumulated to 16×16 pixel blocks, with 8 gradient orientation bins, see [1]. LRF features (in 24 branches, see [23]) were extracted at a 5×5 pixel scale. Identical feature/classifier parameters are used for intensity and depth. The dimension of the resulting feature spaces is 1760 for HOG/linSVM and 3312 for NN/LRF.

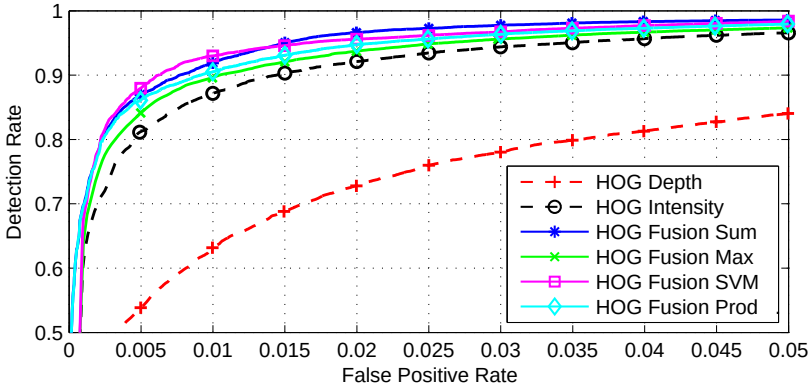
We apply a three-fold cross-validation to our dataset: The dataset is split-up into three parts of the same size, see Table 1. In each cross-validation run, two parts are used for training and the remaining part for testing. Results are visualized in terms of mean ROC curves across the three cross-validation runs.

5.2 Experimental Results

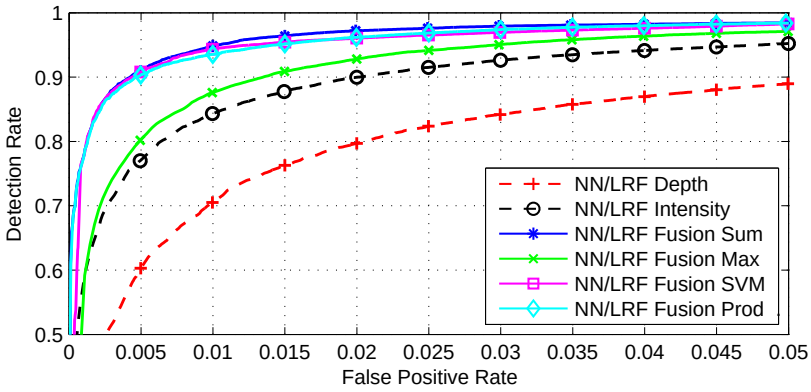
In our first experiment, we evaluate the performance of classifiers for depth and intensity separately, as well as using different fusion strategies. Results are given in Fig. 5(a-b) for the HOG/linSVM and NN/LRF classifier, respectively.

The performance of features derived from intensity images (black ∘) is better than for depth features (red +), irrespective of the actual feature/classifier approach. Furthermore, all fusion strategies between depth and intensity clearly improve performance (Fig. 5(a-b), solid lines). For both HOG/linSVM and NN/LRF, the sum rule performs better than product rule, which in turn outperforms the maximum rule. However, performance differences among fusion rules are rather small. Only for NN/LRF, the maximum rule performs significantly worse. By design, maximum selection is more susceptive to noise and outliers. Using a non-linear RBF SVM as a fusion classifier does not improve performance over fusion by the sum rule, but is far more computationally expensive. Hence, we only employ the sum rule for fusion in our further experiments.

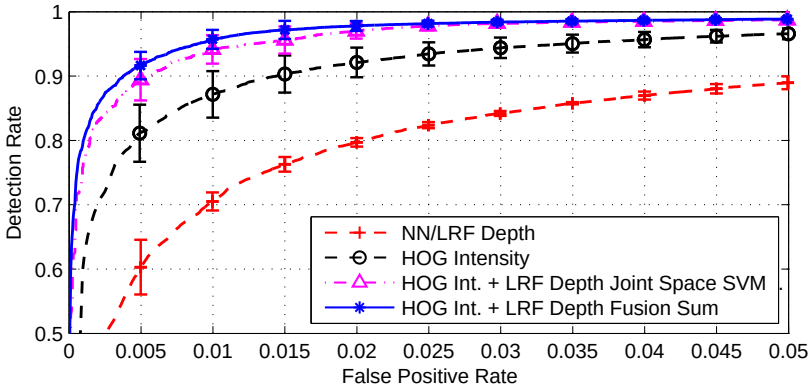
Comparing absolute performances, our experiments show that fusion of depth and intensity can reduce false positives over intensity-only features at a constant detection rate by approx. a factor of two for HOG/linSVM and a factor of four for NN/LRF: At a detection rate of 90%, the false positive rates for HOG/linSVM (NN/LRF) amount to 1.44% (2.01%) for intensity, 8.92% (5.60%) for depth and



(a) HOG/linSVM classifier



(b) NN/LRF classifier



(c) Best performing classifiers and joint feature space with 1-σ error bars.

Fig. 5. Pedestrian classification performance using spatial depth and intensity features. (a) HOG/linSVM, (b) NN/LRF, (c) best performing classifiers.

0.77% (0.43%) for sum-based fusion of depth and intensity. This clearly shows that the different strengths of depth and intensity can indeed be exploited, see Sect. 3. An analysis of correlation between the classifier outputs for depth and intensity confirms this: For HOG/linSVM (NN/LRF), the correlation coefficient between depth and intensity is 0.1068 (0.1072). For comparison, the correlation coefficient beween HOG/linSVM and NN/LRF on intensity images is 0.3096.

In our third experiment, we fuse the best performing feature/classifier for each modality, i.e. HOG/linSVM for intensity images (black $\circ$) and NN/LRF for depth images (red +). See Fig. 5(c). The results of fusion using the sum-rule (blue *) outperforms all previously considered variants. More specifically, we achieve a false positive rate of 0.35% (at 90% detection rate) which is a reduction by a factor of four, compared to the state-of-the-art HOG/linSVM classifier on intensity images (black $\circ$; 1.44% false positive rate). We additionally visualize 1-σ error bars computed from the different cross-validation runs. The non-overlapping error bars of the various system variants underline the statistical significance of our results.

We further compare the proposed *high-level* fusion (Fig. 5(c), blue *) with *low-level* fusion (Fig. 5(c), magenta Δ). For this, we construct a joint feature space combining HOG features for intensity and LRF features for depth (normalized to $[-1; +1]$ per dimension). A linear SVM is trained in the joint space to discriminate between pedestrians and non-pedestrians. A non-linear SVM was computationally not feasible, given the increased dimension of the joint feature space (5072) and our large datasets. Results show, that low-level fusion using a joint feature space is outperformed by the proposed high-level classifier fusion, presumable because of the higher dimensionality of the joint space.

6 Conclusion

This paper presented a novel framework for pedestrian classification which involves a high-level fusion of spatial features derived from dense stereo and intensity images. Our combined depth/intensity approach outperforms the state-of-the-art intensity-only HOG/linSVM classifier by a factor of four in reduction of false positives. The proposed classifier-level fusion of depth and intensity also outperforms a low-level fusion approach using a joint feature space.

References

1. N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. ECCV*, pages 428–441, 2006.
2. R. P. W Duin and D. M. J. Tax. Experiments with classifier combining rules. In *Proc. of the First Int. Workshop on Multiple Classifier Systems*, pages 16–29, 2000.
3. M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE PAMI, available online: IEEE Computer Society Digital Library, http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.260*, 17. Oct. 2008.
4. M. Enzweiler, P. Kanter, and D. M. Gavrila. Monocular pedestrian recognition using motion parallax. In *IEEE IV Symp.*, pages 792–797, 2008.

5. A. Ess, B. Leibe, and L. van Gool. Depth and appearance for mobile scene analysis. In *Proc. ICCV*, 2007.
6. U. Franke et al. Towards optimal stereo analysis of image sequences. In *Robot Vision*, pages 43–58, 2008.
7. T. Gandhi and M. M. Trivedi. Image based estimation of pedestrian orientation for improving path prediction. In *IEEE IV Symp.*, pages 506–511, 2008.
8. D. M. Gavrila. A Bayesian, exemplar-based approach to hierarchical shape matching. *IEEE PAMI*, 29(8):1408–1421, 2007.
9. D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, 2007.
10. A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE PAMI*, 22(1):4–37, 2000.
11. B. Leibe et al.. Dynamic 3d scene analysis from a moving vehicle. In *Proc. CVPR*, 2007.
12. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
13. K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. ECCV*, pages 69–81, 2004.
14. A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE PAMI*, 23(4):349–361, 2001.
15. C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38:15–33, 2000.
16. J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances In Large Margin Classifiers*, pages 61–74, 1999.
17. M. Rapus et al. Pedestrian recognition using combined low-resolution depth and intensity images. In *IEEE IV Symp.*, pages 632–636, 2008.
18. E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. In *Proc. CVPR*, 2007.
19. O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on Riemannian manifolds. In *Proc. CVPR*, 2007.
20. W. Van der Mark and D. M. Gavrila. Real-time dense stereo for intelligent vehicles. *IEEE PAMI*, 7(1):38–50, 2006.
21. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
22. P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005.
23. C. Wöhler and J. K. Anlauf. A time delay neural network algorithm for estimating image-pattern shape and motion. *IVC*, 17:281–294, 1999.
24. B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247 – 266, 2007.
25. L. Zhang, B. Wu, and R. Nevatia. Detection and tracking of multiple humans with extensive pose articulation. In *Proc. ICCV*, 2007.
26. Q. Zhu et al.. Fast human detection using a cascade of histograms of oriented gradients. In *Proc. CVPR*, pages 1491–1498, 2006.