2014 22nd International Conference on Pattern Recognition

Center-surround Contrast Features for Pedestrian Detection

Shanshan Zhang*, Dominik A. Klein*, Christian Bauckhage†‡, and Armin B. Cremers*‡
*Institute of Computer Science III, University of Bonn
Römerstraße 164, 53117 Bonn, Germany
†Fraunhofer IAIS
Schloss Birlinghoven, 53757 Sankt Augustin, Germany
‡Bonn-Aachen International Center for Information Technology (B-IT)
Dahlmannstraße 2, 53113 Bonn, Germany
{zhangs,kleind,abc}@iai.uni-bonn.de, christian.bauckhage@iais.fraunhofer.de

***Abstract*—Inspired by the human vision system, in this paper we propose a specifically organized kind of center-surround contrast features and show their suitability for pedestrian detection. These contrasts are computed from a novel combination of both local color and gradient statistics aggregated quickly for arbitrary sized square cells. We exploit our contrast features in a rich multi-scale and -direction fashion between each central cell and its neighbors and boost the significant ones for pedestrian detection. Experimental results on the INRIA and Caltech pedestrian datasets show that our method achieves state-of-the-art performance.**

I. Introduction

Pedestrian detection enjoys increasing popularity in the computer vision community, since in this area academic interests meet industrial needs. Numerous literature emerged about the topic (see [1] for a survey), where contributions are often specialized to some part of the chief aim enabling computers to discover pedestrians around, e.g. feature extraction [2], classifier design [3], or regions of interest selection [4]. Although significant improvements have been achieved in the last decade, the detection precision still lies far behind human vision, which is capable of rapidly localizing small scale pedestrians even with low contrast and severe occlusion. Therefore, we are encouraged to look into how the human visual system processes incoming stimuli and draw conclusions for the design of our basic features. We believe that the employment of biologically inspired mechanisms aids recognition and yields an effective and efficient pedestrian detector.

Early processing of visual information starts in the human retinal tissue immediately after light has been transduced into electric signals by photoreceptive cells. At a first layer of bipolar cells electrical membrane potentials are locally aggregated. Grouped bipolar cells report to different types of ganglion cells, which convert analog potentials into electric pulse rates. At the transitional synapses between photoreceptive and bipolar, but also from bipolar to ganglion cells, there is a lateral wiring of so called horizontal respectively amacrine cells modulating the signals to enhance contrasts in a center-surround fashion. It was found that the output of certain ganglion cells agree with DoG-filter responses [5] while some are also oriented and agree with Gabor-filters [6]. A survey about retinal cell types and wiring is given by [7].

Center-surround mechanisms also affect the later processing in the brain, guiding human *attention* and thus how we recognize objects of interest. This psychophysical theory has been widely used in computational approaches to generate saliency maps of the environment [8]. However, while attention is about bottom-up, model-free discovery of the environment, visual search for specific entities requires top-down saliency, which tunes the scoring of basic features to the expected appearance.

Our main contribution in this paper is to design center-surround contrast features motivated by the human visual system and boosting them to characterize the appearance of pedestrians. In order to reduce computational costs when calculating our features for arbitrary sized local cells, we combined the fast acquisition of *continuous* Gaussian feature distributions from [9] with the integral histograms technique from [10]. Each cell is represented by one statistical descriptor, and contrasts are computed between a central cell and its surrounding neighbors.

Statistical multi-channel cell descriptors: we collect multi-channel information for each cell area not only w.r.t. lightness and colors, but also w.r.t. gradients, which both can complement each other under the challenging variations of clothing or articulation of the body respectively. In order to summarize the underlying, unknown distribution of each cell's channel values, we fit a normal distribution, which is the type of continuous distribution with maximum entropy given a known mean and variance [11].

Multi-direction and -scale contrast vectors: aiming to incorporate more specific information between central and surrounding cells, we treat neighbors in different directions individually, rather than together as a single surrounding region, thus we obtain multi-direction contrast descriptors; we compute statistical features at different cell-sizes so as to build a contrast pyramid, which is in accordance with the general architecture of most visual saliency systems.

We evaluate our approach in extensive experiments on several benchmark datasets and demonstrate that by employing our novel center-surround contrast features, our pedestrian detector achieves state-of-the-art performance.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.398

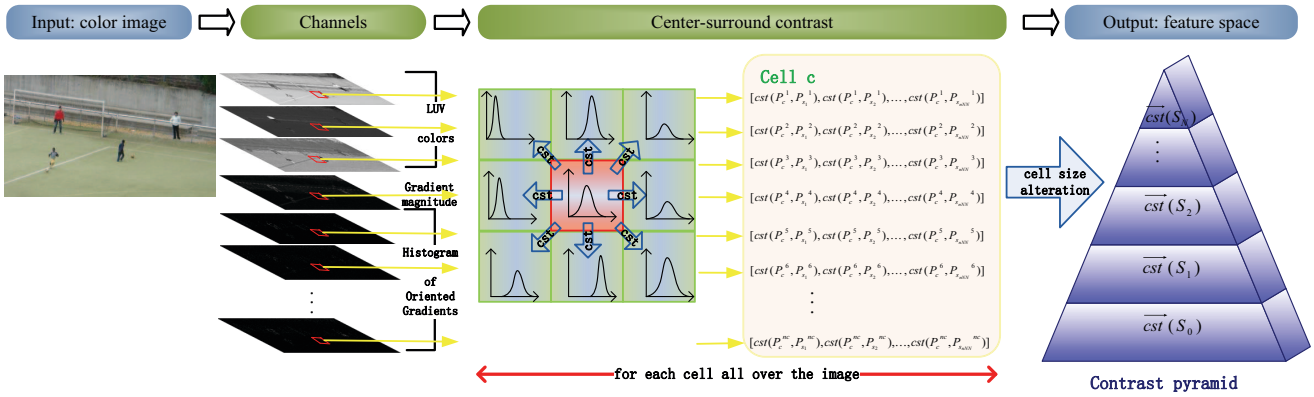


Fig. 1: Flow chart of our feature extraction procedure.

II. Related Work

Since we employ center-surround human visual attention mechanism to design features for pedestrian detection, we concentrate our review in the above two fields respectively.

A. Features for pedestrian detection

In the field of pedestrian detection, almost all kinds of features are extracted by fusing local or global differences in various forms, except those who used intensity or color values directly. Those difference based features can be categorized into two classes: pixel-wise and patch-wise, according to how the differences are computed.

Gradients are a kind of basic pixel-wise local difference, on which the arguably most popular features *Histograms of Oriented Gradients* (HOGs) [12] for pedestrian detection are built; *Local Binary Pattern* (LBP) features [13] incorporated the relationships between neighboring pixels by binary codes, and were combined with HOG features by Wang et al. [14] in order to cope with occlusions.

On contrast, Haar-like features [15] are considered as patch-wise local differences, as they compute abstraction of sums over rectangular regions. Viola and Jones [16] used Haar-like features on both intensity and motion information for pedestrian detection; Walk et al. [2] proposed *Color Self Similarity* (CSS) features to describe global difference between each cell pair on color histograms; Zhang et al. [17] designed informed Haar-like features based on the prior shape of upright human body.

Montabone et al. [18] proposed new features derived from a visual saliency mechanism, which computed differences between central and surrounding regions. Those features are most relevant to ours, but we extend their work in three aspects: we compute the center-surround contrast in multiple channels (not only colors but also gradients); we fit a normal distribution for each cell's channel values; we compute the directional contrasts between the central and eight nearest surrounding regions individually, so as to incorporate more detailed information regarding local difference.

B. Center-surround contrast measurements

Most computational visual attention approaches determine the center-surround contrast by DoG-filters or approximations of these [19]. Recently, some researchers represented the central and surrounding areas by feature distributions to capture more information about the areas. The distributions were in either discrete forms, e.g. histograms [20] or continuous forms, e.g. normal distributions [9]. Various distance measurements can be computed between central and surrounding distributions to describe the local contrast.

III. Center-surround contrast features

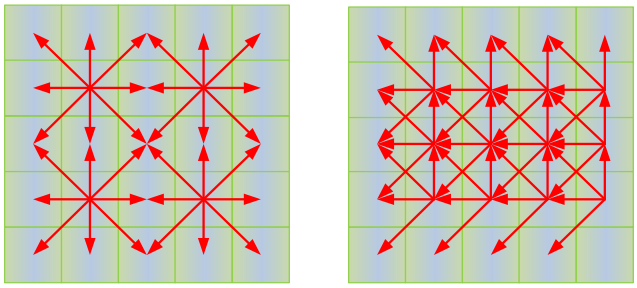
The flow chart of our feature extraction is illustrated in Fig. 1. First, we compute multiple channels for each pixel all over the image, to integrate both color and gradient information; second, we divide the whole image into square cells with a fixed size and fit a normal distribution for each cell's channel values; third, we define the neighborhood for each cell, and compute the differences between one central cell and its neighboring cells, respectively; finally, we repeat the second and third step at several scales with different cell sizes, yielding a multi-scale, multi-direction and multi-channel contrast vector for the whole image.

A. Statistical multi-channel cell descriptors

We consider a total of 10 different channels: 3 channels for LUV colors, 1 channel for gradient magnitude information, and 6 channels for histograms of oriented gradients. Note that all the above channels are computed pixel by pixel. Histograms of oriented gradients are usually computed for a group of pixels inside some region, but we do it for each pixel, which means we simply quantize the gradient magnitudes into orientation bins. For each pixel, two neighboring bins are filled as we employ bilinear interpolation w.r.t. orientation bins.

In order to remove noises, before channel computation, input images are smoothed with a binomial filter of radius 1, i.e. $\sigma \approx 0.87$. Note that we explicitly do not smooth channel data as we observed this to lead to a decrease in performance, as it seems to inhibit characteristic local variations.

We fit a normal distribution to summarize the distribution of each cell's channel values. Assume that we have channel



(a) Sparse map (b) Dense map

Fig. 2: Two structures of neighborhood maps. Each red arrow points from the central cell to its neighboring cells.

values for the whole input image along channel i, denoting as a channel image P^i, and one cell c with its channel vector $P_c^i = [v_1^i, v_2^i, ..., v_p^i]$ along channel i, we implement ML-estimation of normal distribution for P_c^i:

$$\hat{\mu}_c^i = \frac{1}{p}\sum_{k=1}^{p} v_k^i = \overline{P_c^i}, \tag{1}$$

and

$$\hat{\Sigma}_c^i = \frac{1}{p}\sum_{k=1}^{p}(v_k^i - \overline{P_c^i})^2 = \overline{(P_c^i)^2} - \overline{P_c^i}^2. \tag{2}$$

Now the estimation is narrowed down to compute two local averages: $\overline{P_c^i}$ and $\overline{(P_c^i)^2}$ according to Eq. 1 and Eq. 2. For efficiency, along each channel, we employ two integral images: one for the original channel image P^i; another one for the squared channel image $(P^i)^2$. The two integral images avoid exhaustive summing up of channel values inside each cell.

B. Multi-direction and multi-scale contrast vectors

To more accurately describe how the central cell differs from its neighbors, we treat its neighboring cells in different directions individually, rather than together as a single region. Therefore, we compute the contrast between the central cell and each neighboring cell respectively, to yield a multi-direction contrast vector, which integrates the differences in multiple directions.

The first issue comes to defining neighbors for each cell to form center-surround cell pairs. One simple way is to find the 8 nearest neighbors for each cell, but then redundancy emerge significantly, as each neighboring relationship is counted for twice. To get rid of this redundancy, we propose two neighborhood structures: sparse and dense neighborhood maps, as shown in Fig. 2. The sparse map is generated using the simple 8 nearest neighbors principle but with a step size of two cells instead of one; the dense map is generated through finding only 4 out of 8 nearest neighbors for each cell, as shown in Fig. 2b. The advantage of the dense map is that it is capable of bridging each cell with its 8 nearest neighbors, except those cells along the image borders, without any redundancy.

We introduce two different contrast measurements to compute the difference between each two cells' channel values, denoted as P_c^i and P_s^i for the central and its surrounding cell along channel i, respectively. We compare the results of the two measurements in Sec. V.

$\mathcal{W}_2$ distance

The $\mathcal{W}_2$ distance (2nd Wasserstein distance) was first introduced as a measurement for center-surround contrast by Klein et al. [9] and achieved reasonable results for saliency detection. Its definition in our case can be written as:

$$\mathcal{W}_2(P_c^i, P_s^i) = \left[\inf_{\gamma\in\Gamma(P_c^i,P_s^i)}\int_{\mathbb{R}\times\mathbb{R}} |x-y|^2 \,\mathrm{d}\,\gamma(x,y)\right]^{\frac{1}{2}}, \tag{3}$$

with $\Gamma(P_c^i, P_s^i)$ denoting the set of all couplings of P_c^i and P_s^i meets this requirements, if the underlying feature space is defined reasonably.

It would be intractable to compute the integral in Eq. 3 in case of arbitrary distributions. Fortunately, it can be solved algebraically for normal distributions, as established by Givens and Shortt [21]. The contrast vector between one central cell distribution P_c^i and its neighboring cell distribution P_s^i along channel i can be computed in a closed form as follows:

$$\mathcal{W}_2(P_c^i, P_s^i) = \left[||\mu_c^i - \mu_c^i||_2^2 + \Sigma_c^i + \Sigma_s^i - 2\sqrt{\Sigma_c^i\Sigma_s^i}\right]^{\frac{1}{2}}. \tag{4}$$

Gradient matrix

For each center-surround cell pair, we compute the gradient matrix for the mean and variance vector (μ^i, Σ^i), resulting in a contrast vector. The contrast vector between one central cell distribution P_c^i and its neighboring cell distribution P_s^i along channel i can be expressed as follows:

$$\vec{cst}(P_c^i, P_s^i) = \left(\mu_c^i - \mu_s^i, \Sigma_c^i - \Sigma_s^i\right). \tag{5}$$

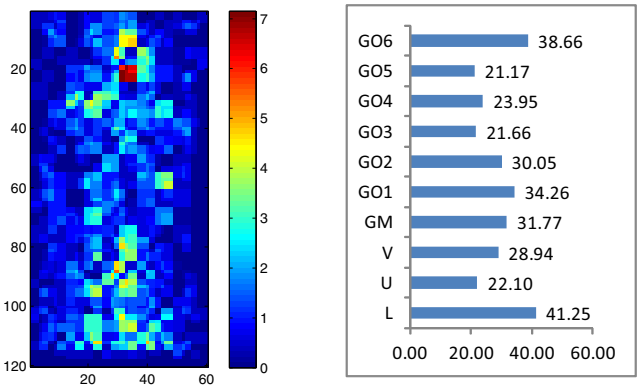
In the feature space, the contrast vector in Eq. 5 is treated as two separate values, which enables convinent training procedure.

In accordance with the general architecture of most visual saliency systems, we build a multi-scale contrast pyramid by varying the cell size at each scale. The final contrast feature space for the whole image consists of all the contrast values from each cell at each scale and along multiple channels.

IV. Feature Selection for Pedestrian Detection

For baseline comparison, we consider [ChnFtrs] [22] which reaches state-of-the-art performance and, in addition, is fast to train and test. It first computes multiple channel information for an image and then considers sums over rectangular channel regions as features which can be computed very efficiently when using integral images.

Our own detector employs the center-surround contrast features proposed in Sec. III. Note that these features, too, are built on channel features but interpret local differences between central rectangular regions and their eight nearest neighboring regions over multiple channels rather than over channel values themselves.



(a) (b)

Fig. 3: Illustration of representative features. (a) Body parts weight map: different colors are used to indicate the accumulative weight of each pixel after boosting. (b) Channel weight bars: accumulative weight of each channel are indicated by one bar (best viewed in color).

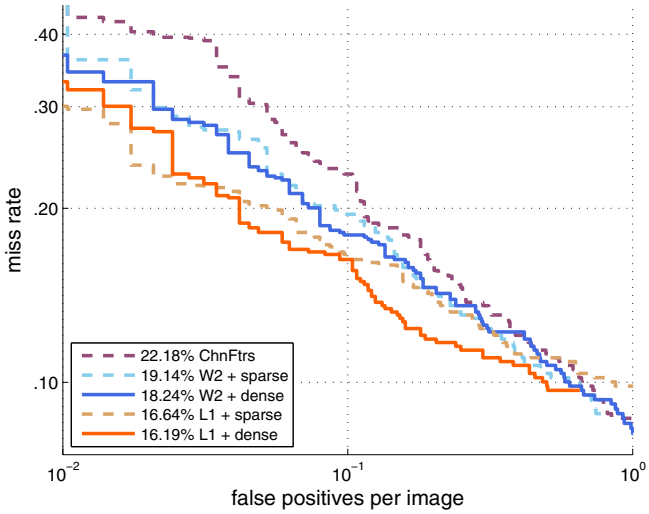


Fig. 4: Results on INRIA dataset under different configurations.

We apply a fast version of AdaBoost [23] for feature selection since it offers a convenient and fast approach to select from a large number of candidate features. We choose decision trees of depth 2 as our weak classifiers and build our final strong classifier with 4096 weak classifiers in total. Initial negative training samples are randomly generated from natural images that do not contain pedestrians and, afterwards, hard negative samples are searched for three rounds over all negative sample images so as to collect 20,000 negative samples in total. This multi-round training strategy is pivotal as it leads to a better performance than a simple one round training procedure with the same number of negative samples. From our experiments, three rounds of retraining were observed to yield optimal performance; additional rounds did not show significant improvements.

In order to look into which features are more informative, we observe in terms of: body parts and channels. First, we collect the weights of the top 1000 features. Then we add the weight of each feature to the pixels it covers, including those inside the central and surrounding cells, and use different colors to indicate the accumulative weight of each cell after boosting as shown in Fig. 3a. As expected, the head-shoulder and feet area of the human body show to be more discriminative for pedestrian detection than other body parts. Moreover, we also add the weight of each feature to the channels it goes through to observe which channels are more representative and use bars to illustrate the accumulative weight of each channel as shown in Fig. 3b. We find that all the channels we choose contribute relatively even to the final classifier, indicating no redundancy in channels.

The most discriminative features determined by the boosting algorithm are then used for pedestrian detection in still images. To this end, we slide a window over the whole image and consider multiple scales. The spatial step size is set to be half of the smallest cell size according to the Nyquist-Shannon sampling theorem, and the scale step is set to be $2^{\frac{1}{8}}$ so that there are 8 scales in each octave. Finally, we use a simplified non-maximal suppression (NMS) procedure [22] to suppress nearby detections.

V. Experiments

We conducted experiments on two public benchmark datasets: the INRIA pedestrian dataset [12] and the Caltech pedestrian dataset [1]. The INRIA data is arguably the most popular dataset for people detection and comes along with predefined subsets for training and testing. The Caltech data is the largest and most challenging dataset for pedestrian detection and we consider subsets set00 - set05 for training and subsets set06 - set10 for testing.

A. Discussions on configurations

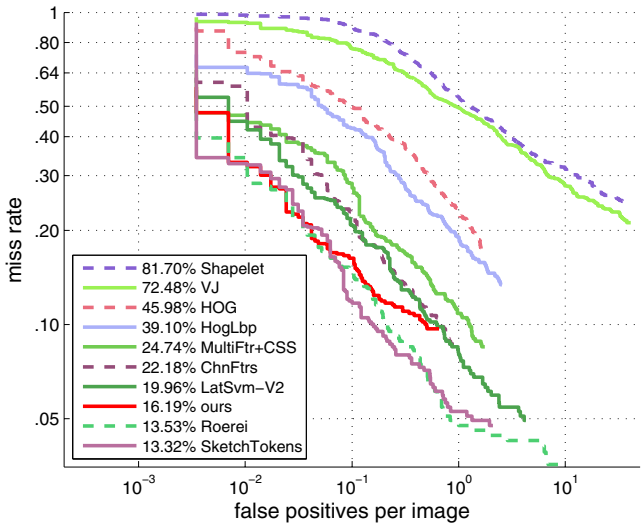
We proposed two neighborhood structures and two distance measurements in Sec. III. Experimental results on the INRIA dataset under different configurations are shown in Fig. 4.

We find that the results under different configurations are similar to each other, which indicates that our feature extraction scheme is robust. Still, $\mathcal{W}_2$ distance leads to a little worse results than gradient matrix because the latter one incorporates more specific differences on mean and variance values separately, yielding a higher dimensional feature space; sparse neighborhood structure results in better results on both contrast measurements, due to richer neighboring differences integrated.

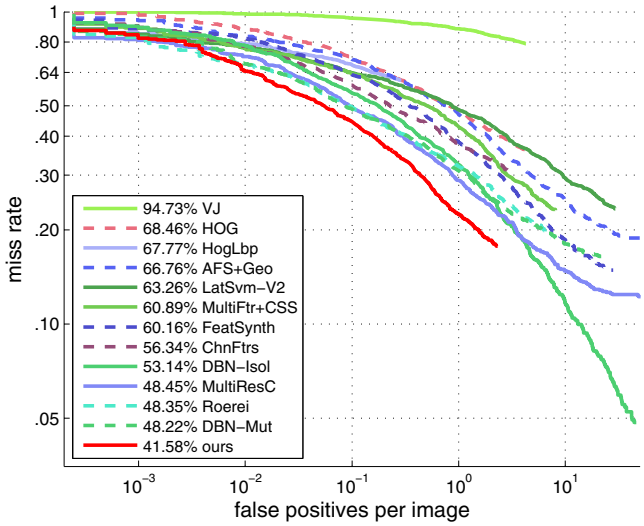
B. Comparisons with state-of-the-art detectors

In this section, we compare our detector to other state-of-the-art detectors whose results are publicly available[1]. Notably, some detectors are not considered here because of use of motion information or semantic analysis of scene which requires elaborate preprocessing.

[1] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/



(a) INRIA



(b) Caltech test

Fig. 5: Results of different detectors on different datasets under standard evaluation settings (best viewed in color).

We use the same experimental protocol as in [1]: first, the overall results are produced on the *reasonable* subset of each test set which shows pedestrians at a resolution of over 50 pixels in height and a visibility of at least 65%; second, one ground truth annotation and one detection bounding box are considered to be matched if and only if their overlap region exceed 50% of their union region.

To compare the performance of state-of-the-art detectors, we plot miss rate against false positives per image (FPPI) curves in logarithmic scales by varying the threshold on the detection confidence. To summarize the overall performances of different detectors, a numerical measurement of *average miss rate* is computed by averaging the miss rates at nine FPPI rates evenly sampled in log-space in the range of $[10^{-2}, 10^{0}]$.

Our pedestrian model is of 96×48 pixels, while our detection window is of 120×60 pixels, including borders of 6 pixels on the left and right, and 12 pixels on the top and bottom as context. In this section, we use the following configurations for the experiments: three scales with cell sizes of 3×3, 4×4 and 6×6 pixels; dense map; gradient matrix contrast measurement.

The results on the INRIA dataset in Fig. 5a show that our detector outperforms the baseline detector [ChnFtrs] by about 6% and reaches the state-of-the-art performance; on the Caltech pedestrian dataset, our detector outperforms not only the baseline detector [ChnFtrs] by about 15% but also yields the overall best performance as shown in Fig. 5b. More extensive comparisons are shown in Tab. I.

VI. Conclusion

Humans are able to efficiently locate what they are looking for, because characteristic visual features are tuned so that those entities become salient. This is called top-down saliency or visual search. We tried to mimic early human visual processing by local distribution contrast features and boosted them to respond to the appearance of pedestrians, so we would call our detector a computational top-down saliency system. Our features are very efficient to compute combining a fast integral method for local averaging and clever arrangement of additional image layers for quick maximum likelihood estimates of normal distributions. We tested different patterns for organizing the center-surround structure as well as different ways to estimate the cell-contrasts.

Experimental results showed that our detector achieves state-of-the-art performance on the INRIA pedestrian dataset and, for the Caltech pedestrian dataset, we found it to outperform all other recent approaches considered.

Given these results, it appears promising to further explore feature design driven by human visual mechanisms. Immediate extensions of the approach presented in this paper could be to incorporate information from additional modalities such as motion and depth.

References

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Trans. PAMI*, vol. 34, no. 4, pp. 743–761, 2011. 1, 4, 5

[2] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *CVPR*, 2010. 1, 2, 6

[3] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *CVPR*, 2008. 1, 6

[4] S. Zhang, C. Bauckhage, D. Klein, and A. Cremers, "Moving pedestrian detection based on motion segmentation," in *IEEE Workshop on Robot Vision (WoRV)*, 2013. 1

[5] R. W. Rodieck, "Quantitative analysis of cat retinal ganglion cell response to visual stimuli," *Vision Research*, vol. 5, no. 12, pp. 583–601, Dec 1965. 1

[6] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, no. 6, pp. 1233–1258, Dec 1987. 1

[7] B. B. Lee, P. R. Martin, and U. Grünert, "Retinal connectivity and primate vision," *Progress in Retinal and Eye Research*, vol. 29, no. 6, pp. 622–639, Nov 2010. 1

Detector	Features	Classifier	Average miss rate	
			INRIA	Caltech
VJ[24]	Haar	AdaBoost	72.48%	94.73%
HOG[12]	HOG	linear SVM	45.98%	68.46%
Shapelet[25]	gradients	AdaBoost	81.70%	91.37%
MultiFtr+CSS [2]	HOG + CSS	AdaBoost	24.74%	60.89%
HikSvm [3]	HOG	HIK SVM	42.82%	73.39%
HogLbp [14]	HOG + LBP	linear SVM	39.10%	67.77%
LatSvm-V1 [26]	HOG	latent SVM	43.83%	79.78%
LatSvm-V2 [27]	HOG	latent SVM	19.96%	63.26%
ChnFtrs [22]	channels	AdaBoost	22.18%	56.34%
FeatSynth [28]	HOG + texture	linear SVM	30.88%	60.16%
MultiResC [29]	HOG	latent SVM	/	48.45%
CrossTalk [30]	channels	AdaBoost	18.98%	53.88%
VeryFast [31]	channels	AdaBoost	15.96%	/
SketchTokens [32]	channels	AdaBoost	13.32%	/
Roerei [33]	channels	AdaBoost	13.53%	48.35%
AFS+Geo [34]	HOG + texture	linear SVM	/	66.76%
DBN-Isol [35]	HOG	DeepNet	/	53.14%
DBN-Mut [36]	HOG	DeepNet	/	48.22%
ours	center-surround contrast	AdaBoost	16.19%	41.58%

TABLE I: Performance comparisons for state-of-the-art pedestrian detectors. Each row in this table summarizes information as to features and classifiers used in a particular approach, and displays the corresponding average performance in terms of miss rates. The approach proposed in this paper yields state-of-the-art performance on the INRIA dataset and consistently better results than previously reported on the Caltech dataset.

[8] S. Frintrop, E. Rome, and C. H. I., "Computational Visual Attention Systems and their Cognitive Foundation: A Survey," *ACM Trans. on Applied Perception*, vol. 7, no. 1, 2010. 1

[9] D.A. Klein and S. Frintrop, "Salient Pattern Detection using W2 on Multivariate Normal Distributions," in *DAGM*, 2012. 1, 2, 3

[10] F. Porikli, "Integral histogram: A fast way to extract histograms in Cartesian spaces." in *Proc. of CVPR*, 2005. 1

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, Sep 2006. 1

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005. 2, 4, 6

[13] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996. 2

[14] X. Wang and T. X. Han, "An HOG-LBP human detector with partial occlusion handling," in *ICCV*, 2009. 2, 6

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001. 2

[16] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. of CV*, vol. 63, no. 2, pp. 153–161, 2005. 2

[17] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *CVPR*, 2014. 2

[18] S. Montabone and A. Soto, "Human detection using a mobile platform and novel features derived from a visual saliency mechanism," *Image and Vision Computing*, vol. 28, no. 3, pp. 391–402, 2010. 2

[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998. 2

[20] D. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *ICCV*, 2011. 2

[21] C. Givens and R. Shortt, "A class of wasserstein metrics for probability distributions," *Michigan Math. J.*, vol. 2, no. 31, 1984. 3

[22] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *BMVC*, 2009. 3, 4, 6

[23] R. Appel, T. Fuchs, P. Dollár, and P. Perona, "Quickly boosting decision trees-pruning underachieving features early," in *ICML*, 2013. 4

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. of CV*, vol. 57, no. 2, pp. 137–154, 2004. 6

[25] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *CVPR*, 2007. 6

[26] P. F. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008. 6

[27] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010. 6

[28] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," in *ECCV*, 2010. 6

[29] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *ECCV*, 2010. 6

[30] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *CVPR*, 2012. 6

[31] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool, "Pedestrian detection at 100 frames per second," in *CVPR*, 2012. 6

[32] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: a learned mid-level representation for contour and object detection," in *CVPR*, 2013. 6

[33] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool, "Seeking the strongest rigid detector," in *CVPR*, 2013. 6

[34] D. Levi, S. Silberstein, and A. Bar-Hillel, "Fast multiple-part based object detection using kd-ferns," in *CVPR*, 2013. 6

[35] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *CVPR*, 2012. 6

[36] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship with a deep model in pedestrian detection," in *CVPR*, 2013. 6