Optical Engineering

SPIEDigitalLibrary.org/oe

Pair normalized channel feature and statistics-based learning for high-performance pedestrian detection

Bobo Zeng
Guijin Wang
Zhiwei Ruan
Xinggang Lin
Long Meng

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Pair normalized channel feature and statistics-based learning for high-performance pedestrian detection

Bobo Zeng
Tsinghua University
Department of Electronic Engineering
Beijing 100084, China
and
Tsinghua University
Graduate School at Shenzhen
Shenzhen 518055, China

Guijin Wang
Zhiwei Ruan
Xinggang Lin
Tsinghua University
Department of Electronic Engineering
Beijing 100084, China
E-mail: wangguijin@tsinghua.edu.cn

Long Meng
Beijing Mingrui Automatic Technology Company
Beijing 100872, China

Abstract. High-performance pedestrian detection with good accuracy and fast speed is an important yet challenging task in computer vision. We design a novel feature named pair normalized channel feature (PNCF), which simultaneously combines and normalizes two channel features in image channels, achieving a highly discriminative power and computational efficiency. PNCF applies to both gradient channels and color channels so that shape and appearance information are described and integrated in the same feature. To efficiently explore the formidably large PNCF feature space, we propose a statistics-based feature learning method to select a small number of potentially discriminative candidate features, which are fed into the boosting algorithm. In addition, channel compression and a hybrid pyramid are employed to speed up the multiscale detection. Experiments illustrate the effectiveness of PNCF and its learning method. Our proposed detector outperforms the state-of-the-art on several benchmark datasets in both detection accuracy and efficiency.
© 2012 Society of Photo-Optical Instrumentation Engineers (SPIE). [DOI: 10.1117/1.OE.51.7.077206]

Subject terms: pedestrian detection; pair-normalized channel features; high-performance detection; statistics based learning.

Paper 120399 received Mar. 15, 2012; revised manuscript received May 23, 2012; accepted for publication May 31, 2012; published online Jul. 17, 2012.

1 Introduction

Pedestrian detection is an important task modified with many applications in computer vision, such as intelligent visual surveillance and advanced driver assistance systems.[1] However, high performance pedestrian detection with good accuracy and fast speed is still a challenge due to the large variations within each class caused by varying poses, illuminations, clothes and so on.

Feature design and its learning method is the key factor for object detection performance. Viola and Jones (VJ)'s seminal face detector[2] benefits much from Haar feature's high discriminative power and low computation cost. In pedestrian detection, many features have been proposed. They can be categorized as shape features [histogram of oriented gradients (HOG),[3] Edgelet,[4] Covariance,[5] etc.], texture features (LBP,[6], etc.), color features (color similarity,[7], etc.) and motion features (HOF,[7], etc.). Among them, HOG is the most influential with its high discriminative power and moderate computational cost.

Because these features are multi-dimensional, the computational cost is high. Scalar features[8–12] compute much faster, while maintaining the discriminative power with a boosting learning algorithm.[13] The HOG bin feature[8,9] is a direct variant of HOG that uses only its bins. Oriented granular feature[10] is also a similar variant. Recently, channel feature[11] is defined on multiple image channels (gradient, color, etc.), but it has no sophisticated normalization, which may be deficient under nonuniform illumination. Scalar features can also be composed by joining several individual scalar features through addition, subtraction or multiplication.[10,12] Joint features are more discriminative, but the computational burden is increased.

In feature learning, most previous works[2,5,11,14] conduct an exhaustive search to explore the feature space, which is a slow process. Some proposed methods intend to learn more efficiently by sampling the feature space. Heuristic grow, combine and cut operations are used in Ref. 10 to generate the features to be searched. Nested AdaBoost is employed in Ref. 12 to select a few seed features and then combine the seed features with many other candidate features. Huang et al.[15] utilize a collaborative learning method for feature selection.

In detection, classifier evaluation and sliding window search determine the detection speed. For scalar features, evaluation with cascade of boosted classifiers[2] is fast and widely employed. However, sliding window search in multiple positions and scales is slow, thus two kinds of complementary approaches are proposed to speed up the search. The first is reducing the number of searched windows, including branch and bound search,[16] pre-filtering with other cues (e.g., foreground extraction,[17] geometry and perspective,[18,19] depth[20]), visual search[21] and coarse-to-fine search.[22,23] The second is reducing the time in constructing image pyramid and pre-processing for feature extraction (e.g., gradient calculation, integral image computing), which is the bottleneck in multiscale detection. Viola et al.[2] build a detector pyramid by directly scaling Haar features to avoid constructing an image pyramid, but most features are not scale invariant. At the cost of reducing detection accuracy, Dollár et al.[24] use gradients in one scale to approximate gradients in multiple nearby scales.

This work has the following contributions. First, we design a novel feature named pair normalized channel

0091-3286/2012/$25.00 © 2012 SPIE

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

feature (PNCF) as the ratio of two channel features (see Fig. 1). The ratio operation accomplishes feature-normalization and combination simultaneously without additional computation. Normalization helps PNCF to be robust to illumination variations. Combination further enhances PNCF's discriminative power. PNCF is versatile as the joint features lie either inside the same image channel or across multiple channels. In addition, PNCF applies to multiple channel types including gradient channels and color channels, so shape and appearance information can be described and integrated in the same feature. Second, we propose a statistics-based PNCF feature-learning method, which selects features efficiently to speed up the detector-training process. The basic idea is that statistics of channel features help to generate discriminative PNCF features more efficiently, thus allowing for the statistics themselves to be computed fast.[25] Third, we combine a detector and image pyramid for fast multiscale detection. Comparisons with state-of-the-art detectors demonstrate that our detector improves the accuracy considerably and runs significantly faster.

The rest of the paper is organized as follows: Sec. 2 introduces the image channels and the novel PNCF feature. Section 3 describes the statistics-based feature learning method, with two variants for combining it with boosting learning algorithm. Section 4 presents the multiscale detection method. Section 5 provides extensive experimental results on several challenging datasets. Section 6 provides the conclusion.

2 Pair Normalized Channel Feature

2.1 *Image Channels*

Inspired by Ref. 11 multiple image channels $C_n(x,y)$ are extracted from the input image $I(x,y)$. We use two types of channels, gradient channels (GC) and color channels (CC), to describe the pedestrian's shape and appearance, respectively.

GC channels are inspired by the HOG feature.[3] The image I's gradient magnitude $G(x,y)$ and orientation $O(x,y)$ are computed as

$$\begin{cases} G(x,y) = \sqrt{I_x(x,y)^2 + I_y(x,y)^2} \\ O(x,y) = \arctan[I_y(x,y)/I_x(x,y)] + \frac{\pi}{2} \end{cases}, \tag{1}$$

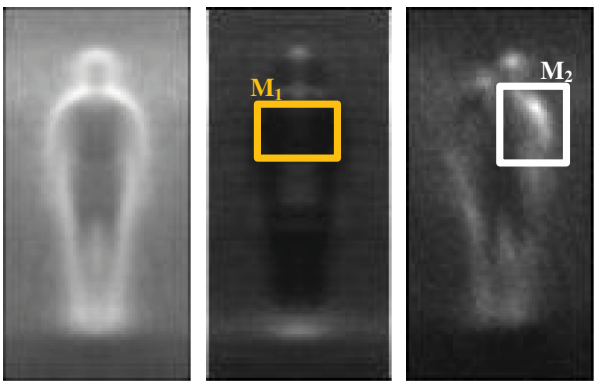


Fig. 1 Illustration of PNCF feature. Discriminative channel features M_2 and M_1 are selected in two image channels. They are combined by ratio operation as $\frac{M_2}{M_1}$, resulting in a normalized and more discriminative feature.

where $I_x(x,y)$, $I_y(x,y)$ are the x, y derivatives obtained by a $[-1 \quad 0 \quad 1]$ mask. The orientation's range $(0,\pi)$ is quantized into N_b bins (e.g., $N_b = 9$) with bin centers θ_n, $n = 1, \ldots, N_b$. Gradient orientation channel $G_n(x,y)$ contains the gradient magnitude within bin n's orientation range. $G(x,y)$ is allocated into two neighboring channels $G_n(x,y)$ and $G_{n+1}(x,y)$ with linear interpolation according to its orientation $O(x,y) \in (\theta_n, \theta_{n+1})$. The N_b gradient orientation channels $G_n(x,y)$, $n = 1, \ldots, N_b$, GC contains one additional gradient magnitude channel $G_0(x,y) = G(x,y)$.

Color channels are three color components of image I in one particular color space, such as LUV [LUV channel (LC)]; red, green and blue [RGB (RC)] or hue, saturation, and value [HSV (HC)]. The best color space will be identified in the experiments.

The average GC and LC channels of 2000 pedestrian samples are shown in Fig. 2. The brighter or darker than the average regions in the channels serves as informative pedestrian cues, such as the shoulder positions in the GC or the head positions in the LC. The figure illustrates that GC and LC are complementary, as they have different informative regions.

2.2 *PNCF Feature*

Features are designed in the image channels. A channel feature $M_n(R)$ is defined as the sum of a local rectangular region $R(x,y,w,h)$ in the channel $C_n(x,y)$:

$$M_n(R) = \sum_{(u,v)\in R(x,y,w,h)} C_n(u,v), \tag{2}$$

where n is the channel index. The feature is fast to compute by integral image,[2] with only four table lookup and three addition/subtraction operations. The channel feature set $\mathcal{M} = \{M_n(R)\}$ is generated with all the possible rectangles R on all the N_c channels. To reduce $\mathcal{M}$ size, parameters x, y, w and h of R are restricted to be divisible by 4. In this case, we propose to use a channel compression technique. The channel $C_n(x,y)$ with size $(W_{\text{ori}}, H_{\text{ori}})$ is compressed to a quarter as $(W,H) = (\frac{W_{\text{ori}}}{4}, \frac{H_{\text{ori}}}{4})$, by summing each nonoverlapping 4×4 patch's values sequentially in $C_n(x,y)$. The total feature number

$$card(\mathcal{M}) = \frac{(W+1)W}{2} \times \frac{(H+1)H}{2} N_c = \quad O(W^2H^2N_c).$$

Apart from feature number reduction, this channel compression technique is also advantageous to the subsequent feature-learning and multi-scale detection module, which will be explained later.

Directly using the channel features for classification is not appropriate since it is sensitive to illumination variations. We propose the PNCF defined as

$$P_{n_1,n_2}(R_1,R_2) = \frac{M_{n_1}(R_1)}{M_{n_2}(R_2)}, \tag{3}$$

where numerator $M_{n_1}(R_1)$ and denominator $M_{n_2}(R_2)$ constitute the channel feature pair. n_1 and n_2 belongs to the same channel type. We constrain R_1 and R_2 to be adjacent in

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

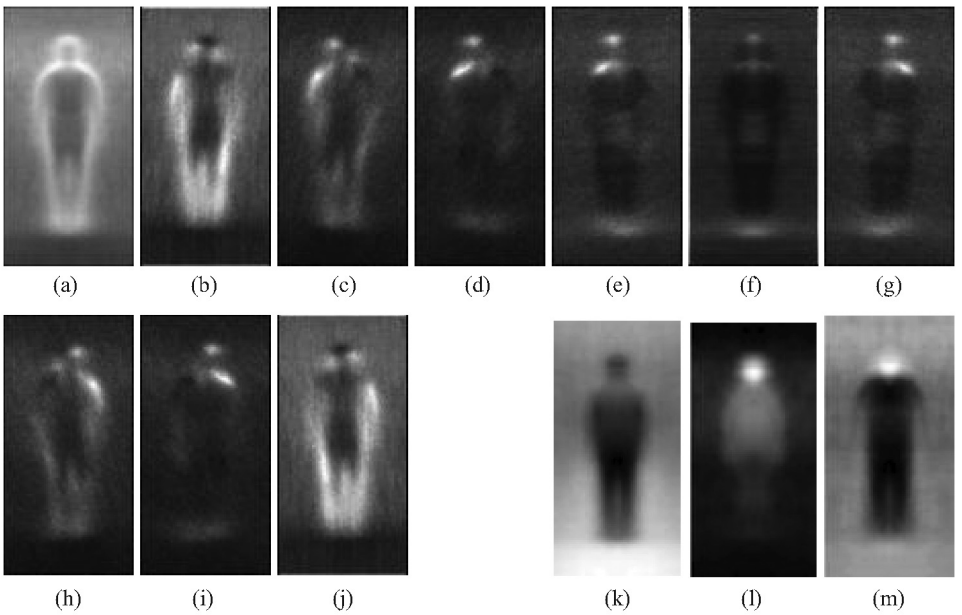


Fig. 2 Average GC and LC channels of two thousands pedestrian samples. The regions brighter or darker than average are informative. (a) Gradient magnitude channel. (b)–(j) 9 gradient orientation channels. (k)–(m) 3 LUV color channels.

position and scale to have approximately uniform illumination. This local constraint is

$$\begin{aligned} |R_1(x) - R_2(x)| &\le (r_a - 1)R_1(w) \\ |R_1(y) - R_2(y)| &\le (r_a - 1)R_1(h) \\ \frac{1}{r_a} \le \frac{R_1(w)}{R_2(w)} &\le r_a \\ \frac{1}{r_a} \le \frac{R_1(h)}{R_2(h)} &\le r_a \quad , \end{aligned} \tag{4}$$

where r_a is the adjacent rate.

With the ratio operation, PNCF is normalized to gain illumination invariance. The pedestrian is supposed to be a Lambertian surface, so the image is

$$I(x, y) = \rho(x, y)n(x, y) \cdot s, \tag{5}$$

where $\rho(x, y)$ is the albedo, $n(x, y)$ is the surface normal direction, and $\boldsymbol{s}$ is the illumination source intensity with direction. Note that under uniform illumination, $\boldsymbol{s}$ is constant. If the illumination s changes to $s' = \alpha s$ by a factor α, the corresponding image is $I'(x, y) = \alpha I(x, y)$. Then the channel and channel feature has $C'_n(x, y) = \alpha C_n(x, y)$, $M'_n(R) = \alpha M_n(R)$, so it changes with illumination variation. But after normalization,

$$P'_{n_1,n_2}(R_1, R_2) = \frac{M'_{n_1}(R_1)}{M'_{n_2}(R_2)} = \frac{\alpha M_{n_1}(R_1)}{\alpha M_{n_2}(R_2)} = P_{n_1,n_2}(R_1, R_2). \tag{6}$$

The ratio operation ensures PNCF's invariance to illumination variation.

The above derivation applies to the gradient channels, as well as color channels of linear color spaces, such as RGB space. For nonlinear color spaces, it does not necessarily hold. But for LUV space, since both U and V channel are obtained by multiplying L channel with an illumination invariant term, the ratio of PNCF can also remove the illumination variations similarly.

The ratio also makes PNCF a joint feature by combining two channel features, making the PNCF feature versatile and enhances its discriminative power. Equation (3) illustrates that the joint features can either lie inside the same image channel ($n_1 = n_2$) or across multiple channels ($n_1 \neq n_2$), so both inter- and intra- channel information are utilized and that the rectangles R_1 and R_2 do not need to be the same.

If exhaustively combining all the possible channel feature pairs from $\mathcal{M}$ under the local constraint, the generated PNCF set $\mathcal{P}$ has $\text{card}(\mathcal{P}) = \text{card}(\mathcal{M})^2 = O(W^4H^4N_c^2)$. For typical parameters ($W = 16$, $H = 32$, $N_c = 13$), $\mathcal{P}$ contains billions of features. The huge PNCF feature space provides plenty of potential discriminative features, but it is computationally formidable to be fully explored. Therefore, an efficient PNCF feature-learning method is necessary.

3 Statistics-Based PNCF Feature Learning

We propose a statistics-based feature-learning method that contains two modules, with the goal of reducing the feature set $\mathcal{P}$. The first module is a seed channel features selection as illustrated in Fig. 3(a). The basic principle is $\mathcal{M}$ can be refined by only selecting discriminative features as the seeds. The feature's discriminative power is measured by its statistics (i.e., mean and covariance), which can be calculated efficiently with an integral image. The channel feature is then categorized into two types (Sec. 3.2), given the feature statistics. The second module is the PNCF generation and selection with boosting algorithm. PNCF features are generated by combining seed features pairwisely according to its type. Since the number of seed features is reduced, PNCF set $\mathcal{P}$ is reduced greatly. Then PNCF generation is fed into boosting learning algorithms.

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

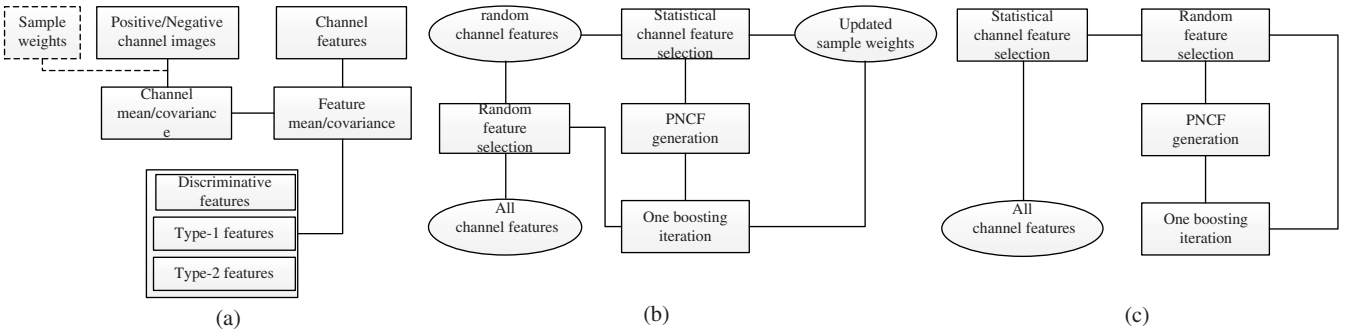


Fig. 3 The process of PNCF feature learning. (a) Statistical channel feature selection. Discriminative channel features are selected as seed features based on channel statistics. Dashed module means it is optional. (b) IBS method. Seed channel features are selected with the updated sample weights in each boosting round. (c) PBS method. Seed channel features are selected only once before all the boosting rounds.

3.1 *Fast Statistics Calculation*

For a channel image $C_n(x, y)$ with size $W \times H$, its integral image $IC_n(x, y)$ is calculated and represented as a $d \times 1 (d = WH)$ column vector $\boldsymbol{x}$. The value of a channel feature $M_n(R)$ (denoted as m) is computed by

$$v_m = \mathbf{g}_m^T \mathbf{x}, \tag{7}$$

where $\mathbf{g}_m$ is a sparse vector with only four nonzeros determined by four vertices of R. The sparsity of $\mathbf{g}_m$ explains the extremely fast computation of the feature. We denote the expectation operator as $\langle \cdot \rangle$. Given the samples which are weighted in the latter boosting learning algorithm (Sec. 3.3), the feature mean μ_m and feature covariance σ_m^2 are computed as

$$\mu_m = \langle v_m \rangle = \langle \omega \mathbf{g}_m^T \mathbf{x} \rangle = \mathbf{g}_m^T \boldsymbol{\mu}_\mathbf{x}, \tag{8}$$

where $\boldsymbol{\mu}_\mathbf{x} = \langle \omega \mathbf{x} \rangle$ is the channel mean vector, and ω is the normalized sample weights.

$$\begin{aligned} \sigma_m^2 &= \langle v_m^2 \rangle - \langle v_m \rangle^2 = \langle \omega \mathbf{g}_m^T \mathbf{x} \mathbf{x}^T \mathbf{g}_m \rangle - \mathbf{g}_m^T \boldsymbol{\mu}_\mathbf{x} \boldsymbol{\mu}_\mathbf{x}^T \mathbf{g}_m \\ &= \mathbf{g}_m^T (\langle \omega \mathbf{x} \mathbf{x}^T \rangle - \boldsymbol{\mu}_\mathbf{x} \boldsymbol{\mu}_\mathbf{x}^T) \mathbf{g}_m = \mathbf{g}_m^T \boldsymbol{\sigma}_\mathbf{x}^2 \mathbf{g}_m, \end{aligned} \tag{9}$$

where $\boldsymbol{\sigma}_\mathbf{x}^2 = \langle \omega \mathbf{x} \mathbf{x}^T \rangle - \boldsymbol{\mu}_\mathbf{x} \boldsymbol{\mu}_\mathbf{x}^T$ is the channel covariance matrix. The above channel statistics $\boldsymbol{\mu}_\mathbf{x}$ and $\boldsymbol{\sigma}_\mathbf{x}^2$ are calculated with sample weights, and the unweighted statistics are obtained by simply removing the ω term. With Eqs. (8) and (9), the feature's statistics can be calculated fast given the channel statistics due to the sparsity of $\mathbf{g}_m$. The time complexity for calculating the channel statistics is $O(N_s N_c d^2)$, where N_s and N_c are the number of samples and channels. With the channel compression technique (Sec. 2.2), the space and time complexity drop to 1/16 and 1/256, respectively.

For a two-class classification problem with $c = \{0, 1\}$, we denote the positive and negative feature and channel statistics as $\mu_{m,c}$, $\sigma_{m,c}^2$, $\boldsymbol{\mu}_{\mathbf{x},c}$ and $\boldsymbol{\sigma}_{\mathbf{x},c}^2$. They can be calculated efficiently with the positive and negative samples, respectively.

3.2 *PNCF Features Generation*

Given all the channel features and their statistics, we propose a discriminative power measure to select the seed features. We use the feature's classification error on the samples as an indicator. The lower the error, the higher the discriminative power.

Assuming the feature's values v_m in both classes obey the normal distribution as $v_m|c \sim \mathcal{N}(\mu_{m,c}, \sigma_{m,c}^2)$ for $c = \{0, 1\}$, a simple bayesian thresholding classifier with optimal threshold θ_m and parity $t_m = \{-1, 1\}$ can be obtained in a closed form solution (see chapter 2 in Ref. 26). The classification error e_m is

$$e_m = P(t_m v_m > t_m \theta_m | c = 0) + P(t_m v_m < t_m \theta_m | c = 1). \tag{10}$$

The channel features are sorted according to their errors and features with low errors are selected.

A PNCF feature p is a ratio of a numerator feature m_n and denominator feature m_d with normal distribution, but this ratio distribution is too complicated.[27] Fortunately, precise statistical characteristic of a PNCF feature is not necessary for the PNCF feature generation. Thus we propose a rough rule based on mean value. Feature p's truncated mean is approximated as

$$\mu_{p,c} = \left\langle \frac{v_{m_n}}{v_{m_d}} \right\rangle | c \approx \tilde{\mu}_{p,c} = \frac{\mu_{m_n,c}}{\mu_{m_d,c}}. \tag{11}$$

Truncated mean excludes outliers caused by small denominators, which are too rare to influence the classification. We define μr as ratio of a feature's positive and negative class mean, for reflecting the feature's classification margin of two classes. Then we can derive

$$\mu r_p = \frac{\mu_{p,1}}{\mu_{p,0}} \approx \tilde{\mu r}_p = \frac{\mu_{m_n,1}/\mu_{m_d,1}}{\mu_{m_n,0}/\mu_{m_d,0}} = \frac{\mu_{m_n,1}/\mu_{m_n,0}}{\mu_{m_d,1}/\mu_{m_d,0}} = \frac{\mu r_{m_n}}{\mu r_{m_d}}. \tag{12}$$

To make μr_p large for good classification, μr_{m_n} should be large and μr_{m_d} be small. In this case, the channel features are categorized into two types named type-1 ($\mu r > 1$) and type-2 ($\mu r < 1$). We restrict m_n to be type-1 and m_d to be type-2. Then $\mu r_p > \mu r_{m_n} > 1$, so PNCF feature p has improved discriminative power compared with m_n. Table 1 gives two real examples. From the perspective of classification error, the approximation in Eqs. (11) and (12) is good enough. When compared with m_n and m_d, p has a larger μr and thus lower classification error e_p indeed, achieving better performance. As for PNCF's physical significance, type-1 feature lies in the channel's brighter region and type-2 lies in the darker region (see Fig. 2). Both regions

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 1 Two real examples of statistics and classification errors of numerator, denominator and resulting PNCF feature. "•" denotes the feature.

Example	Feature	$(\mu_{\bullet,1}, \mu_{\bullet,0})$, $(\tilde{\mu}_{\bullet,1}, \tilde{\mu}_{\bullet,0})$	$(\mu r_\bullet, \tilde{\mu r}_\bullet)$	$e_\bullet$
#1	m_n	$(4.31 \times 10^4, 2.17 \times 10^4)$, (-,-)	(1.98, -)	0.121
	m_d	$(4.08 \times 10^3, 6.05 \times 10^3)$, (-,-)	(0.675, -)	0.193
	$p = \frac{m_n}{m_d}$	(13.1, 4.98), (10.6, 3.59)	(2.94, 2.63)	0.104
#2	m_n	$(3.55 \times 10^4, 1.71 \times 10^4)$, (-,-)	(1.98, -)	0.118
	m_d	$(2.67 \times 10^3, 5.49 \times 10^3)$, (-,-)	(0.485, -)	0.148
	$p = \frac{m_n}{m_d}$	(18.0, 4.61), (13.3, 3.11)	(4.28, 3.91)	0.0851

are informative and discriminative. Also, the ratio of brighter versus darker region makes it brighter, hence more discriminative.

We select S_1 type-1 features and S_2 type-2 features with small classification error as the seeds, and generate $S_1 \times S_2$ PNCF features by combining each pair of type-1 and type-2 features.

3.3 *PNCF Feature Selection with Boosting*

The Gentle AdaBoost type boosting learning algorithm[13] is utilized to select the features from the generated PNCF feature pool. We randomly sample a small set of features for boosting with sampling rate r_s, since randomness can improve the diversity of the ensemble classifier and reduce the training time complexity.[28] Decision stump is employed as the weak classifier.

To integrate the PNCF generation and random selection into boosting framework, we consider two methods called InBoost selection (IBS) and PreBoost selection (PBS). InBoost is similar to Ref. 25. In IBS [see Fig. 3(b)], channel features are randomly sampled and then seed features are selected based on channel and feature statistics. With the seed features, the PNCF features are generated and then the most discriminative one is selected by boosting. Since the sample weights are updated in each boosting round, the channel statistics should be updated correspondingly in every boosting round, resulting in a large computational cost. To make it more efficient, we propose the PBS [see Fig. 3(c)] as an alternative. Before boosting, channel and features statistics are calculated without sample weights [see dashed module in Fig. 3(b)]. Then the seed features are selected. In each boosting learning round, features are randomly sampled from the seed features and PNCF features are generated. Since channel statistics calculation does not involve updated sample weights in boosting rounds, time complexity drops greatly compared with PBS.

4 Fast Multi-scale Detection

The detector is a cascade of boosted classifiers for classification. Even though the PNCF feature and the cascade structure make the classification fast, preprocessing channels and corresponding integral images is still a bottleneck in multi-scale detection. To increase speed, the channel compression technique in Sec. 2.2 is applied to reduce the time in computing integral images. A channel that is not compressed with size $(W_{\text{ori}}, H_{\text{ori}})$ costs $2\,W_{\text{ori}}H_{\text{ori}}$ additions to compute the integral image, while a compressed one only costs $W_{\text{ori}}H_{\text{ori}}$ plus $(W_{\text{ori}}H_{\text{ori}})/8$ additions, which is a 40% decrease.

Another speedy technique is to utilize a hybrid pyramid (HPy) by combining the detector pyramid and image pyramid. With a detector pyramid, the preprocessing is performed only once, saving lots of time. To construct the detector pyramid, Ref. 2 directly scales the detector to different sizes. But it is not feasible for PNCF feature as it is sensitive to scale. We propose to train the detector pyramid at the cost of increased training time. The detector pyramid includes seven pedestrian detectors of sizes 32×64, 36×72, 44×88, 48×96, 56×112, 64×128, 72×144. For pedestrians larger than 72×144, we use the 72×144 detector to detect them by image pyramid. With this HPy less preprocessing is needed, so overall detection time is reduced significantly. On the other hand, most previous detectors use a fixed size of 64×128, which may cause performance damage when detecting smaller pedestrians by upsampling.

5 Experiments

We first decide the methods and parameter settings by training and testing on the Institut National de Recherche en Informatique et Automatique (INRIA) pedestrian dataset.[3] Then the final detectors are trained and compared with other methods on various datasets. In detection, we use mean shift mode estimation[29] for merging multiple detections of the same pedestrian. The detection results are evaluated with false positive per-image (FPPI) measure[30] (Pascal overlap threshold is 0.5). The miss rate at reference point 0.1 FPPI is given for comparison, the FPPI evaluation tool and other detector's results from the public benchmark.[30]

5.1 *Method and Parameter Settings*

Excepting those being studied, the parameters are kept. By default, 10 GC channels are employed including one gradient magnitude channel and nine gradient orientation channels; the adjacent rate $r_a = 2$, feature random sampling rate $r_s = 0.02$, type-1 and type-2 feature number $S_1 = 1000$,

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

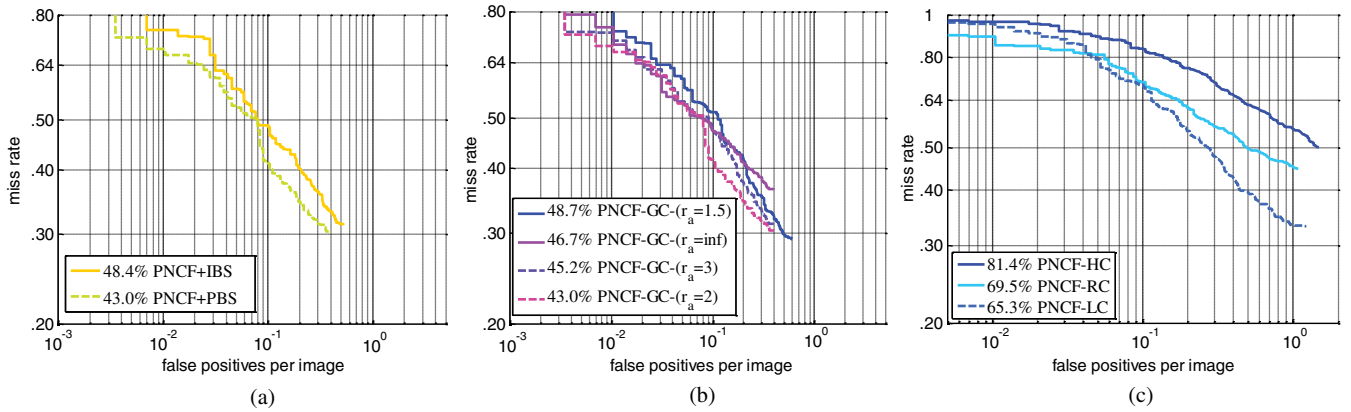


Fig. 4 Some method and parameter setting comparisons of our detector.

$S_2 = 40$; PNCF generation is integrated with boosting in PBS method.

First, the two feature selection methods PBS and IBS are compared, as illustrated in Fig. 4(a). It shows PBS feature selection has 5.4% improvement in detection accuracy over IBS, so it is better to calculate the statistics without sample weights in the boosting round. We infer that as the sample weights are highly biased to the hard samples in the boosting training, the generated PNCF features with IBS are biased and become less discriminative. Also, the advantage in performance, the training time of PBS method (140 min) is only half of IBS (260 min). In Fig. 4(b), performance curves with different normalization adjacent rates r_a are illustrated. The best result is achieved by $r_a = 2$, which is consistent with HOG feature's 2×2 cell/block structure.[3] Smaller value ($r_a = 1.5$) has less PNCF candidate features, and bigger values ($r_a = 3$, $r_a = \inf$) are sensitive to nonuniform illumination, so their performances drop by 2.2% to 5.7%. Figure 4(c) shows the performances of PNCF features extracted on color channels with different color spaces, including RGB (RC), HSV (HC) and LUV (LC). LUV channels gain the best performance, which is consistent with Ref. 11.

In Fig. 5, HOGBin-enriched, PNCF-GC and PNCF-GCLC are compared. HOGBin-enriched is an enriched version of HOG bin feature,[8,9] by containing $1*1$, $1*2$, $2*1$ and $2*2$ HOG feature's cell/block structures, as well as multiple aspect ratios for the cells or blocks. The feature is similar to PNCF-GC, but the numerator and denominator are fixed in the cell/block structure. Also, the denominator must lie in the gradient magnitude channel. Therefore, it is far less versatile than PNCF-GC, and the traditional exhaustive search method is sufficient for learning. The training is 240 minutes, slower than PNCF-GC's 140 minutes. PNCF-GCLC uses additional LC channels for combining gradient and color information. From the results, PNCF-GCLC is the best with 4.9% improvement over PNCF-GC, which is better than HOGBin-enriched with 3.8%. Apart from the detection accuracy, the feature numbers are listed in Table 2. Having fewer features makes the feature evaluation time drop and allows for fast detection. Compared with HOGBin-enriched, PNCF-GCLC and PNCF-GC have fewer features, especially PNCF-GCLC. ChnFtrs[11] has more feature numbers, though it is similar to PNCF-GC by integrating gradient and color. The above results illustrate the discriminative power of PNCF feature and the effectiveness of the statistics-based learning method. Thus, 47% of PNCF-GCLC features are LC features, which proves the importance of color information. The percentage of the cross channel feature is 73%, so the utilization of intra-channel information in PNCF is also justified.

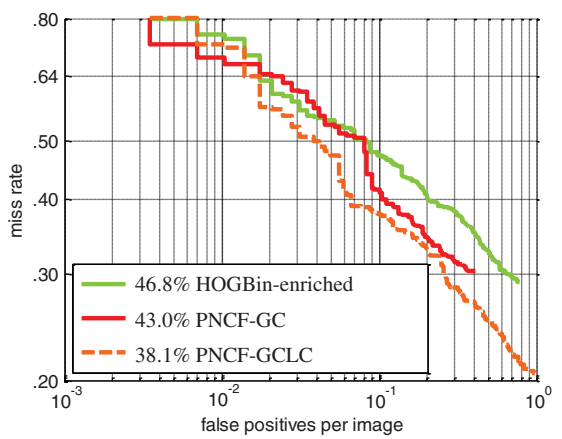


Fig. 5 FPPI evaluations of three different methods.

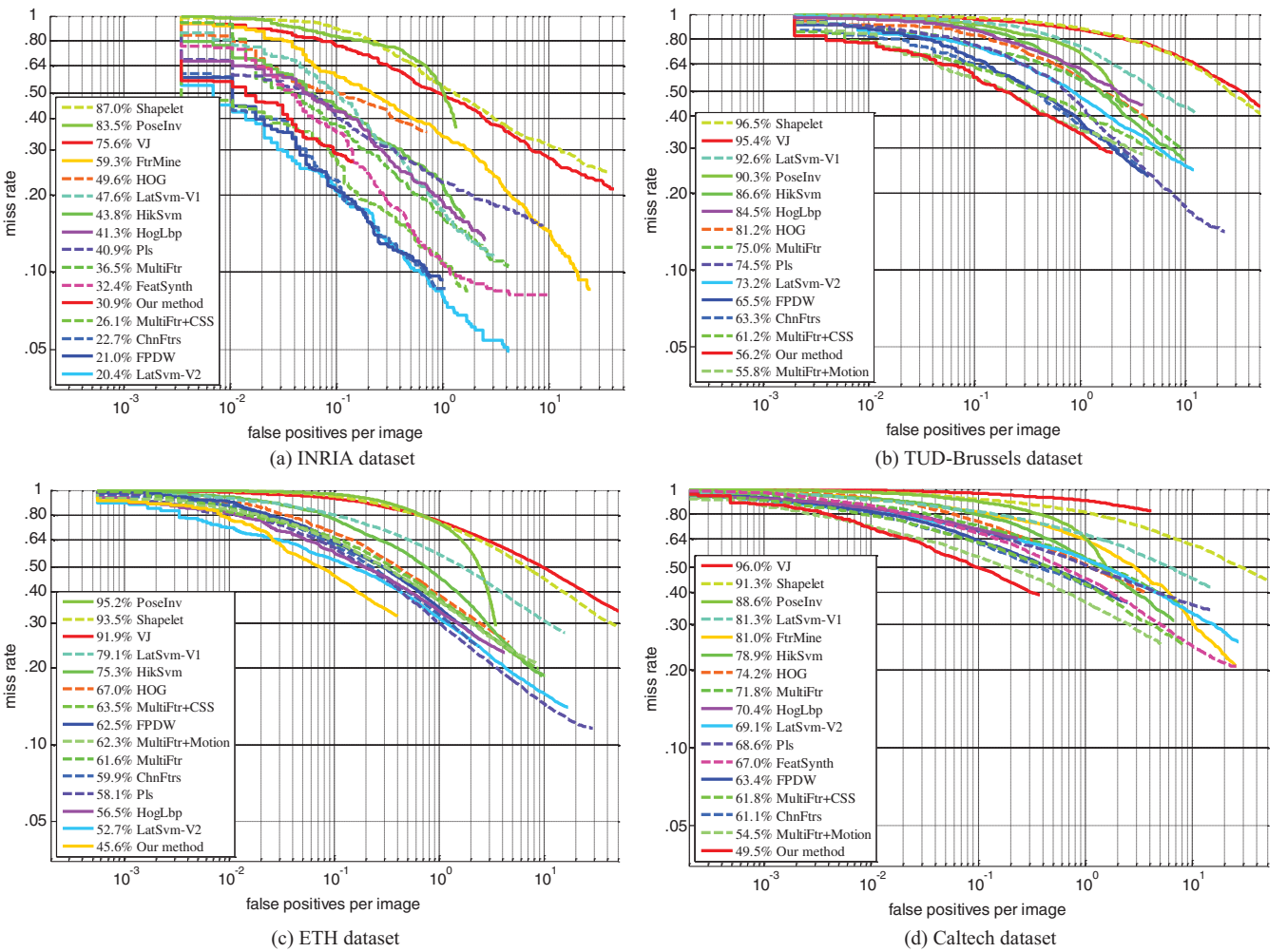
5.2 *Comparison with State-of-the-Art Methods*

We compare our method with all the available methods in Ref. 30 in the reasonable setting,[30] which evaluates pedestrians with at least 50 pixels in height under no or partial occlusion. We use both GC and LC channels together for training the detector. We train a detector pyramid on a

Table 2 Feature numbers of three different methods.

Method	# Features
ChnFtrs[11]	2000
HOGBin-enriched	1307
PNCF-GC	681
PNCF-GCLC	436

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use



(a) INRIA dataset (b) TUD-Brussels dataset

(c) ETH dataset (d) Caltech dataset

Fig. 6 FPPI evaluations of our detectors in the reasonable setting, with the other detectors' data being provided in the benchmark.[30]

much bigger dataset with 10,000 pedestrians, including INRIA training set and samples we collected.

Apart from INRIA dataset, the trained detector is evaluated on three additional benchmark datasets, TUD-Brussels,[31] ETH[32] and Caltech.[30] The three datasets are captured on a moving platform in the streets. TUD-Brussels contains 508 frames with 1498 pedestrians. ETH contains three videos with 1804 frames and 12,000 pedestrians. Caltech is the biggest dataset with precise annotation. It contains 347,000 pedestrians in total. While INRIA dataset has big pedestrians with high resolution, these datasets have a large portion of small pedestrians range from 50 to 100 pixels in height, which is a frequent resolution for many surveillance or driver assistance applications.

Figure 6 gives the comparison results on the four datasets. Among the state-of-the-art methods, ChnFtrs[11] is similar to our method by utilizing features in different image channels, but its features have no careful normalization. FPDW[24] approximates ChnFtrs to gain significant speedup. LatSvm-v2[33] is a well-known part-based method. MultiFtr+Motion[7] uses additional motion feature for better accuracy, but it is time-consuming. Overall the results show our method gains better accuracy than the state-of-the-art methods, especially on TUD-Brussels, ETH and Caltech whose pedestrians has adequate height. On INRIA, ChnFtrs and FPDW utilize carefully tuned post-processing parameters (the result without tuning is presented in Ref. 11 and is inferior to our method). LatSvm-v2 is more suitable for big pedestrians, so it is not so good on TUD-Brussels and Caltech. Our method is comparable to MultiFtr + Motion on TUD-Brussels, but outperforms it in all the other datasets, especially on ETH. On ETH and the much bigger Caltech dataset, our method is better than the runner-up by 7.1% and 5.0%, respectively.

Figure 7 gives the accuracy versus running speed of our detectors on Caltech dataset. We use the same time normalization as Ref. 30 We perform two kinds of multi-scale detection schemes including image pyramid (IPy) and HPy to illustrate the attained speed. In Fig. 7, our IPy and HPy generate the best performance in accuracy and efficiency. Though IPy does not apply fast multi-scale detection, it is still faster than other methods including FPDW due to the reduced feature number. As expected, HPy is better than IPy in accuracy and efficiency with three times the speed. The efficiency improvement is better than visual search[21] or coarse-to-fine search,[22] which has only twice the speed. If we further integrate these quicker, compatible techniques, our method can be even faster.

Figure 8 depicts some detection results on the datasets given by our detector.

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

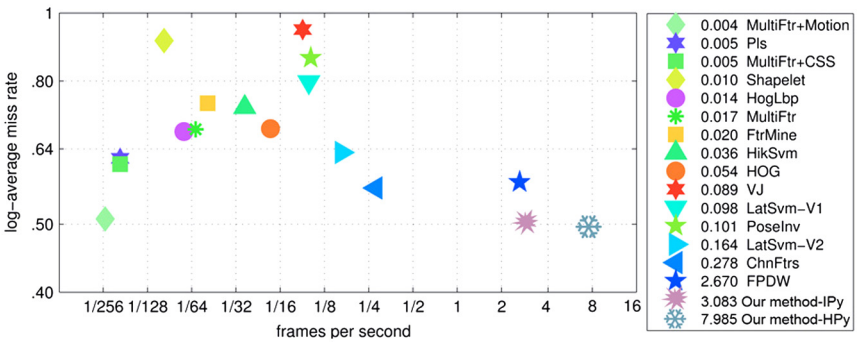


Fig. 7 Accuracy versus running speed for pedestrians over 50 pixels. Other detectors' data is provided in the benchmark.[30]

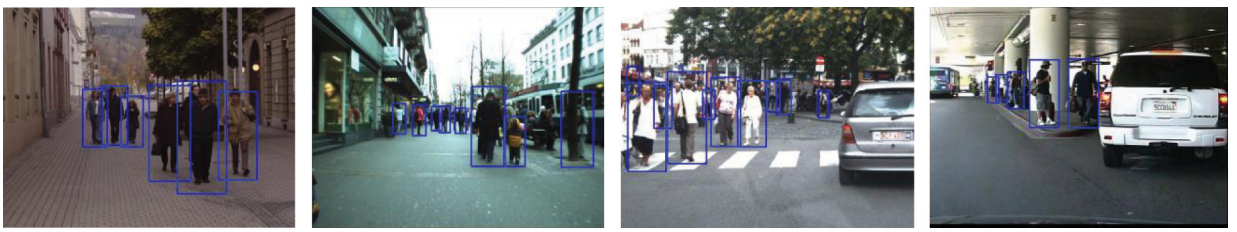


Fig. 8 Some detection results of our detector.

6 Conclusion

In this paper, we have proposed a novel PNCF feature type and described a statistics-based PNCF feature learning method. We employ channel compression and HPy technique to increase the multi-scale detection. Experiments demonstrate the proposed PNCF feature has high discriminative power and low computational cost. Extensive comparisons with other methods show that our detector achieves state-of-the-art accuracy and a much faster running speed. Although the PNCF feature and the feature learning are proposed for pedestrian detection, they can also be used in other object detection. In future, we will extend the PNCF detector to multiple views and parts-based pedestrian detection for further accuracy improvement and occlusion handling.

Acknowledgments

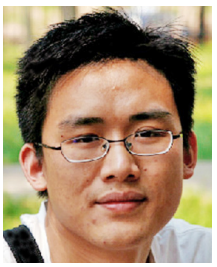
This work is partially supported by the National Natural Science Foundation of China (No. 61132007). We want to thank Piotr Dollár, Christian Wojek, Bastian Leibe, and Pietro Perona for publicly providing the useful evaluation benchmark. Also, further we thank Piotr Dollár for his nice, helpful discussions.

References

1. M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(12), 2179–2195 (2009).
2. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Conference on Computer Vision and Pattern Recognition*, IEEE, Kauai, HI (2001).
3. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, Vol. **1**, pp. 886–893, IEEE (2005).
4. B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.* **75**(2), 247–266 (2007).
5. O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on riemannian manifolds," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, pp. 1–8, IEEE (2007).
6. X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *IEEE International Conf. on Computer Vision*, Kyoto, Japan, pp. 32–39, IEEE (2009).
7. S. Walk et al., "New features and insights for pedestrian detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, pp. 1030–1037, IEEE (2010).
8. C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," in *DAGM-Symposium*, Munich, Germany, pp. 82–91, Springer (2008).
9. Z. Lin, G. Hua, and L. Davis, "Multiple instance ffeature for robust part-based object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, pp. 405–412, IEEE (2009).
10. W. Gao, H. Z. Ai, and S. H. Lao, "Adaptive contour features in oriented granular space for human detection and segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, Vol. **1–4**, pp. 1786–1793, IEEE (2009).
11. P. Dollár et al., "Integral channel features," *BMVC 2009*, British Machine Vision Association, London, England (2009).
12. Y. Yamauchi et al., "Feature co-occurrence representation based on boosting for object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, pp. 31–38, IEEE (2010).
13. J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Ann. Statist.* **28**(2), 337–374 (2000).
14. K. Levi and Y. Weiss, "Learning object detection from a small number of examples: the importance of good features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, DC, Vol. **2**, pp. 53–60, IEEE (2004).
15. H. Chang and R. Nevatia, "High performance object detection by collaborative learning of joint ranking of granules features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 41–48 (2004).
16. C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: a branch and bound framework for object localization," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(12), 2129–2142 (2009).
17. J. Yao and J.-M. Odobez, "Fast human detection from videos using covariance features," in *The Eighth International Workshop on Visual Surveillance—VS2008*, Marseille, France, IEEE (2008).
18. D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *IEEE Conf. on Computer Vision and Pattern Recognition*, New York, NY, Vol. **2**, pp. 2137–2144, IEEE (2006).
19. G. Gualdi, A. Prati, and R. Cucchiara, "Covariance descriptors on moving regions for human detection in very complex outdoor scenes," in *ACM/IEEE International Conf. on Distributed Smart Cameras*, Como, Italy, pp. 1–8, IEEE/ACM (2009).
20. A. Ess et al., "Robust multiperson tracking from a mobile platform," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(10), 1831–1846 (2009).
21. N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, pp. 2751–2758, IEEE (2009).

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

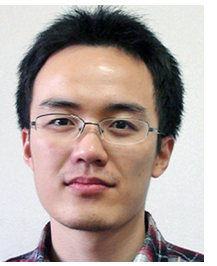
22. G. Gualdi, A. Prati, and R. Cucchiara, "Multi-stage particle windows for fast and accurate object detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(8), 1589–1604 (2012).
23. M. Pedersoli et al., "Recursive coarse-to-fine localization for fast object detection," *Lecture Notes in Computer Science*, pp. 280–293, Springer, Berlin/Heidelberg (2010).
24. P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west,"*British Machine Vision Conference (BMVC)*, Aberystwyth , Wales, UK, pp. 68.1–68.11, British Machine Vision Association (2010).
25. M. T. Pham and T. J. Cham, "Fast training and selection of haar features using statistics in boosting-based face detection," in *IEEE International Conf. on Computer Vision*, Rio de Janeiro, Brazil, Vol. **1–6**, pp. 1634–1640, IEEE (2007).
26. R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley-Interscience Publication, Singapore (2001).
27. D. V. Hinkley, "On the ratio of two correlated normal random variables," *Biometrika* **56**(3), 635–639 (1969).
28. L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.* **33**(1), 1–39 (2010).
29. N. Dalal, "Finding people in images and videos," Ph.D. Thesis, Institut National Polytechnique de Grenoble (July 2006).
30. P. Dollar et al., "Pedestrian detection: an evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(4), 743–761 (2012).
31. C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, FL, Vol. **1–4**, pp. 794–801, IEEE (2009).
32. A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *IEEE International Conf. on Computer Vision*, Rio de Janeiro, Brazil, **1–6**, 2065–2072, IEEE (2007).
33. P. Felzenszwalb et al., "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1627–1645 (2010).



Bobo Zeng is a PhD student in the Department of Electronics Engineering, Tsinghua University, China. He received his BS degree in the Department of Electronics and information Engineering, Huazhong University of Science & Technology, Wuhan in 2007. His research interests include human detection and tracking, surveillance video analysis, etc.



Guijin Wang received the BS and PhD degree in signal and information processing (with honors) from the department of Electronics Engineering, Tsinghua University, China, in 1998 and 2003, respectively. From 2003 to 2006, he was with Sony Information Technologies Laboratories as a researcher. Since 2006, he has been with the department of Electronics Engineering at, Tsinghua University, China, as an associate professor. He has published over 50 International journal and conference papers and holds several patents. He is the session chair of IEEE CCNC'06. His research interests are focused on wireless multimedia, image and video processing, depth imaging, pose recognition, intelligent surveillance, industry inspection, object detection and tracking and online learning.



Zhiwei Ruan is a PhD student in the Department of Electronics Engeering, Tsinghua University, China. He received his BS in the Department of Electronics Engineering, Tsinghua University, China in 2009. His research interests are focused on object tracking and detection and machine learning.



Xinggang Lin received his BS in electronics engineering, Tsinghua University, China in 1970; an MS in 1986 and a PhD in 1982, both in information science from Kyoto University, Japan. He joined the Department of Electronics Engineering at Tsinghua University in 1986 where he has been a full professor since 1990. He received a "Great Contribution Award" from the Ministry of Science and Technology of China, and "Promotion Awards of Science and Technology" from Beijing Municipality. He was a General co-chair of the second IEEE Pacific-Rim Conference on Multimedia, an associate editor of IEEE T. on CSVT, and a technical/organizing committee member of many international conferences. He is a fellow of the China Institute of Communications, and has published over 140 referred conference and journal papers in diversified research fields.



Long Meng was born in 1979. She received a BS from the Department of Telecommunication Engineering, Beijing Posts and Telecommunications University, China, in 2001 and an MS in signal and information processing from the Department of Electronics Engineering, Tsinghua University, China, in 2004. From 2004 to 2006 she was with Sony Company as a developer for digital cameras in Tokyo, Japan. From 2007 to 2011, she has been with Sony China Research Lab as a researcher. In 2011, she joined a startup company named Beijing Mingrui Automatic Technology Company as vice president. Her research interests are focused on image and video processing, intelligent surveillance, industry inspection, object detection and tracking and online learning.

Downloaded From: https://www.spiedigitallibrary.org/journals/Optical-Engineering on 9/29/2018
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use