Implicit Color Segmentation Features for Pedestrian and Object Detection

Patrick Ott and Mark Everingham
School of Computing
University of Leeds
{ott|me}@comp.leeds.ac.uk

Abstract

We investigate the problem of pedestrian detection in still images. Sliding window classifiers, notably using the Histogram-of-Gradient (HOG) features proposed by Dalal and Triggs are the state-of-the-art for this task, and we base our method on this approach. We propose a novel feature extraction scheme which computes implicit 'soft segmentations' of image regions into foreground/background. The method yields stronger object/background edges than gray-scale gradient alone, suppresses textural and shading variations, and captures local coherence of object appearance. The main contributions of our work are: (i) incorporation of segmentation cues into object detection; (ii) integration with classifier learning cf. a post-processing filter; (iii) high computational efficiency.

We report results on the INRIA person detection dataset, achieving state-of-the-art results considerably exceeding those of the original HOG detector. Preliminary results for generic object detection on the PASCAL VOC2006 dataset also show substantial improvements in accuracy.

1. Introduction

Detection and localization of pedestrians in images has attracted much recent attention [7, 12, 9, 4, 18, 16, 15, 6, 11] not least because of important applications in image understanding and autonomous vehicles. One of the most successful recent methods is that proposed by Dalal & Triggs [4], which combines an image descriptor capturing local gradient orientation with a linear support vector machine (SVM) classifier in a sliding window framework. We extend the method of Dalal & Triggs by proposing a novel feature extraction scheme which incorporates segmentation cues in the feature descriptor. The main idea is to compute *window-specific* features which adapt to local image characteristics – color of foreground (pedestrian) and background. This differs from conventional feature extraction methods *e.g*. Histogram of Oriented Gradients (HOG) [4] which compute fixed descriptors independent of the local image statistics. Using local estimates of foreground and background statistics, a 'soft segmentation' of an image region is computed. Gradients in this segmentation image amplify edges between object and background, while reducing clutter from variation in texture and shading. The soft segmentation is obtained in a computationally efficient manner by local linear projections of the original image, such that the approach is suitable for use in a sliding window scheme. While exploiting color information, the proposed method does not assume any *class*-level color model, so it is applicable to object categories *e.g*. people/vehicles which vary greatly in color. Compared to other methods which have exploited color-based segmentation to aid detection [14, 13], our method incorporates segmentation cues directly in the feature extraction and learning process, rather than as a separate post-processing step.

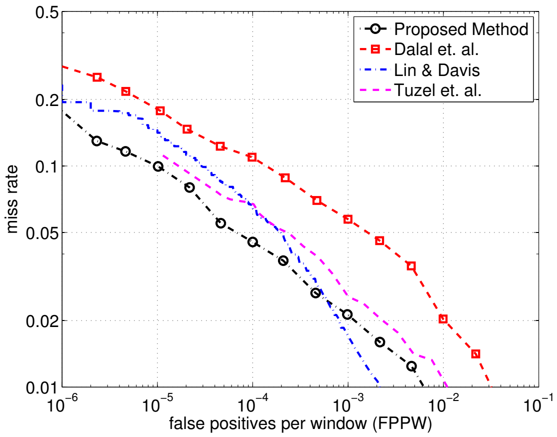


Figure 1. Comparison of the proposed detector to other state-of-the-art methods. DET curves are shown for our proposed method, HOG [4], the method of Lin *et al*. [10] and Tuzel *et al*. [16].

Related work. Most, and indeed the most successful, approaches to pedestrian detection can be considered sliding window classifiers or template matchers. A classifier is applied to every window of an image over multiple scales, yielding a positive (pedestrian) or negative (non-pedestrian)

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

output for each window. Early work by Papageorgiou and Poggio [12] used a wavelet representation of a window and a support vector machine classifier. Jones *et al*. [9] applied AdaBoost learning and Haar-like features, adding features computed over several frames to capture dynamic cues. Gavrila's pioneering work on chamfer matching [7] can also be viewed as sliding window classification, using a threshold on distance to a hierarchy of edge templates as the classifier.

The Histogram of Oriented Gradients (HOG) feature descriptor proposed by Dalal & Triggs [4] has proven particularly successful, achieving perfect performance on the MIT database used in earlier work [12], and this has sparked new interest in the pedestrian detection problem. We review the method fully in Section 2 since we build upon it. The computational efficiency of the detector has been improved by applying the cascade architecture [18] popularized in vision by Viola & Jones [17]. Tuzel *et al*. [16] have reported improved results using a method which describes an image window in terms of the covariance of features within the window. Lin & Davis [10] have recently proposed a scheme which extracts instance-specific features based on edge template matching. Okada & Soatto [11] proposed to divide the pedestrian class into disjoint sub-classes by pose, and train individual classifiers for each. They report modest improvements on the detection problem, and apply the method to improve regression-based pose estimation. Work by Tran & Forsyth [15] incorporates detailed 2D pose estimation in the detection process, estimating pose for every image window (positive or negative) to reach a classification decision. McAllester *et al*. [6] also estimate the position of parts, but use semi-automatically learnt parts rather than manually defined limbs. This method gives very promising results on the challenging PASCAL VOC database [5].

Motivation. The proposed method is inspired by two pieces of work which incorporate segmentation cues for object detection and tracking.

Ramanan [13] proposed a scheme for 'verification' of detections from a sliding window detector by segmenting the window and verifying if the segmentation resembles the class of interest *e.g*. a pedestrian. A window is segmented using a weak prior on object shape learnt from unsegmented training images, representing foreground/background color distributions using a color histogram, and applying a graph-cut segmentation method [2]. A linear classifier is applied to the resulting binary segmentation mask to verify the hypothesized detections. The method works as a 'post-processing' filter on the detections – training of the sliding window detector is performed independently of the segmentation process. We use the idea of incorporating instance-specific appearance into the detection process, but integrate this into the classifier learning.

Collins *et al*. [3] propose a method for extracting discriminative features for tracking which are adapted to the surroundings of the object to be tracked in each frame. The approach is particularly elegant in its simplicity: given the bounding box of the object and a larger bounding box capturing the background appearance in its immediate neighborhood, the method picks between a set of pre-defined linear transformations of the RGB color space, choosing the transformation which maximizes discrimination between the object and background region in the current frame. This transformation is then applied to pixels in the next frame (assuming some coherence across frames), obtaining better foreground/background discrimination with which to drive the tracker. We exploit this idea of using different transformations of the color space to 'pull out' stronger foreground/background features, applying the approach to image windows.

Outline. Section 2 reviews the HOG descriptor on which we base our method. Section 3 describes the proposed approach. We report experiment results in Section 4, and offer conclusions and directions for future work in Section 5.

2. Histogram of Oriented Gradients approach

We first briefly review the HOG descriptor and detection scheme proposed by Dalal & Triggs [4] since our method builds on these.

HOG descriptor. At each pixel within the detection window the image gradient is computed, and accumulated into bins over (i) orientation, and (ii) spatial regions ('cells'). Within each cell, a histogram of gradient orientation is computed. Although many schemes for dividing the image into cells *e.g*. log-polar schemes have been investigated [4], using simple square cells is shown to be effective and is computationally efficient. The intuition to binning (quantizing) orientation and spatial position is to introduce some invariance to local image deformation.

A second stage of spatial accumulation groups contiguous ranges of cells into 'blocks'. The descriptor for each block is then independently normalized to have constant norm. The blocks typically overlap by one or more cells such that each cell is represented multiple times (with different normalizations) in the final descriptor formed by concatenating all blocks. This gives some contrast invariance over a larger scale than that of the individual cells.

Dalal & Triggs [4] investigated various schemes of cell size and shape, spacing, block arrangement, orientation binning and normalization. For the task of pedestrian detection in the INRIA dataset [4], image windows are 128×64 pixels, a reasonable scheme is to use blocks of 2×2 cells overlapping by one cell in each direction, square cells of width

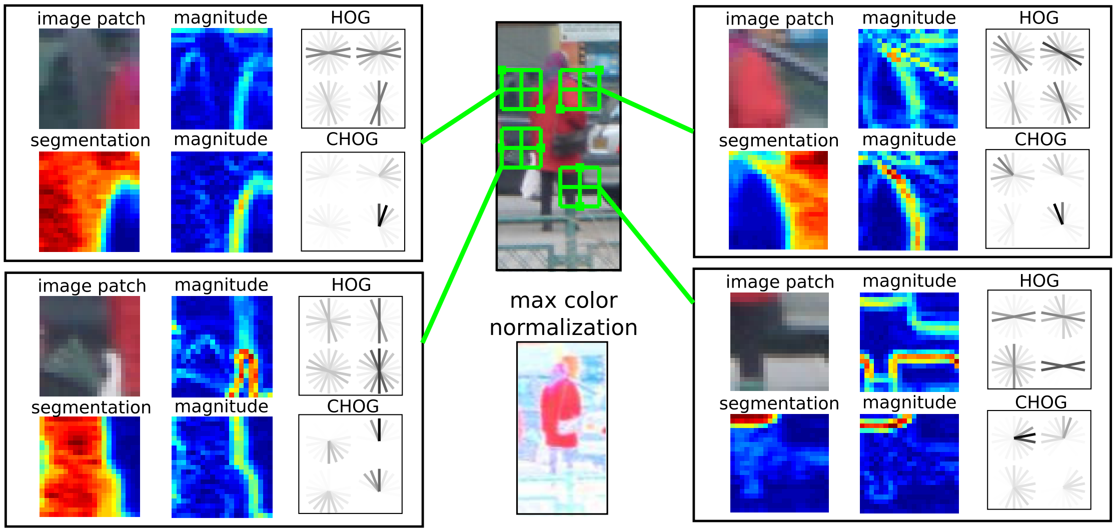


Figure 2. Overview of feature extraction. For each block of the window (center top) a HOG descriptor and CHOG descriptor is computed (insets, right), each consisting of a histogram over orientation and spatial position. CHOG descriptors are computed on a soft segmentation of pixels in the block into foreground and background (insets, bottom left). For each block, hypothesized foreground and background samples are taken at reference points relative to the block (center top). Color normalization (center bottom) is used to enhance color edges in the CHOG descriptor complementary to the intensity edges in the HOG descriptor. Note that the CHOG features emphasize the edges of the pedestrian while attenuating shading on the clothing and clutter edges in the background.

6–8 pixels, 9 orientation bins (discarding gradient sign) and L2 normalization with a constant term added to avoid amplifying noise in regions with very low variance. For more detailed description of the HOG features we refer the reader to [4].

Gradient computation. The HOG scheme can be applied to gray-scale images since it makes use of the image gradient alone. However, when applying the descriptor to color images, Dalal & Triggs' implementation computes the gradient at a pixel in each of the red, green and blue channels, and selects the response with greatest magnitude. This can potentially increase the sensitivity to edges in color images which correspond to changes in color with no associated change in luminance. However, because the decision is made independently at every pixel it can also have the effect of increasing noise, and does not enforce any 'coherence' in the edges. Additionally, because the unsigned gradient is typically used, the method is agnostic to dark/light *vs.* light/dark transitions, and again cannot capture the likelihood that nearby edges are likely to be of the same sign. Our proposed method (Section 3) addresses some of these limitations.

Classification. Dalal & Triggs [4] paired the HOG descriptor with an SVM classifier in the sliding window framework and investigated the use of nonlinear and linear kernels. Competitive results were obtained using a linear SVM which is appealing because of its computational efficiency, and we adopt the same classifier. However, as reported in Section 4, we found that results of both HOG and our proposed method could be improved by use of a quadratic kernel.

3. Method

This section describes the proposed feature extraction scheme. The essence of the scheme is to extract a descriptor which captures the shape of the foreground object (if any) in a given window. There are two steps: (i) soft segmentation into foreground/background; (ii) describing edges in the soft segmentation using a HOG descriptor. Figure 2 gives an overview of the method. We refer in the following to our features as 'CHOG' to emphasize the extension of the HOG descriptor with color information.

3.1. Soft segmentation by Fisher discriminant

Given an image region our aim is to segment pixels in the region into foreground and background. We reason that such a segmentation *i.e.* the object silhouette gives strong cues to recognition [13].

Let us assume that we are given a sample of RGB pixel values $\mathbf{x}_i$ where $i \in \mathcal{C}_1$ denotes background pixels and $i \in \mathcal{C}_2$ denotes foreground pixels. Samples are taken from a particular image window, so the foreground distribution captures instance-specific properties (this person is wearing blue trousers) rather than class-level properties (sheep are white). We seek a linear projection of the pixel values

$y = \mathbf{w}^T\mathbf{x}$ which maximizes the separation between foreground and background.

Assuming that the distributions of foreground and background pixels are Gaussian, the Fisher criterion (see [1]) gives a suitable means for choosing $\mathbf{w}$:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathtt{S}_B \mathbf{w}}{\mathbf{w}^T \mathtt{S}_W \mathbf{w}} \tag{1}$$

where $\mathtt{S}_B$ is the *between-class* covariance matrix, given by

$$\mathtt{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \tag{2}$$

where $\mathbf{m}_1$ and $\mathbf{m}_2$ are the sample means of the background and foreground classes respectively, and $\mathtt{S}_W$ is the total *within-class* covariance matrix, given by

$$\mathtt{S}_W = \sum_{i \in \mathcal{C}_1} (\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{i \in \mathcal{C}_2} (\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)^T \tag{3}$$

The value $\hat{\mathbf{w}}$ maximizing $J(\mathbf{w})$ can readily be shown to be:

$$\hat{\mathbf{w}} \propto \mathtt{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1) \tag{4}$$

Projecting the original image pixels by $y = \hat{\mathbf{w}}^T\mathbf{x}$ gives an image which can be considered a 'soft segmentation' – large values indicate high probability of foreground, and small values high probability of background. Gradients in this image give strong evidence for edges between foreground and background.

Simplified model. Eqn. 4 could directly be applied to obtain soft segmentations, but is somewhat computationally expensive because of the need to compute and invert the covariance matrix $\mathtt{S}_W$ for each window. If we assume that the covariance of both foreground and background distributions is isotropic *i.e.* $\mathtt{S}_W$ proportional to the identity matrix, then the optimal $\mathbf{w}$ simplifies to the difference between the class means:

$$\hat{\mathbf{w}} \propto (\mathbf{m}_2 - \mathbf{m}_1) \tag{5}$$

In the following we use this simplified model, and justify its use in Section 3.2.

Projection of gradients. Given a soft segmentation, the image is described by the distribution of gradient position and orientation, in the same manner as the HOG descriptor. This requires computation of gradient orientation and magnitude at each pixel. Since computation of the gradient is a linear operation, the gradient of the segmentation image can simply be computed by linear transformation of the original RGB gradients. Denoting the original image I and the projected image S:

$$\frac{\partial}{\partial x} S = \mathbf{w}^T \left\langle \frac{\partial}{\partial x} I_R, \frac{\partial}{\partial x} I_G, \frac{\partial}{\partial x} I_B \right\rangle^T \tag{6}$$

and similarly for the gradient in the y-direction.

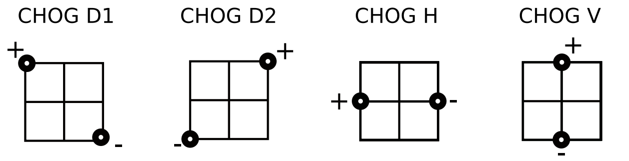


Figure 3. Potential reference points for a single block. The 'foreground' color distribution is estimated from pixels in the neighborhood of the +ve point, and 'background' from the -ve point. The resulting projection is used for all pixels within the block.

The relevance of this identity is that the gradient of the image under any projection can be computed without needing to apply the convolution operator to the projected image. Since, as described below, a large number of projections are used for every image window, this enables the feature extraction to be performed in a computationally efficient manner. We refer to the CHOG features as using 'implicit' segmentation since the gradients of the soft segmentation used to construct the descriptor are computed without explicitly computing the segmentation.

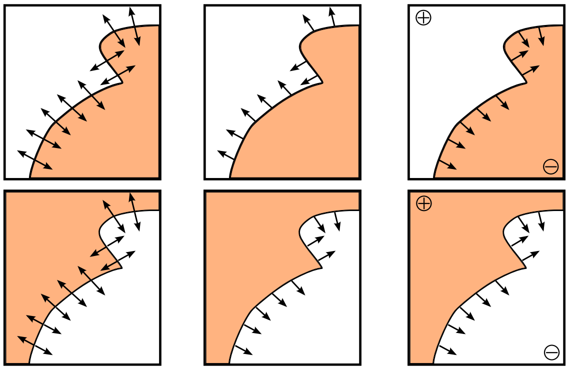
3.2. Semi-local segmentation

Thus far we have assumed that we are given a set of sample pixels labeled as foreground or background. Such samples could be obtained by using an average person mask [13] to be applied to each window. However, for negative windows this can result in 'hallucinating' pedestrian regions. Additionally, the Gaussian assumption made in the Fisher discriminant will be a poor model when there is significant variation in color within *e.g.* the background – Ramanan [13] uses a color histogram, which prevents applying the method to every window because of computational expense.

We therefore use a number of *semi-local* projections rather than a single global sample of foreground and background. For all HOG blocks, a set of reference pairs of image points chosen at training time are defined as positive (potential foreground) and negative (potential background). For a given window the means of the 'foreground' and 'background' distributions are estimated from pixels in the neighborhood of these reference points. Since a HOG scheme is used to describe the soft segmentations, we base the selection of reference points on HOG blocks. Figure 3 shows the selected potential reference points for a block.

For each block, the discriminant projection is computed using Eqn. 5 and the gradients of the soft segmentation within the block are computed using Eqn. 6. These gradients are represented in the descriptor by a conventional HOG block *i.e.* histograms over orientation in the 2 × 2 cells of the block.

Using semi-local projections in this way has two advantages over attempting a global segmentation of the window:



(a) HOG unsigned (b) HOG signed (c) CHOG (signed)

Figure 4. Additional modeling capacity of CHOG features. (a) Unsigned HOG features give invariance to light object/dark background but cannot represent coherence of edge direction since gradient sign is discarded; (b) Signed HOG features distinguish between both cases but given unknown brightness relationship between object and background the (linear) classifier must assign equal weight to both gradient signs; (c) Signed CHOG gradient consistently captures background-to-foreground transitions, allowing coherence of gradients to be learnt by weights for a single gradient sign.

(i) the Gaussian assumption is likely to be well-satisfied given the local color measurements, enabling use of the simple linear projection scheme for implicit segmentation and gradient computation; (ii) by choosing projections at training time the method can also capture selected salient internal gradients, *e.g*. between shirt and trousers, which is lost in the object silhouette. Section 3.5 offers further discussion.

3.3. Color normalization

When using RGB pixels to form the soft segmentation, a significant component of the discriminant projections (Eqn. 5) is difference in intensity. We found that results were improved by using intensity-invariant features, which give complementary cues to the intensity gradients used by the HOG descriptor. Prior to computation of projections and gradients, the RGB image pixels $\mathbf{x} = \langle x_r, x_g, x_b \rangle$ are normalized thus:

$$\mathbf{x}' = \frac{\mathbf{x}}{\max(x_r, x_g, x_b) + \epsilon} \tag{7}$$

where ϵ is a small constant, avoiding unstable results for pixels of low intensity. Figure 5 shows the effect of color normalization on two example images.

3.4. Feature selection

The computational expense of computing the CHOG features is modest, and it is possible to compute features for all block-wise pairs of reference points (Figure 3) for every block. However, improvements in speed (and reduction in dimensionality) can be obtained by selecting a subset of salient reference points at training time. We found that a simple feature selection scheme gave considerable speedup with negligible loss of accuracy.

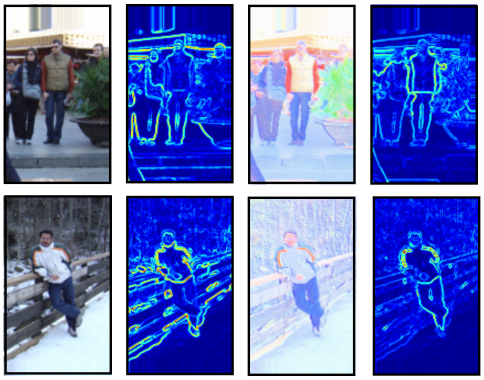


Figure 5. Color normalization. Columns from left to right show: (i) original image; (ii) gradient magnitude of original image; (iii) color normalized image; (iv) gradient magnitude of normalized image. Note how the color normalization is able to suppress background clutter due to intensity invariance and 'pull out' the different gradients of the pedestrian.

At training time, for each block the distance between the means of the color distributions around the reference points (Figure 3) is computed (Eqn. 4) for all four potential pairs of reference points. Then for each block we count the number of times a specific pair of reference points gives the largest difference in color space (compared to the other three reference points) over all positive training images. We keep those pairs of reference points in the feature set that give the largest difference in means for more than 25% of the positive images.

Empirically the selected projections are mainly diagonal D1 and D2 (see Figure 3) for the shoulder and lower-leg regions, horizontal for the sides of the torso, and vertical for the head and feet.

3.5. Discussion

Figure 2 shows example output of the feature extraction method for both HOG and CHOG. We show the conventional HOG descriptor using max-RGB gradient (see Section 2), and the CHOG descriptor using the gradient computed on the soft segmentation including max-norm color normalization (Section 3.3). Note, for example, how in the top left inset CHOG gives a strong response to the boundary of the coat, and suppresses responses to shading and background clutter. In the top right inset CHOG substantially attenuates the gradients introduced by background clutter and emphasizes the pedestrian's edge.

Figure 4 shows the modeling capacities of HOG and CHOG descriptors for the common case where the fore-

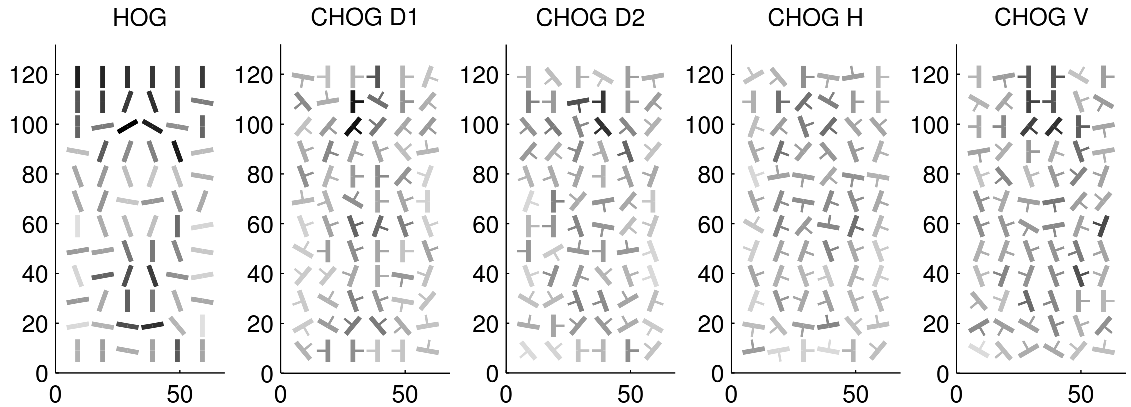


Figure 6. Visualization of the positive weights of the final SVM. For each cell the orientation of the bar shown denotes the orientation bin with maximum weight. For CHOG cells, which use signed gradient, orthogonal lines indicate the sign (foreground/background *vs.* background/foreground) of the orientation bin. The labels correspond to the choice of projection reference points (see Figure 3).

ground object may be brighter or darker than the background. By using consistent projection of the image gradients in a foreground-to-background direction CHOG features combine the invariance of HOG to the bright/dark relationship between object and background with the ability to capture coherence of gradient sign given locally coherent background intensity. By use of consistent signed gradient CHOG is also able to distinguish between noisy edges in background regions and coherent gradients corresponding to true object/background edges.

4. Experimental results

We have principally evaluated our proposed method for pedestrian detection on the INRIA person dataset [4]. We report results here, comparing HOG, CHOG, and published results of other methods. Section 4.2 additionally reports preliminary results of applying the method to generic object detection on the PASCAL VOC2006 dataset.

4.1. Pedestrian detection on the INRIA dataset

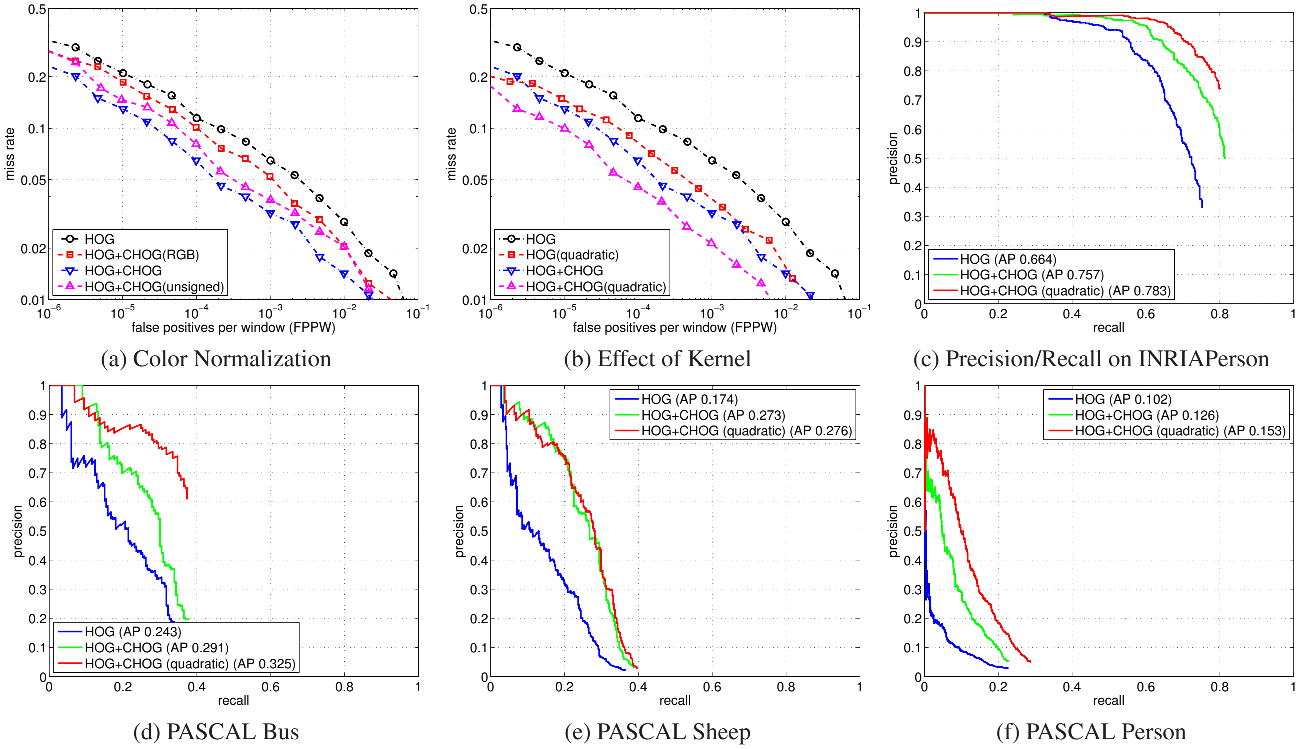
Training. We adopt the dataset and training and test protocol used by Dalal & Triggs [4]. The SVMLight package [8] was used for SVM training. As in previous work we perform several rounds of bootstrapping to collect false positive examples during training.

Testing. Results on per-window classification accuracy are reported using the conventional Detection Error Tradeoff (DET) curve [4]. The sliding window scheme in this case uses a scale factor of 1.2 between pyramid levels [4]. We also report a comparison using the PASCAL VOC methodology [5] – precision/recall curve with a bounding box overlap threshold of 50%, as adopted by Ramanan [13]. In this case, as in previous work, a scale factor of 1.05 is used. In both cases a window step of 8 pixels was used [4].

Implementation details. A composite descriptor is formed by concatenating HOG and CHOG blocks (see Figure 2). Combining the two is effective since the gradient information from the original and intensity-invariant soft segmentation images is complementary. For HOG cells, 9 unsigned orientation bins were used. For CHOG cells, 18 signed orientation bins were used – as noted in Section 3.5 the gradient sign carries 'inside/outside' information for CHOG. CHOG features were selected using the method described in Section 3.4. For both descriptors we use a cell size of 6×6 pixels and a block size of 2×2 cells, with blocks overlapping by one cell in horizontal/vertical directions [4]. For CHOG the neighborhood around each reference point (Figure 3) is a 3×3 square – the means (Eqn. 5) are efficiently computed using the integral image [17].

Learnt features. Figure 6 shows a visualization of the positive SVM weights for a complete set of HOG+CHOG features (no feature selection). For each cell the orientation with maximum weight is shown. It can be seen that the HOG channel models the complete outline of a pedestrian while the individual CHOG channels assign high weight to the parts that are emphasized by the corresponding block-specific projection. For CHOG D1 and D2 (diagonal reference points – see Figure 3) these are the shoulder regions. CHOG H (horizontal reference points) assigns high weights to the side regions of the pedestrian. Note that the maximally-weighted orientations for the horizontal projections have the same sign – since the reference points for a block are fixed this indicates learning of consistent inside-to-outside gradient.

Comparison of descriptors. Figure 7 (a) shows a comparison of our proposed method (HOG+CHOG) to the original HOG method (HOG). It is conventional to report results



(a) Color Normalization

(b) Effect of Kernel

(c) Precision/Recall on INRIAPerson

(d) PASCAL Bus

(e) PASCAL Sheep

(f) PASCAL Person

Figure 7. (a) Comparison of descriptors. DET curves are shown for HOG, HOG+CHOG without max color normalization (RGB), HOG+CHOG with unsigned gradients, and the proposed method HOG+CHOG. Curves nearer to the southwest corner are better; (b) Comparison of linear/kernel SVM classifiers. DET curves are shown for HOG and HOG+CHOG descriptors with linear and quadratic kernels; (c) Comparison of HOG and the proposed HOG+CHOG method using the PASCAL VOC methodology. Precision/recall curves are shown for the two methods. Curves nearer the northeast corner are better. (d)-(f) Precision/recall curves for three PASCAL VOC2006 classes, comparing HOG to HOG+CHOG (linear kernel) and HOG+CHOG using a quadratic kernel.

at a false positive per window (FPPW) rate of 10^{-4} [4]; at this FPPW rate our method decreases the miss rate from 0.115 to 0.065, a relative improvement of $\sim$44%. Figure 7(c) shows the corresponding comparison of the two methods using the PASCAL VOC methodology [5]. In this case our method increases the average precision (AP) from 0.664 to 0.757, a relative improvement of $\sim$14% (HOG+CHOG compared to HOG). As the curves show our proposed method increases both recall and precision, with no loss of precision at lower recall. Note that DET and precision/recall curves are not directly correlated since the latter requires suppression of multiple detections which are counted as false positives [5].

Figure 7(a) also shows the effect of using subsets of components in the proposed method. HOG+CHOG(RGB) shows the combination of the original HOG descriptor with CHOG descriptors using the un-normalized RGB image, giving an improvement of 12% at 10^{-4} FPPW. This establishes the effectiveness of the proposed color normalization (Section 3.3). As can be seen, the use of the normalized color gradients gives a substantial improvement compared to HOG+CHOG on the standard RGB image. HOG+CHOG(unsigned) shows results of the proposed method using *unsigned* gradients for the CHOG descriptor – this verifies the claim that the CHOG descriptor can exploit ‘inside/outside’ relations by use of signed gradients. As expected, performance is worse than the signed gradients (HOG+CHOG), in contrast to the original HOG descriptor, where comparable results for signed and unsigned gradients have been reported [4]. The use of signed gradients accounts for $\sim$30% of the overall reduction in miss rate at 10^{-4} FPPW (HOG+CHOG compared to HOG).

Comparison of classifiers. We also compared the performance of the proposed descriptor using both linear and kernel SVM classifiers. Dalal & Triggs [4] reported very modest improvements in accuracy using a radial basis function (RBF) kernel. We ran comparisons using an inhomogeneous quadratic kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^2$. The quadratic kernel is a natural choice here since it can represent dependencies between pairs of features; this is particularly relevant for the CHOG descriptor where gradient sign is meaningful.

Figure 7(b) shows results for HOG and HOG+CHOG with linear and quadratic kernel. Use of the quadratic kernel improved results for both descriptors – at 10^{-4} FPPW the miss rate decreased from 0.115 to 0.08 for HOG (relative improvement $\sim$31%) and from 0.065 to 0.045 for HOG+CHOG (relative improvement $\sim$31%). On the precision/recall curve using the PASCAL VOC methodology [5] (Figure 7(c)) using the kernel gives an additional 3.5% improvement in average precision. These are substantial improvements, suggesting it may be fruitful to re-consider using kernels with these descriptors if computational expense of the kernel evaluation can be reduced.

Comparison to state-of-the-art. The combination of HOG+CHOG descriptors with a quadratic kernel forms our final method. Figure 1 compares the results obtained with this method to the original HOG method and recent work by Lin *et al*. [10] and Tuzel *et al*. [16]. As shown, our method gives, to our knowledge, the best reported results on the INRIA dataset. Comparing at the standard FPPW rate of 10^{-4} our method reduces the miss rate by $\sim$30% compared to the next best method [16]. For FPPW rates less than around 10^{-3} our method gives consistently lower miss rate than all other results reported.

4.2. Object detection on PASCAL VOC2006

We also present results of initial experiments on generic object detection in the PASCAL VOC2006 dataset [5]. Parameters were unchanged from the INRIA tests, except a cell size of 8×8 pixels is used, and the window size is set to match the average object aspect ratio.

Figures 7(d)–(f) show precision/recall curves for three classes – 'bus', 'sheep' and 'person' – having significant differences in the level of intra-class appearance variation. For the 'bus' class (d) using HOG+CHOG increases Average Precision (AP) from 0.243 (HOG) to 0.291; additionally using a quadratic classifier further improves the results to AP of 0.325, a total relative improvement over HOG of $\sim$34%. For the 'sheep' class (e) the relative improvement using HOG+CHOG and a quadratic kernel is $\sim$57%, though in this case the linear kernel performs equivalently. For the 'person' class (f) the total relative improvement is $\sim$50% (AP of 0.153 *vs.* 0.102). These substantial improvements demonstrate the applicability of the proposed method to non-pedestrian classes.

5. Conclusions and future work

We have proposed a scheme for incorporating 'soft' segmentation cues directly into sliding-window object detection, obtaining substantial improvements in accuracy on pedestrian and object detection tasks. In future work we intend to investigate further the incorporation of segmentation cues in the form of instance-specific models of appearance. Both the proposed models of foreground and background, and the simple method for feature selection can be improved. Challenges remain in incorporating more advanced models of segmentation without excessively compromising computational efficiency.

Acknowledgements. Financial support for this work was provided by EPSRC project HAML and an RCUK Academic Fellowship.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Y. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *Proc. ICCV*, 2001.

[3] R. T. Collins and Y. Liu. On-line selection of discriminative tracking features. In *Proc. ICCV*, 2003.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[5] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

[7] D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *Proc. ICCV*, 1999.

[8] T. Joachims. Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, 1999.

[9] M. Jones, P. Viola, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. ICCV*, 2003.

[10] Z. Lin and L. S. Davis. A pose-invariant descriptor for human detection and segmentation. In *Proc. ECCV*, 2008.

[11] R. Okada and S. Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. In *Proc. ECCV*, 2008.

[12] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *Proc. ICIP*, 1999.

[13] D. Ramanan. Using segmentation to verify object hypotheses. In *Proc. CVPR*, 2007.

[14] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. ECCV*, 2006.

[15] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *NIPS*, 2007.

[16] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on Riemannian manifolds. In *Proc. CVPR*, 2007.

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, 2001.

[18] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *Proc. CVPR*, 2006.