**DECLARATION**

I declare that this written submission represents my ideas in my own words and where others’ ideas have been include, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature

**Asif Imam Mulla**

**Human Detection And Counting**

# Abstract

*We study the question of feature sets for robust visual object recognition, adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.*

# Introduction

Detecting humans in images is a challenging task owing to their variable appearance and the wide range of poses that they can adopt. The first need is a robust feature set that allows the human form to be discriminated cleanly, even in cluttered backgrounds under difficult illumination. We study the issue of feature sets for humandetection, showingthat locally normalized Histogram of Oriented Gradient (HOG) descriptors provide excellent performance relative to other existing feature sets including wavelets [17,22]. The proposed descriptors are reminiscent of edge orientation histograms [4,5], SIFT descriptors [12] and shape contexts [1], but they are computed on a dense grid of uniformly spaced cells and they use overlapping local contrast normalizations for improved performance. We make a detailed study of the effects of various implementation choices on detector performance, taking “pedestrian detection” (the detection of mostly visible people in more or less upright poses) as a test case. For simplicity and speed, we use linear SVM as a baseline classifier throughoutthe study. The new detectors give essentially perfect results on the MIT pedestriantest set [18,17], so we have created a more challenging set containing over 1800 pedestrian images with a large range of poses and backgrounds. Ongoing work suggests that our feature set performs equally well for other shape-based object classes.

We briefly discuss previous work on human detection in §2, give an overview of our method §3, describe our data sets in §4 and give a detailed description and experimental evaluation of each stage of the process in §5–6. The main conclusions are summarized in §7.

# Previous Work

There is an extensive literature on object detection, but here we mention just a few relevant papers on human detection [18,17,22,16,20]. See [6] for a survey. Papageorgiou *et al* [18] describe a pedestrian detector based on a polynomial SVM using rectified Haar wavelets as input descriptors, with a parts (subwindow) based variant in [17]. Depoortere *et al* give an optimized version of this [2]. Gavrila & Philomen [8] take a more direct approach, extracting edge images and matching them to a set of learned exemplars using chamfer distance. This has been used in a practical real-time pedestrian detection system [7]. Viola *et al* [22] build an efficient moving person detector, using AdaBoost to train a chain of progressively more complex region rejection rules based on Haar-like wavelets and space-time differences. Ronfard *et al* [19] build an articulated body detector by incorporating SVM based limb classifiers over 1st and 2nd order Gaussian filters in a dynamic programmingframework similar to those of

Felzenszwalb & Huttenlocher [3] and Ioffe & Forsyth [9].

Mikolajczyk *et al* [16] use combinations of orientationposition histograms with binarythresholdedgradient magnitudes to build a parts based method containing detectors for faces, heads, and front and side profiles of upper and lower body parts. In contrast, our detector uses a simpler architecture with a single detection window, but appears to give significantly higher performance on pedestrian images.

# Overview of the Method

This section gives an overview of our feature extraction chain, which is summarized in fig. 1. Implementation details are postponed until §6. The method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. Similar features have seen increasing use over the past decade [4,5,12,15]. The basic idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or

1

|  |
| --- |
| **tsWeighted voteint LinearSVM** |

**Input Normalizegamma &Computegradieno spatial &orientation cellsContrast normalizeover overlappingspatial blocksCollect HOG’sover detectionwindow Person /non−personclassification** **image colour**

Figure 1. An overview of our feature extraction and object detection chain. The detector window is tiled with a grid of overlapping blocks in which Histogram of Oriented Gradient feature vectors are extracted. The combined vectors are fed to a linear SVM for object/nonobject classification. The detection window is scanned across the image at all positions and scales, and conventional non-maximum suppression is run on the output pyramid to detect object instances, but this paper concentrates on the feature extraction process.

edge directions, even without precise knowledge of the little difference if they are much smaller that the local spatial corresponding gradient or edge positions. In practice this is or orientation bin size. For human detection, rather coarse implemented by dividing the image window into small spatial spatial sampling, fine orientation sampling and strong local regions (“*cells*”), for each cell accumulating a local 1-D photometric normalizationturns out to be the best strategy, histogram of gradient directions or edge orientations over the presumably because it permits limbs and body segments to pixels of the cell. The combined histogram entries form the change appearance and move from side to side quite a lot representation. For better invariance to illumination, provided that they maintain a roughly upright orientation.

shadowing, *etc*., it is also useful to contrast-normalize the

local responses before using them. This can be done by **4 Data Sets and Methodology** accumulating a measure of local histogram “energy” over somewhat larger spatial regions (“*blocks*”) and using the **Datasets.** We tested our detector on two different data sets. results to normalize all of the cells in the block. We will refer The first is the well-established MIT pedestrian database [18], to the normalized descriptor blocks as *Histogram of Oriented* containing 509 training and 200 test images of pedestrians in *Gradient (HOG)* descriptors. Tiling the detection window city scenes (plus left-right reflections of these). It contains with a dense (in fact, overlapping) grid of HOG descriptors only front or back views with a relatively limited range of and using the combined feature vector in a conventional SVM poses. Our best detectors give essentially perfect results on based window classifier gives our human detection chain (see this data set, so we produced a new and significantly more fig. 1). challenging data set, ‘INRIA’, containing 1805 64×128

The use of orientation histograms has many precursors images of humans cropped from a varied set of personal [13,4,5], but it only reached maturity when combined with photos. Fig. 2 shows some samples. The people are usually local spatial histogramming and normalization in Lowe’s standing, but appear in any orientation and against a wide *Scale Invariant Feature Transformation (SIFT)* approach to variety of background image including crowds. Many are wide baseline image matching [12], in which it provides the bystanders taken from the image backgrounds, so there is no underlying image patch descriptor for matching scaleinvariant particular bias on their pose. The database is available from keypoints. SIFT-style approaches perform remarkably well in *http://lear.inrialpes.fr/data* for research purposes.

this application [12,14]. The *Shape Context* work [1] studied **Methodology.** We selected 1239 of the images as positive alternative cell and block shapes, albeit initially using only training examples, together with their left-right reflections edge pixel counts without the orientation histogramming that (2478 images in all). A fixed set of 12180 patches sampled makes the representation so effective. The success of these randomly from 1218 person-free training photos provided the sparse feature based representations has somewhat initial negative set. For each detector and parameter overshadowedthe power and simplicity of HOG’s as dense combination a preliminary detector is trained and the 1218 image descriptors. We hope that our study will help to rectify negative training photos are searched exhaustively for false this. In particular, our informal experiments suggest that even positives (‘hard examples’). The method is then re-trained the best current keypoint based approaches are likely to have using this augmented set (initial 12180 + hard examples) to false positive rates at least 1–2 orders of magnitude higher produce the final detector. The set of hard examples is than our dense grid approach for human detection, mainly subsampled if necessary, so that the descriptors of the final because none of the keypoint detectors that we are aware of training set fit into 1.7 Gb of RAM for SVM training. This detect human body structures reliably. retraining process significantly improves the performance of The HOG/SIFT representation has several advantages. It captures edge or gradient structure that is very characteristic each detector (by 5% at 10−4 False Positives Per Window of local shape, and it does so in a local representation with an tested (FPPW) for our default detector), but additional rounds easily controllable degree of invariance to local geometric and of retraining make little difference so we do not use them. photometric transformations: translations or rotations make



Figure 2. Some sample images from our new human detection database. The subjects are always upright, but with some partial occlusions and a wide range of variations in pose, appearance, clothing, illumination and background.

To quantify detector performance we plot *Detection Error* performance (the values selected were somewhat variable, in *Tradeoff (DET)* curves on a log-log scale, *i.e*. miss rate the region of 20–50 graylevels). **Results.** Fig. 3 shows the

(1−Recall or  TruePos+FalseNegFalseNeg ) versus performance of the various detectors on the MIT and INRIA FPPW. Lower values are better. DET plots are used data sets. The HOG-based detectors greatly outperformthe extensively in speech and in NIST evaluations. They present wavelet, PCA-SIFT and Shape Context ones, giving nearthe same information as Receiver Operating Characteristics perfect separation on the MIT test set and at least an order of

(ROC’s) but allow small probabilities to be distinguished magnitude reduction in FPPW on the INRIA one. Our Haar-

−4FPPW as a like wavelets outperform MIT wavelets because we also use more easily. We will often use miss rate at 10 reference point for results. This is arbitrary but no more so 2nd order derivatives and contrast normalize the output vector.

than, *e.g*. Area Under ROC. In a multiscale detector it Fig. 3(a) also shows MIT’s best parts based and monolithic corresponds to a raw error rate of about0*.*8 false positives per detectors (the points are interpolated from [17]), however 640×480 image tested. (The full detector has an even lower beware that an exact comparison is not possible as we do not false positive rate owing to nonmaximum suppression). Our know how the database in [17] was divided into training and DET curves are usually quite shallow so even very small test parts and the negative images used are not available. The improvements in miss rate are equivalent to large gains in performances of the final rectangular (R-HOG) and circular FPPW at constant miss rate. For example, for our default (C-HOG) detectors are very similar, with C-HOG having the detector at 1e-4 FPPW, every 1% absolute (9% relative) slight edge. Augmenting R-HOG with primitive bar detectors reduction in miss rate is equivalent to reducing the FPPW at (oriented 2nd derivatives – ‘R2-HOG’) doubles the feature constant miss rate by a factor of 1.57. dimension but further improves the performance (by 2% at

**5 Overview of Results** 10−4 FPPW). Replacing the linear SVM with a Gaussian

Before presenting our detailed implementation and kernel one improves performance by about 3% at 10−4 FPPW, performance analysis, we compare the overall performance of at the cost of much higher run times[[1]](#footnote-1). Using binary edge our final HOG detectors with that of some other existing voting (ECHOG) instead of gradient magnitude weighted methods. Detectors based on rectangular (R-HOG) or circular voting (C- log-polar (C-HOG) blocks and linear or kernel SVM are compared with our implementations of the Haar wavelet, HOG) decreases performance by 5% at 10−4 FPPW, while PCA-SIFT, and shape context approaches. Briefly, these omitting orientation information decreases it by much more, approaches are as follows: even if additional spatial or radial bins are added (by 33% at

10−4 FPPW, for both edges (E-ShapeC) and gradients **Generalized Haar Wavelets.** This is an extended set of (GShapeC)). PCA-SIFT also performs poorly. One reason is oriented Haar-like wavelets similar to (but better than) that that, in comparison to [11], many more (80 of 512) principal used in [17]. The features are rectified responses from 9×9 and vectors have to be retained to capture the same proportion of

12×12 oriented 1st and 2nd derivative box filters at 45◦ the variance. This may be because the spatial registration is weaker when there is no keypoint detector. intervals and the corresponding 2nd derivative *xy* filter.

**PCA-SIFT.** These descriptors are based on projecting gradient images onto a basis learned from training images using PCA [11]. Ke & Sukthankar found that they outperformed SIFT for key point based matching, but this is controversial

[14]. Our implementation uses 16×16 blocks with the same derivative scale, overlap, *etc*., settings as our HOG descriptors. The PCA basis is calculated using positive training images.

**Shape Contexts.** The original Shape Contexts [1] used binary edge-presence voting into log-polar spaced bins, irrespective of edge orientation. We simulate this using our CHOG descriptor (see below) with just 1 orientation bin. 16 angular and 3 radial intervals with inner radius 2 pixels and outer radius 8 pixels gave the best results. Both gradientstrength and edge-presence based voting were tested, with the edge threshold chosen automatically to maximize detection **6 Implementation and Performance Study** be the best. We tested gradients computed using Gaussian smoothing followed by one of several discrete derivative We now give details of our HOG implementations and systematically study the effects of the various choices on masks. Several smoothing scales were tested including *σ*=0

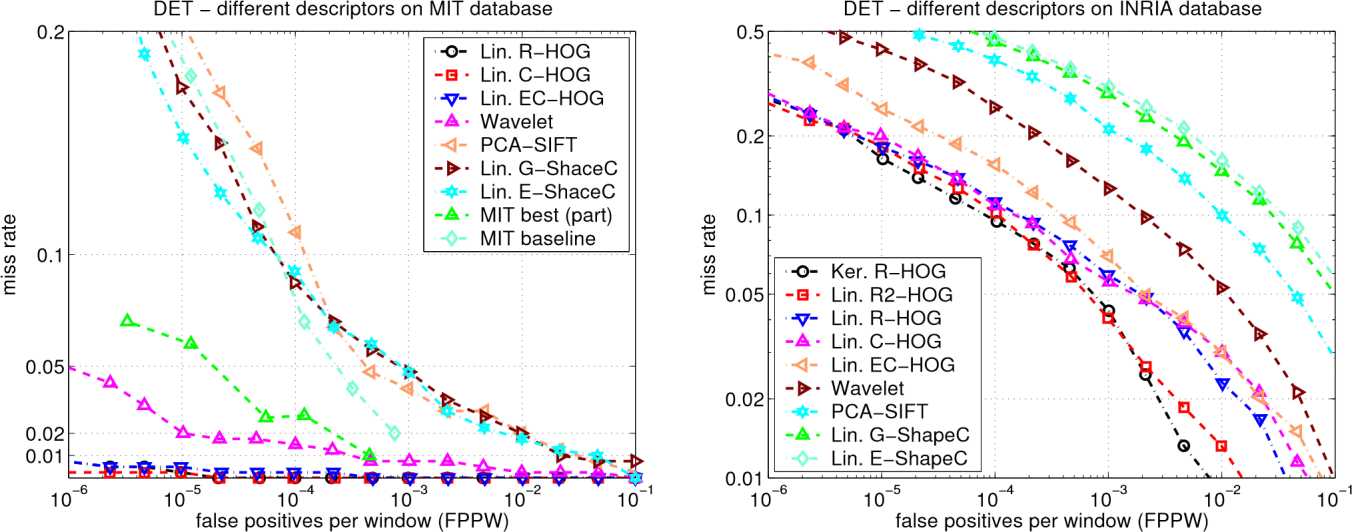


Figure 3. The performance of selected detectors on (left) MIT and (right) INRIA data sets. See the text for details.

detector performance. Throughout this section we refer results (none). Masks tested included various 1-D point derivatives to our default detector which has the following properties, (uncentred [−1*,*1], centred [−1*,*0*,*1] and cubiccorrected described below: RGB colour space with no gamma [1*,*−8*,*0*,*8*,*−1]) as well as 3×3 Sobel masks and correction; [−1*,*0*,*1] gradient filter with no smoothing; linear 2×2 diagonal ones (the most compact centred 2gradient voting into 9 orientation bins in 0◦–180◦; 16×16

D derivative masks). Simple 1-D [−1*,*0*,*1] masks at *σ*=0 work

pixel blocks of four 8×8 pixel cells; Gaussian spatial window best. Using larger masks always seems to decrease with *σ* = 8 pixel; *L2-Hys* (Lowe-styleclipped L2 norm) block performance, and smoothing damages it significantly: for normalization; block spacing stride of 8 pixels (hence 4-fold Gaussian derivatives, moving from *σ*=0 to *σ*=2 reduces the coverage of each cell); 64×128 detection window; linear recall rate from 89% to 80% at 10−4 FPPW. At *σ*=0, cubic SVM classifier. corrected 1-D width 5 filters are about 1% worse than [−1*,*0*,*1]

Fig. 4 summarizes the effects of the various HOG at 10−4 FPPW, while the 2×2 diagonal masks are 1.5% worse. parameters on overall detection performance. These will be Using uncentred [−1*,*1] derivative masks also decreases examined in detail below. The main conclusions are that for performance (by 1.5% at 10−4 FPPW), presumably because good performance, one should use fine scale derivatives

(essentially no smoothing), many orientation bins, and orientation estimation suffers as a result of the *x* and *y* filters moderately sized, strongly normalized, overlapping descriptor being based at different centres.

blocks. For colour images, we calculate separate gradients for each **6.1 Gamma/Colour Normalization** colour channel, and take the one with the largest norm as the

We evaluated several input pixel representations including pixel’s gradient vector.

grayscale, RGB and LAB colour spaces optionally with **6.3 Spatial / Orientation Binning**

powerlaw (gamma)equalization. These normalizationshave The next step is the fundamental nonlinearity of the only a modest effect on performance, perhaps because the descriptor. Each pixel calculates a weighted vote for an edge subsequent descriptor normalization achieves similar results. orientation histogram channel based on the orientation of the We do use colour information when available. RGB and LAB gradient element centred on it, and the votes are accumulated colour spaces give comparable results, but restricting to into orientation bins over local spatial regions that we call grayscale reduces performance by 1.5% at 10−4 FPPW. Square *cells*. Cells can be either rectangular or radial (log-polar root gamma compression of each colour channel improves sectors). The orientation bins are evenly spaced over 0◦– performance at low FPPW (by 1% at 10−4 FPPW) but log 180◦ (“unsigned” gradient) or 0◦–360◦ (“signed” gradient).

compression is too strong and worsens it by 2% at 10−4 FPPW. To reduce aliasing, votes are interpolated bilinearly between the neighbouring bin centres in both orientation and position.

**6.2 Gradient Computation** The vote is a function of the gradient magnitude at the pixel,

Detector performance is sensitive to the way in which either the magnitude itself, its square, its square root, or a gradients are computed, but the simplest scheme turns out to

clipped form of the magnitude representing soft contrast, so effective local contrast normalization turns out to presence/absence of an edge at the pixel. In practice, using the be essential for good performance. We evaluated a num-

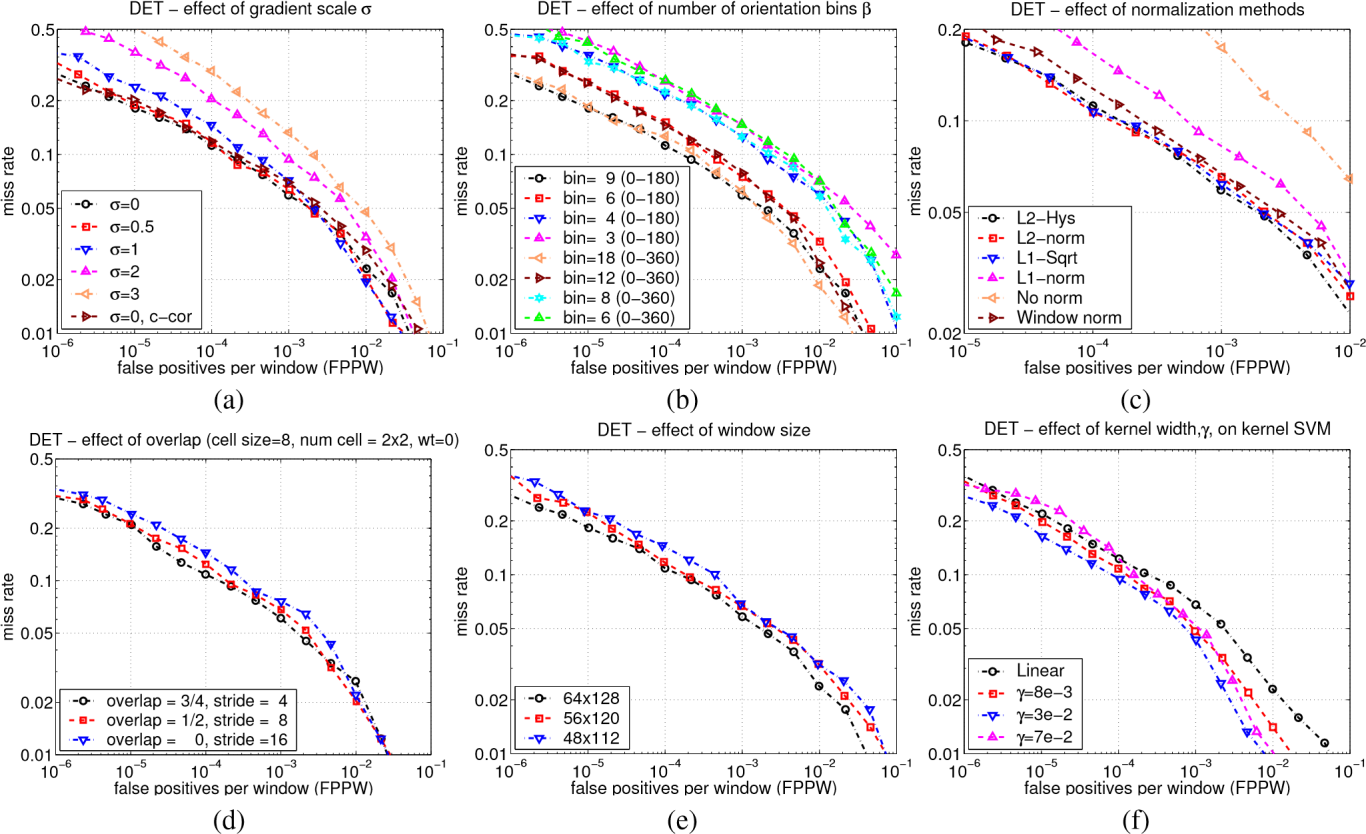
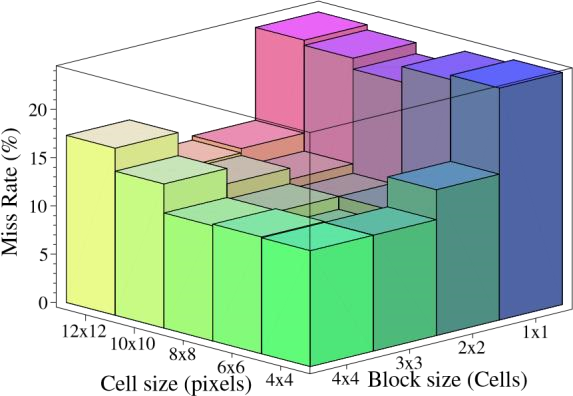


Figure 4. For details see the text. (a) Using fine derivative scale significantly increases the performance. (‘c-cor’ is the 1D cubic-corrected point derivative). (b) Increasing the number of orientation bins increases performance significantly up to about 9 bins spaced over 0◦– 180◦. (c) The effect of different block normalization schemes (see §6.4). (d) Using overlapping descriptor blocks decreases the miss rate by around 5%. (e) Reducing the 16 pixel margin around the 64×128 detection window decreases the performance by about 3%. (f) Using

a Gaussian kernel SVM, exp(−*γ*k**x1**−**x2**k2), improves the performance by about 3%.

magnitudeitself gives the best results. Taking the square root reduces performance slightly, while using binary edge presence votingdecreases it significantly (by 5% at 10−4 FPPW).

Fine orientation coding turns out to be essential for good performance, whereas (see below) spatial binning can be rather coarse. As fig. 4(b) shows, increasing the number of orientation bins improves performance significantly up to about 9 bins, but makes little difference beyond this. This is for bins spaced over 0◦–180◦, *i.e*. the ‘sign’ of the gradient is ignored. Including signed gradients (orientation range

0◦–360◦, as in the original SIFT descriptor) decreases the performance, even when the number of bins is also doubled to

preserve the original orientation resolution. For humans, the Figure 5. The miss rate at 10−4 FPPW as the cell and block sizes wide range of clothing and background colours presumably change. The stride (block overlap) is fixed at half of the block size. makes the signs of contrasts uninformative. However note that 3×3 blocks of 6×6 pixel cells perform best, with 10.4% miss rate.

including sign information does help substantially in some otherobject recognitiontasks, *e.g*.cars, motorbikes. ber of different normalization schemes. Most of them are

**6.4 Normalization and Descriptor Blocks** based on grouping cells into larger spatial blocks and contrast

Gradient strengths vary over a wide range owing to local normalizing each block separately. The final descriptor is then variations in illumination and foreground-background

the vector of all components of the normalized cell responses pairs are significantly better than horizontal pairs alone, but

from all of the blocks in the detection window. not as good as 2×2 blocks (1% worse at 10−4 FPPW).

In fact, we typically overlap the blocks so that each scalar cell

response contributes several components to the final **C-HOG.** Our circular block (C-HOG) descriptors are descriptor vector, each normalized with respect to a different reminiscent of Shape Contexts [1] except that, crucially, each block. This may seem redundant but good normalization is spatial cell contains a stack of gradient-weighted orientation critical and including overlap significantly improves the cells instead of a single orientation-independent edgepresence performance. Fig. 4(d) shows that performanceincreases by count. The log-polar grid was originally suggested by the idea that it would allow fine coding of nearby structure to be

4% at 10−4 FPPW as we increase the overlap from none (stride combined with coarser coding of wider context, and the fact

16) to 16-fold area / 4-fold linear coverage (stride 4). that the transformation from the visual field to the V1 cortex We evaluated two classes of block geometries, square or in primates is logarithmic[21]. However small descriptors rectangular ones partitioned into grids of square or rectangular with very few radial bins turn out to give the best performance, spatial cells, and circular blocks partitioned into cells in log- so in practice there is little inhomogeneity or context. It is polar fashion. We will refer to these two arrangements as R- probably better to think of C-HOG’s simply as an advanced HOG and C-HOG (for rectangular and circular HOG). form of centre-surround coding.

We evaluated two variants of the C-HOG geometry, ones

**R-HOG.** R-HOG blocks have many similarities to SIFT with a single circular central cell (similar to the descriptors [12] but they are used quite differently. They are GLOH feature of [14]), and ones whose central computed in dense grids at a single scale without dominant cell is divided into angular sectors as in shape orientation alignment and used as part of a larger code vector contexts. We present results only for the that implicitly encodes spatial position relative to the circularcentre variants, as these have fewer spatial detection window, whereas SIFT’s are computed at a sparse cells than the divided centre ones and give the set of scale-invariant key points, rotated to align their same performance in practice. A technical report dominant orientations, and used individually. SIFT’s are will provide further details. The C-HOG layout has four optimized for sparse wide baseline matching, R-HOG’s for parameters: the numbers of angular and radial bins; the radius dense robust coding of spatial form. Other precursors include of the central bin in pixels; and the expansion factor for the edge orientation histograms of Freeman & Roth [4]. We subsequent radii. At least two radial bins (a centre and a usually use square R-HOG’s, *i.e*. *ς*×*ς* grids of *η*×*η* pixel cells surround) and four angular bins (quartering) are needed for each containing *β* orientation bins, where *ς,η,β* are parameters. good performance. Including additional radial bins does not Fig. 5 plots the miss rate at 10−4 FPPW w.r.t. cell size and change the performance much, while increasing the number

block size in cells. For human detection, 3×3 cell blocks of of angular bins decreases performance (by 1.3% at 10−4 FPPW when going from 4 to 12 angular bins). 4 pixels is the best 6×6 pixel cells perform best, with 10.4% miss-rate at 10−4 radius for the central bin, but 3 and 5 give similar results. FPPW. In fact, 6–8 pixel wide cells do best irrespective of the Increasing the expansion factor from 2 to 3 leaves the block size – an interesting coincidence as human limbs are performance essentially unchanged. With these parameters, neither Gaussian spatial weighting nor inverse weighting of

about 6–8 pixels across in our images. 2×2 and 3×3 blocks cell votes by cell area changes the performance, but work best. Beyond this, the results deteriorate: adaptivity to combining these two reduces slightly. These values assume local imaging conditions is weakened when the block fine orientation sampling. Shape contexts (1 orientation bin) becomes too big, and when it is too small (1×1 block / require much finer spatial subdivision to work well.

normalization over orientations alone) valuable spatial **Block Normalization schemes.** We evaluated four different information is suppressed. block normalization schemes for each of the above HOG

As in [12], it is useful to downweightpixels near the edges geometries. Let **v** be the unnormalized descriptor vector, k**v**k*k* of the block by applying a Gaussian spatial window to each be its *k*-norm for a small constant. The pixel before accumulating orientation votes into cells. This schemes are: (a) *L2-norm*, **v** 2



improves performance by 1% at 10−4 FPPW for a Gaussian *L2-Hys*, L2-norm followed by clipping (limiting the with *σ* = 0*.*5 block width. maximum values of **v** to 0.2) and renormalizing, as in [12]; (c) We also tried including multiple block types with different *L1-norm*, **v** ; and (d) *L1-sqrt*, L1-norm cell and block sizes in the overall descriptor. This slightly followed by square root **v** , which amounts



−4 FPPW), at the to treating the descriptor vectors as probability distriband using the Bhattacharya distance between them. Fig. 4(c) utions improves performance (by around 3% at 10

cost of greatly increased descriptor size. Besides square R- shows that L2-Hys, L2-norm and L1-sqrt all perform equally HOG blocks, we also tested vertical (2×1 cell) and horizontal well, while simple L1-norm reduces performance by 5%, and (1×2 cell) blocks and a combined descriptor including both omitting normalization entirely reduces it by 27%, at 10−4 vertical and horizontal pairs. Vertical and vertical+horizontal FPPW. Some regularization is needed as we evaluate descriptors densely, including on empty patches, but the **6.7 Discussion** results are insensitive to ’s value over a large range. Overall, there are several notable findings in this work. The **Centre-surround normalization.** We also investigated an fact that HOG greatly out-performs wavelets and that any alternative centre-surround style cell normalization scheme, in significant degree of smoothing before calculating gradients which the image is tiled with a grid of cells and for each cell damages the HOG results emphasizes that much of the the total energy in the cell and its surrounding region (summed available image information is from *abrupt edges at fine* over orientations and pooled using Gaussian weighting) is *scales*, and that blurring this in the hope of reducing the used to normalize the cell. However as fig. 4(c) (“*window* sensitivity to spatial position is a mistake. Instead, gradients *norm*”) shows, this decreases performance relative to the should be calculated at the finest available scale in the current corresponding block based scheme (by 2% at 10−4 FPPW, for pyramid layer, rectified or used for orientation voting, and only then blurred spatially. Given this, relatively coarse

pooling with longer any overlapping blocks so each cell is coded only once *σ*=1 cell widths). One reason is that there are no spatial quantization suffices (8×8 pixel cells / one limb width). in the final descriptor. Including several normalizationsfor On the other hand, at least for human detection, it pays to each cell based on differentpoolingscales *σ* providesno sample orientation rather finely: both wavelets and shape perceptible change in performance,so it seems that it is the contexts lose out significantly here. existence of several pooling regions with *different* spatial Secondly, strong *local* contrast normalization is essential offsets relative to the cell that is important here, not the for good results, and traditional centre-surround style schemes pooling scale. are not the best choice. Better results can be achieved by

To clarify this point, consider the R-HOG detector with normalizing each element (edge, cell) *several times* with overlapping blocks. The coefficients of the trained linear respect to different local supports, and treating the results as SVM give a measure of how much weight each cell of each independent signals. In our standard detector, each HOG cell block can have in the final discriminationdecision. Close appears four times with different normalizations and including examination of fig. 6(b,f) shows that the most important cells this ‘redundant’ information improves performance from 84% are the ones that typically contain major human contours to 89% at 10−4 FPPW.

(especially the head and shoulders and the feet), normalized

w.r.t. blocks lying *outside* the contour. In other words — **7 Summary and Conclusions** despite the complex, cluttered backgrounds that are common We have shown that using locally normalized histogram of in our training set — the detector cues mainly on the contrast gradient orientations features similar to SIFT descriptors [12] of silhouette contours against the background, not on internal in a dense overlapping grid gives very good results for person edges or on silhouette contours against the foreground. detection, reducing false positive rates by more than an order Patterned clothing and pose variations may make internal of magnitude relative to the best Haar wavelet based detector regions unreliable as cues, or foreground-to-contour from [17]. We studied the influence of various descriptor transitions may be confused by smooth shading and parameters and concluded that fine-scale gradients, fine shadowing effects. Similarly, fig. 6(c,g) illustrate that orientation binning, relatively coarse spatial binning, and gradients inside the person (especially vertical ones) typically high-quality local contrast normalization in overlapping count as negative cues, presumably because this suppresses descriptor blocks are all important for good performance. We false positives in which long vertical lines trigger vertical head also introduced a new and more challenging pedestrian and leg cells. database, which is publicly available.

**6.5 Detector Window and Context Future work:** Although our current linear SVM detector is

Our 64×128 detection window includes about 16 pixels of reasonably efficient – processing a 320×240 scale-space margin around the person on all four sides. Fig. 4(e) shows image (4000 detection windows) in less than a second – there that this border provides a significant amount of context that is still room for optimization and to further speed up helps detection. Decreasing it from 16 to 8 pixels (48×112 detections it would be useful to develop a coarse-to-fine or detection window) decreases performance by 6% at 10−4 rejectionchain style detector based on HOG descriptors. We FPPW. Keeping a 64×128 window but increasing the person are also working on HOG-based detectors that size within it (again decreasing the border) causes a similar incorporatemotioninformation using block matching or loss of performance, even though the resolution of the person optical flow fields. Finally, although the current fixed-

is actually increased. template-style detector has proven difficult to beat for fully

**6.6 Classifier** visible pedestrianbelieve that including a parts based model with a greater s, humans are highly articulated and we

By default we use a soft (*C*=0*.*01) linear SVM trained with degree of local spatial invariance

SVMLight [10] (slightly modified to reduce memory usage for problems with large dense descriptor vectors). Using a Gaussian kernel SVM increases performance by about 3% at 10−4 FPPW at the cost of a much higher run time.



(a) (b) (c) (d) (e) (f) (g)

Figure 6. Our HOG detectors cue mainly on silhouette contours (especially the head, shoulders and feet). The most active blocks are centred on the image background just *outside* the contour. (a) The average gradient image over the training examples. (b) Each “pixel” shows the maximum positive SVM weight in the block centred on the pixel. (c) Likewise for the negative SVM weights. (d) A test image. (e) It’s computed R-HOG descriptor. (f,g) The R-HOG descriptor weighted by respectively the positive and the negative SVM weights. would help to improve the detection results in more general [11] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive situations. representation for local image descriptors. *CVPR, Washington,*

**Acknowledgments.** This work was supported by the *DC, USA*, pages 66–75, 2004.

European Union research projects ACEMEDIA and PASCAL. [12] D. G. Lowe. Distinctive image features from scale-invariant We thanks Cordelia Schmid for many useful comments. keypoints. *IJCV*, 60(2):91–110, 2004.

SVMLight [10] provided reliable training of large-scale [13] R. K. McConnell. Method of and apparatus for pattern

SVM’s. **References** recognition, January 1986. U.S. Patent No. 4,567,610.

1. S. Belongie, J. Malik, and J. Puzicha. Matching shapes. *The 8th* [14] K. Mikolajczyk and C. Schmid. A performance evaluation of *ICCV, Vancouver, Canada*, pages 454–461, 2001. local descriptors. *PAMI*, 2004. Accepted.
2. V. de Poortere, J. Cant, B. Van den Bosch, J. de Prins, F. [15] K. Mikolajczyk and C. Schmid. Scale and affine invariant Fransens, and L. Van Gool. Efficient pedestrian detection: a interest point detectors. *IJCV*, 60(1):63–86, 2004.

test case for svm based categorization. *Workshop on Cognitive* [16] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human *Vision*, 2002. Available online: detection based on a probabilistic assembly of robust part http://www.vision.ethz.ch/cogvis02/. detectors. *The 8th ECCV, Prague, Czech Republic*, volume I, [3] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pages 69– 81, 2004.

pictorial structures. *CVPR, Hilton Head Island, South* [17] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based

*Carolina, USA*, pages 66–75, 2000. object detection in images by components. *PAMI*, 23(4):349– [4] W. T. Freeman and M. Roth. Orientation histograms for hand 361, April 2001.

gesture recognition. *Intl. Workshop on Automatic Faceand* [18] C. Papageorgiou and T. Poggio. A trainable system for object *Gesture- Recognition, IEEE Computer Society, Zurich,* detection. *IJCV*, 38(1):15–33, 2000.

*Switzerland*, pages 296–301, June 1995. [19] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse

1. W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma. Computer pictures of people. *The 7th ECCV, Copenhagen, Denmark*, vision for computer games. *2nd International Conference on* volume IV, pages 700–714, 2002.

*Automatic Face and Gesture Recognition, Killington, VT, USA*, [20] Henry Schneiderman and Takeo Kanade. Object detection pages 100–105, October 1996. using the statistics of parts. *IJCV*, 56(3):151–177, 2004.

1. D. M. Gavrila. The visual analysis of human movement: A [21] Eric L. Schwartz. Spatial mapping in the primate sensory survey. *CVIU*, 73(1):82–98, 1999. projection: analytic structure and relevance to perception. [7] D. M. Gavrila, J. Giebel, and S. Munder. Vision-based *Biological Cybernetics*, 25(4):181–194, 1977.

pedestrian detection: the protector+ system. *Proc. of the IEEE* [22] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using

*Intelligent Vehicles Symposium, Parma, Italy*, 2004. patterns of motion and appearance. *The 9th ICCV, Nice,*

1. D. M. Gavrila and V. Philomin. Real-time object detection for *France*, volume 1, pages 734–741, 2003.

smart vehicles. *CVPR, Fort Collins, Colorado, USA*, pages 87– 93, 1999.

1. S. Ioffe and D. A. Forsyth. Probabilistic methods for finding people. *IJCV*, 43(1):45–68, 2001.
2. T. Joachims. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. The MIT Press, Cambridge, MA, USA, 1999.

1. We use the hard examples generated by *linear* R-HOG to train the kernel R-HOG detector, as kernel R-HOG generates so few false positives that its hard example set is too sparse to improve the generalization significantly. [↑](#footnote-ref-1)