Pedestrian Detection
Histograms of Oriented Gradients for Human Detection
Navneet Dalal and Bill Triggs
Challenges of pedestrian detection

- Wide variety of articulated poses
- Variable appearance/clothing
- Complex backgrounds
- Unconstrained illumination
- Occlusions
- Different Scales
• Histogram of Oriented Gradient descriptor assumes that the local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions.

• The implementation of these descriptors can be achieved by dividing the image into small connected regions (cells), and for each cell computing a histogram of gradient directions (i.e. edge orientations) for the pixels within the cell. The combination of these histograms then represents the descriptor.

• The Histogram of Oriented Gradients descriptor has some key advantages over other descriptor methods.
  – Since it operates on localized cells, it shows invariance to geometric and photometric transformations such changes (the would only appear in larger spatial regions).
  – Coarse spatial sampling, fine orientation sampling, and strong local photometric normalization permits the individual body movement of pedestrians to be ignored so long as they maintain a roughly upright position.
  – The HOG descriptor is thus particularly suited for human detection in images.

• Essential in contextually critical environments: surveillance of pedestrians, vehicles, luggages and groups of unknown objects. Performance limited by
  • the occlusion problem often occurring in surveillance applications
  • noise occurring in e.g. large illumination variations, persistent shadows
Person detection with HOG descriptors

In the Dalal and Triggs human detection experiment, the optimal parameters were found to be 3x3 cell blocks of 6x6 pixel cells with 9 histogram channels.
In the Dalal and Triggs experiment tests were performed with different color spaces:
- RGB
- LAB
- Grayscale

Gamma Normalization and Compression
- Square root
- Log

This step can be omitted in HOG descriptor computation, as the descriptor normalization essentially achieves the same result.
Dalal and Triggs tested several masks, such as the 1-D centered mask, 3x3 Sobel mask or diagonal masks. The 1-D centered point discrete derivative mask in one of or both the horizontal and vertical directions (filtering the color or intensity data of the image with the \([-1, 0, 1]\) filter kernel) resulted the best performance.

They also experimented Gaussian smoothing before applying the derivative mask, but found that omission of any smoothing performed better in practice.
• HOG blocks typically overlap: each cell contributes more than once to the final descriptor.

• Two main block geometries exist.
  • rectangular R-HOG blocks
  • circular C-HOG blocks

• Some minor improvement in performance can be gained by applying a Gaussian spatial window within each block before tabulating histogram votes in order to weight pixels around the edge of the blocks less.
• R-HOG blocks are generally square grids, represented by three parameters:
  − the number of cells per block,
  − the number of pixels per cell,
  − the number of channels per cell histogram.

The R-HOG blocks are different from the scale-invariant feature transform descriptors; R-HOG blocks are computed in dense grids at some single scale without orientation alignment, whereas SIFT descriptors are computed at sparse, scale-invariant key image points and are rotated to align orientation.

The R-HOG blocks are used in conjunction to encode spatial form information, while SIFT descriptors are used singly.

• C-HOG blocks can be found in two variants: a) With one single, central cell b) With an angularly-divided central cell. C-HOG blocks can be described with four parameters:
  − the number of angular and radial bins,
  − the radius of the center bin,
  − the expansion factor for the radius of additional radial bins.

C-HOG blocks appear similar to Shape Contexts, but differ strongly in that C-HOG blocks contain cells with several orientation channels, while Shape Contexts only make use of a single edge presence count in their formulation.
Dalal and Triggs found that:
- the two main variants provided equal performance
- two radial bins with four angular bins, a center radius of 4 pixels, and an expansion factor of 2 provided the best performance
- Gaussian weighting provides no benefit when used in conjunction with the C-HOG blocks.
• For improved accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination or shadowing.

• Dalal and Triggs explored four different methods for block normalization:
  - L1-norm
    \[ L1 - \text{norm} : v \rightarrow \frac{v}{\|v\|_1 + \epsilon} \]
  - L2-norm
    \[ L2 - \text{norm} : v \rightarrow \frac{v}{\sqrt{\|v\|_2^2 + \epsilon^2}} \]
  - L1-sqrt
    \[ L1 - \text{sqrt} : v \rightarrow \sqrt{v/(\|v\|_1 + \epsilon)} \]
  - L2-Hys
    \[ L2 - \text{hys} : \text{L2-norm, plus clipping at .2 and renomalizing} \]

• In their experiments, Dalal and Triggs found the L2-Hys, L2-norm, and L1-sqrt schemes provide similar performance, while the L1-norm provides slightly less reliable performance. All four methods showed very significant improvement over the non-normalized data.
Input image → Normalize gamma & colour → Compute gradients → Weighted vote into spatial & orientation cells → Contrast normalize over overlapping spatial blocks → Collect HOG’s over detection window → Linear SVM → Person / non-person classification
HOG descriptors are fed into a recognition system based on SVM supervised learning which looks for an optimal hyperplane as a decision function.

In the Dalal and Triggs human recognition tests, they used the freely available SVMLight software package.
Movie example