Histogram of Oriented Gradients for Human Detection

Article by Navneet Dalal and Bill Triggs

All images in presentation is taken from article
Presentation by Inge Edward Halsaunet
Introduction

What:

- Detect humans in images

Problem:

- Detection of humans in images is a challenging task
  - Cluttered backgrounds
  - Difficult illumination

Main Focus:

- "Pedestrian detection"
  - Mostly visible
  - More or less upright pose
Contributions

A novel method of detecting humans in images.

A new, significantly more challenging dataset of images for detection of humans in images which is publicly available
HOG - The Overview

The main idea is to use “well-normalized local histograms of image gradient orientation in a dense grid.”

Algorithm overview:

- Divide the image into small spatial cells
- For each cell: Accumulate a local 1-D histogram of gradient directions
- Contrast normalize blocks of cells
- Blocks of cells known as HOG descriptors
- Use overlapping grid of HOG descriptors with the combined feature vector in a conventional Support Vector Machine (SVM) based window classifier
HOG - A Closer Look

First thing to do: Calculate gradients.

- Several derivative masks were tested, with and without smoothing.
- Simple works best; centered, unsmoothed 1D point derivative mask
  - \([-1, 0, 1]\) and \([-1, 0, 1]^T\)
- For color images: Separate gradients, use the one with the largest norm
HOG - A Closer Look

Second step: Gradient binning

- Each pixel throws a weighted vote for an edge orientation histogram channel.
  - based on the gradient element centered here
- Votes are collected into orientation bins over local spatial regions
  - These local spatial regions are hereby known as “cells”
  - Cells are either rectangular or radial (log-polar sectors)
  - Bins are evenly spaced -
    - over 0° - 180°: “unsigned” gradient (best choice for humans)
    - over 0° - 360°: “signed” gradient (best choice for e.g. cars, motorbikes)
  - Votes bilinearly interpolated between bin centres (both orientation and position)
  - Using gradient magnitude as vote give the best result
  - Increase number of orientation bins up about 9 -> increase performance
HOG - A Closer Look

Third step: Descriptor blocks

There are two main ways:

- Use rectangular blocks (R-HOG)
- Use circular blocks (C-HOG)
HOG - A Closer Look

R-HOG

- Similar to SIFT descriptors, though R-HOG’s are optimized for “dense robust coding of spatial form” whilst SIFT’s are optimized for “sparse wide baseline matching.”
- Mostly tested with square blocks (N x N), although vertical (2 x 1) and horizontal (1 x 2) were also tested. 2 x 2 or 3 x 3 cell blocks were found to be best.
- Cells of between 6 x 6 and 8 x 8 are found to be superior to other. Coincidentally, in training and test images human limbs are about 6 - 8 pixels across.
HOG - A Closer Look

C-HOG

- Reminiscent of Shape Contexts, but each spatial cell contain a stack of gradient-weighted orientation cells
- Think of it as advanced form of centre-surround coding
- Log-polar grid originally suggested because
  - it would allow fine coding of nearby structure combined with coarser context.
  - “the fact that the transformation from the visual field to the V1 cortex in primates is logarithmic.”
- Small descriptors with very few radial bins give best performance
- Evaluated two forms, which in practice give same performance:
  - one circular center and
  - one where also center is divided in angular sections.
- Minimum two radial bins (center + surround) and four angular bins required for good performance
- More radial bins: little change. More angular bins: decrease performance
- 4 is the best radius for central bin, increase expansion factor from 2 to 3: little change
HOG - A Closer Look

Fourth step: Normalization

● Block normalization
  ○ Normalize over blocks found to be the best approach
  ○ Four different schemes, here $v$ is the unnormalized description vector, $||v||_k$ is its $k$-norm, $k \in \{0,1\}$ and $e$ is a small constant
    ■ L2-norm: $v \rightarrow v/\sqrt{(||v||_2^2 + e^2)}$
    ■ L2-hys: L2-norm followed by clipping, limit max values of $v$ to 0.2 followed by renormalization
    ■ L1-norm: $v \rightarrow v/(||v||_1 + e)$
    ■ L1-sqrt: L1-norm followed by square root, $v \rightarrow \sqrt{(v/(||v||_1 + e))}$. Same as treating descriptor vectors as probability distributions and taking the Bhattacharyya distance between them

● Centre-surround normalization
  ○ Image tiled with grid of cells
  ○ For each cell, normalize energy based on surrounding cells based on Gaussian weighting
HOG - *Moving to results*

Datasets used in training and testing:

- the well-established MIT pedestrian database
- the new, more challenging INRIA dataset with 1805 64x128 images of humans

Two units of measure chosen

- False Positives Per Window tested; FPPW
- miss rate: $1 - \text{Recall or } \frac{\text{FalseNeg}}{\text{TruePos + FalseNeg}}$

Detection Error Tradeoff (DET) is plotted on a log-log scale, that is FPPW against miss rate. In this scheme, lower values are better.
HOG - *Moving to results*

To compare the results with some other methods:

- **Generalized Haar Wavelets**
  - an extended set of oriented Haar like wavelets
  - features are rectified responses from 9x9 and 12x12 1\textsuperscript{st} and 2\textsuperscript{nd} derivative box filters at 45° intervals and the corresponding 2\textsuperscript{nd} derivative xy filter

- **PCA-SIFT**
  - based on projecting gradient images onto a basis learned from training images using PCA
  - implementation use 16x16 blocks with the same derivative scale, overlap etc. as the HOG descriptors
  - PCA basis calculated using positive training images

- **Shape Contexts**
  - originally use binary edge-presence voting into log-polar spaced bins irrespective of edge orientation
  - simulated using C-HOG descriptor with 1 orientation bin, 16 angular and 3 radial intervals with inner radius 2 pixels and outer radius 8 pixels for best results
HOG - *Moving to results*

Other methods presented by others, where insufficient data is present to give exact comparisons:

- MIT’s best parts based detector
- MIT’s best monolithic detector

Results for these are interpolated
HOG - Results

A detail about color

- Color information is used when available
  - Grayscale reduce performance by 1.5% at $10^{-4}$ FPPW

- Normalization in color space has modest performance impact
  - Perhaps because the later descriptor normalization
  - Square root gamma compression improve performance at low FPPW by 1% at $10^{-4}$ FPPW
  - log compression is too strong and decrease performance by 2% at $10^{-4}$ FPPW
Results - *MIT dataset*

Details:

- **EC-HOG:** binary edge voting C-HOG
- **E-ShapeC:** Shape Context (edge)
- **G-ShapeC:** Shape Context (gradient)

**DET – different descriptors on MIT database**

- Lin. R-HOG
- Lin. C-HOG
- Lin. EC-HOG
- Wavelet
- PCA-SIFT
- Lin. G-ShaceC
- Lin. E-ShaceC
- MIT best (part)
- MIT baseline
Results - *INRIA dataset*

Details:

- Ker. R-HOG: Linear SVM replaced by Gaussian kernel
- R2-HOG: R-HOG with primitive bar detectors
Results

Effects of parameter change
Results - Gradient smoothing

Details:

- c-cor is the 1D cubic-corrected derivative
Results - Orientation bins

DET – effect of number of orientation bins $\beta$

- $\text{bin} = 9$ (0–180)
- $\text{bin} = 6$ (0–180)
- $\text{bin} = 4$ (0–180)
- $\text{bin} = 3$ (0–180)
- $\text{bin} = 18$ (0–360)
- $\text{bin} = 12$ (0–360)
- $\text{bin} = 8$ (0–360)
- $\text{bin} = 6$ (0–360)
Results - Overlapping cells

DET – effect of overlap (cell size=8, num cell = 2x2, wt=0)

- overlap = 3/4, stride = 4
- overlap = 1/2, stride = 8
- overlap = 0, stride = 16

miss rate

false positives per window (FPPW)
Results - *Normalization method*

![Graph showing DET - effect of normalization methods](image-url)
Results - *Detector cue*

The HOG detector cue mainly on silhouette contours

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.
Results - *Window size*

Note:

Normally there is around 16 pixels margin around the person. Decreasing window size affects this
Results - *Kernel width* (kernel SVM, Ker. R-HOG)
Results - Selecting best cell and block size

For R-HOG, given the images in the tested datasets, the choices for cell and block size summarized:

Figure 5. The miss rate at $10^{-4}$ FPPW as the cell and block sizes change. The stride (block overlap) is fixed at half of the block size. 3×3 blocks of 6×6 pixel cells perform best, with 10.4% miss rate.
Conclusion

The novel HOG descriptor performs good, and outperforms the known methods presented in the paper.

The method performs best using “fine-scale gradients, fine orientation binning, relatively coarse spatial binning and high-quality local contrast normalization”.
Questions

Thank you for listening…

Any questions?
Information Outside the Presentation Follows
Help: ask Wikipedia (or follow references in article)

- Rubbish presentation?
  Also: R. C. Gonzales, R. E. Woods - Digital Image Processing
Known Approaches for Pedestrian Detection

- Papageorgiou et. al.: Based on polynomial SVM using Haar wavelets as input descriptors.
  - Improve this later
  - Optimized by Depoortere et. al.
- Gavrila and Philomen: Extraction of edge images and match against learned exemplars using chamfer distance
  - Used in real life
- Viola et. al.: Moving person detector that use AdaBoost to train increasingly complex region rejection rules based pn Haar-like wavelets and space-time differences.
- Ronfard et. al.: Articulated body detector by incorporating SVM based limb classifiers over Gaussian filters.
- Mikolajczyk et. al.: Combinations of orientation-position histograms with binary-thresholded gradient magnitudes to build parts based method for body parts.