IDA

INSTITUTE FOR DEFENSE ANALYSES

Robust Feature Vector for Efficient Human Detection

Amy E. Bell

October 22, 2013
IDA Document
NS D-5058
Log: H 13-001528
Copy

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract DASW01-04-C-0003, for IDA's Information Technology and Systems Division, Task OTSDPB. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

INSTITUTE FOR DEFENSE ANALYSES

IDA NS D-5058

Robust Feature Vector for Efficient Human Detection

Amy E. Bell

Robust Feature Vector for Efficient Human Detection

Amy E. Bell
Institute for Defense Analyses
Alexandria, VA 22311

***Abstract*— This research presents a method for the automatic detection of a dismounted human at long range from a single, highly compressed image. The histogram of oriented gradients (HOG) method provides the feature vector, a support vector machine performs the classification, and the JPEG2000 standard compresses the image. This work presents an understanding of how HOG for human detection holds up as range and compression increases. The results indicate that HOG remains effective even at long distances: the average miss rate and false alarm rate are both kept to 5% for humans only 12 pixels tall and 4-5 pixels wide in uncompressed images. Next, classification performance for humans at close range (100 pixels tall) is evaluated for compressed and uncompressed versions of the same test images. Using a compression ratio of 32:1 (97% of each image's data is discarded and the image is reconstructed from only the 3% retained), the miss rates for the compressed and uncompressed images are equivalent at 0.5% while the 1.0% false alarm rate for the compressed images is only slightly higher than the 0.5% rate for the uncompressed images. Finally, this work depicts good detection performance for humans at long ranges in highly compressed images. Insights into important design issues—for example, the impact of the amount and type of training data needed to achieve this performance—are also discussed.**

Keywords—automatic human detection, histogram of oriented gradients, image compression

I. INTRODUCTION

This work addresses the specific challenge of automatically and efficiently detecting dismounted humans at long ranges from a single image. It is part of the larger, more general, image processing challenge of automated object detection in imagery. Conventional challenges in object detection include: different image formation processes lead to variations in viewpoint and scale, variations in illumination, partial occlusions, intra-class variation, context, background clutter, etc. Here, the goal is to detect members of the human object class (not the detection of particular humans).

A common, two-part approach to object detection is employed in this study: a feature extraction algorithm generates a descriptive feature vector from the image and a trained classifier analyzes the feature vector and decides human or no-human present. The strategy here is to focus on the first part (feature extraction) and maintain a simple, fast classifier. Moreover, the desire is to achieve an efficient approach by deriving the feature vector from a highly compressed image.

A fundamental problem of automated object detection in images is the type of feature representation or descriptors to employ. Histogram of oriented gradients (HOG) is a dense, overlapping grid of features extracted from a single image at one resolution [1]. HOG descriptors are a collection of neighboring, normalized, weighted histograms of gradient orientations; spatial information is implicitly encoded by its position in the feature vector. HOG is robust to small variations in contour locations and directions; HOG is also robust to large changes in illumination and color. A support vector machine (SVM) is a supervised learning technique that can be used for pattern recognition and classification [2]. SVMs have recently been recognized as a classification tool capable of superior performance—oftentimes with simple, fast architectures. JPEG2000 is the image compression standard based on a two-dimensional discrete wavelet (biorthogonal 9/7) transform followed by a specialized arithmetic encoder (EBCOT). The standard allows the user to manipulate many variables to customize compression performance [3].

This study employs a HOG feature vector and a linear SVM classifier to investigate the detection of dismounted humans at long ranges from a single, JPEG2000 compressed image. HOG+SVM has been shown to be effective for detecting humans at close range; consequently, it is desirable to understand its performance at greater distances. The idea is that if this encoding of images into feature vectors is sufficiently discriminative even at long ranges and high compression levels, and if the training data includes the variations that the classifier needs to separate the two object classes (human/no-human), then this *efficient* HOG+SVM system should be able to accurately detect humans even at long ranges.

II. BACKGROUND

Automatic object detection in images receives considerable attention in the image processing and computer vision literature. This brief background focuses on the state-of-the-art image descriptors (i.e. feature vectors) that are most germane to the goal of the current study.

A. Dense Representations: Haar Wavelet

Haar wavelet based object detection methods rely on dense features extracted at several scales and orientations. Dalal used nine first and second order, horizontal, vertical, and diagonal, generalized Haar wavelet filters at two scales (the over-complete representation arises from the overlapping support of the wavelets). Although the wavelet based approach to feature vectors does not exploit the advantages of the sparse representations based on points, wavelets demonstrated the best performance of all the alternative methods compared to HOG [1].

The alternative to dense representations is sparse representations based on local descriptors of identified image regions. These key point descriptors are assembled into feature vectors for classification. The primary advantage of the key point methods is their compact representation—offering speed with lower memory and power requirements. However, because key point detectors are designed for particular objects, they may be limited in their ability to generalize to object classes. Two key point methods that have received attention recently are shape contexts and scale invariant feature transformation (SIFT).

B. *Sparse Representations: Shape Contexts and SIFT*

Shape contexts compute local histograms of image gradients or edges over log-polar grids. The voting into the histograms is performed without regard for edge orientation; consequently, it compares to a histogram with one orientation bin. In Dalal's comparison with HOG, both shape context methods performed worst among the alternative approaches [1].

SIFT also computes local histograms; however, it uses the local scale and dominant orientation given by the key point detector to vote the weighted gradient magnitudes into the orientation histograms. Consequently, the feature vectors are invariant to scale and orientation. Unlike shape contexts, SIFT computes histograms over rectangular grids. For comparison with HOG, Dalal employed a method that is a combination of principal components analysis (PCA) and SIFT, referred to as PCA-SIFT [4]. PCA-SIFT modifies the way in which the SIFT key point detectors are assembled into feature vectors; the result is a more compact representation that provides faster, more accurate performance. In PCA-SIFT, the computed local image gradient is projected onto a pre-computed eigenspace of image gradients—the closest match provides a more compact feature vector than the standard SIFT feature vector. Dalal implemented PCA-SIFT with 16x16 blocks and compared it to HOG using the same scale and overlap (the PCA eigenspace, or basis, was calculated using only the positive training images). PCA-SIFT outperformed shape contexts, but underperformed the wavelet approach—which did not perform as well as HOG [1].

III. EFFICIENT HUMAN DETECTION WITH HOG

A. *Image Compression*

The JPEG2000 image compression standard offers significantly improved performance over the previous JPEG standard [5]. JPEG2000 has proven capable of reconstructing images from compressed data that are perceptually identical to original, high resolution images up to lossy compression ratios of 20:1 (95% of the original information is discarded; the image is reconstructed from only the 5% retained). The resulting storage and transmission power savings are sizeable: at a 32:1 compression ratio, the compressed image size is 3.13% that of the original image (e.g. a 1MB image compresses to 31K). Figure 1 depicts the inclusion of image compression in the automatic classification system. Removing the components circled in blue would result in an inefficient system with no savings in storage or transmission power.

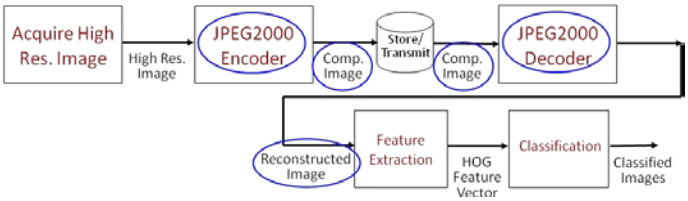

Fig. 1. Image compression for efficient human detection: the components circled in blue are present due to compression. The compressed image can be significantly smaller than the high resolution image, yet look nearly identical, at the expense of the computational cost of encoding and decoding.

However, the inefficient version of the classification system would not incur any encoding/decoding computational costs. In this application, the primary issue to understand is the impact of compression on classification performance.

B. *Histogram of Oriented Gradients (HOG)*

Histogram of oriented gradients (HOG) is a method for automatically detecting people in images [1]. HOG incorporates the dense, overcomplete framework of the wavelet approach and also exploits the local image feature key points of the point detector approaches like shape contexts and SIFT.

HOG builds on these previous methods by proposing a monolithic encoding of dense, overlapping image region features that are derived from the distribution of gradient orientations.

The HOG feature vector of an image is computed as follows.

1) Compute image gradient. For each color channel, convolve the gradient mask, [-1,0,1], with the image in both the vertical and horizontal directions, resulting in dv and dh. Form the gradient vector at each pixel as (dv + i·dh) and compute its gradient magnitude, r, and phase, θ (in degrees). Restrict the phase to only positive values by adding 180 to any negative phase values. For color images, keep only the magnitude and phase of the dominant color channel at each pixel (determined by the largest magnitude).

2) Compute block histogram cubes.

a) Here, the image is divided into cells of size η x η pixels; cells are grouped into blocks of size ζ x ζ cells ($\eta \cdot \zeta$ x $\eta \cdot \zeta$ pixels).

b) For each block, apply a Gaussian window with $\sigma = 0.5 \cdot \eta \cdot \zeta$ to the block gradient magnitudes; this deemphasizes the pixels at the block edges.

c) For each cell in the block, compute a histogram of gradient orientations (i.e. phases) by dividing the [0,180] range into β equal-sized bins. For each cell pixel's gradient phase, determine its two neighboring bin centers, then distribute the pixel's gradient magnitude to the two adjacent bins based on the phase distance from each bin center. If a pixel's phase is to the left of the first bin center, to the right of the last bin center, or exactly equal to a bin center, then the pixel's magnitude will be added to only one of the β histogram bins. Equations (1-3) describe how a pixel magnitude is distributed into two neighboring bins:

$$w = 1 - \left(\frac{\theta - LB}{180 / \beta} \right) \quad (1)$$

$$LB = LB + (r \cdot w) \quad (2)$$

$$UB = UB + r \cdot (1 - w) \quad (3)$$

where w is the distribution weight, LB is the lower neighbor histogram bin center, and UB is the upper neighbor histogram bin center. This approach differs from the original HOG trilinear interpolation for voting the gradient magnitudes into the orientation histograms.

d) Finally, the cell histograms are collected and arranged into block histogram cubes. Blocks are not distinct from one another; instead, they overlap. The stride variable describes the amount of overlap in pixels. Consequently, one cell histogram can make an appearance in multiple block histogram cubes.

3) Block normalization. For each block histogram cube, use the L2 norm to normalize it (independently of all the other blocks). Arrange all of the normalized block histogram cubes into one large HOG feature vector.

C. Efficient HOG+SVM Classification

HOG feature vectors derived from compressed, reconstructed images are input to a support vector machine (SVM) trained for classification of the images: a "positive" classification means that the classifier has decided that a human is present in the image and a "negative" classification is a decision that no human is present. The classifier designed for this study is a linear kernel, one-norm, soft margin SVM [2]; it was implemented in Matlab. The classifier was trained on 1036, 128x64 pixel, positive and negative images in less than one second on a standalone 64-bit desktop machine with a quad core Intel™ Xeon (3.33GHz).

IV. RESULTS

A. Data and Experiments

Image data was derived from the well-known Massachusetts Institute of Technology pedestrian image database [6]. The original 128x64 pixel images were downsampled by a factor of two, four and eight to simulate the effect of the image having twice, four times, and eight times the range of the original image. Figure 2 shows one example of an original image and its factor-2, factor-4, and factor-8 reductions (displayed here at the same size). Table 1 depicts the image size (in pixels), the size of the human height, and an estimated ground sample distance (GSD) based on an average human height of 1.8m and a thigh width of 20cm. The image data contains little variation in: object viewpoint, partial occlusions, context, and background clutter;

(a) (b) (c) (d)

Fig. 2. The positive test image at four scales (displayed at the same size): (a) original size; (b) factor-2 reduction; (c) factor-4 reduction; (d) factor-8 reduction.

TABLE I. OVERALL IMAGE SIZE IN PIXELS; NUMBER OF PIXELS FROM TOP OF PERSON'S HEAD TO BOTTOM OF PERSON'S FEET; CALCULATED GROUND SAMPLE DISTANCE FOR PERSON HEIGHT (BASED ON AN AVERAGE PERSON HEIGHT OF 1.8M); CALCULATED GROUND SAMPLE DISTANCE FOR PERSON THIGH WIDTH (BASED ON AN AVERAGE WIDTH OF 0.2M).

Image	Image Size	Person Height	GSD (Height)	GSD (Width)
Original	128x64 pixels	100 pixels	1.8 cm/pixel	2.0 cm/pixel
Factor-2	64x32 pixels	50 pixels	3.6 cm/pixel	4.0 cm/pixel
Factor-4	32x16 pixels	25 pixels	7.2 cm/pixel	10.0 cm/pixel
Factor-8	16x8 pixels	12 pixels	15.0 cm/pixel	20.0 cm/pixel

however, the data includes significant variations in illumination, scale, and intra-class object appearance.

A rule-of-thumb requirement for stationary target detection is 12 pixels across the target's minimum dimension. If an average human is roughly ½m wide, then the maximum GSD for automatic target recognition of stationary humans is about 4.2cm/pixel. Consequently, it is expected that this HOG+SVM classification system will experience difficulty detecting humans in the factor-4 and factor-8 test images.

The HOG+SVM ATR was first trained with positive (human is present) and negative (no human is present) training images; next, the trained HOG+SVM ATR was tested with positive and negative test images. This train/test procedure was repeated for a variety of conditions—each of which is described below. In all experiments, there is no overlap between the training and test image sets. The positive images include one upright human, centered in the image, from a front or back view, in an urban setting. The negative images are also of urban settings, but no human is present. Figure 3 shows one example of an original negative image and its factor-2, -4, and -8 reductions (displayed here at the same size).

In the computation of the HOG feature vector, $\beta = 9$ histogram bins was used for all experiments. For the original and factor-2 images, the HOG variable settings were: $\eta = 8$, $\zeta = 2$, and the stride was 8 pixels (i.e. 105 blocks for

(a) (b) (c) (d)

Fig. 3. One negative test image (no human present) at four scales (displayed at the same size): (a) original size; (b) factor-2 reduction; (c) factor-4 reduction; (d) factor-8 reduction.

each original image, 21 blocks for each factor-2 image). For the factor-4 images, the HOG variable settings were: $\eta = 4$, $\zeta = 2$, and the stride was 4 pixels (i.e. 21 blocks for each image). For factor-8, the HOG variable settings were: $\eta = 2$, $\zeta = 2$, and stride = 2 pixels (21 blocks for each image).

B. Classification as a Function of Range

The HOG+SVM ATR (without compression) was trained with 518 positive and 518 negative training images at the original scale; next, the HOG+SVM ATR was tested with 200 positive and 200 negative test images at the original scale. This train/test procedure was repeated at the factor-2, factor-4, and factor-8 reduced scales. Figure 4 depicts the detection error tradeoff plot (miss rate vs. false alarms) for the original image and its three scale reductions. The best performing wavelet result, using the original image scale, is shown for the same false alarm rate [1]; its correspondingly higher miss rate is expected. As expected, the miss rate and false alarm rate increase as the image scale is reduced. However, at these ranges, HOG+SVM performance is significantly better than anticipated. For example, the factor-8 image contains a person only 12 pixels high and yet the miss rate is only 5.5% and the false alarm rate is only 4.5%! Although range was not reported for the original images, one could estimate the original image range (with a person height of 100 pixels) to be about 50m. Consequently, the factor-8 image range would be 400m; a 5.5% miss rate and 4.5% false alarm rate for automatic human detection is unprecedented at these distances.

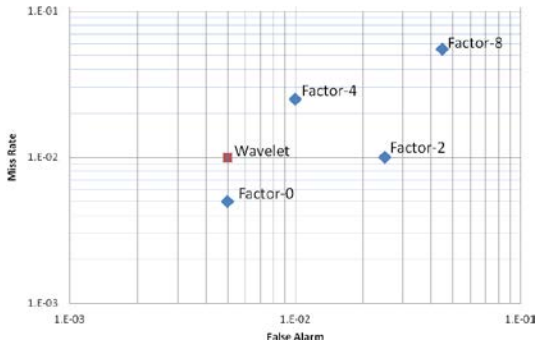

Fig. 4. Detection error tradeoff plot depicts increasing false alarm rate and miss rate as the original image size is reduced (i.e. range is increased). HOG+SVM achieves good performance even at the longest range: factor-8 image (person is approximately 12 pixels high) results in a miss rate of 0.055 and false alarms at 0.045.

Fig. 5. Test image, at original scale: uncompressed (left), compressed at 16:1 (middle), and compressed at 64:1 (right).

C. Classification as a Function of Compression

Next, the *efficient* HOG+SVM ATR (with compression) was trained with 518 positive and 518 negative training images, at the original scale, for four compression ratios: 8:1, 16:1, 32:1, and 64:1. Then testing was performed with 200 positive and 200 negative test images, at the original scale, after compression and subsequent reconstruction (refer to Figure 5 for the perceptual differences due to compression). Table 2 depicts the miss rates and false alarm rates: there is no significant difference in classification performance at the three lowest compression ratios. However, at the most highly compressed level (64:1, in which only 1.5% of the information is retained after compression), both the miss rate and false alarm rate are four times higher than the uncompressed case.

This performance is based on a classifier trained with compressed images. However, when a classifier trained with uncompressed images is then applied to compressed test images, performance suffers even more at the higher compression levels. For example, the false alarm rate for test images compressed at 64:1 jumped to 7.5%.

TABLE II. CLASSIFICATION PERFORMANCE OF THE ORIGINAL SCALE IMAGES, UNCOMPRESSED, AND AT FOUR COMPRESSION LEVELS.

Image Scale	Miss Rate	False Alarm
Uncompressed	0.5%	0.5%
JPEG2000 8:1	0.0%	0.5%
JPEG2000 16:1	1.0%	0.5%
JPEG2000 32:1	0.5%	1.0%
JPEG2000 64:1	2.0%	2.0%

D. Efficient, Long-Range Classification

The original scale images are 128x64 pixels with a person approximately 100 pixels tall. Factor-2 images (64x32, 50 pixel tall human) were compressed using the conventional 5-levels of the discrete wavelet transform in JPEG2000. However, the factor-4 and factor-8 images were too small to compress—even at only 3-levels in JPEG2000. For these images in which the person is only 25 or 12 pixels tall, there isn't enough data in which to isolate and exploit redundancies (the key to effective compression). In other words,

compression is unnecessary for humans at these very far distances.

Since the efficient HOG+SVM ATR performed well at the 32:1 compression ratio for the original scale images, the factor-2 images were examined at this same compression level. For 518 positive and 518 negative training factor-2, compressed images, the miss rate increased to 6.5% and the false alarm rate also rose to 6.0%. Compare these with a 2.0% miss rate and a 3.0% false alarm rate for the factor-2 uncompressed case. Although the classification performance is still acceptable, clearly the combination of range and compression is impacting the results.

The type and amount of training data impacts classification performance; indeed, the results for a classifier trained on uncompressed data, but tested on compressed data, is described above. The more closely the training data resemble the testing data, the better the performance. Moreover, the amount of training data is also an important factor. For example, when the classifier was trained on only 300 positive and 300 negative training, factor-2, compressed images, the miss rate rose even higher to 10.5% and the false alarm rate also increased to 7.0%.

V. CONCLUSIONS AND FUTURE WORK

This investigation revealed that humans can be efficiently detected in a single, highly compressed image using a HOG feature vector and SVM classifier. Classification performance degraded gradually as a function of range and compression. Both the amount and type of training data were demonstrated to impact classification performance. Adapting these methods to more realistic conditions—greater image variation in terms of size, partial occlusion, amount of context, noise, clutter, etc.—presents difficult challenges. However, the performance of these methods, in the conditions presented here, indicate that they are worthwhile candidates for further investigation.

REFERENCES

[1] N. Dalal, "Finding People in Images and Video," Ph.D. dissertation, Institut National Polytechnique de Grenoble, France, 2006.

[2] N. Cristianini, and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, pp. 93-124, 2000.

[3] K. Varma and A. E. Bell, "JPEG2000—Choices and Tradeoffs for Encoders," *IEEE Signal Processing Magazine*, vol. 21, no. 6, pp. 70-75, November 2004.

[4] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," *Proc. of Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, pp. 66-75, 2004.

[5] D. Taubman and M. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice, Springer Publishers, 2002.

[6] Papageorgiou, C. P., Oren, M. and Poggio, T., "A General Framework for Object Detection," Proc. of the 6th International Conference on Computer Vision (ICCV), Bombay, India, 555-562 (1998).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
22-10-13	Conference Paper	

4. TITLE AND SUBTITLE

Robust Feature Vector for Efficient Human Detection

5a. CONTRACT NUMBER

DASW01-04-C-0003

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBERS

6. AUTHOR(S)

Amy E. Bell

5d. PROJECT NUMBER

5e. TASK NUMBER

N/A

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

NS D-5058

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSOR'S / MONITOR'S ACRONYM

N/A

11. SPONSOR'S / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Cleared for open publication.

13. SUPPLEMENTARY NOTES

Project Leader: Amy E. Bell

14. ABSTRACT

This research presents a method for the automatic detection of a dismounted human at long range from a single, highly compressed image. The histogram of oriented gradients (HOG) method provides the feature vector, a support vector machine performs the classification, and the JPEG2000 standard compresses the image. This work presents an understanding of how HOG for human detection holds up as range and compression increases. The results indicate that HOG remains effective even at long distances: the average miss rate and false alarm rate are both kept to 5% for humans only 12 pixels tall and 4-5 pixels wide in uncompressed images. Next, classification performance for humans at close range (100 pixels tall) is evaluated for compressed and uncompressed versions of the same test images. Using a compression ratio of 32:1 (97% of each image's data is discarded and the image is reconstructed from only the 3% retained), the miss rates for the compressed and uncompressed images are equivalent at 0.5% while the 1.0% false alarm rate for the compressed images is only slightly higher than the 0.5% rate for the uncompressed images. Finally, this work depicts good detection performance for humans at long ranges in highly compressed images. Insights into important design issues—for example, the impact of the amount and type of training data needed to achieve this performance—are also discussed.

15. SUBJECT TERMS

automatic human detection, histogram of oriented gradients, image compression

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include Area Code)
Unclassified	Unclassified	Unclassified	Unlimited	12	