HUMAN DETECTION USING SPARSE REPRESENTATION

G. Krishna Vinay [⋆] *S. M. Haque* [◇] *R. Venkatesh Babu* [†] *K. R. Ramakrishnan* [⋆]

[⋆]Department of Electrical Engineering
[◇]Department of Electrical Communication Engineering
[†] Supercomputer Education and Research Centre
Indian Institute of Science, Bangalore
krishnavinay.iisc@gmail.com, mohammadul@gmail.com,
venky@serc.iisc.ernet.in and krr@ee.iisc.ernet.in

ABSTRACT

The problem of human detection is challenging, more so, when faced with adverse conditions such as occlusion and background clutter. This paper addresses the problem of human detection by representing an extracted feature of an image using a sparse linear combination of chosen dictionary atoms. The detection along with the scale finding, is done by using the coefficients obtained from sparse representation. This is of particular interest as we address the problem of scale using a scale-embedded dictionary where the conventional methods detect the object by running the detection window at all scales.

Index Terms— Human Detection, Histogram of Oriented Gradients(HOG), l_1-norm minimization, Sparse representation, Scale-embedded Dictionary

1. INTRODUCTION

Human detection in images is a difficult task owing to variable appearance and wide range of poses that they can adopt. It usually involves two steps, namely, feature extraction and classification. The conventional object detection methods detect the object by scanning the image at several scales. The obvious disadvantage of this approach is that it is computationally expensive. To overcome this, a new object detection algorithm based on scale-embedded dictionary is proposed using sparse representation. Sparse representations [1] represent the test sample as a sparse linear combination of a small number of elementary signals called atoms (positive class atoms, negative class atoms and trivial atoms) of a dictionary ($\boldsymbol{\mathcal{D}}$). The advantage of scale embedded dictionary is that, it reduces the need to run the detector across various scales and hence the computation time for object detection. Also, there is no need of rigorous training as required in learning phase of the conventional classifiers like SVM and AdaBoost. The number of positive and negative class atoms are designed to be less than the number of trivial atoms, leading to a sparse coefficient vector. In the case of occlusion, noise corruption or background clutter, a limited number of trivial coefficients spike, although the coefficient vector remains sparse. The object detection is performed using these coefficients obtained from sparse representation.

The rest of the paper is organized as follows. In the following section we describe the related work on human detection. In section 3, we describe the the proposed method. We discuss the reported results in section 4 along with performance evaluation. Finally we conclude in last section.

2. RELATED WORK

Histogram of Oriented Gradients (HOG) introduced by Dalal and Triggs [2] based on the distribution of intensity gradients uses single detection window approach. Papgeorgiou and Poggio [3] also used single-detection-window approach, with Haar-based features and polynomial SVM for classification. Felzenszwalb *et. al.* [4] used detection by parts and the detection is positive only if the detected parts form a predefined proper human model. Studies with features like edge templates [5], haar features [6] and rectangular differential features [7] etc. and classifiers like SVM [2, 3] and AdaBoost [7], have also been reported.

Recent work on human detection using sparse representation is presented in [8] where the classifier is learned from sparse representation of dense HOG features, but lacks multi-scale feature and also use a high dimensional feature vector. In [9], v-HOG is used with feature dimension of 36 for 4097 blocks and the training is also three step optimization with large training set. The underlying idea of present work evolves from the idea of Wright *et al*., [10]. The novelty of the proposed method lies in integrating HoG features in forming the scale embedded dictionary. Although dictionary based object detection in the context of sparse representation is presented in [11], but here it is used for detection of two different objects occurring simultaneously.

3. OVERVIEW

Initially, p_1 number of clean positive images are selected from MIT pedestrian dataset and p_2 number of negative images are selected randomly from INRIA dataset. These images are resized to a size of 64×32, for computational ease.

3.1. Sparse Representation

Given an over-complete dictionary $\boldsymbol{\mathcal{D}}$ in $\boldsymbol{R}^{m\times p}$ where m is the size of feature vector and p is the number of atoms in the dictionary, the input sample $\mathbf{z}$ in $\boldsymbol{R}^{m\times 1}$ is represented as a sparse linear combination of dictionary atoms *i.e.*, $\mathbf{z} = \boldsymbol{\mathcal{D}}\boldsymbol{\alpha}$. To obtain the sparsest representation, we need to solve the following optimization problem,

$$\min_{\boldsymbol{\alpha}\in\boldsymbol{R}^p} \|\boldsymbol{\alpha}\|_0 \text{ subject to } \mathbf{z} = \boldsymbol{\mathcal{D}}\boldsymbol{\alpha} \tag{1}$$

where $\|\ .\ \|_0$ denotes the l_0-norm, which counts the number of nonzero entries in a vector. But, if the solution $\boldsymbol{\alpha}$ is sparse enough, we can use l_1-norm instead of l_0-norm for convexity. Also, to account

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

for noise, we relax it as, $\mathbf{z} = \boldsymbol{\mathcal{D}}\boldsymbol{\alpha} + \mathbf{n}$ where $\mathbf{n} \in \boldsymbol{R}^m$ is a noise term with bounded energy $\|\mathbf{n}\|_2 \leq \lambda$. The sparse solution $\boldsymbol{\alpha}$ can be approximately recovered by solving the following l_1-norm minimization problem [1]

$$\min_{\boldsymbol{\alpha} \in \boldsymbol{R}^p} \|\boldsymbol{\alpha}\|_1 \text{ subject to } \|\mathbf{z} - \boldsymbol{\mathcal{D}}\boldsymbol{\alpha}\|_2^2 \leq \lambda \tag{2}$$

where λ is a small positive value chosen.

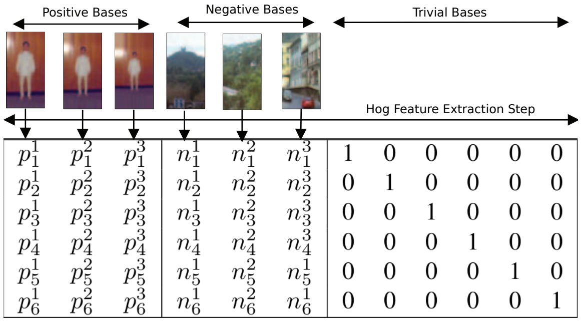
3.2. Feature Extraction

We use HOG features[2] with non-overlapping blocks of size 16×16 to form the dictionary $\boldsymbol{\mathcal{D}}$. For each block in the detection window, 36-dimensional feature vectors are l_1-normalized and concatenated to form the overall feature vector of length 288.

3.3. Dictionary Formulation

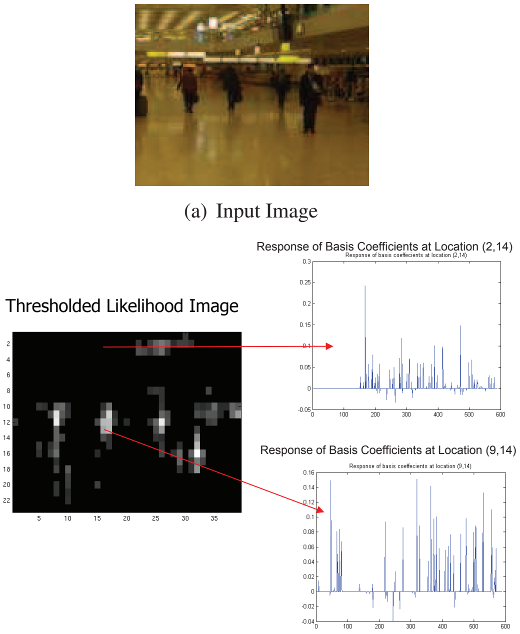
The dictionary is a collection of normalized HOG features extracted from positive class images at different scales, negative class images at original scale and the trivial standard basis atoms, arranged in order. This is called scale embedded dictionary. To obtain the different scales, every positive class image is padded across the borders and then resized to 64×32 which form the first 150 atoms of the dictionary ($\boldsymbol{\mathcal{D}}$). The 150 negative class atoms are extracted from 150 non-human images of size 64×32 and these form the next 150 atoms. Trivial standard basis atoms are included to handle occlusion and clutter.

As an illustration, we consider an example dictionary as in Figure 1. Here, $\mathbf{p}^1$, $\mathbf{p}^2$ and $\mathbf{p}^3$ form the positive class atoms for a single image at three different scales and $\mathbf{n}^1$, $\mathbf{n}^2$ and $\mathbf{n}^3$ form the negative class atoms for three different negative images. The size of feature dimension (m) has been taken as 6, number of positive class atoms (p_1) as 3, number of negative class atoms (p_2) as 3. As the size of feature dimension is 6, the number of trivial atoms should be 6. Hence, the total number of atoms is p where $p = p_1 + p_2 + m$. Here $\mathbf{p}^i$ and $\mathbf{n}^i$ where $i = 1, 2, 3$ form the dictionary atoms for positive class and negative class respectively. The identity matrix appended at the end of negative class atoms forms the trivial atoms.



$$\left[\begin{array}{ccc|ccc|cccccc}
p_1^1 & p_1^2 & p_1^3 & n_1^1 & n_1^2 & n_1^3 & 1 & 0 & 0 & 0 & 0 & 0 \\
p_2^1 & p_2^2 & p_2^3 & n_2^1 & n_2^2 & n_2^3 & 0 & 1 & 0 & 0 & 0 & 0 \\
p_3^1 & p_3^2 & p_3^3 & n_3^1 & n_3^2 & n_3^3 & 0 & 0 & 1 & 0 & 0 & 0 \\
p_4^1 & p_4^2 & p_4^3 & n_4^1 & n_4^2 & n_4^3 & 0 & 0 & 0 & 1 & 0 & 0 \\
p_5^1 & p_5^2 & p_5^3 & n_5^1 & n_5^2 & n_5^1 & 0 & 0 & 0 & 0 & 1 & 0 \\
p_6^1 & p_6^2 & p_6^3 & n_6^1 & n_6^2 & n_6^1 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right]$$

Fig. 1. Example dictionary bases or atoms



(a) Input Image

(b) Likelihood Image

Fig. 2. Detection Result on a big image: The likelihood image gives the confidence measure for the presence of human. The final detection results are obtained after fusing the detections based on detection score and neighbouring detections

3.4. Confidence Measure

For detection, the confidence measure as given in (3) has been used. It is defined as the ratio of sum of coefficients of positive class atoms to sum of coefficients of negative class and trivial atoms and computed as

$$confidence\ value\,(\tau) = \frac{\sum\limits_{i \in 1:p_1} \boldsymbol{\alpha}_i}{\sum\limits_{i \in p_1+1:p} \boldsymbol{\alpha}_i} \tag{3}$$

where p_1 is the number of positive class atoms in the dictionary and p is the total number of atoms. Intuitively, the confidence measure is high for positive class test samples and tends to zero for negative class test samples. This is because when the test sample is positive, the coefficients corresponding to the positive class (numerator) will be higher compared to the case when the testing sample belongs to the negative class. The threshold for confidence measure τ has been fixed experimentally. The value of the scale-factor is determined using the position of the peak coefficients $\boldsymbol{\alpha}$, with respect to the scale-embedded dictionary. Mathematically,

$$scale(\mathbf{z}) = \arg max_i\, \boldsymbol{\alpha}_i \quad \text{where} \quad i \in 1 : p_1 \tag{4}$$

Algorithm

The proposed algorithm is summarized as below.

1. Consider a matrix of training feature atoms in the dictionary

$$\boldsymbol{\mathcal{D}} = [\mathbf{x}^1\ \mathbf{x}^2 \ldots \mathbf{x}^{p_1}\ \mathbf{y}^1\ \mathbf{y}^2 \ldots \mathbf{y}^{p_2}\ \mathbf{I}_{m \times m}] \in R^{m \times p}$$

and a test feature $\mathbf{z} \in \boldsymbol{R}^m$.

2. Normalize the columns of $\boldsymbol{\mathcal{D}}$ to have unit l_2 norm.

3. Solve the l_1 minimization problem

$$\min_{\boldsymbol{\alpha} \in \boldsymbol{R}^p} \|\boldsymbol{\alpha}\|_1 \text{ subject to } \|\mathbf{z} - \boldsymbol{\mathcal{D}}\boldsymbol{\alpha}\|_2^2 \leq \lambda$$

4. Calculate the confidence value using equation 3.
5. The human is detected, if the confidence value is greater than the chosen threshold.
6. If detected, then the scale is determined by the location of maximum magnitude coefficient element in alpha using equation 4

Figure 2 shows a test image and likelihood image obtained by running the detector across the entire image. Figure 2(b)(up) shows the coefficients obtained by running the detector at certain position where human is not present and Figure 2(b)(down) shows the coefficients obtained by running the detector at another position where human is present. It was clear that if human is not present then the positive atoms do not respond at all but if present then both positive and negative atoms respond. Hence the detection can be performed using the confidence value.

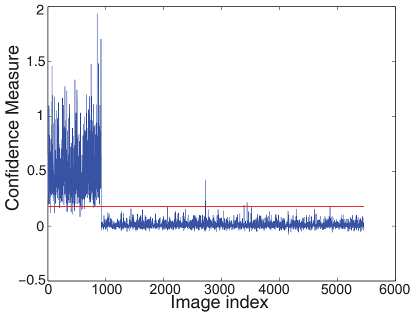


Fig. 3. Confidence measure for positive(924) and negative(4530) class of test images.

4. RESULTS

4.1. Experimental Setup

During training, for dictionary, ten positive images are selectively chosen from MIT pedestrian database[3] and resized to size 64×32 and then 15 images at different scales are generated for each positive image as described in section 3.3. 150 negative images for the dictionary are randomly chosen from INRIA dataset[2] and they are cropped to size 64×32. For testing, 924 positive images are taken from the same MIT pedestrian dataset and resized to 64×32. The negative images for testing are selected from INRIA person dataset and cropped to size 64×32. The HOG features are extracted with non-overlapping blocks.The SPArse Modeling Software [12] has been used for solving the stated l_1 minimization problem.

4.2. Discussion

The proposed algorithm is experimented with different number of positive atoms in dictionary and fixed number of negative and trivial atoms.

4.2.1. Experiment 1

In this experiment, the dictionary contains 150 positive(10×15), 150 negative and 288 trivial atoms. This experiment is performed with two test sets. In case 1, 924 positive and 453 negative images are considered. In case 2, 924 positive and 4530 negative images are considered. The accuracies obtained are 99.79% and 99.94% respectively. Figure 3 shows confidence values for all testing images for the the above cases. The detection performance obtained using SVM are 98.81% and 99.89% respectively, while for the l_1-norm Minimisation Learning [8] 96.10 % and 98.70% respectively.

4.2.2. Experiment 2

In this experiment, the dictionary contains 15 positive from single image, 150 negative and 288 trivial atoms. For the same two test sets as in Experiment 1, the accuracy obtained are 99.13% and 99.5416% respectively. Accuracies using SVM are 46.76% and 89.81 % respectively, while for the l_1-norm Minimisation Learning [8] 86.27% and 93.11% respectively. It is clear that even with features extracted from one image at all scales gives quite a good results. This is because the confusion increases when large number of positive atoms are included.

4.2.3. Experiment 3

In this experiment, the same dictionary as in experiment 1 is used, but the test set is created by randomly resizing the 924 positive images and 453 negative image between 50% and 100% of their full size. The accuracy obtained in our case is 99.57% . For the SVM classifier, the accuracy is 98.47%, while for the l_1-norm Minimisation Learning, the accuracy is 96.19%.
The proposed method is evaluated using Precision-Recall curves and Detection Error Trade off (DET) curves which are shown below in Figures 4, 6, 5 and 7. It is clear that the proposed method (sparse representation by scale embedded dictionary) performs better than SVM classifier and the l_1-norm minimisation learning. The experimental results are shown below.

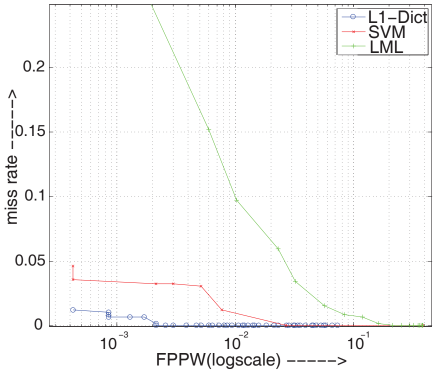


Fig. 4. Miss-rate vs FPPW: Blue legend (L1-Dict) is using scale embedded dictionary, Red legend (SVM) using SVM classifier with the same training and testing features, while the green one (LML) using l_1-norm Minimisation Learning. FPPW are plotted on log scale.

Figure 8 shows one result obtained by proposed algorithm. The algorithm is able to find the scale exactly for all the detected persons.

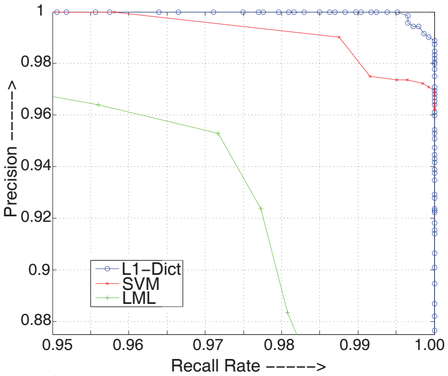


Fig. 5. Precision vs Recall: Blue legend (L1-Dict) is using scale embedded dictionary, Red legend (SVM) using SVM classifier with the same training and testing features, while the green one (LML) using l_1-norm Minimisation Learning.

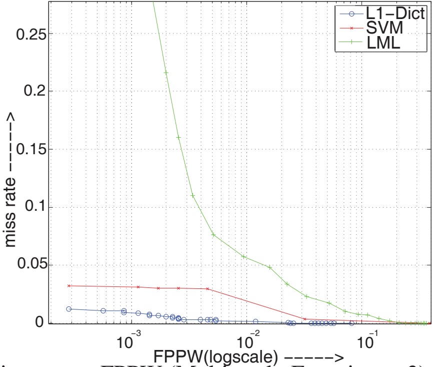


Fig. 6. Miss-rate vs FPPW (Multi-scale-Experiment 3): Blue legend (L1-Dict) is using scale embedded dictionary, Red legend (SVM) using SVM classifier with the same training and testing features, while the green one (LML) using l_1-norm Minimisation Learning. FPPW are plotted on log scale.

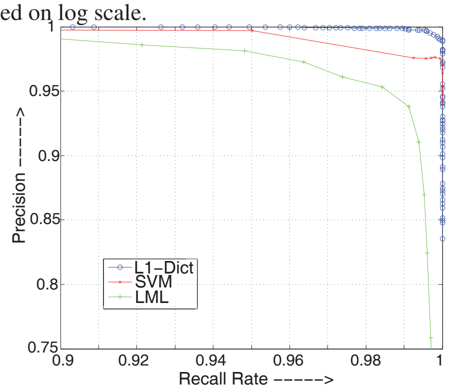


Fig. 7. Precision vs Recall (Multi-scale-Experiment 3): Blue legend (L1-Dict) is using scale embedded dictionary, Red legend (SVM) using SVM classifier with the same training and testing features, while the green one (LML) using l_1-norm Minimisation Learning.

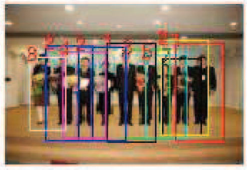


Fig. 8. Detection result using proposed algorithm

5. CONCLUSION

We proposed an efficient method for human detection which addresses the multi-scale detection problem using the theory of sparse representation which is computationally efficient due to multi-scale embedding in one single dictionary. This approach can be used for detection of other object classes likes cars, cycles and bikes *etc.*

6. REFERENCES

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal of Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[2] N. Dalal and B. Triggs, “Histogram of oriented gradients for human detection,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, June 2005.

[3] A. Mohan, C. Papageorgiou, and T. Poggio, “Example based object detection in images by components,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 349–361, 2001.

[4] P. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, September 2010.

[5] D. Gavrila and V. Philomin, “Real-time object detection for smart vehicles,” in *Proceedings of the Seventh IEEE International Computer Vision*, 1999, vol. 1, pp. 87–93.

[6] C. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” in *Proceedings of Sixth International Conference on Computer Vision*, January 1998, pp. 555–562.

[7] P. Viola and M. Jones, “Robust real-time object detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2001.

[8] R. Xu, B. Zhang, Q. Ye, and J. Jiao, “Human detection in images via l1-norm minimization learning,” in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 2010.

[9] R. Xu, B. Zhang, Q. Ye, and J. Jiao, “Cascaded l1-norm minimization learning CLML for human detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 89–96.

[10] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[11] R. Sivalingam, G. Somasundaram, V. Morellas, N. Papanikolopoulos, O. Lotfallah, and Y. Park, “Dictionary learning based object detection and counting in traffic scenes,” in *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, August 2010, pp. 42–48.

[12] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *Journal of Machine Learning Research*, vol. 11, pp. 19–60, March 2010.