Selection of Histograms of Oriented Gradients Features for Pedestrian Detection

Takuya Kobayashi[1], Akinori Hidaka[1], and Takio Kurita[2]

[1] University of Tsukuba,
Tennoudai 1-1-1, Tsukuba, Ibaraki, 305-8577 Japan
{taku-kobayashi,hidaka.akinori}@aist.go.jp
[2] Institute of Advanced Industrial Science and Technology (AIST),
Neuroscience Research Institute,
Umezono 1-1-1, Tsukuba, Ibaraki, 305-5868, Japan
takio-kurita@aist.go.jp

Abstract. Histograms of Oriented Gradients (HOG) is one of the well-known features for object recognition. HOG features are calculated by taking orientation histograms of edge intensity in a local region. N.Dalal *et al.* proposed an object detection algorithm in which HOG features were extracted from all locations of a dense grid on a image region and the combined features are classified by using linear Support Vector Machine (SVM). In this paper, we employ HOG features extracted from all locations of a grid on the image as candidates of the feature vectors. Principal Component Analysis (PCA) is applied to these HOG feature vectors to obtain the score (PCA-HOG) vectors. Then a proper subset of PCA-HOG feature vectors is selected by using Stepwise Forward Selection (SFS) algorithm or Stepwise Backward Selection (SBS) algorithm to improve the generalization performance. The selected PCA-HOG feature vectors are used as an input of linear SVM to classify the given input into pedestrian/non-pedestrian. The improvement of the recognition rates are confirmed through experiments using MIT pedestrian dataset.

1 Introduction

Pedestrian detection in images could be used in video surveillance systems and driver assistance systems. It is more challenging than detecting other object such as faces and cars because appearance of people has lots of fluctuations such as clothing, pose, or illumination.

So far, many algorithms for pedestrian detection have been proposed. For a practical real-time pedestrian detection system, Gavrila [9] employed hierarchical template matching to find pedestrian candidates from incoming images. His method provide multiple templates and they are matched by using Chamfer distance. Papageorgiou *et al.* [7] proposed a pedestrian detection algorithm based on a polynomial SVM using Haar wavelets as input features. Mohan *et al.* [6] extended this algorithm by combining the results of the component detectors. Nishida *et al.* [8] automated the selection process of the components by using

M. Ishikawa et al. (Eds.): ICONIP 2007, Part II, LNCS 4985, pp. 598–607, 2008.

AdaBoost. These researches show that the selection of the components and the combination of them are important to get a good pedestrian detector.

Recently many local descriptors are proposed for object recognition and image retrieval. Mikolajczyk *et al.* [11] compared the performance of the several local descriptors and showed that the best matching results were obtained by the Scale Invariant Feature Transform (SIFT) descriptor [2].

Dalal *et al.* [1] proposed a human detection algorithm using histograms of oriented gradients (HOG) which are similar with the features used in the SIFT descriptor. HOG features are calculated by taking orientation histograms of edge intensity in a local region. They are designed by imitating the visual information processing in the brain and have robustness for local changes of appearances, and position. Dalal *et al.* extracted the HOG features from all locations of a dense grid on a image region and the combined features are classified by using linear SVM. They showed that the grids of HOG descriptors significantly outer-performed existing feature sets for human detection. Ke *et al.* [12] applied Principal Components Analysis (PCA) to reduce the dimensionality of the feature vectors and tested them in an image retrieval application.

On the other hand, Kurita *et al.* [13] showed that the performance of face detection could be improved by selecting a subset of local Gabor features. Viola *et al.* [10] selected Haar-like local features using AdaBoost. These studies show the importance of the selection of a proper subset of the local features in image recognition.

In this paper, we employ HOG features extracted from all locations of a grid on the image as candidates of the feature vectors. Principal Component Analysis (PCA) is applied to the HOG feature vectors to obtain the score vectors. This process reduces the dimensionality of the feature vectors. We call the score vectors PCA-HOG feature vectors. Then a proper subset of PCA-HOG feature vectors is selected by using Stepwise Forward Selection (SFS) algorithm or Stepwise Backward Selection (SBS) algorithm to improve the generalization performance. The selected PCA-HOG feature vectors are used as the input of the linear SVM to classify the given input into pedestrian/non-pedestrian. The improvement of the recognition rates are confirmed through experiments using MIT pedestrian dataset.

In the next section, the proposed algorithm is shown. Then the experimental results are described in section 3. The conclusion and the future works are given in section 4.

2 Feature Selection for Pedestrian Detection

In the works by Dalal *et al.* [1], the HOG features are extracted from all points on a dense grid. In this paper we use the grids of HOG features as the primitive features because they significantly outer perform existing feature sets for human detection as shown in [1]. To improve the recognition performance and reduce the computation cost, we also apply PCA to the HOG features. In this paper, we call them PCA-HOG features.

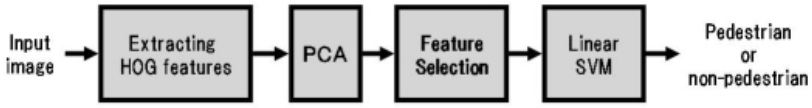


Fig. 1. The overview of our pedestrian detection algorithm. The HOG features are extracted from all locations in an image grid. Then PCA is applied to the extracted HOG features to reduce the dimensionality (PCA-HOG features). A proper subset of these PCA-HOG features are selected to improve the generalization. The selected PCA-HOG features are used as an input vector of the linear SVM.

It is well known that feature selection is effective for pattern classification as shown in [14]. It is expected that the recognition performance can be further improved by selecting a proper subset of the PCA-HOG features because some local regions are irrelevant to pedestrian detection. For example, textures in clothes are not relevant to person detection. Then the selected PCA-HOG features are used as an input vector of linear SVM for person/non-person classification. The overview of our pedestrian detection algorithm is shown in Fig.1.

2.1 Histograms of Oriented Gradients (HOG) Features

Local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge derection. HOG features are calculated by taking orientation histograms of edge intensity in local region. HOG features are used in the SIFT descriptor proposed by Lowe [2]. Mikolajczyk *et al.* reported in [11] that the best matching results were obtained by the SIFT descriptor.

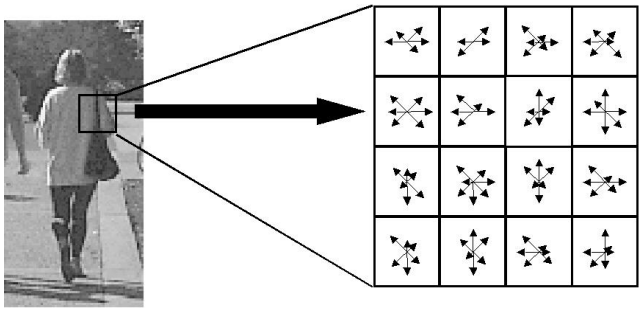


Fig. 2. Extraction Process of HOG features. The HOG features are extracted from local regions with 16×16 pixels. Histograms of edge gradients with 8 orientations are calculated from each of 4×4 local cells. The edge gradients and orientations are obtained by applying Sobel filters. Thus the total number of HOG features becomes $128 = 8 \times (4 \times 4)$.

In this paper, we extract HOG features from 16×16 local regions as shown in Fig.2. At first, edge gradients and orientations are calculated at each pixel in this local region. Sobel filters are used to obtain the edge gradients and orientations. The gradient magnitude $m(x, y)$ and orientation $\theta(x, y)$ are calculated using the x- and y-directional gradients $dx(x, y)$ and $dy(x, y)$ computed by Sobel filter as

$$m(x,y) = \sqrt{dx(x,y)^2 + dy(x,y)^2}$$
$$\theta(x,y) = \begin{cases} \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) - \pi \text{ if } dx(x,y) < 0 \text{ and } dy(x,y) < 0 \\ \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) + \pi \text{ if } dx(x,y) < 0 \text{ and } dy(x,y) > 0 \\ \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) \qquad \text{otherwise} \end{cases} \tag{1}$$

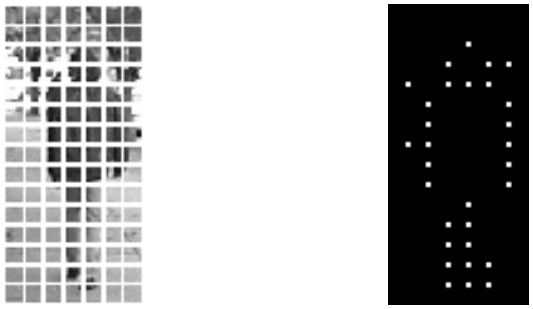
This local region is divided into small spatial area called "cell". The size of the cell is 4×4 pixels. Histograms of edge gradients with 8 orientations are calculated from each of the local cells. Then the total number of HOG features becomes $128 = 8 \times (4 \times 4)$ and they constitute a HOG feature vector. To avoid sudden changes in the descriptor with small changes in the position of the window, and to give less emphasis to gradients that are far from the center of the descriptor, a Gaussian weighting function with σ equal to one half the width of the descriptor window is used to assign a weight to the magnitude of each pixel.

A HOG feature vector represents local shape of an object, having edge information at plural cells. In flatter regions like a ground or a wall of a building, the histogram of the oriented gradients has flatter distribution. On the other hand, in the border between an object and background, one of the elements in the histogram has a large value and it indicates the direction of the edge. Although the images are normalized to position and scale, the positions of important features will not be registered with same grid positions. It is known that HOG features are robust to the local geometric and photometric transformations. If the translations or rotations of the object are much smaller than the local spatial bin size, their effect is small.

Dalal *et al.* [1] extracted a set of HOG feature vectors from all locations in an image grid and are used for classification. In this paper, we extract the HOG features from all locations on a 6×14 grid of a given input image with 56×120 pixels as shown in Fig.3 (a).

2.2 Principal Component Analysis of HOG (PCA-HOG) Features

The total number of features becomes over ten thousands when the HOG features extracted from all locations on the grid. These features are probably too many and are redundant. Ke *et al.* [12] applied Principal Components Analysis (PCA) to reduce the dimensionality of the feature vectors. In this paper, we utilize this idea but we have to take the properties of HOG features into account. The HOG features extracted from regions without edges are not effective for classification because they are based on the information on edges. We have to gather training samples for PCA from effective regions. To select such regions, we use Difference



(a) (b)

Fig. 3. (a) HOG features are extracted from all locations on a 6×14 grid of a given input image with 56×120 pixels. (b) Selected points for PCA.

of Gaussian (DOG). In the SIFT descriptor, the DOG filter is used to detect the key-points for image matching [2]. DOG can be define as the difference of two images smoothed with different Gaussian filters as

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y), \tag{2}$$

where $I(x, y)$ is an input image and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x^2 + y^2)}{2\sigma^2}\right). \tag{3}$$

The absolute value of outputs of the DOG filter became large at the point with large variations. This means that DOG filter has the effect to emphasize the contrast. When we select points from a pedestrian image whose absolute value of DOG filter is greater than a threshold, a silhouette of the human appears as shown Fig.3 (b). The HOG features of the selected points are used as the training samples for PCA.

PCA is one of the well known techniques for dimensionality reduction. It has been applied to several computer vision problems. PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear subspace, known as the principal subspace, such that the variance of the projected samples is maximized. Equivalently, it can be defined as the liner projection that minimizes the mean squared distance between the data points and their projections.

Let $\{\boldsymbol{x}_i | i = 1, \ldots, N\}$ be a set of M-dimensional vectors. This is a given training samples for PCA. Then the principal scores are defined by using the projection matrix U as

$$\boldsymbol{y} = U^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{4}$$

where the mean vector of the training samples are defined as $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$. The optimum projection matrix U is obtained by solving the eigen equations of the covariance matrix Σ

$$\Sigma U = U\Lambda, \quad (UU^T = I) \tag{5}$$

where the covariance matrix Σ is defined as

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T. \tag{6}$$

After the training of PCA, we can compute PCA scores for any HOG features by using the equation (4). We call this new features PCA-HOG features.

To determine the number of principal components, we use the rate of cumulative contribution. In the following experiments, we use the principal scores whose cumulative contribution is less than 90%.

2.3 Selection of PCA-HOG Features

It is well known that the selection of a proper subset of the features can improve the recognition performance in pattern recognition. Kurita *et al.* [13] showed that the performance of face detection could be improved by selecting a subset of local Gabor features. Viola *et al.* [10] selected Haar-like local features using AdaBoost. These studies show the importance of the selections of local features in image recognition. In this paper, we select a subset of locations on a 6×14 grid and PCA-HOG features extracted on the subset of the locations are used as input of the SVM classifier. To evaluate the goodness of the subset, a set of sample images of pedestrians and non-pedestrians is prepared. The goodness of the subset is evaluated by the recognition rate to this evaluation samples.

To find the optimal subset we have to evaluate the all possible combinations of $M = 84 = 6 \times 14$ locations on the grid. But it is not feasible because the number of combinations becomes 2^M. Several sub-optimal methods have been proposed. Two of the simplest methods are Stepwise Forward Selection (SFS) and Stepwise Backward Selection (SBS). SFS starts from the subset with the empty set and repeatedly adds the best feature vector in terms of the goodness of the subset. On the other hand, SBS starts from the set with all feature vectors and repeatedly removes the most unnecessary feature vector in terms of the goodness of the subset.

Let $\{\boldsymbol{y}_i | i = 1, \ldots, M\}$ be a set of PCA-HOG features extracted from all M locations on the grid. In SFS algorithm, the feature vector F is initialized as empty set $F^{(0)} = \emptyset$. Then the best feature vector $\boldsymbol{y}^*$ is searched in terms of the goodness of the subset of feature vectors $F^{(k-1)} + \boldsymbol{y}^*$. The selected feature vector is added to the feature vector as $F^{(k)} = F^{(k-1)} + \boldsymbol{y}^*$. This process is repeated until all feature vectors are included in the feature vector F. Similarly in SBS algorithm, the feature vector F is initialized as $F^{(0)} = \{\boldsymbol{y}_i | i = 1, \ldots, M\}$. Then the best feature vector $\boldsymbol{y}^*$ is searched in terms of the goodness of the subset of feature vectors $F^{(k-1)} - \boldsymbol{y}^*$. The selected feature vector is removed from the feature vector as $F^{(k)} = F^{(k-1)} - \boldsymbol{y}^*$. This process is repeated until no vector is left in the set F.

2.4 Linear SVM Classifier

In the human detection algorithm proposed by Dalal *et al.* [1], the HOG features are extracted from all locations of a dense grid and the combined features are classified by linear Support Vector Machine (SVM). They showed that this HOG features significantly outer-performed existing feature sets for human detection. In this paper, we also use the linear SVM because the dimension of the selected PCA-HOG features is enough high.

SVMs were proposed by Vapnik [5] and have yielded excellent results in various data classification tasks. Let $\{\boldsymbol{f}_i, t_i\}_{i=1}^N$ ($\boldsymbol{f}_i \in R^D, t_i \in \{-1, 1\}$) be the given training samples in D-dimensional feature space. The classification function is given as

$$z = \mathrm{sign}(\boldsymbol{w}^T \boldsymbol{f}_i - h) \tag{7}$$

where $\boldsymbol{w}$ and h are the parameters of the model. For the case of soft-margin SVM, the optimal parameters are obtained by minimizing

$$L(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N} \xi_i \tag{8}$$

under the constraints

$$\xi_i \geq 0,\ t_i(\boldsymbol{w}^T \boldsymbol{f}_i - h) \geq 1 - \xi_i\ (i = 1, \ldots, N) \tag{9}$$

where $\xi_i (\geq 0)$ is the error of the i-th sample measured from the separating hyperplane and C is the hyper-parameter which controls the weight between the errors and the margin. The dual problem of Eq.(8) is obtained by introducing Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N), \alpha_k \geq 0$ as

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j t_i t_j \boldsymbol{f}_i^T \boldsymbol{f}_j \tag{10}$$

under the constraints

$$\sum_{i=1}^{N} \alpha_i t_i = 0, \quad 0 \leq \alpha_i\ \ (i = 1, ..., N). \tag{11}$$

By solving Eq.(10), the optimum function is obtained as

$$z = \mathrm{sign}(\sum_{i \in S} \alpha_i^* t_i \boldsymbol{f}_i^T \boldsymbol{f} - h^*) \tag{12}$$

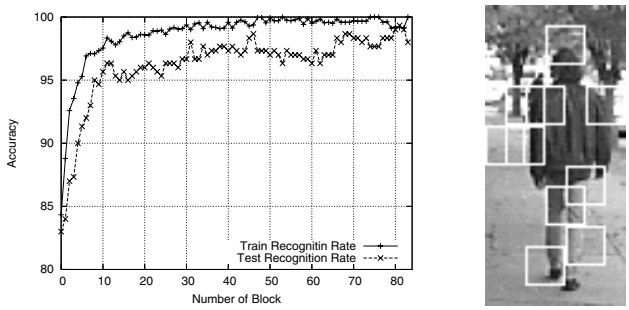
where S is the set of support vectors.

To get a good classifier, we have to search the best hyper-parameter C. The cross-validation is used to measure the goodness of the linear SVM classifier.

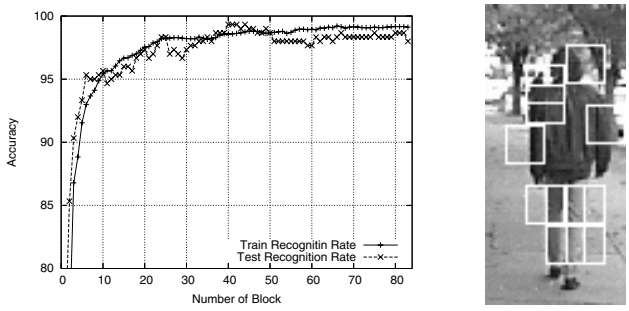
3 Experiments

The proposed algorithm was evaluated by using MIT CBCL pedestrian database which contains 924 images of pedestrians in city scenes [18]. It contains only front or back views with relatively limited range of poses and the position and the hight of human in the image are almost adjusted. The size of the image is 64×128 pixels. These images were used for positive samples in the following experiments. The negative samples were originally collected from images of sky, mountain, airplane, building, etc. The number of negative images is 2000. From these images, 800 pedestrian images and 1600 negative samples were used as training samples to determine the parameters of the linear SVM. The remaining 100 pedestrian images and 200 negative samples were used as test samples to evaluate the recognition performance of the constructed classifier. When we implemented Dalal algorithm using that dataset, the recognition rate for test dataset is 98.3%. We applied PCA and feature selection to improve its result.

PCA-HOG feature vectors were extracted from all locations of the grid for each training sample. Then subsets of PCA-HOG feature vectors were selected by using SFS or SBS algorithms. The selected feature vectors were used as input of the linear SVM. The goodness of the selected subsets were evaluated by cross



(a) Results obtained by SFS algorithm.



(b) Results obtained by SBS algorithm.

Fig. 4. Experimental results on feature selecting. The left graphs show the relation between the recognition rates and the number of selected PCA-HOG feature vectors. The Right images represent locations of the selected features.

Fig. 5. Examples of pedestrian detection. In upper image pedestrians are correctly detected. Lower images show example of false detection.

validation. Also we evaluated the recognition rates of the constructed classifier using test samples.

The left graphs of Fig.4 show the relation between the recognition rates and the number of selected PCA-HOG feature vectors. The graph in Fig.4 (a) was obtained by using SFS algorithm. Similarly SBS algorithm was used to obtain the graph in Fig.4 (b).

When we used SFS algorithm, the best recognition rate 99.3 % for test dataset was obtained at 82 PCA-HOG feature vectors. This is 1.3% better than the recognition rate obtained by using all 84 feature vectors. When SFS and SBS algorithms were compared, SBS algorithm gave better results. The best recognition rate 99.3 % was obtained at 41 PCA-HOG feature vectors. This means that we can reduce the number of features less than half.

The white squares in the right images of Fig4 show the 10 locations of the selected feature vectors by using SFS or SBS. It is noticed that the proposed algorithm succeeded to select some of reasonable regions such as the head, the shoulder, the leg, the arms, etc.Again SBS algorithm seems give better results.

Finally we applied to proposed algorithm to detect pedestrians in images of INRIA person dataset [1]. The final detector for this experiment was produced to retrain using an augmented dataset (initial 2400 sample + 1161 false positive samples). The results are shown in Fig5.

4 Conclusion

We evaluated the effect of the selection of PCA-HOG feature vectors for pedestrian detection.As a result, we could reduce the number of features less than half without lowering the performance.

References

1. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005)
2. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. IJCV 60(2), 91–110 (2004)
3. Swain, M.J., Ballard, D.H.: Color Indexing. Int'l j. Computer Vision 7(1), 11–32 (1991)
4. Daugman, J.: Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. Trans. on Biomedical Engineering 36(1), 107–114 (1989)
5. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons, Chichester (1998)
6. Mohan, A., Papageorgiou, C., Poggio, T.: Example-Based Object Detection in Images by Components. PAMI 23(4), 349–361 (2001)
7. Papageorgiou, C., Oren, M., Poggio, T.: A General Framework for Object Detection. In: Proc. Int'l Conf. Computer Vision (January 1998)
8. Nishida, K., Kurita, T.: Boosting Soft-Margin SVM with Feature Selection for Pedestrian Detection. In: Proc. of International Workshop on Multiple Classifier Systems, vol. 13, pp. 22–31 (2005)
9. Gavrila, D.M.: Pedestrian Detection form a Moving Vehicle. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 37–49. Springer, Heidelberg (2000)
10. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: Proc of the 9th International Conf. of Computer Vision, Nice, vol. 1, pp. 734–741 (2003)
11. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: Proc. of Computer Vision and Pattern Recognition (2003)
12. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proc. of Computer Vision and Pattern Recognition, Washington, pp. 66–75 (2004)
13. Kurita, T., Hotta, K., Mishima, T.: Feature ordering by cross validation for face detection. In: Proc. of IAPR Workshop on Machine Vision Applications, The University of Tokyo, Japan, November 28-30, pp. 211–214 (2000)
14. Tanaka, K., Kurita, T., Meyer, F., Berthouze, L., Kawabe, T.: Stepwise feature selection by cross validation for EEG-based Brain Computer Interface. In: Proc. of Inter. Joint Conf. on Neural Networks, Vancouver, July 16-21, pp. 9422–9427 (2006)
15. Wu, B., Nevatia, R.: Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors
16. Zhu, Q., Avidan, S., Yeh, M.-C., Cheng, K.-T.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: CVPR 2006 (2006)
17. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2001)
18. MIT CBCL: `http://cbcl.mit.edu/software-datasets/PedestrianData.html`