HUMAN DETECTION IN IMAGES VIA L1-NORM MINIMIZATION LEARNING

Ran Xu[1], Baochang Zhang[2], Qixiang Ye[1], Jianbin Jiao[1]
[1]Graduate School of Chinese Academy of Sciences, Beijing, China
[2]School of Automation Science and Electrical Engineering, Beihang University, Beijing,China
+Corresponding Author: Fax: +86-10-88256278, Email: jiaojb@gucas.ac.cn

ABSTRACT

In recent years, sparse representation originating from signal compressed sensing theory has attracted increasing interest in computer vision research community. However, to our best knowledge, no previous work utilizes L1-norm minimization for human detection. In this paper we develop a novel human detection system based on L1-norm Minimization Learning (LML) method. The method is on the observation that a human object can be represented by a few features from a large feature set (sparse representation). And the sparse representation can be learned from the training samples by exploiting the L1-norm Minimization principle, which can also be called feature selection procedure. This procedure enables the feature representation more concise and more adaptive to object occlusion and deformation. After that a classifier is constructed by linearly weighting features and comparing the result with a calculated threshold. Experiments on two datasets validate the effectiveness and efficiency of the proposed method.

Index Terms—Human detection, L1-norm, feature selection, sparse representation

1. INTRODUCTION

Feature representation and classifier are two basic elements in a typical object detection algorithm. In the aspect of the feature representation, various global and local methods are widely investigated on human detection.

In [1], the global shape-based features are exploited for body detection, the classification rule behind which is actually based on the Chamfer distance. Compared to global ones, the local features achieved much more attention in recent years. In [2] the well-known overlapped and dense local descriptor, histogram of oriented gradient (HOG), is introduced for feature representation and trained by a SVM classifier. Serre et al [3] utilize the cortex features for object contour representation using the multi-scale features of Gabor filters. In [4], the co-variance feature is recently proposed and classified on a Riemannian manifolds and achieves reasonable performance. Mu et al. [5], employ improved LBP features, which have good tolerance to color variance, for human detection. In addition, some researchers detect human parts and combine these features to form the overall human model [6-9]. Although these features have succeeded in some detection tasks by fusing with various classifiers, feature selection process, which can further improve the representation effectiveness and efficiency, is not fully investigated.

For the issue of constructing the classifier for human detection, popular methods are SVM, Adaboost, etc. Mohan et al. [10] adopt silhouette information to representing human, exploiting SVM for final classification. Viola et al. [11] employ Adaboost for face and human classification based on the Haar-like features. In [12], individual detectors based on the Shapelet features are trained for each part using AdaBoost. However, in accordance with above methods, SVM is a little complex and not very effective for reducing time consuming. And Adaboost needs extensive time to adjust every weak learner as the number of samples and dimension of feature increase [11] and extremely depends on large training set.

The proposed method in this paper is an effective way to extract the compact feature representation, meanwhile designing a linear classifier in a harmonious way for human detection via L1-norm minimization. Sparse representation using L1 minimization has been widely applied in to the field on compression of signals [13-14]. And it has been successfully used in the filed of face recognition [15]. Intuition lies in that the sparse representation is naturally discriminative by L1-norm minimization which selects the subset most compactly expressing the input signals. To verify the performance of the proposed method, we exploit the simple HOG descriptors to extract features. We firstly compute blocks of HOG features on training samples and use L1-minimization to obtain weight and the sparse representation. Then, we design a simple but effective linear classifier on these weighted features. It is also investigated that the proposed method is robust to the occlusion and multi-posture to some extent.

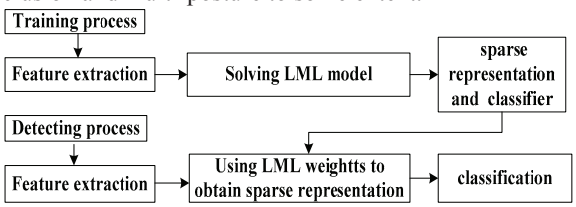


Fig.1. Framework of the proposed method.

The framework of the learning procedure is described in Fig.1. In the training procedure, we firstly perform feature extraction, and then use LML to calculate sparse representation, meanwhile on which the classifier for human detection is built. During the detection procedure, a linear classifier achieved in training procedure is used to classify whether the input object is human or not.

The rest of the paper is organized as follows. In section 2, we describe the LML method in details. Experiments are presented in section 3 and conclusions in section 4.

2. L1-MINIMIZATION LEARNING (LML) METHOD

This section presents the LML method with feature extraction, sparse representation and human detection.

2.1. Feature extraction

The feature extraction is based on the well-known R-HOG descriptor, which is a local contour representation of objects. A 64x128 image is divided into blocks with size of 16x16, which consist of 2x2 cells with size of 8x8. Gradient orientations of pixels in a cell are projected into discrete 9 orientation feature bin. Each block contains a 36 dimension concatenated vector of all its cells. And finally 3780 dimension features are extracted and normalized. Details of the feature extraction procedure can be referred to [2].

2.2. Sparse representation

A compact object representation plays a significant role on designing an efficient pattern recognition system. In our human detection approach, we use L1-norm Minimization Learning (LML) to learn sparse representation from a group of dense HOG features. LML aims to find a subset of dominant features with large weights, which are also incorporated into the final classifier construction. This is formulated as an optimization model as

$$\min \quad \| w \|_1 \tag{1}$$

$$\text{s.t. } y_i \cdot h_w(x_i) \geq \alpha, \quad i = 1, \ldots, N \tag{2}$$

where $\|\bullet\|_1$ denotes L1-Norm, Eq.2 is the constraint to Eq.1, which ensures that training samples should be correctly classified. $h_w(x_i) = w^T x_i$, $w \in R^n$ is the feature weights, x_i is the feature vector of i^{th} sample, $x_i \in R^n$ and y_i is the class label of i^{th} sample, $y_i \in \{-1, 1\}$. N is training sample number. α together with different y_i can guarantee that the shortest distance of different classes is 2α .

The optimization model is a disciplined convex program. It is known that L1-norm is not differentiable, which make it difficult to be solved with a direct method. There is, however, a simple and relatively common transformation that allows this problem to be solved effectively.

We introduce vectors, $u \in R^n, v \in R^n$ and make the substitution $w = u - v$, $u \geq 0$, $v \geq 0$. These relationships are satisfied by $u^j = (w^j)_+$ and $v^j = (-w^j)_+$ $j = 1, 2....n$, j denotes dimension of feature vector, where $(\bullet)_+$ denotes the *positive-part operator* defined as $(w^j)_+ = \max\{0, w^j\}$ .We thus have $\| w \|_1 = I_n^T u + I_n^T v$, where $I_n = [1, 1, 1, \ldots 1, 1]^T$ is a n-dimension unit vector. And (1) and (2) can be rewritten as the following disciplined convex program model:

$$\min \quad I_n^T u + I_n^T v \tag{3}$$

$$\text{s.t. } \begin{cases} y_i \cdot (u - v)^T x_i \geq \alpha \\ u \geq 0 \\ v \geq 0 \end{cases} \tag{4}$$

where u and v are two new variables of the model. The optimization model shown in Eq.3 and Eq.4 is equivalent to Eq. 1 and Eq. 2, and we can solve the new model using the Interior Point method. Details converting the optimization model refer to [14].

When solving the above model, training samples are iteratively inserted. If there are some samples that can't be classified with a linear classifier, they will be removed and some new samples are re-selected and inserted. In this process, it is ensured that most of samples can be optimized. It is not necessary that all of the training samples participate in optimization, which is similar with SVM training process.

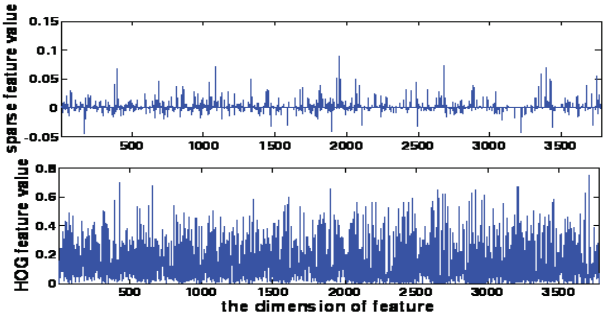


Fig.2. The comparison between sparse representation and the original dense feature representation for positives.

After solving Eq.3-4, we obtain weights and the weighted feature vector of sample. When setting a threshold at 1×10^{-3} , about 75% weighted feature values are less than the given threshold .The other 25% are dominant features which can be regarded as a spare representation. Fig.2 shows the comparison between sparse representation by LML and the HOG feature representation for human samples. The sparse representation via LML is intuitively

robust to occlusion and multi-posture to some extent. Since the weighted feature vector is at least 75% sparse by experimental observation. When the occlusion or multi-posture appears in the sparse weights, they have little influence on human body representation. In other words, sparse representation mitigates the effect of the occlusion or multi-posture. Therefore, the LML method is insensitive to occlusion and multi-posture to some extent.

2.3. Classifier for human detection

We choose linear method to construct classifier in our system. According to Occam's razor theory, a simple classifier, like linear one, can be more effective for object classification. One can also refer to the state-of-art classifiers, such as SVM, Adaboost, to understand the principle of building an effective classifier. In addition, the efficiency of linear classifiers is higher.

In the LML procedure, given the learned sparse representation, a final classifier $g(x)$ is designed as follows: $g(x) = sign(h_w(x) - \theta) = sign(\sum_{j=1}^{n} w^j x^j - \theta)$ (5)

where x^j denotes j^{th} dimension feature of test sample x, $x \in R^n$, $j = 1, 2, ..n$. θ is the threshold value calculated by the linear combination of $h_w(x_i)$ between nearest training positives and training negatives , that is

$$\theta = \eta \left(\min_{x_i \in positives} h_w(x_i) \right) + (1-\eta) \left(\max_{x_i \in negatives} h_w(x_i) \right) \quad (6)$$

where $i = 1, 2...N$, $\eta \in [0,1]$. And as a summary the LML method is described as follows.

Input: Training-set $\{(x_i, y_i)\}_{i=1..N}$, $y_i \in \{-1, 1\}$

- Solve optimization model as shown in Eq. (3) and Eq.(4), set $\alpha = 1.0$
- **Repeat** for $k = 1...K$, $K \le N/2$
  - Randomly select a positive sample of feature vector x_k^+ and a negative sample of feature x_k^- and insert them into the model.
  - Repeat for $l = 1...L$,
    $u_{l+1} = u_l + \lambda_l d_l$, $v_{l+1} = v_l + \gamma_l t_l$,where λ_l, γ_l are iteration steps. d_l, t_l are iteration directions.
  - Remove trained samples x_k^+ and x_k^-
- Obtaining weight vector: $w = u - v$

Output: The final classifier is

$$g(x) = sign(h_w(x) - \theta) = sign(\sum_{j=1}^{n} w^j x^j - \theta)$$

Fig.3. Learning procedure of the LML method.

When conducting human detection, we classify the image window by window in multi-scales by the learned classifier $g(x)$. And all positives in all scales are regarded as the detection results.

3. EXPERIMENTS

There are more than 3000 training positives and about 3700 negatives from MIT and SDL [16] for frontal view. We choose the training positives nearly coming from frontal view and the obtained model can handle multi-posture and occlusion problem which experiment results demonstrate. We perform the experiments on two different test sets. One is our SDL human test set with 59 images [16]. Another is the challenging INRIA test set of 288 images [2]. In this test set, humans are mostly in standing position, but it covers more diverse body poses and complex backgrounds in comparison to the SDL set.

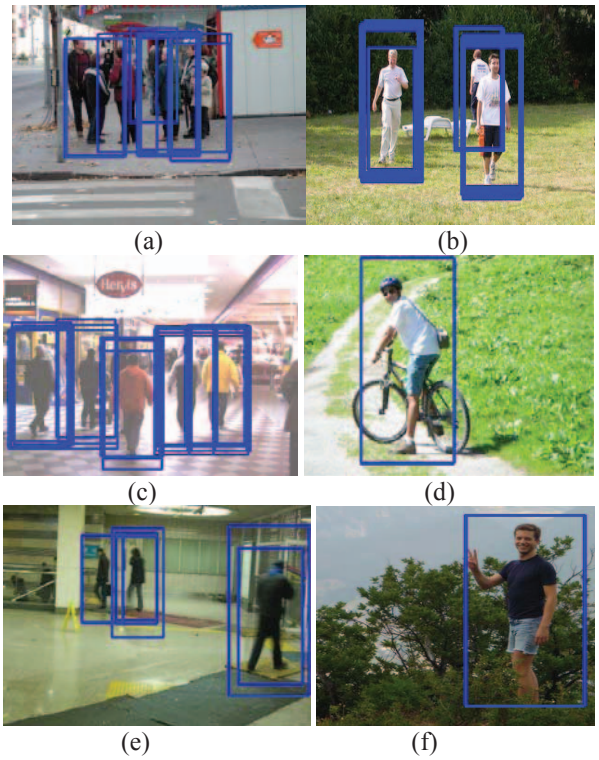
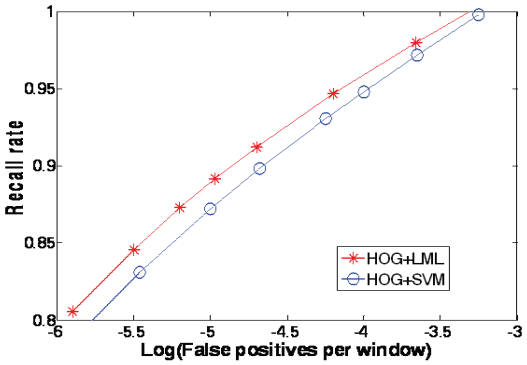


Fig.4 Detection examples, without multi-scale integration.

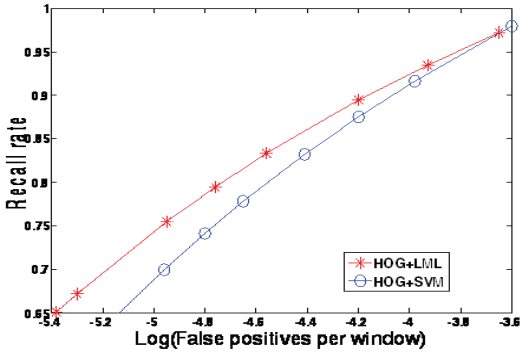
When learning the classifier, we empirically determine $\eta = 0.87$ and $\theta = 1.1$ in Eq.(6) to pursuit a good tradeoff between recall rate and False Positives Per Window(FPPW). In Fig.4, we show some detection examples. From Fig.4a to Fig.4c, all humans are correctly located in spite of some occlusion. In Fig.4a, the person in black jacket occluded can be detected correctly. In Fig.4c, the third person from the left of picture, who is occluded, also can be found. In Fig. 4d to Fig.4f, note that examples covers the subject's

unusual pose (e.g. riding a bicycle) which are also be correctly detected. Fig.4b and Fig.4e are from SDL test dataset, others from INRIA dataset.

Recall rate and False Positives Per Window (FPPW) are used to quantitatively evaluate our method and compared it with SVM method. It is defined as a correct detection if the overlapping between the predicted region and the ground-truth region is more than 90 percent. Fig.5a shows results on SDL test set, and Fig.5b on INRIA test set. It can be seen on Fig.5a-b that the proposed method outperforms the SVM classifier on both of the two test sets.



(a)



(b)

Fig.5. Performance and comparisons. (a) SDL frontal test set [16], (b) on INRIA test set [2]

Efficiency is also compared with SVM classifier. It takes about 8 seconds to process a 320x240 scale-space image with C++ on Pentium IV 3.0 GHZ CPU with our method (without program optimization and do not use Integral image when calculating the HOG features). This speed is 6~10 times faster than that of SVM in the same context.

4. CONCLUSIONS AND FUTURE WORKS

We propose the LML method integrating feature selection with classifier construction for a pedestrian detection in images, which achieves a much better performance than the existing detection method both on speed and accuracy. Furthermore, the proposed method is robust to occlusion and multi-posture to some extent. Both the experiment results and theory support our conclusion.

In the future work, we will justify the performance of proposed method by comparing it with more representative methods and applying it into other objects e.g. vehicles etc.

5. ACKNOWLEDGMENT

This work is supported by Major State Basic Research Development Program of China (973 Program) with No. 2010CB731800, National Natural Science Foundation of China with No.60872143, No.60903065. This work is supported by a grant from the Ph.D. Programs Foundation of Ministry of Education of China (No.20091102120001)

6. REFERENCES

[1] Gavrila, D.M. and Giebel, J. "Shape-based pedestrian detection and tracking," *IEEE Intelligent Vehicle Symposium*, vol.1, pp.8-14, 2002.

[2] Dalal, N, Triggs, B., "Histograms of Oriented Gradients for Human Detection," *IEEE CVPR*, vol.1, pp.886-893, 2005.

[3] Serre, T., Wolf L., Bileschi S., Riesenhuber M. and Poggio T.., "Object Recognition with Cortex-like Mechanisms," *IEEE Transactions on PAMI*, vol.29(3), pp.411-426, 2007.

[4] Tuzel O., Porikli F., Meer P., "Pedestrian detection via Classification on Riemannian manifolds," *IEEE Transactions on PAMI,* vol. 30(10), pp.1713-1727, 2008.

[5] Mu Y. Yan S. Liu Y., Huang T., Zhou B., "Discriminative Local Binary Patterns for Human Detection in Personal Album," *IEEE CVPR*, issue 23-28, pp.1-8, 2008

[6] P.F. Felzenszwalb and D.P. Huttenlocher. "Pictorial structures for object recognition," *IJCV*, vol. 61(1), pp.55–79, 2005.

[7] S. Ioffe and D.A. Forsyth. "Probabilistic methods for finding people," *IJCV*,vol. 43(1), pp.45–68, 2001.

[8] B. Leibe, E. Seemann, and B. Schiele. "Pedestrian detection in crowded scenes," *IEEE CVPR*, vol. 1, pp.878-885, 2005.

[9]D. Vinay, J. Neumann, V. Ramesh, and L.S. Davis, "Bilattice-based logical reasoning for human detection," *IEEE CVPR* , 2007.

[10]A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Transactions on PAMI*, vol.23(4) pp.349-360, 2001

[11] P.Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE CVPR*, 2001.

[12] Sabzmeydani P. and Mori G., "Detecting Pedestrians by Learning Shapelet Features," *IEEE CVPR*, 2007.

[13] K Huang, SAviyente, "Sparse representation for signal classification," *Advances in Neural Information Processing Systems, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, 2007

[14] A.T. Mario, D. Nowak, J.Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems," *IEEE Selected Topics in Signal Processing*, vol.1(4),pp.586-597,2007

[15] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on PAMI*, vol.31(2), 2009.

[16] http://coe.gucas.ac.cn/SDL-Homepage/resource.html