NONLINEAR L1-NORM MINIMIZATION LEARNING FOR HUMAN DETECTION

Ran Xu, Jianbin Jiao+, Qixiang Ye

Graduate University of Chinese Academy of Sciences, Beijing, China
+Corresponding Author: Fax: +86-10-88256278, Email: jiaojb@gucas.ac.cn

ABSTRACT

View, appearance and pose variations make it difficult to detect human objects only by using linear classification methods. Inspired by the successful applications of L1-norm minimization learning (LML) for human detection, we propose a new nonlinear L1-norm minimization learning method (NL-LML). It integrates a nonlinear transformation with an LML optimization model for human detection. The NL-LML method first maps the samples into a space based on the kernel function, and then combines the reformulated samples in the transformed space with the LML model to learn a classifier. Histograms of orientated gradient (HOG) features are used as the feature descriptors, and the sliding window scheme is adopted to detect humans in images. Experiments on two human datasets validate the efficiency and effectiveness of the proposed method.

Index Terms— Human detection, L1-norm minimization, Nonlinear classification, Kernel function

1. INTRODUCTION

Detecting humans in images is a very challenging task owing to the various views, appearances and poses of a human body, together with cluttered background under different illumination. A robust solution to this problem has extensive applications, such as video surveillance, image retrieval and some driver assistant systems etc.

In the design of a typical human detection algorithm, feature representation and classification model are two basic aspects to be considered. In this paper, our work focuses more on developing an effective classification method.

In recent years, the L1-norm has been employed to solve some important problems of image processing and computer vision. Its successful applications in the fields of compressed sensing of signals [9, 20] and face recognition [10] show its powerful ability to cope with problems. In addition, L1-norm minimization learning (LML) has been proposed to design linear classification methods for human detection [12-13], aiming to achieve the sparseness of features and simultaneously perform classification from the perspective of directly minimizing VC-dimension. The sparseness, which highlights the difference among features, can be viewed as an effective feature selection scheme [13].

On the success of the LML, we employ it to construct a nonlinear method in this paper. Kernel technique--without engendering a high computational cost--has become a powerful tool to generalize classification ability. Furthermore, the application of kernel functions in Support Vector Machine (SVM) method has shown its ability of coping with nonlinear classification. However, the kernel inner product (kernelization), adopted by the SVM method, does not exist in the dual programming of the LML model. Therefore, it is infeasible to directly perform kernelization in the LML model to reach a nonlinear classification.

In this paper, the proposed NL-LML method combines a nonlinear transformation induced by kernel functions with the LML model. Although the transformation is also dependent on the kernel function, it is essentially different from the kernelization technique used by SVM. The nonlinear classification and feature selection ability of the NL-LML method make it competitive to human object detection.

Related work on human detection. Most state-of-the-art human detection systems need to extract the discriminative features from available image data and apply efficient classification techniques.

In the aspect of feature representation, various features are proposed to represent a human body. The Haar-like wavelet features [8, 16], HOG features [1], Local Binary Pattern (LBP) features [4], Covariance (COV) features [3], Shape Context descriptors [19], the combination of LBP and HOG features [14], and Edgelet features [15] have been employed as descriptors. Lately, many variants of HOG are presented in [2, 21].

For the issue of designing a classifier for human detection, popular methods are SVM and Adaboost, etc. Mohan et al. [7] use SVM to classify a human body. Viola et al. [8] employ the Adaboost to detect pedestrians. In [5], the authors combine local and global cues via a probabilistic segmentation. In [6], the logical reasoning method is used to discriminate whether a region covers a human body. In [1, 4, 7] linear or kernel SVM is utilized for classification. In [2], the authors use linear SVM to train weak classifiers and then build an Adaboost cascade mechanism for human detection. In [17], Partial Least Squares (PLS) method is introduced to cope with the features in a high dimensional space and SVM is used to classify a human body. In [12-13], the LML and a cascaded LML are proposed to obtain a

sparse representation and make classification for human detection.

2. NONLINEAR L1-NORM MINIMIZATION LEARNING

2.1. The LML optimization model

The LML optimization model aims to learn a sparse representation from a large amount of dense features. Furthermore, previous work has shown that the LML model pursues the VC-dimension minimization and further guarantees minimization of the upper bound on test error [12-13]. The LML optimization model is formulated as follows:

$$\begin{aligned} &\min_{w,\xi_i} && \| w \|_1 + C_1 \sum_{i=1}^{N} \xi_i \\ &s.t. && \begin{cases} y_i \cdot h_w(x_i) \geq \alpha - \xi_i, \\ h_w(x_i) = w^T \cdot x_i \\ \xi_i \geq 0, \qquad i = 1, \ldots\ldots, N \end{cases} \end{aligned} \tag{1}$$

where $\| w \|_1 = \sum_{j=1}^{n} |w_j|$ denotes the L1-norm and w_j is the jth dimension of the weight vector w. C_1 is a predefined penalty parameter to balance the minimization of the misclassification degree and the weight vector. ξ_i is a slack variable that is used to measure the misclassification degree of the ith training sample. x_i is the transformed feature vector of the ith sample and y_i is its class label. N is the training sample number. In addition, α is a predefined parameter, which can guarantee the separability between the positives and negatives together with ξ_i.

2.2. NL-LML

According to the form of Model (1), the LML optimization model can be converted into a linear programming problem [9]. The dual form of the linear programming is also a linear one. Therefore, an elegant inner-product and kernelization do not exist in the dual programming of LML model due to the property of the L1-norm. In other words, it is infeasible to directly substitute the inner-product into the nonlinear kernel functions in the LML optimization to achieve the nonlinear classification.

The inspiration of the nonlinear transformation based on the kernel function derives from the work of [18], in which a kernel function is viewed as a mapping to a low-dimensional space. Meanwhile, this mapping can guarantee that the samples are linearly separable at a smaller margin, with a probabilistic error upper-bound in the transformed space.

Following their work, we present a nonlinear method by combining the nonlinear transformation with the LML model. First, the samples are transformed into a new feature space induced by the kernel function. To eliminate the high correlation and redundancy of transformed features, it is necessary to find the principle components of the kernel matrix. Then, the samples are projected onto the principle components to obtain a new representation. It equals that each sample is reformulated in the new feature space. These reformulated samples are optimized by the LML model to solve the weight vector of the classifier.

Mathematically, the feature vectors form a training set $S = \{x_1, x_2 \cdots x_N\}$ and a nonlinear mapping is constructed:

$$\begin{aligned} &\phi : x \to R^N \\ &\phi(x) = (K(x_1, x), K(x_2, x) \cdots K(x_N, x))^T \end{aligned} \tag{2}$$

where $K(\cdot,\cdot)$ is an arbitrary nonlinear kernel function. Generally, the form of kernel function between the sample x_i and the sample x_j determines the elements of kernel matrix. A general form of a kernel matrix can be expressed as follow:

$$K = (K_{ij}) = \begin{pmatrix} K(x_1, x_1), K(x_1, x_2), \cdots K(x_1, x_N) \\ K(x_2, x_1), K(x_2, x_2), \cdots K(x_2, x_N) \\ \vdots \\ K(x_N, x_1), K(x_N, x_2), \cdots K(x_N, x_N) \end{pmatrix} \tag{3}$$

The matrix V consists of the top d normalized eigenvectors of K, which are also the principle components of the matrix K. Then, we can obtain a new representation of x_i by projecting the sample onto the matrix V. To be more specific, the new representation is like this:

$$\Psi(\phi(x)) = V^T \cdot \phi(x) \tag{4}$$

After these operations, we use the LML optimization model to solve a weight vector in the new feature space. The process of solving LML model is equivalent to finding the sparse representation of reformulated samples. The threshold θ of the classifier is determined by using the min-max penalty function model in [13]. The final classifier is as follows:

$$g(x) = sign(h_w(x) - \theta) = sign(\sum_{j=1}^{d} w_j \cdot \Psi(\phi(x)) - \theta) \tag{5}$$

3. EXPERIMENTS

The NL-LML method is mainly designed from the perspective of pure classification techniques, without concerning the other complex detection techniques, such as the cascade mechanism. Hence, the main experiments are compared with some similar methods, such as Linear and Kernel SVMs. To demonstrate the performance of this

method, we carry out experiments on the INRIA human dataset [1] and the Pri-SDL human dataset [11] respectively.

3.1. Feature representation

We use the simple and effective HOG descriptors as the original features to represent a human body. The feature extraction is based on the well-known R-HOG descriptor, which captures a local contour of an object. A 64x128 training image is divided into blocks of size 16x16, which consist of 2x2 cells of size 8x8. Gradient orientations of pixels in a cell are projected onto discrete 9-orientation feature bins. Each block contains a 36-dimension concatenated vector of all its cells. Finally, a 3780-dimension grayscale feature vector is extracted and normalized. Details of the feature extraction procedure can be found in [1].

3.2. Evaluation and comparison

There are more than 1300 training positives from MIT and Pri-SDL [11] for frontal view. Our negative training set consists of about 3000 images from the INRIA big training pictures. We perform the experiments on two test sets. One is the challenging the INRIA dataset with 288 images [1], in which humans are mostly in a standing position while it also covers some diverse body poses and complex backgrounds. The other is our Pri-SDL human dataset with 140 images [11], which is also challenging owing to incorporating the view variations of humans.

During the training, the predefined penalty parameter C_1 is set between [30, 60], which is related to the range of the feature vector value. In Model (1), α is set to 1.0. The Radial Base function (RBF) is chosen to form the kernel matrix. When the parameter of RBF σ ranges from 0.01 to 10, we have found 10 is better for the human detection with respect to training error. The Singular Value Decomposition (SVD) is employed to obtain the principle components of the kernel matrix. The dimension of the transformed space d is determined by the ratio of the maximum eigenvalue to the minimum eigenvalue. d is chosen empirically when the ratio ranges from 10 to 100. This equals that as few as possible eigenvectors are selected on the condition of meeting the positive definite property of RBF kernel matrix.

When conducting human detection, we classify the image with a sliding window approach in multi-scales by the learned classifier $g(x)$. Recall Rate and False Positives Per Window (FPPW) are used to quantitatively evaluate the NL-LML method (see the Eq.(6) and (7)).

It is defined as a correct detection if the overlapping between the predicted region and the ground-truth region is more than 50 percent, which is the criterion of [1]. We implement SVM methods by using the open source codes LibSVM. Fig.1 shows the results on the INRIA dataset, and Fig.2 on the Pri-SDL one. It can be seen that the proposed NL-LML method outperforms the LML method [12], Linear SVM and kernel SVM on both of the two datasets. Although the NL-LML method is not as fast as the Kernel SVM method, it achieves a better result for human detection owing to the sparseness. We will improve the NL-LML method in terms of speed.

$$\text{Recall Rate} = \frac{\#\,RightPositiveDetections}{\#\,TotalPositive} \tag{6}$$

$$FPPW = \frac{\#\,FalsePositiveDetections}{\#\,Total\ \text{Im}\,ageWindows} \tag{7}$$

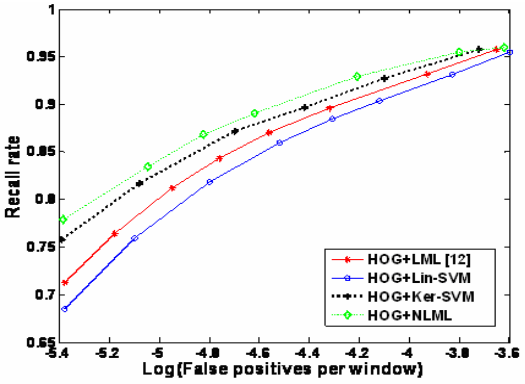

Fig.1 Performance and comparisons on INRIA dataset

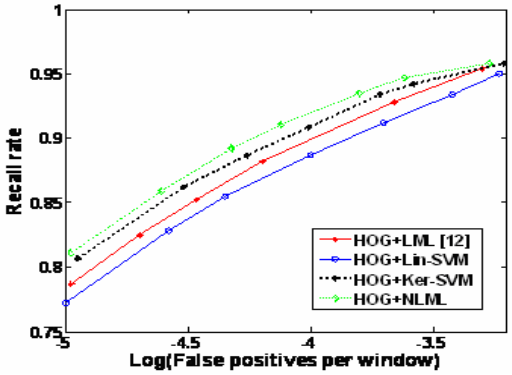

Fig.2 Performance and comparisons on Pri-SDL dataset

3.3 Detection examples

In Figure 3, we show some detection examples without merging results from multi-scales. From Fig.3 (a) to Fig.3 (f), most humans are correctly located in spite of the variation of background, occlusion and views. In Fig.3 (a) all of the people are correctly located in spite of their multi-views. In Fig.3 (b), the example covers the subject's unusual pose (i.e. riding a bicycle), which is also correctly detected. In Fig.3 (c), the girl with the green T-shirt is occluded by the front car and can be found via our method. In Fig.3 (f), there are some false detection windows covering the part of a human body and a baby carriage. The

false detections may be brought out by the complex structure of the objects, which can be avoided by integrating more powerful feature representations in the future work.

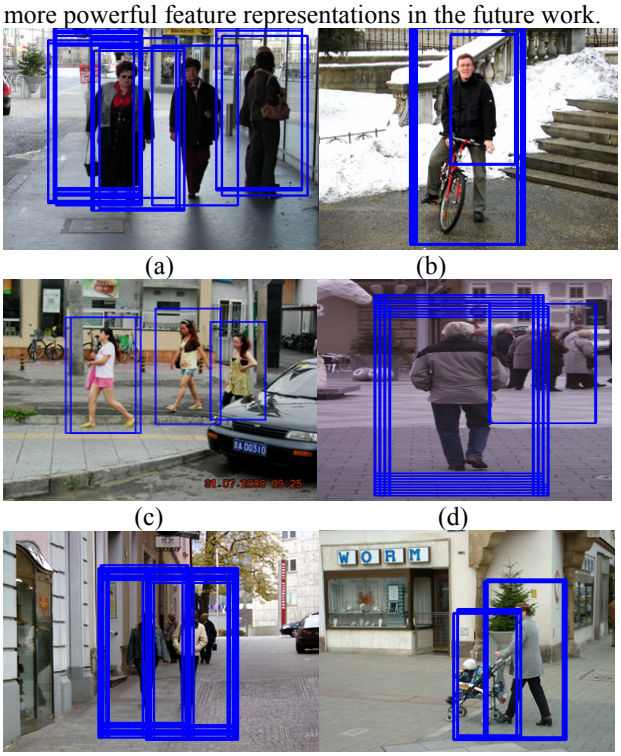

Fig.3 Detection examples, without multi-scale integration.

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a new nonlinear method for human detection in images, which integrates feature selection in transformed space with classifier construction to achieve a nonlinear classification. Experiments validate the effectiveness of the NL-LML approach. In the future, the cascade mechanism and other detail techniques, such as segmentation, will be added into our NL-LML method to attain higher efficiency. Furthermore, we will justify the performance of the proposed method by comparing it with more representative methods and applying it to other objects, e.g., vehicles.

5. ACKNOWLEDGMENT

This work is supported by National Basic Research Program of China (973 Program) with Nos. 2011CB706900, 2010CB731800, and National Science Foundation of China with Nos. 60872143 and 61039003.

6. REFERENCES

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

[2]. Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng, "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients," in Proc. IEEE CVPR, vol. 2, pp.1491-1498, 2006.

[3]. O. Tuzel, F. Porikli, P.Meer, "Pedestrian Detection via Classification on Riemannian Manifolds," IEEE Trans. on PAMI, vol. 30(10), pp.1713-1727, 2008.

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

[5]. B. Leibe, E. Seemann, and B. Schiele. "Pedestrian Detection in Crowded Scenes," in Proc. IEEE CVPR, vol. 1, pp.878-885, 2005.

[6]. D. Vinay, J. Neumann, V. Ramesh, and L.S. Davis, "Bilattice-Based Logical Reasoning for Human Detection," Proc. IEEE CVPR, 2007.

[7]. A. Mohan, C. Papageorgiou, and T. Poggio, "Example-Based Object Detection in Images By Components," IEEE Trans. on PAMI, vol. 23(4) pp.349-360, 2001

[8]. P. Viola, M. Jones, and D. Snow, 2005, "Detecting Pedestrians Using Patterns of Motion and Appearance," IJCV, vol. 63(2), pp. 153-161.

[9]. 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

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

[11]. http://coe.gucas.ac.cn/SDL-Homepage/resource.asp

[12]. R. Xu, B. Zhang, Q. Ye, J. Jiao, "Human Detection in Images via L1-norm Minimization Learning", IEEE International Conference on Acoustics, Speech and Signal Processing, pp.3566-3569, 2010.

[13]. R. Xu, B. Zhang, Q. Ye, J. Jiao, "Cascaded L1-norm Minimization Learning (CLML) Classifier for Human Detection", in Proc. IEEE CVPR, pp.89-96, 2010.

[14]. X. Wang, T. Han, S. Yan, "An HOG-LBP Human Detector with Partial Occlusion Handling," in Proc. IEEE ICCV, Kyoto, 2009.

[15]. B. Wu, and R. Nevatia. "Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors," in Proc. IEEE ICCV, 2005.

[16]. C. Papageorgiou and T. Poggio, "A Trainable System for Object Detection," IJCV, vol. 38, pp. 15-33, 2000.

[17]. W. R. Schwartz , A. Kembhavi , D. Harwood and L. S. Davis, "Human Detection Using Partial Least Squares Analysis," in Proc. IEEE ICCV, 2009.

[18]. M. Balcan, A. Blum, and S. Vempala, "Kernels as Features: On kernels, Margins, and Low-dimensional Mappings", Machine Learning Journal, 65(1):79 94, 2004.

[19]. M. Andriluka, S. Roth, B. Schiele., "Pictorial Structures Revisited: People Detection and Articulated Pose Estimation," in Proc. IEEE CVPR, 2009.

[20]. D. L. Donoho, "Compressed Sensing," IEEE Transactions on Information Theory, vol. 52(4), pp: 1289-1306, 2006.

[21]. L. Zhang, B. Wu, and R. Nevatia, "Detection and Tracking of Multiple Humans with Extensive Pose Articulation," in Proc. IEEE CVPR, 2007.