An Accurate Framework for Arbitrary View Pedestrian Detection in Images

Y Fan¹*, G Wen¹ and S Qiu¹
¹Science and Technology on Automatic Target Recognition Laboratory (ATR), National University of Defense Technology, Changsha, China

*E-mail: fanyaxiang@126.com

Abstract. We consider the problem of detect pedestrian under from images collected under various viewpoints. This paper utilizes a novel framework called locality-constrained affine subspace coding (LASC). Firstly, the positive training samples are clustered into similar entities which represent similar viewpoint. Then Principal Component Analysis (PCA) is used to obtain the shared feature of each viewpoint. Finally, the samples that can be reconstructed by linear approximation using their top- \( k \) nearest shared feature with a small error are regarded as a correct detection. No negative samples are required for our method. Histograms of oriented gradient (HOG) features are used as the feature descriptors, and the sliding window scheme is adopted to detect humans in images. The proposed method exploits the sparse property of intrinsic information and the correlations among the multiple-views samples. Experimental results on the INRIA and SDL human datasets show that the proposed method achieves a higher performance than the state-of-the-art methods in form of effect and efficiency.

1. Introduction

Inspired by robot vision, intelligent transportation system, security monitoring and other applications, detecting object from the image and video is a hot topic in computer vision research. As one of the typical sub-problem, because of its different applications, pedestrian detection has attracted a surge of interest from both academia and industry. Despite the fact that the research of pedestrian detection has made significant progress, where the camera has a fixed viewpoint and captures a static background, it is still far away to solve this problem under more complex challenge such as viewpoint changes.

Most of the current methods regard pedestrian detection as a pedestrian/non-pedestrian two categories of classification problem. Current methods contain two steps [1]: feature extraction and classification. In [2], the histograms of oriented gradients (HOG) descriptors are proposed as the feature extraction method and linear SVM is used as the classification. However, the continuous change viewpoints of pedestrians constitute a Manifold structure, which resulting in poor detection performance for the traditional linear classifier. To deal with this problem, cascade classifiers [3][4] are proposed. In [3], an effective and efficient feature selection method based on Gentle Adaboost (GAB) cascade and the Four Direction Feature (FDF) are proposed. As a result, the irrelevant and redundant features are removed to improve the performance. Ye et al. [4] combined cascade Adaboost classifier and linear SVM for multi-view pedestrian detection. The first four layers of the classifier used the Adaboost and the last layer used the linear SVM. These methods have achieved good results, but the relevant information among the various viewpoints is ignored. In addition, the viewpoint of the
test sample is not limited to the training set of several discrete viewpoints in practical applications, and one of the most difficult problems is to detect the unseen view samples. Joint multi-view model [5] is proposed to address this issue by establishing a certain relationship among the various viewpoints. Another way to deal with this problem is to use the shared feature (or called joint feature) [6][7] by finding common features that can be shared across the viewpoints. In this way, the latent information among different viewpoints is utilized and the compute cost is reduced.

Recently, by combining the advantages of both sparse representation based classification (SRC) and joint sparse representation based classification (JSRC), Zhang et al. [8] proposed a mixed-norm sparse representation model for multi-view face recognition. Their work assumed that the sample of unseen view can be regarded as the sparse representation of samples from all viewpoints. However, this method tried to use the samples of all the viewpoints to identify the arbitrary view sample, which ignored the shared feature among the training sample. In light of this, we propose a new method for detecting pedestrian under arbitrary view from images collected under various views based on the locality-constrained affine subspace coding (LASC) [9]. In the training stage, we first cluster the positive training samples into similar entities according to the viewpoints. Then the Principal Component Analysis (PCA) is utilized to learn the shared features of each viewpoint. In the testing stage, sliding window scheme is adopted to find the correct detection window that can be reconstructed by linear approximation using their top- \( k \) nearest shared features with a small error. No negative samples are required for our method. We conduct experiments on two widely available public datasets. The experiment results demonstrate the superior performance of the proposed approach compared to several state-of-the-art algorithms.

The remainder of this paper is organized as follows. Section II is a detailed presentation of the proposed algorithm. Experimental results and discussions are given in Section III. In the end, Section IV concludes this paper.

2. Method

An overview of the proposed method is illustrated in Figure 1. First, the positive training samples are clustered into several clusters according to the similarity of viewpoints and PCA is employed to obtain the subspace of each cluster which referred to the shared feature. Then, given the query window, the reconstruction cost is computed by locality-constrained affine subspace coding (LASC) with the shared feature, and then, the detection is determined with the small reconstruction cost.

2.1. Feature Extracting

In our work, HOG features [2] are extracted to represent pedestrian. Due to the ability to characterize the edge feature and insensitivity to light and small offset, HOG features have achieved excellent results in pedestrian detection task. A \( 64 \times 128 \) training image is divided into \( 16 \times 16 \) overlapped image blocks, which contain \( 2 \times 2 \) cells of size \( 8 \times 8 \) pixels. In each small cell, gradient is calculated according to calculating the nine-bin histogram of gradient orientations in the range of \( 0^\circ \) to \( 360^\circ \).
a result, each sample is represented by 105 blocks and a 3780-dimensional feature vector is extracted and normalized.

2.2. Clustering
Suppose the positive training samples can be represented as \( D = [d_1, d_2, \ldots, d_l] \in \mathbb{R}^{m \times l} \) where \( m=3780 \) is the dimension of feature vector and \( l \) is the number of training samples. The training sample contains various viewpoints and it can be clustered into \( n \) similar entities \( C = \{C_1, C_2, \ldots, C_n\} \) based on the similarity of viewpoint, which can be performed by employing K-means clustering algorithm as follow

\[
\arg \min \sum_{i=1}^{n} \sum_{d \in C_i} \|d - \mu_i\|^2 \tag{1}
\]

where \( \mu_i \) is the mean vector of cluster \( C_i \).

2.3. Shared Feature Obtaining
After clustering the positive training samples into several similar viewpoint entities, we try to find the shared feature of each viewpoint. It is assumed that each cluster has the geometry of a low-dimensional linear subspace, which is referred to the shared feature. Principal Component Analysis (PCA) is utilized to preserve the directions that have larger variances for each cluster. Suppose that \( u_{i,t} \) is the \( t \)-th orthogonal of eigenvalues of covariance matrix \( \Sigma_i \), then the shared feature of cluster \( C_i \) can be represented as:

\[
A_i = U_i = [u_{i,1}, \ldots, u_{i,p}] \tag{2}
\]

where \( i=1,2,\ldots,n \) and \( p \) is the dimension of the shared feature.

2.4. Pedestrian Detection via Locality-constrained Affine Subspace Coding
With the shared feature mentioned above, we present the detail of pedestrian detection utilizing locality-constrained affine subspace coding (LASC). LASC [9] is an improved model from the locality-constrained linear coding (LLC) method [10], which leverages locality constraints to achieve both high efficiency and local smooth sparsity. The LASC scheme focuses on the local geometric structure of various viewpoints samples and uses top- \( k \) nearest shared feature to encode the test sample. It is obvious that a normal pedestrian of arbitrary view can be linearly represented by its neighbouring subspaces of various viewpoints samples while the reconstruction error is small. In contrast, the negative sample is different from the normal pedestrian of arbitrary view. As a result, it is often difficult to find the linear representation of the neighbouring subspaces around the negative sample.

In detail, suppose \( A \in \mathbb{R}^{n \times p} \) is the linear subspace projection matrix, the geometry of the feature space \( D = [d_1, d_2, \ldots, d_l] \in \mathbb{R}^{m \times l} \) can be represented as a new local coordinate scheme: \( S_i = \{d_i + A \beta_i, \beta_i \in \mathbb{R}^p\} \), \( i=1,\ldots,n \). With a test sample \( y \in \mathbb{R}^m \), LASC attempts to represent it by making a linear combination of the top- \( k \) nearest affine subspaces which is referred to the shared feature. Specifically, LASC can be formulated based on the following optimization:

\[
\min_{\forall \beta_i} \sum_{S_i \in \mathcal{N}_k(y)} \left\{ \|y - d_i\|^2 + \lambda \text{dis}(y, S_i) \right\} \tag{3}
\]

where \( \mathcal{N}_k(y) \) is the neighbourhood of \( y \), which is defined by the \( k \) closest shared features of \( y \). Here, \( \text{dis}(y, S_i) \) is the approximate distance of \( y \) to the shared feature \( S_i \). Use the Gaussian-
distribution as a kernel to measure the similarity between the point $y$ and cluster centroid $\mu_i$, dis($y, S_i$) can be represented as:

$$
\text{dis}(y, S_i) = \frac{\exp\left(-\frac{1}{2\sigma^2}\|y - \mu_i\|^2\right)}{\sum_{j=1}^{c}\exp\left(-\frac{1}{2\sigma^2}\|y - \mu_j\|^2\right)}
$$

(4)

where $\mu_i$ is the mean vector of the cluster $C_i$.

According to [9], the solution of (3) is as follows:

$$
\beta_i = (A_i^T A_i + \lambda \text{dis}(y, S_i) I)^{-1} A_i^T (y - d_i)
$$

(5)

where $I$ is the unit matrix of order $p$. If $S_i \notin N^c_i(y)$, then $\beta_i$ is equal to zero. The code result of $y$ can be represented as $\hat{\beta} = [\beta_1^T, \ldots, \beta_i^T, \ldots, \beta_n^T]^T$.

Then we detect the pedestrian that is based on the reconstruction error by

$$
LRC(y) = \sum_{S_i \in N^c_i(y)} \| (y - d_i) - A_i \beta_i \|^2
$$

(6)

where $\hat{\beta}_i$ is the solution of (3) according to (4) and (5), $A_i$ is obtained from (2). Here, $y$ is detected as a correct detection if the following criterion is satisfied:

$$
LRC(y) > \hat{\epsilon}
$$

(7)

where $\hat{\epsilon}$ is the threshold that decides the sensitivity of the detection method.

3. Experiments

In this section, we describe the datasets used in our experiments, and then compare the detection results with the state-of-the-arts on these datasets. Details are given as follows.

3.1. Datasets and Evaluation

We collected 8000 positive samples of different viewpoints for training from INRIA dataset [1], SDL dataset [11] and Multi-view Pedestrian Recognition (MVPR) dataset [12]. All the samples are trimmed to size of $64 \times 128$. Some of training samples of different viewpoints are shown in Figure 2.

![Figure 2. Training samples of different viewpoints](image)

The INRIA test set and the SDL test set are chose for testing. For the INRIA test set, there are 288 images with 589 multi-view pedestrian samples. And for the SDL test set, there are 258 images with 1688 pedestrian samples for testing, in which pedestrians are almost situated in multi-view appearance. Both of the two datasets contain a variety of viewpoints, to demonstrate that our method could detect the pedestrian of arbitrary view.
An overlap threshold [1] for determining a correct detection for can be represented as

\[
\frac{\text{area}(BB_{det} \cap BB_{ann})}{\text{area}(BB_{det} \cup BB_{ann})} \geq \theta ,
\]

where \(BB_{det}\) and \(BB_{ann}\) stand for the detected bounding box and annotated bounding box, respectively. \(\theta\) is often set to 0.5.

Miss rate against False Positives Per Window (FPPW) is chose for evaluation, which are defined as

\[
\text{MissRate} = \frac{\#\text{Missed positive detection}}{\#\text{Total positives}} \quad \text{and} \quad \text{FPPW} = \frac{\#\text{False positive detections}}{\#\text{Total image windows}}.
\]

3.2. Experiments result

The number \(k\) of shared feature and the shared feature dimension \(p\) were set empirically to 3 and 64, respectively. The parameter \(\sigma\) in (4) and the regularization parameter \(\lambda\) in (5) are set to 0.1 and 1, respectively. We first experimented with the INRIA dataset in order to evaluate the effect of the number \(n\) of cluster centroids. In detail, we set \(n = 4, 6, 8, 10, 12\) and check the experiment result test on the INRIA test set. The results are shown in Figure 3 (a), given different parameter settings in terms of Miss rate against False Positives Per Window (FPPW). The performance increases with the number of cluster centroids when the latter is less than 8. This is due to the fact that if the number of cluster centroids is too few, the samples of different viewpoints are confused. Equally, performance drops quickly with increasing number of cluster centroids when the latter is beyond 8. Here the reason is that more cluster centroids are more likely to produce fewer characteristic shared features. Based on the above, we chose \(n = 8\) as the parameters for the succeeding experiments.

![Graphs showing performance results](image)

Figure 3. The parameters analysis and comparison of ours method with recent state-of-the-art methods.

When evaluating the detection performance on the INRIA Dataset, four algorithms specifically designed for Multi-View pedestrian-detection are chose for comparable, including HOG+SVM [2], v-HOG+CAadaboost [13], HOG-LBP+MVPPE method [14] and v-HOG+CLML [15]. The results of the contrast methods are obtained in their respective papers. It can be observed in Figure 3 (b) that the Miss rate decreases with the increase if the False Positive Per-Window (FPPW), and our method achieves superior performance compared with other four algorithms in all situations. In detail, our method achieves 3.5% miss rate at FPPW of \(10^{-4}\), which is about 9% lower than the HOG+SVM method, about 5% lower than the v-HOG+CAadaboost method, about 3.5% lower than HOG-LBP+MVPPE method and 1.5% lower than v-HOG+CLML method.

For the SDL dataset, which is a new dataset, we compare the results of our method with the HOG+SVM [2] and v-HOG+CLML [14]. The results of the contrast methods are obtained in [14]. As shown in Figure 3.(C), our method is superior to the other state-of-the-arts methods, with 5.5% miss rate at FPPW of \(10^{-5}\) and 2.5% miss rate at FPPW of \(10^{-4}\). Therefore, these findings demonstrate that...
our algorithm achieves better results than state-of-the-art methods for arbitrary view pedestrian detection task.

4. Conclusions
We have presented a new method for arbitrary view pedestrian detection in images by employing locality-constrained affine subspace coding (LASC). Notwithstanding that the method only uses the positive samples for training, the learning model is capable of detecting pedestrian under arbitrary view. Experimental results on two datasets prove that the proposed method achieves a higher performance than the state-of-the-art methods in form of effect and efficiency. In the future, we will extend the approach to other objects detection task.

References
[1] Nguyen D T, Li W, Ogunbona P O. Human detection from images and videos: A survey. Pattern Recognition, 2015; 51:148-175.
[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, 886–893.
[3] Kuang H, Chong Y, Li Q, Zheng C. Mutual cascade method for pedestrian detection. Neurocomputing, 2014, 137: 127–135.
[4] Ye Q, Jiao J, Jiang S, Fast and robust pedestrian detection algorithm with multi-Scale orientation features. Journal of Software, 2011, 22(12): 3004–3014.
[5] Xue Z, Li G, Huang Q, Joint multi-View representation learning and image tagging. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[6] Torralba A, Murphy K P, Freeman W T, Sharing visual features for multiclass and multiview object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007, 29(5): 854-869.
[7] Zhu Y, Wang S, Yue L, Ji Q, Multiple-Facial action unit recognition by shared feature learning and semantic relation modeling. International Conference on Pattern Recognition, 2014.
[8] Zhang X, Pham D, Venkatesh S, Liu W and Phung D. Mixed-norm sparse representation for multi view face recognition. Pattern Recognition. 2015, 48(9): 2935-2946.
[9] Li P, Lu X, Wang Q. From dictionary of visual words to subspaces: locality-constrained affine subspace coding. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[10] Wang J, Yang J, Yu K, Lv F, Huang T S, and Gong Y, Locality-constrained linear coding for image classification. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.
[11] Ye Q, Liang J, Jiao J, Pedestrian detection in video images via error correcting output code classification of manifold subclasses, IEEE Transaction on Intelligent Transportation Systems, 2012, 13(1): 193-202.
[12] Wu L, Sun H, Ji K, Fan Y, Zhang Y, Multi-view pedestrian recognition via non-negative least-squares. Proceedings of the Chinese Intelligent Automation Conference, 2015, 199-208.
[13] Zhu Q, Avidan S, Yeh M C, Cheng K T. Fast human detection using a cascade of histograms of oriented gradients. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[14] Liu W, Yu B, Duan C, Cha L, Yuan H, Zhao H, A pedestrian-detection method based on heterogeneous features and ensemble of multi-view–pose parts, IEEE Transaction on Intelligent Transportation Systems, 2015, 16(2): 813-823.
[15] Xu R, Jiao J, Zhang B, Ye Q. Pedestrian detection in images via cascaded L1-norm minimization learning method. Pattern Recognition. vol. 2012, 45(7). 2573–2583.