FUSED DISCRIMINATIVE METRIC LEARNING FOR LOW RESOLUTION PEDESTRIAN DETECTION

Xinzhao Li* *Yuehu Liu** *Zeqi Chen** *Jiahuan Zhou*† *Ying Wu*†

* Xi'an Jiaotong University, IAIR, China † Northwestern University, EECS, USA.

ABSTRACT

Low resolution (LR) is one of the most challenging factor in pedestrian detection. In this paper, we propose a fused discriminative metric learning (F-DML) approach for low resolution pedestrian detection without explicit super resolution. We firstly learn a discriminative high resolution (HR) feature space as target space. Then, an optimal Mahanalobis metric is learned to transform the LR feature space into a new LR classification space, which largely preserves the discriminative structure of the HR feature space. Finally, a weighted K-nearest neighbors classifier is applied in the LR classification space which inherits good discrimination from HR feature space. A new training strategy is proposed to find the fewest and most representative LR-HR exemplars. In addition, we build a new dataset for the evaluation of low resolution pedestrian detection methods. Extensive experimental results demonstrate that the proposed approach performs favorably against the state-of-the-art methods.

Index Terms— Pedestrian detection, Low resolution, Metric learning

1. INTRODUCTION

Among the various domains of pedestrian detection, low resolution (LR) pedestrian detection is the most imperative and challenging. In large-scale open scenes, the size of pedestrian is very small (approximate 10-20 pixels tall). For such LR images, the details of the visual appearances are lost. Since the visual features cannot be reliably extracted from such few pixels, it is very difficult to detect such LR pedestrian.

The performance of existing pedestrian detection approaches drops with the resolution decreasing, no matter the traditional approaches such as HOG [1], DPM [2], ACF [3] and Checkerboards [4], but also hot deep learning based approaches such as JDN [5], DeepParts [6] and RPN-BP [7]. In several detailed surveys [8, 9, 10, 11], it is shown that all the existing methods failed on LR pedestrian images. However, only a few studies paid attention to low resolution issue. MT-DPM [12] learned a resolution-aware DPM model and RACNN [13] proposed a multiple CNN-based architecture.

The loss of discriminative details and heavy noises of the LR images restricts most LR pedestrian detection methods. To solve the former problem, we extract prior knowledge which is implicitly included in a set of training instances of LR-HR pairs. There is a well-known assumption that the discrimination of HR feature space is much better than LR feature space. If we can learn a metric to transform the chaotic LR feature space into a new feature space which has similar structure and discrimination with HR feature space, we can directly classify the pedestrian from background in this new feature space without explicit super resolution (SR). To solve the latter issue, we present a new attempt to use the difference of LR feature vectors to extract feature. It is assumed that the noises in LR images are identically distributed in the same scene. The difference between LR feature vectors may greatly reduce the impact of the identically distributed noises.

In this paper, we propose a fused discriminative metric learning method (F-DML) for low resolution pedestrian detection. Firstly, the F-DML learns a discriminative HR feature space as target space. Then, an optimal Mahanalobis metric is learned to transform the LR feature space into a new space, which largely preserves the structure of the discriminative HR feature space. Finally, a weighted KNN classifier is applied in the new projected space for classification.

The main contributions of this paper are summarized as follows: (1) we propose a new metric learning method, which uses the discriminative HR feature space to guide the classification of LR feature space. This makes it possible to detect LR pedestrian without explicit SR. (2) We use the difference in Mahalanobis distance to represent the LR features to weaken the negative impact of heavy noises. (3) We design a small instance based training strategy to choose the fewest and most representative LR-HR exemplars pairs. (4) We built a more challenging dataset for low resolution pedestrian detection problem.

2. FUSED DISCRIMINATIVE METRIC LEARNING

Firstly, we will give an overview of our discriminative metric learning approach, which is summarized in Figure 1.

In the training phase, we need to learn a metric to preserve the discriminative structure of HR feature space when transforming the chaotic LR feature space. Firstly, a HR-LR training dataset is built by resizing the HR training images into LR images by Gaussian pyramid. Then we obtain a dis-

978-1-4799-7061-2/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: Northwestern University. Downloaded on April 06,2022 at 09:41:11 UTC from IEEE Xplore. Restrictions apply.

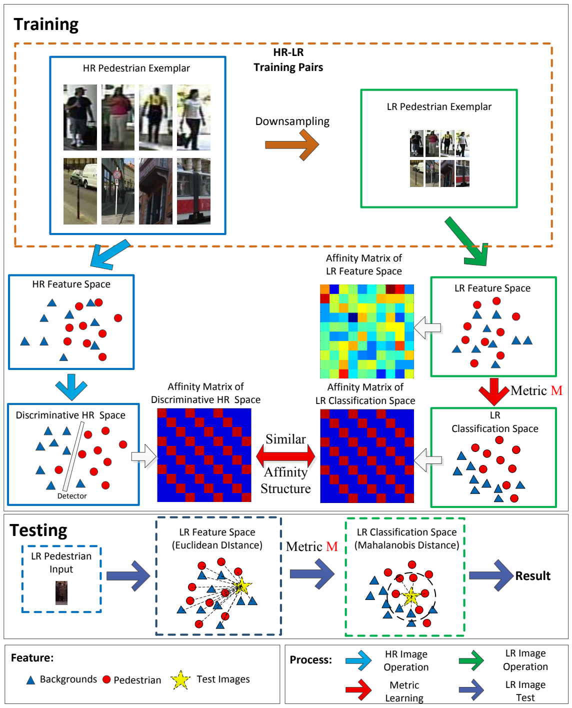


Fig. 1. The framework of the Discriminative Metric Learning for LR pedestrian detection.

criminative HR feature space in Euclidean distance by Partial Least Squares Regression, which can contribute to the classification performance in LR feature space. Next, we construct affinity matrices to represent the structure of LR and HR feature spaces. Finally, we put forward a metric learning method to learn a metric to preserve the discriminative structure of HR feature space when transforming the chaotic LR feature space into LR classification space.

Assume $D_l = \{I_l\}$ and $D_h = \{I_h\}$ are the LR-HR training datasets. $\boldsymbol{L} = \{\boldsymbol{l}_i, i \subset D_l\}$ and $\boldsymbol{H} = \{\boldsymbol{h}_i, i \subset D_h\}$ denote the corresponding feature vectors. Let $\boldsymbol{M}$ be the learning Mahalanobis distance metric, the objective function is to minimize the distance between the affinities of discriminative HR feature space and LR classification space as follows:

$$\boldsymbol{M}^* = \arg\min_{\boldsymbol{M}} Dis(Aff(\boldsymbol{H}, \boldsymbol{W}), Aff(\boldsymbol{L}, \boldsymbol{M})) \qquad (1)$$

where $\boldsymbol{W}$ represents the class-aware discriminative projection matrix. $Aff()$ denotes the function from feature space to the affinity matrix which reflects the relative structure of corresponding feature space. $Dis()$ means the distance measure function of two affinities matrices.

In the testing phase, the input LR test images are classified by weighted K-nearest neighbors classifier in LR classification space by using the learned metric $\boldsymbol{M}$.

2.1. Discrimination of High Resolution Feature Space

As the target space used to guide the transformation of LR feature space, HR feature space needs to have good discrimination. Therefore, we obtain a more discriminative HR feature space in Euclidean distance by Partial Least Squares Regression (PLSR) [14]. Here we want to project the HR feature vectors into a new discriminative space to ensure the correlation between projected features and labels in each dimension.

Let $\widetilde{\boldsymbol{H}}$ be the zero-mean HR feature matrix and $\widetilde{\boldsymbol{y}}$ denote the zero-mean class label vector. Let $\boldsymbol{W} = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ...\boldsymbol{\omega}_p\}$ be the discriminative projection matrix. We want to maximize both the variance of projected HR vectors and the correlation between the projected HR vectors and labels as follows:

$$\begin{aligned}\boldsymbol{\omega}_i^* &= \arg\max_{|\boldsymbol{\omega}_i|=1} [Cov(\widetilde{\boldsymbol{H}}\boldsymbol{\omega}_i, \widetilde{\boldsymbol{y}})]^2 \\ &= \arg\max_{|\boldsymbol{\omega}_i|=1} Var(\widetilde{\boldsymbol{H}}\boldsymbol{\omega}_i) Var(\widetilde{\boldsymbol{y}}) [Corr(\widetilde{\boldsymbol{H}}\boldsymbol{\omega}_i, \widetilde{\boldsymbol{y}})]^2\end{aligned} \qquad (2)$$

where $Cov()$ denotes the sample covariance, $Var()$ means the variance and $Corr()$ indicates the correlation.

The $\boldsymbol{\omega}_i$ can be solved by Lagrange multipliers method [15]. After obtaining the projected vector $\boldsymbol{t}_i = \widetilde{\boldsymbol{H}}\boldsymbol{\omega}_i$ in the ith dimension, the regression function is constructed based on the nonlinear iterative partial least squares (NIPALS) [16]. Finally, the HR feature space $\mathcal{H}$ is projected to a more discriminative space $\mathcal{Z}$ in Euclidean distance, namely $\boldsymbol{Z} = \{\boldsymbol{z}_i = \boldsymbol{W}\boldsymbol{h}_i,\ \boldsymbol{z}_i \subset \mathcal{Z}\}$.

2.2. Metric Learning for Similar Space Structure

In this section, we seek for an optimal transformation which projects the LR feature space to a LR classification space, so that the structure and distribution of discriminative HR feature space can be preserved in the LR classification space.

Firstly, we need to represent the structure of these two feature spaces. For discriminative HR feature space, the distance matrix is computed in Euclidean distance as $\mathcal{D}_p(\boldsymbol{z}_i, \boldsymbol{z}_j) = (\boldsymbol{z}_i - \boldsymbol{z}_j)^T(\boldsymbol{z}_i - \boldsymbol{z}_j)$. For LR feature space, the distance matrix is defined by Mahalanobis distance: $\mathcal{D}_q(\boldsymbol{l}_i, \boldsymbol{l}_j) = (\boldsymbol{l}_i - \boldsymbol{l}_j)^T\boldsymbol{M}(\boldsymbol{l}_i - \boldsymbol{l}_j)$. $\boldsymbol{M}$ is a positive semi-definite (PSD) matrix which changes the structure distribution of LR feature space.

The structure of discriminative HR feature space is defined by the normalized affinity matrix $\boldsymbol{P} = \{p_{ij}\}$, where

$$u_{ij} = exp(-\frac{\mathcal{D}_p(\boldsymbol{z}_i, \boldsymbol{z}_j)}{2\sigma_{hr}}),\ p_{ij} = \frac{u_{ij}}{\sum_{k\neq i} u_{ik}},\ p_{ii} = 0 \qquad (3)$$

so that $\boldsymbol{P}$ is a distribution that represents the nearest neighbor probability from vector $\boldsymbol{z}_i$ to $\boldsymbol{z}_j$.

Similarly, we construct the normalized affinity matrix $\boldsymbol{Q} = \{q_{ij}\}$ of the classification LR feature space:

$$v_{ij} = exp(-\frac{\mathcal{D}_q(\boldsymbol{l}_i, \boldsymbol{l}_j)}{2\sigma_{lr}}),\ q_{ij} = \frac{v_{ij}}{\sum_{k\neq i} v_{ik}},\ q_{ii} = 0 \qquad (4)$$

Authorized licensed use limited to: Northwestern University. Downloaded on April 06,2022 at 09:41:11 UTC from IEEE Xplore. Restrictions apply.

Choosing the KL divergence as distance measure for two distributions, we get the objective function as follows:

$$\boldsymbol{M}^* = \arg\min_{\boldsymbol{M}} \sum_{ij} KL[p_{ij}|q_{ij}],\ s.t.\ M \preceq PSD \tag{5}$$

Denoted by $f(\boldsymbol{M}) \triangleq \sum_{ij} KL[p_{ij}|q_{ij}]$, we can use the gradient-decent technique to minimize the objective function:

$$\nabla f(\boldsymbol{M}) = \frac{1}{2\sigma_{lr}} \sum_{ij} (p_{ij} - q_{ij})(\boldsymbol{l}_i - \boldsymbol{l}_j)(\boldsymbol{l}_i - \boldsymbol{l}_j)^T \tag{6}$$

Metric $\boldsymbol{M}$ is updated by $\boldsymbol{M}^{t+1} = \boldsymbol{M}^t - \epsilon\nabla f(\boldsymbol{M}^t)$, where ϵ is the step length of gradient descent.

Since $\boldsymbol{M}$ has to be PSD, we need to eliminate its negative components by equation 7. λ_k is the eigenvalue of $\boldsymbol{M}$ and $\boldsymbol{v}_k$ is the corresponding eigenvector.

$$\hat{\boldsymbol{M}} = \sum_k max(\lambda_k, 0)\boldsymbol{v}_k\boldsymbol{v}_k^T \tag{7}$$

2.3. Classifier and Multi-Channel Fusion

Denote the test image I_t and the corresponding LR feature vector l_t. The confidence score S is computed by weighted voting in K-nearest neighbors of the training LR exemplars.

$$S = \frac{\sum_i^K g(\boldsymbol{l}_t, \boldsymbol{l}_i)c_i}{\sum_i^K g(\boldsymbol{l}_t, \boldsymbol{l}_i)},\ c_i \in \{0, 1\} \tag{8}$$

where $C = \{c_i\}$ denotes the class label. The voting weight is defined as $g(\boldsymbol{l}_t, \boldsymbol{l}_i) = exp(-\frac{(\boldsymbol{l}_t - \boldsymbol{l}_i)^T \boldsymbol{M}(\boldsymbol{l}_t - \boldsymbol{l}_i)}{2\sigma})$. The σ represents the variance of the difference between l_t and l_i in the projected LR classification space.

We believe that different features can preserve different discriminative characteristics of HR feature space. Using N kinds of feature descriptors in HR feature space, we can learn N matrices. The weighted summation of the confidence scores is used to fuse all the N classifiers as follows:

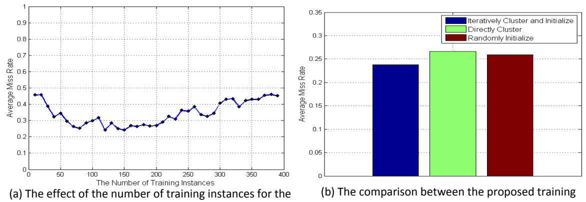
$$S_{final} = \sum_{i=1}^{N} \lambda_i S_i \tag{9}$$

where λ_i denotes the weight of ith channel.

3. UNDERSTANDING PROPERTIES

In this section, we present several experiments to analyze the properties of the proposed F-DML method.

Datasets. Although pedestrian detection is a well-defined problem with rich datasets, none of them is suitable for LR pedestrian detection. The pedestrian under 21 pixels tall are not annotated in the traditional datasets. Therefore, we create a comprehensive dataset for LR pedestrian detection task by gathering three famous datasets together (INRIA [1], ETH [17] and Caltech [18]). For training dataset, we use the INRIA training dataset with a set of labeled LR-HR patch pairs, which includes 2416 positive pairs and 12180 background pairs. The resolutions of HR-LR patches are 128×64 and 20×10. For testing dataset, we crop 5614 positive exemplars and 58732 negative exemplars with the resolution of 20×10.



(a) The effect of the number of training instances for the proposed DML

(b) The comparison between the proposed training strategy and other training methods without iteration

Fig. 2. The contribution of the new training strategy. (a) find the optimal number of LR-HR training instances; (b) find a better local minimum in the solution space.

Feature Extraction. For LR exemplars, valid pixels are too few for traditional feature extraction approaches. So the rawpixel is a good choice. For HR exemplars, we use HOGLUV [10], VGG16 (layer fc7) [19] and SketchTokens [20] to express the HR feature space respectively.

Evaluation Metric. The detection error tradeoff (DET) curve and average miss rate (AMR) [8] are the main evaluation metrics. The x-axis corresponds to false positive per window (FPPW) and the y-axis shows the miss rate.

3.1. Find Better Training Data and Local Minimum

Because of the heavy noises in LR pedestrian images, it is not wise to use all the training LR-HR exemplars pairs to train the Mahalanobis kernel $\boldsymbol{M}$. We propose a iterative training strategy to choose the fewest and most representative LR-HR exemplars pairs. In each iteration, we cluster the false negative and false positive HR-LR patches by Gaussian Mixture Model and redivide the training dataset. In addition, the metric $\boldsymbol{M}$ is initialized with the last trained result.

Figure 2(a) shows that the proposed DML reaches the best performance when using 150 training exemplars. Too many training exemplars will flat the metric $\boldsymbol{M}$. Too few training exemplars will not cover the entire solution space. Compared to the strategies of directly clustering the training exemplars and randomly initializing metric $\boldsymbol{M}$, figure 2(b) shows that the iterative strategy will help to find a better local minimum in the solution space by choosing the training HR-LR exemplars and iteratively initializing the metric $\boldsymbol{M}$.

3.2. Discriminative Analysis of HR Feature Space

We explore the effect of PLSR on improving the discrimination of HR feature space in Euclidean distance. Figure 3 shows the visualization of the discrimination of LR and HR

Authorized licensed use limited to: Northwestern University. Downloaded on April 06,2022 at 09:41:11 UTC from IEEE Xplore. Restrictions apply.

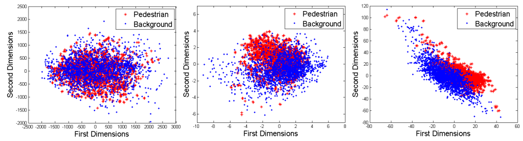
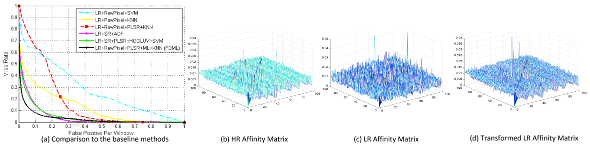


Fig. 3. Visualization for the discrimination of feature spaces. (Left: LR+PCA; Mid: HR+PCA; Right: HR+PLSR+PCA)



(a) Comparison to the baseline methods (b) HR Affinity Matrix (c) LR Affinity Matrix (d) Transformed LR Affinity Matrix

Fig. 4. The advantages of proposed metric learning method. (a) Comparison to the benchmark methods. (b-d) Visualization of affinity matrices (HR, LR and transformed LR spaces).

feature space. Principal Component Analysis (PCA) [21] is used to extract the first two dimensions of the training exemplars. Treating the HR feature space as baseline, the discrimination of LR feature space is totally lost. On the contrary, the projected HR feature space has better discrimination.

3.3. Why Choosing Metric Learning

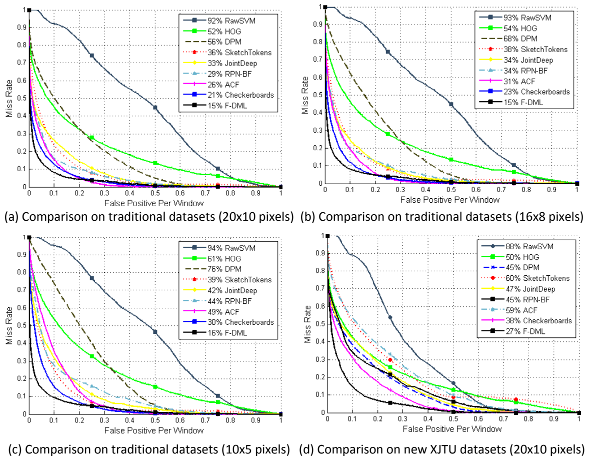
To evaluate the effect of proposed metric learning approach, we set up two benchmarks for comparison: 1) directly extract feature from LR images and apply classifier; 2) rebuild HR images from LR images by SR technique and extract feature. The comparison with these two traditional solutions are shown in Figure 4(a). It is shown that KNN is more suitable for LR images than SVM (LR+KNN vs. LR+SVM). Directly extracting feature from LR images (LR+RawPixel+X) has the worst performance. SR process (LR+SR+X) significantly improves the performance. However, the proposed method outperforms others in a large margin. Figure 4(b-d) show the visualization of affinity matrices in three feature spaces.

4. COMPARATIVE EXPERIMENTS

In this section, we compare the proposed method with several state-of-the-art pedestrian detection methods.

Comparative Methods. Since the Caltech pedestrian dataset doesn't annotate the pedestrians lower than 21 pixels tall, the benchmarks in Caltech pedestrian dataset are not available for LR pedestrian detection. Therefore, we download all the open-source pedestrian detection methods for comparison, including HOG [1], DPM [2], SketchTokens [20], JDN [5], ACF [3], Checkerboards [4] and RPN-BF [7].

Training. To ensure a fair comparison, all the comparative methods are trained on INRIA. All the comparative methods are fine tune based on the LR pedestrian detection task.



(a) Comparison on traditional datasets (20x10 pixels) (b) Comparison on traditional datasets (16x8 pixels)

(c) Comparison on traditional datasets (10x5 pixels) (d) Comparison on new XJTU datasets (20x10 pixels)

Fig. 5. The comparisons with the state-of-the-art methods.

Multi-resolution Comparisons. The testing exemplars are resized to the resolutions of 20×10, 16×8, 10×5. Figure 5(a-c) shows the comparison with state-of-the-art methods on multi-resolution traditional datasets. We can see that the proposed method outperforms others on all these LR resolutions.

Comparisons on New Challenging Dataset. Existing pedestrian datasets can not fully test an approach for LR pedestrian detection. Therefore, we collect a new dataset for the evaluation of LR pedestrian detection methods, which is called XJTU (see supplemental files). Figure 5(d) shows the comparison with state-of-the-art methods on the new dataset.

5. CONCLUSION

In this paper, we propose a fused discriminative metric learning method for LR pedestrian detection. We use the discriminative HR feature space as prior information to guide the classification in LR feature space. A new training strategy is designed to find better training exemplars and model parameters. In addition, we build a new LR pedestrian dataset with the pedestrians lower than 21 pixels tall. The proposed method outperformed the state-of-the-art methods on LR pedestrian detection task on different resolutions and datasets.

Acknowledgment

The authors would like to thank all the experts who have contributed to this paper. This work is supported by National Natural Science Foundation of China (NSFC) under Grant No.91520301 and Grant No.61328303. What's more, this work was supported in part by National Science Foundation grant IIS-1217302, IIS-1619078, and the Army Research Office ARO W911NF-16-1-0138.

Authorized licensed use limited to: Northwestern University. Downloaded on April 06,2022 at 09:41:11 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] Navneet Dalal and Bill Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[2] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[3] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, “Fast feature pyramids for object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[4] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, “Filtered channel features for pedestrian detection,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 1751–1760.

[5] Wanli Ouyang and Xiaogang Wang, “Joint deep learning for pedestrian detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2056–2063.

[6] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Deep learning strong parts for pedestrian detection,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1904–1912.

[7] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He, “Is faster r-cnn doing well for pedestrian detection?,” in *European Conference on Computer Vision*. Springer, 2016, pp. 443–457.

[8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona, “Pedestrian detection: An evaluation of the state of the art,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[9] Markus Enzweiler and Dariu M Gavrila, “Monocular pedestrian detection: Survey and experiments,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[10] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele, “Ten years of pedestrian detection, what have we learned?,” in *European Conference on Computer Vision*. Springer, 2014, pp. 613–627.

[11] Duc Thanh Nguyen, Wanqing Li, and Philip O Ogunbona, “Human detection from images and videos: A survey,” *Pattern Recognition*, vol. 51, pp. 148–175, 2016.

[12] Junjie Yan, Xucong Zhang, Zhen Lei, Shengcai Liao, and Stan Z Li, “Robust multi-resolution pedestrian detection in traffic scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3033–3040.

[13] Rakesh Nattoji Rajaram, Eshed Ohn-Bar, and Mohan Manubhai Trivedi, “Looking at pedestrians at different scales: A multiresolution approach and evaluations,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3565–3576, 2016.

[14] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S Davis, “Human detection using partial least squares analysis,” in *Computer vision, 2009 IEEE 12th international conference on*. IEEE, 2009, pp. 24–31.

[15] Dimitri P Bertsekas, *Nonlinear programming*, Athena scientific Belmont, 1999.

[16] Herman Wold, “Partial least squares,” *Encyclopedia of statistical sciences*, 1985.

[17] A. Ess, B. Leibe, K. Schindler, , and L. van Gool, “A mobile vision system for robust multi-person tracking,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR’08)*. June 2008, IEEE Press.

[18] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, “Pedestrian detection: A benchmark,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 304–311.

[19] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[20] Joseph J Lim, C Lawrence Zitnick, and Piotr Dollár, “Sketch tokens: A learned mid-level representation for contour and object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3158–3165.

[21] Svante Wold, Kim Esbensen, and Paul Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

Authorized licensed use limited to: Northwestern University. Downloaded on April 06,2022 at 09:41:11 UTC from IEEE Xplore. Restrictions apply.