Pedestrian Detection Based on Kernel Discriminative Sparse Representation

Keyang Cheng[1,2], Qirong Mao[2], and Yongzhao Zhan[2]

[1] School of Computer Science & Technology, Nanjing University of Aeronautics
& Astronautics, Nanjing, Jiangsu, China, 210016
[2] School of Computer Science & Telecommunications Engineering,
Jiangsu University, Zhenjiang, Jiangsu, China, 212013
kycheng@ujs.edu.cn

Abstract. This article puts forward a novel framework for pedestrian detection tasks, which proposing a model with both sparse reconstruction and class discrimination components, jointly optimized during dictionary learning. We present an efficient pedestrian detection system using mixing sparse features of HOG, FOG and CSS to combine into a Kernel classifier. Results presented on our data set show competitive accuracy and robust performance of our system outperforms current state-of-the-art work.

Keywords: Kernel Discriminative Sparse Representation, Pedestrian Detection.

1 Introduction

Pedestrian counting in public places plays a key role in many applications, such as evacuating from a dense region to a sparse one when an emergency happens, or optimizing the design of traffic infrastructures to provide better transportation services. Furthermore, social security and surveillance strongly depend on the effectiveness of pedestrian counting. A wide variety of pedestrian detection methods have been proposed [1-6].

Sparse representations have recently drawn much interest in signal, image, and video processing. Under the assumption that natural images admit a sparse decomposition in some redundant basis (or so-called dictionary), several such models have been proposed, e.g., curve lets, wedge lets, band lets and various sorts of wavelets [7]. Interestingly, while discrimination is the main goal of these papers, the optimization (dictionary design) is purely generative, based on a criterion which does not explicitly include the actual discrimination task, which is one of the key contributions of our work. In [8], a discriminative method is introduced for various classification tasks, learning one dictionary per class; the classification process itself is based on the corresponding reconstruction error, and does not exploit the actual decomposition coefficients. In [9], a generative model for documents is learned at the same time as the parameters of a deep network structure. In [10], multi-task learning is performed by learning features and tasks are selected using a sparsity criterion. The framework we present in this paper extends these approaches by learning

Z. Pan et al. (Eds.): Transactions on Edutainment IX, LNCS 7544, pp. 184–195, 2013.

simultaneously a single shared dictionary as well as models for different signal classes in a mixed generative and discriminative formulation (see also [11], where a different discriminative term is added to the classical reconstructive one). Similar joint generative/discriminative frameworks have started to appear in probabilistic approaches to learning, e.g., [12, 13, 14, 15, 16], and in neural networks [17], but not, to the best of our knowledge, in the sparse dictionary learning framework.

The remainder of this paper is organized as follows. In Section 2, we describe the procedure of feature extraction, and in Section 3, we present a formulation for learning a dictionary tuned for a classification task, which we call discriminative sparse learning. Section4 gives the optimization procedure of discriminative sparse learning .Experimental results are provided and analyzed in Section 5. Finally, Section 6 concludes this work.

2 Feature Extraction

Obviously, the choice of features is the most critical decision when designing a detector, and finding good features is still largely an empirical process with few theoretical guidelines. We evaluate different combinations of features, and introduce a new feature based on the similarity of colors in different regions of the detector window, which significantly raises detection performance. The pedestrian region in our detection window is of size 48*96 pixels.

Histograms of oriented gradients (HOG) are a popular feature for object detection, first proposed in [18]. They collect gradient information in local cells into histograms using trilinear interpolation, and normalize overlapping blocks composed of neighboring cells. Interpolation, local normalization and histogram binning make the representation robust to changes in lighting conditions and small variations in pose. HOG was recently enriched by Local Binary Patterns (LBP), showing a visible improvement over standard HOG on the INRIA Person data set [24]. In our experiments we compute histograms with 9 bins on cells of 8*8 pixels. Block size is 2*2 cells overlapping by one cell size.

HOF Histograms of flow were initially also proposed by Dalal et al. [19]. We have shown that using them (e.g. in [19]'s IMHwd scheme) complementary to HOG can give substantial improvements on realistic datasets with significant ego motion. Here, we introduce a lower-dimensional variant of HOF, IMHd2, which encodes motion differences within 2*2 blocks with 4 histograms per block, while matching the performance of IMHwd (3*3 blocks with 9 histograms). Fig. 2(d) schematically illustrates the new coding scheme: the 4 squares display the encoding for one histogram each. For the first histogram, the optical flow corresponding to the pixel at the ith row and jth column of the upper left cell is subtracted from the one at the corresponding position of the lower left cell, and the resulting vector votes into a histogram as in the original HOF scheme. IMHd2 provides a dimensionality reduction of 44% (2520 instead of 4536 values per window), without changing performance significantly. We used the publicly available flow implementation of [20]. In this work we show that

HOF continues to provide a substantial improvement even for flow fields computed on JPEG images with strong block artifacts (and hence degraded flow fields).

Several authors have reported improvements by combining multiple types of low-level features [21, 22, 23]. Still, it is largely unclear which cues should best be used in addition to the now established combination of gradients and optic flow. Intuitively, additional features should be complementary to the ones already used, capturing a different part of the image statistics. Color information is such a feature enjoying popularity in image classification [24] but is nevertheless rarely used in detection. Furthermore, second order image statistics, especially co-occurrence histograms, are gaining popularity, pushing feature spaces to extremely high dimensions [25, 22].

We propose to combine these ideas and use second order statistics of colors as additional feature. Color by itself is of limited use, because colors vary across the entire spectrum both for people (respectively their clothing) and for the background, and because of the essentially unsolved color constancy problem. However, people do exhibit some structure, in that colors are locally similar—for example (see Fig. 1) the skin color of a specific person is similar on their two arms and face, and the same is true for most people's clothing. Therefore, we encode color self similarities within the descriptor window, i.e. similarities between colors in different sub-regions. To leverage the robustness of local histograms, we compute D local color histograms over 8*8 pixel blocks, using trilinear interpolation as in HOG to minimize aliasing. We experimented with different color spaces, including 3*3*3 histograms in RGB, HSV, HLS and CIE Luv space, and 4*4 histograms in normalized rg, HS and uv, discarding the intensity and only keeping the chrominance. Among these, HSV worked best, and is used in the following.

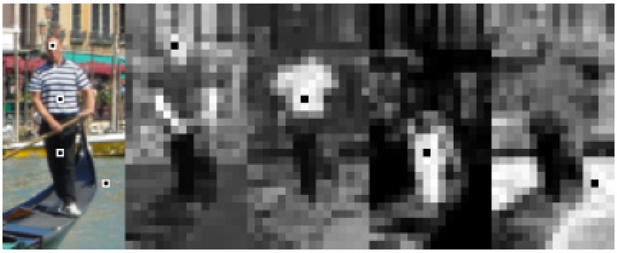


Fig. 1. Self-similarity encodes relevant parts

3 Supervised Dictionary Learning

We present in this section the core of the proposed model. In classical sparse coding tasks, one considers a signal x in R^n and a fixed dictionary D = [d_1, . . . , d_k] in $R^{n\times k}$ (allowing k>n, making the dictionary over complete). In this setting, sparse coding with an ℓ1 regularization amounts to computing

$$R^*(x, D) = \min_{\alpha \in R^k} \| x - D\alpha \|_2^2 + \lambda_1 \| \alpha \|_1 \tag{1}$$

It is well known in the statistics, optimization, and compressed sensing communities that the ℓ1 penalty yields a sparse solution, very few non-zero coefficients in α, although there is no explicit analytic link between the value of λ1 and the effective sparsity that this model yields. Other sparsity penalties using the ℓ0 regularization can be used as well. Since it uses a proper norm, the ℓ1 formulation of sparse coding is a convex problem, which makes the optimization tractable with algorithms such as those introduced in [26, 27], and has proven in practice to be more stable than its ℓ0 counterpart, in the sense that the resulting decompositions are less sensitive to small perturbations of the input signal x. Note that sparse coding with an ℓ0 penalty is an NP-hard problem and is often approximated using greedy algorithms.

In this paper, we consider a setting, where the signal may belong to any of p different classes. We first consider the case of p = 2 classes and later discuss the multiclass extension. We consider a training set of m labeled signals (x_i)m_i=1 in R^n, associated with binary labels (y_i ∈{−1, +1})m_i=1.Our goal is to learn jointly a single dictionary D adapted to the classification task and a function f which should be positive for any signal in class +1 and negative otherwise. We consider in this paper two different models to use the sparse code α for the classification task:

(i) linear in α: $f(x,\alpha,\theta)=w^T\alpha+b$, where θ = {w∈$R^k$, b∈R} parametrizes the model.

(ii) bilinear in x and α: $f(x,\alpha,\theta)=x^T w\alpha+b$, where θ= {W∈$R^{n\times k}$, b ∈R}. In this case, the model is bilinear and f acts on both x and its sparse code α.

The number of parameters in (ii) is greater than in (i), which allows for richer models. Note that one can interpret was a linear filter encoding the input signal x into a model for the coefficients, which has a role similar to the encoder in [28] but for a discriminative task. A classical approach to obtain for (i) or (ii) is to first adapt D to the data, solving

$$\min_{D,\alpha}\sum_{i=1}^{m}\| x_i-D\alpha_i\|_2^2+\lambda_1\|\alpha_i\|_1 \tag{2}$$

Note also that since the reconstruction errors $\| x_i-D\alpha_i\|_2^2$ are invariant to scaling simultaneously D by a scalar and αi by its inverse, we need to constrain the ℓ2 norm of the columns of D. Such a constraint is classical in sparse coding [29]. This reconstructive approach provides sparse codes αi for each signal xi, which can be used a posteriori in a regular classifier such as logistic regression, which would require to solve

$$\min_{\theta}\sum_{i=1}^{m}c(y_i f(x_i,\alpha_i,\theta))+\lambda_2\|\theta\|_2^2 \tag{3}$$

where C is the logistic loss function (C(x) = log(1 + e^{-x})), which enjoys properties similar to that of the hinge loss from the SVM literature, while being differentiable, and λ2 is a regularization parameter, which prevents over fitting. This is the approach chosen in [30] (with SVMs). However, our goal is to learn jointly D and the model parameters θ. To that effect, we propose the formulation

$$\min_{D,\theta,\alpha}(\sum_{i=1}^{m}c(y_i f(x_i,\alpha_i,\theta))+\lambda_0\|\| x_i-D\alpha_i\|_2^2+\lambda_1\|\alpha_i\|_1)+\lambda_2\|\theta\|_2^2 \tag{4}$$

where λ0 controls the importance of the reconstruction term, and the loss for a pair (x_i,y_i) is

$$S^*(x_i, D, \theta, y_i) = \min_{\alpha} S(\alpha, x_i, D, \theta, y_i) \tag{5}$$

Where $S(\alpha, x_i, D, \theta, y_i) = c(y_i f(x_i, \alpha_i, \theta)) + \lambda_0 \| x_i - D\alpha_i \|_2^2 + \lambda_1 \| \alpha_i \|_1$ In this setting, the classification procedure of a new signal x with an unknown label y, given a learned dictionary D and parameters θ, involves supervised sparse coding:

$$\min_{y \in \{-1;+1\}} S^*(x, D, \theta, y) \tag{6}$$

The learning procedure of Eq. (4) minimizes the sum of the costs for the pairs (x_i,y_i)m_i=1 and corresponds to a generative model. We will refer later to this model as SDL-G (supervised dictionary learning, generative). Note the explicit incorporation of the reconstructive and discriminative component into sparse coding, in addition to the classical reconstructive term (see [31] for a different classification component).

However, since the classification procedure from Eq. (6) compares the different costs $S^*(x, D, \theta, y)$ of a given signal for each class y = −1, +1, a more discriminative approach is to not only make the costs $S^*(x_i, D, \theta, -y_i)$ small, as in (4), but also make the value of $S^*(x_i, D, \theta, -y_i)$ greater than $S^*(x_i, D, \theta, -y_i)$, which is the purpose of the logistic loss function C. This leads to:

$$\min_{D,\theta} (\sum_{i=1}^{m} c(S^*(x_i, D, \theta, -y_i) - S^*(x_i, D, \theta, y_i))) + \lambda_2 \| \theta \|_2^2 \tag{7}$$

As detailed below, this problem is more difficult to solve than (4), and therefore we adopt instead a mixed formulation between the minimization of the generative Eq. (4) and its discriminative version (7), (see also [32])—that is,

$$(\sum_{i=1}^{m} \mu c(S^*(x_i, D, \theta, -y_i) - S^*(x_i, D, \theta, y_i)) + (1-\mu) S^*(x_i, D, \theta, y_i)) + \lambda_2 \| \theta \|_2^2 \tag{8}$$

where μ controls the trade-off between the reconstruction from Eq. (4) and the discrimination from Eq. (7). This is the proposed generative/discriminative model for sparse signal representation and classification from learned dictionary D and model θ. We will refer to this mixed model as SDL-D, (supervised dictionary learning, discriminative). Note also that, again, we constrain the norm of the columns of D to be less than or equal to one.

All of these formulations admit a straightforward multiclass extension, using softmax discriminative cost functions $c_i(x_1, ..., x_p) = \log(\sum_{i=1}^{p} e^{x_j - x_i})$, which are multiclass versions of the logistic function, and learning one model θi per class. Other possible approaches such as one-vs-all or one-vs-one are of course possible, and the question of choosing the best approach among these possibilities is still open. Compared with earlier work using one dictionary per class [33], our model has the advantage of letting multiple classes share some features, and uses the coefficients of the sparse representations as part of the classification procedure, thereby following the works from[34, 35, 30], but with learned representations optimized for the classification task similar to [31, 36].

Our bilinear model with $f(x,\alpha,\theta)=x^T w\alpha+b$ does not admit a straightforward probabilistic interpretation. On the other hand, it can easily be interpreted in terms of kernels: Given two signals x$_1$ and x$_2$, with coefficients α$_1$ and α$_2$, using the kernel $K(x_1,x_2)=\alpha_1^T\alpha_2 x_1^T x_2$ in a logistic regression classifier amounts to finding a decision function of the same form as f. It is a product of two linear kernels, one on the α's and one on the input signals x. Interestingly, Raina et al. [30] learn a dictionary adapted to reconstruction on a training set, then train an SVM a posteriori on the decomposition coefficients. They derive and use a Fisher kernel, which can be written as $K'(x_1,x_2)=\alpha_1^T\alpha_2 r_1^T r_2$ in this setting, where the r's are the residuals of the decompositions. In simple experiments, which are not reported in this paper, we have observed that the kernel K, where the signals x replace the residuals r, generally yields a level of performance similar to K′ and often actually does better when the number of training samples is small or the data are noisy.

4 Optimization Procedure

Classical dictionary learning techniques (e.g., [30, 37, 38]), address the problem of learning a reconstructive dictionary D in R$^{n×k}$ well adapted to a training set, which is presented in Eq. (3). It can be seen as an optimization problem with respect to the dictionary D and the coefficients. Although not jointly convex in D, it is convex with respect to each unknown when the other one is fixed. This is why block coordinate descent on D and performs reasonably well [30, 37, 38], although not necessarily providing the global optimum. Training when μ = 0 (generative case), i.e., from Eq. (4), enjoys similar properties and can be addressed with the same optimization procedure. Equation (4) can be rewritten as:

$$\min_{D,\theta,\alpha}\left(\sum_{i=1}^{m} S(x_j,\alpha_j,D,\theta,y_i)\right)+\lambda_2 \| \theta \|_2^2, s.t.\forall j=1,...,k, \| d_j \|_2 \leq 1 \tag{9}$$

Block coordinate descent consists therefore of iterating between supervised sparse coding, where D and θ are fixed and one optimizes with respect to the α's and supervised dictionary update, where the coefficients α$_i$'s are fixed, but D and θ are updated. Details on how to solve these two problems are given in sections 4.1 and 4.2. The discriminative version SDL-D from Eq.(7) is more problematic. To reach a local minimum for this difficult non-convex optimization problem, we have chosen a continuation method, starting from the generative case and ending with the discriminative one as in [33]. The algorithm is presented in Figure 2.

```
Input: n (signal dimensions); $(x_i, y_i)$ $m_i$=1 (training signals); k (size of the dictionary); $\lambda_0$, $\lambda_1$, $\lambda_2$(parameters); $0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_m \leq 1$ (increasing sequence).
Output: $D \in R^{nxk}$ (dictionary); $\theta$ (parameters).
Initialization: Set D to a random Gaussian matrix with normalized columns. Set $\theta$ to zero.
Loop: For $\mu = \mu_1, \ldots, \mu_m$,
Loop: Repeat until convergence (or a fixed number of iterations),
• Supervised sparse coding: Solve, for all i = 1, . . . ,m,
  $\begin{cases} \alpha_{i,-}^* = \arg\min_\alpha S(\alpha, x_i, D, \theta, -1) \\ \alpha_{i,+}^* = \arg\min_\alpha S(\alpha, x_i, D, \theta, +1) \end{cases}$
• Dictionary and parameters update: Solve
  $\min_{D,\theta} (\sum_{i=1}^{m} \mu c((S(\alpha_{i,-}^*, x_i, D, \theta, -y_i) - S(\alpha_{i,+}^*, x_j, D, \theta, y_i)))$
  $+ (1-\mu) S(\alpha_{i,y_i}^*, x_i, D, \theta, y_i) + \lambda_2 \| \theta \|_2^2) s.t. \forall j, \| d_j \|_2 \leq 1$
```

Fig. 2. SDL: Supervised dictionary learning algorithm

4.1 Supervised Sparse Coding

The supervised sparse coding problem from Eq. (6) (D and θ are fixed in this step) amounts to minimizing a convex function under a ℓ1 penalty. The fixed-point continuation method (FPC) from [27] achieves good results in terms of convergence speed for this class of problems. For our specific problem, denoting by g the convex function to minimize, this method only requires ∇g and a bound on the spectral norm of its Hessian Hg. Since the we have chosen models g which are both linear in α, there exists, for each supervised sparse coding problem, a vector a in Rk and a scalar c in R such that

$$\begin{cases} g(\alpha) = c(a^T \alpha + c) + \lambda_0 \| x - D\alpha \|_2^2 \\ \nabla g(\alpha) = \nabla c(a^T \alpha + c) a - 2\lambda_0 D^T (x - D\alpha) \end{cases}$$

and it can be shown that, if$\| U \|_2$denotes the spectral norm of a matrix U(which is the magnitude of its largest Eigen value), then we can obtain the following bound, $\| H_g(\alpha) \|_2 \leq | H_c(a^T \alpha + c) | \| a \|_2^2 + 2\lambda_0 \| D^T D \|_2$

4.2 Dictionary Update

The problem of updating D and θ in Eq. (11) is not convex in general (except when μ is close to 0), but a local minimum can be obtained using projected gradient descent (as in the general literature on dictionary learning, this local minimum has experimentally been found to be good enough in terms of classification performance). Denoting E(D,θ) the function we want to minimize in Eq. (11), we just need the partial derivatives of E with respect to D and the parameters θ. When considering the linear model for the α's, $f(x, \alpha, \theta) = w^T \alpha + b$, and θ = {w∈Rk, b∈R}, we obtain

$$\begin{cases} \dfrac{\partial E}{\partial D} = -2\lambda_0 (\sum_{i=1}^{m} \sum_{z=\{-1,+1\}} \omega_{i,z} (x_i - D\alpha_{i,z}^{*}) \alpha_{i,z}^{*T}), \\ \dfrac{\partial E}{\partial w} = \sum_{i=1}^{m} \sum_{z=\{-1,+1\}} \omega_{i,z} z \nabla c(w^T \alpha_{i,z}^{*} + b) \alpha_{i,z}^{*}, \\ \dfrac{\partial E}{\partial b} = \sum_{i=1}^{m} \sum_{z=\{-1,+1\}} \omega_{i,z} z \nabla c(w^T \alpha_{i,z}^{*} + b), \end{cases}$$

Where $\omega_{i,z} = -\mu z \nabla c(S(\alpha_{i,-}^{*}, x_i, D, \theta, -y_i) - S(\alpha_{i,+}^{*}, x_i, D, \theta, y_i) + (1-\mu) 1_{z=y_i}$

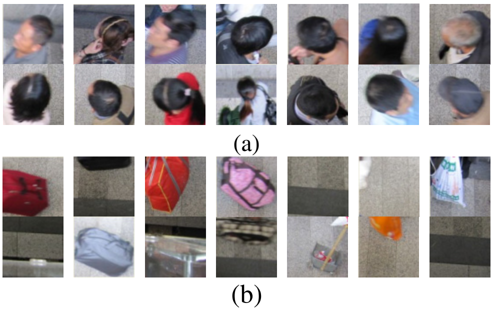
Partial derivatives when using our model with multiple classes or with the bilinear models $f(x,\alpha,\theta) = x^T w\alpha + b$ are not presented in this paper due to space limitations

5 Experiments

To evaluate the performance of the proposed algorithm, we carry out a series of experiments on a dataset extracted 500 images of size 48*96 from a video. If the image is contain a pedestrian, the label of it will be 1, otherwise -1. Fig. 3(a) shows several images with label 1. Fig. 3(b) shows several images with label -1. 100 images from the dataset are selected as the test examples. Different number images of the dataset are selected as the training examples to compare the accuracy rate.

Fig.4 shows the compare results of recognition between with HOG, HOF and Color features respectively and with the corresponding sparse features.Fig.5 shows the result of using mixing features to compare the two methods. As shown in the graph, our method performs better than the method directly using HOG, HOF and Color features to recognition. In addition, with the increasing number of training samples, our method performs better.

Fig.6 shows the result of these two methods using shading images to test. Compared with the traditional method, our method has better recognition accuracy and shows good robustness.



(a)

(b)

Fig. 3. Images with label 1(a) and images with label -1(b)

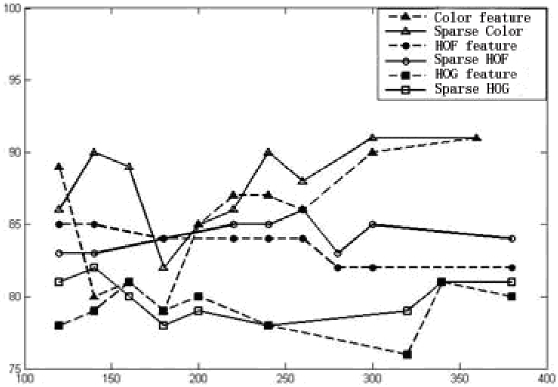


Fig. 4. The compare results of recognition between with HOG, HOF and Color features respectively and with the corresponding sparse features

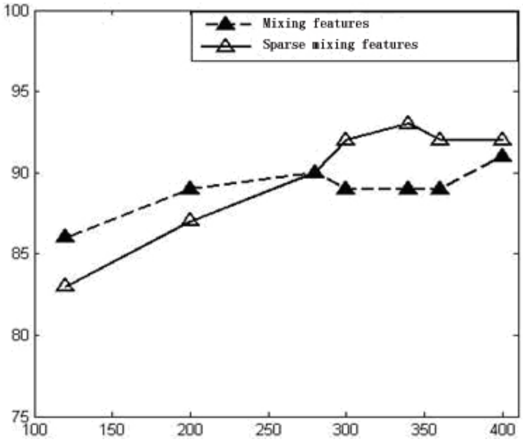


Fig. 5. The result of using mixing features to compare the two methods

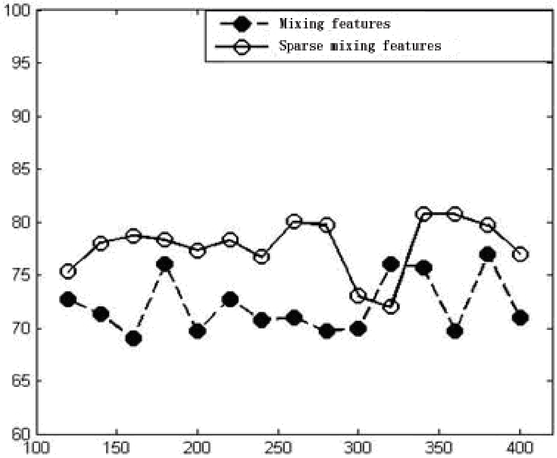


Fig. 6. The result of these two methods using shading images to test

6 Conclusion

We proposed a system for pedestrian detection with very good accuracy. To achieve good classification performance, we put forward a novel framework for pedestrian detection tasks, which proposing a model with both sparse reconstruction and class discrimination components, jointly optimized during dictionary learning. We present an efficient pedestrian detection system using mixing sparse features of HOG, FOG and CSS to combine into this a Kernel classifier. Results presented on our data set show competetive accuracy and robust performance of our system outperforms current state-of-the-art work. Although we use the system for the detection of pedestrians, the general idea can be applied to the detection of other object classes as well.

Acknowledgments. This research is supported by the national science foundation of China (NFSC) No. 61170126, 60273040, provincial universities natural science foundation of Jiangsu province (11KJD520004).

References

1. Andreas, E., Konrad, S., Bastian, L., Luc, G.: Object detection and tracking for autonomous navigation in dynamic environments. International Journal of Robotics Research 14, 1707–1725 (2010)
2. Anthony, C., Yin, J., Sergio, A.: Crowd monitoring using image processing. Electronics and Communication Engineering Journal 1, 37–47 (1995)

3. Kim, C., Human, B.: Gait analysis using Self Organizing Map. In: The 2009 Chinese Conference on Pattern Recognition and the 1st CJK Joint Workshop on Pattern Recognition, pp. 888–891. IEEE Press, Piscataway (2009)
4. Ma, G., Ioffe, A., Stefan, M., Anton, K.: A real time object detection approach applied to reliable pedestrian detection. In: IEEE Intelligent Vehicles Symposium, pp. 755–760. IEEE Press, Piscataway (2007)
5. Marana, A., Cavenaghi, M., Ulson, R., Drumond, F.: Real-time crowd density estimation using images. In: Bebis, G., Boyle, R., Koracin, D., Parvin, B. (eds.) ISVC 2005. LNCS, vol. 3804, pp. 355–362. Springer, Heidelberg (2005)
6. Kong, C., Yang, J., Nie, J.: A study on pedestrian detection models based on the analysis on real accident scenarios. Qiche Gongcheng/Automotive Engineering 11, 977–983 (2010)
7. Mallat, S.: A wavelet tour of signal processing, 2nd edn. Academic Press, New York (1999)
8. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Learning discriminative dictionaries for local image analysis. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, Piscataway (2008)
9. Ranzato, M., Szummer, M.: Semi-supervised learning of compact document representations with deep networks. In: 25th International Conference on Machine Learning, pp. 792–799. ACM Press, New York (2008)
10. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-Task Feature Learning. In: 20th Annual Conference on Neural Information Processing Systems, pp. 41–48. Neural Information Processing System Foundation, Vancouver (2006)
11. Rodriguez, F., Sapiro, G.: Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries. In: IMA Preprint, p. 16 (2008)
12. Blei, D., McAuliffe, J.: Supervised topic models. In: Advances in Neural Information Processing Systems, pp. 208–213. Curran Associates Inc., New York (2007)
13. Holub, A., Perona, P.: A discriminative framework for modeling object classes. In: Conference on Computer Vision and Pattern Recognition, pp. 664–671. IEEE Press, Piscataway (2005)
14. Lasserre, J., Bishop, C., Minka, T.: Principled hybrids of generative and discriminative models. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 87–94. IEEE Press, Piscataway (2006)
15. Raina, R., Shen, Y., Ng, A., McCallum, A.: Classification with hybrid generative/discriminative models. In: Advances in Neural Information Processing Systems, pp. 109–113. MIT Press, British Columbia (2004)
16. Salakhutdinov, R., Hinton, G.: Learning a non-linear embedding by preserving class neighbourhood structure. In: The 11th International Conference on Artificial Intelligence and Statistics, pp. 412–419. Microtome Publishing, Brookline (2007)
17. Larochelle, H., Bengio, Y.: Classification using discriminative restricted boltzmann machines. In: The 25th International Conference on Machine Learning, pp. 536–543. ACM Press, New York (2008)
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893. IEEE Press, Piscataway (2005)
19. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)

20. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 optical flow. In: The British Machine Vision Conference, pp. 123–128. Elsevier Ltd., Oxford (2009)
21. Doll, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: The British Machine Vision Conference, pp. 777–780. Elsevier Ltd., Oxford (2009)
22. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.: Human detection using partial least squares analysis. In: 12th International Conference on Computer Vision, pp. 24–31. IEEE Press, Piscataway (2009)
23. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 794–801. IEEE Press, Piscataway (2009)
24. Sande, K., Gevers, T., Snoek, C.: Evaluation of color descriptors for object and scene recognition. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1582–1596. IEEE Press, Piscataway (2008)
25. Watanabe, T., Ito, S., Yokoi, K.: Co-occurrence histograms of oriented gradients for pedestrian detection. In: Wada, T., Huang, F., Lin, S. (eds.) PSIVT 2009. LNCS, vol. 5414, pp. 37–47. Springer, Heidelberg (2009)
26. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics 32, 407–451 (2004)
27. Hale, E., Yin, W., Zhang, Y.: A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. In: CAAM Tech. Report,TR07-07 (2007)