2014 22nd International Conference on Pattern Recognition

Robust Object Recognition via Visual Pathway Feedback

Chong Wang, Junge Zhang, PeiPei Yang, Kaiqi Huang
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
Email:{cwang, jgzhang, ppyang, kqhuang}@nlpr.ia.ac.cn

***Abstract*—Object recognition, which consists of classification and detection, has two important attributes for robustness: (1) *Closeness*: detection windows should be close to object locations, and (2) *Adaptiveness*: object matching should be adaptive to object variations in classification. It is difficult to satisfy both attributes by considering classification and detection separately, thus recent studies combine them based on confidence contextualization and foreground modeling. However, these combinations neglect feature saliency and object structure, which are important for recognition. In fact, object recognition originates in the mechanism of "what" and "where" pathways in human visual systems, and more importantly, these pathways have feedback to each other, which provides a probable way to improve closeness and adaptiveness. Inspired by the feedback, we propose a robust object recognition framework by designing a computational model of the feedback mechanism. In the "what" feedback, the feature saliency from classification is exploited to rectify detection windows for better closeness; while in the "where" feedback, object parts from detection are used to model object matching of object structure for better adaptiveness. Experiments show that the "what" and "where" feedback can be effective to improve closeness and adaptiveness for robust object recognition, and encouraging results are obtained on the challenging PASCAL VOC 2007 dataset.**

I. Introduction

Object recognition is a fundamental problem in computer vision. It has two basic tasks: classification and detection, which aim to identify object category and location. However, due to large object variations, it is challenging to achieve robust object recognition. Empirical studies propose two important attributes for the robustness [1], [2]: (1) *Closeness*: detection windows should be as close to object locations as possible [1], and (2) *Adaptiveness*: object matching should be adaptive to objects with large variations in classification [2]. In the past decade, most studies consider classification and detection separately [3]–[5], while they are difficult to satisfy both attributes. Based on some biological evidence which shows the dependence between them [6], researchers enhance the robustness by combining them with two primary methods.

The first method is the confidence contextualization, which concentrates on closeness and rectifies detection windows by taking the confidence (score) of classification as context [7], [8]. Harzallah *et al.* [7] score each detection window by combining the confidence of both tasks based on each individual category. Given the fact that other categories provide co-occurrence context [8], Song *et al.* [8] use the confidence of all categories. The second method is the foreground modeling, which focuses on adaptiveness and exploits foregrounds to model object matching [2], [9]. Russakovsky *et al.* [2] divide the foreground into rigid object partitions for matching, and Zhang *et al.* [9] consider spatial arrangements of local features in the foreground to overcome object variations. On some challenging datasets, these two primary methods have achieved improvements.

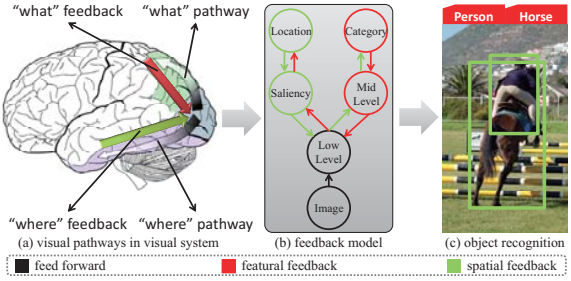

Fig. 1. An illustration of the feedback between visual pathways. (a) the feedback mechanism in the human visual system. (b) the proposed computational model of the feedback mechanism in this paper.

However, the previous methods neglect feature saliency and object structure, which can be important for recognition. Recently, researchers study object recognition from the mechanism of human visual systems [6], which include "what" and "where" pathways, as shown in Fig.1(a). These pathways can obtain object category and location in high-level cortical areas, thus the "what" and "where" pathways have the similar functionality to classification and detection respectively. More importantly, Fig.1(a) shows that they have feedback to each other at the low-level areas, i.e., the "what" and "where" feedback, and these two feedback can carry category and location information to the low-level areas. This feedback mechanism provides a probable way to discover feature saliency and obtain object structure [6].

In this paper, we propose a robust object recognition framework by designing a computational model of the feedback mechianism. Particularly, the bag-of-words (BoW) and the deformable part model (DPM) [5] are used for classification and detection. The feedback model is given in Fig.1(b), in which the *low-level* represents local features, and the *mid-level* denotes the image representation in BoW. Specifically, the feature saliency is obtained from the "what" feedback to rectify detection windows. Then, the object parts obtained from the "where" feedback are used to model object matching of object structure. Finally, they are processed iteratively to achieve robust object recognition. Evaluation on the PASCAL VOC 2007 dataset shows that the "what" and "where" feedback can be effective to improve closeness and adaptiveness, and encouraging improvements are obtained.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.114

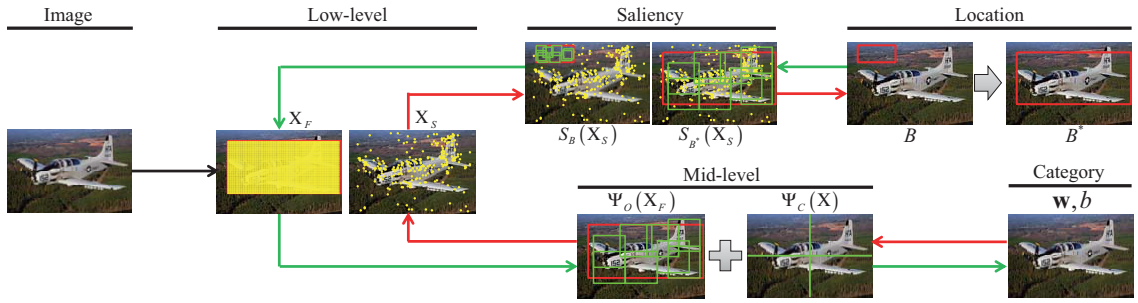

Fig. 2. An illustration of the computational model of the what and where feedback. The red and green arrows represent the what and where feedback respectively, and the image is from the PASCAL VOC 2007 dateset. Best viewed in color.

II. Feedback Model

In this section, we elaborate the feedback model for object recognition. We first give the formulation, and then present the algorithm of the what and where feedback.

A. Formulation

Given N data pairs $\{I_i, y_i\}_{i=1}^N$, wherein I_i is the i^{th} image and $y_i \in \{+1, -1\}$ is the binary label, we formulate object recognition as an optimization problem [2]:

$$\begin{aligned} &\min_{\mathbf{w},b} \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^N \xi_i \\ &s.t.\ \ y_i \max_{B\in \mathrm{BB}(i)} \left[\mathbf{w}^T \Psi_{B(\mathbf{w},b)}(I_i) + b\right] \geq 1 - \xi_i\ , \\ &\quad \xi_i \geq 0\ \ \forall i \end{aligned} \tag{1}$$

wherein $\mathbf{w}$ is the weight vector and b is the bias term. $\Psi_{B(\mathbf{w},b)}(I_i)$ is the image representation of I_i given the detection window B, which belongs to a set of candidate windows $\mathrm{BB}(i)$ generated after detection [5]. Particularly, $B(\mathbf{w}, b)$ denotes that B is dependent on classification.

For the optimization of Eq.1, it is non-convex because of the maximization operation and the unknown image representation in the constraint. Therefore, we propose an iterative procedure to solve the problem. $\mathbf{w}$ and b are first fixed to optimize B, then B is fixed to optimize $\mathbf{w}$ and b. Finally, these two steps are processed iteratively to find the solution. For the optimization of each step, we use the feedback of classification and detection.

B. Algorithm

For classification and detection, the bag-of-words (BoW) and deformable part model (DPM) are used. The *low-level* in Fig.1(b) represents local features, and we denote $X = \{x_j | j = 1, ..., |X|\}$ as a set of $|X|$ local features. For the *mid-level*, it represents the image representation in BoW [3]. Let V be a vocabulary with $|V|$ words and $\phi(x_j)$ be the encoding of x_j on V, then $\Psi(X)$ is the image representation by pooling $\phi(x_j)$ on V. Besides, *category* is represented by $\mathbf{w}$ and b, and *location* is denoted as B that belongs to $\mathrm{BB}(i)$.

*1) **What Feedback**:* Based on the fixed $\mathbf{w}$ and b, the what feedback optimizes B, and the basic idea is to exploit feature saliency. According to Fig.1(b), the what feedback has two main steps: *category* to *low-level* and *low-level* to *location*.

The first step goes from *category* to *low-level*. Based on the maximization in Eq.1, it can be transformed into

$$\begin{aligned} &\max_{B\in \mathrm{BB}(i)} \left(\mathbf{w}^T \Psi_{B(\mathbf{w},b)}(I_i) + b\right) \\ &= \max_{\mathrm{X}} \left(\mathbf{w}^T \Psi(\mathrm{X}) + b\right) \end{aligned}\ , \tag{2}$$

in which $\Psi_{B(\mathbf{w},b)}(I_i)$ becomes $\Psi(X)$ because in each iteration, $\Psi(X)$ uses $B(\mathbf{w}, b)$ from the previous iteration. Maximum pooling is used to construct $\Psi(X)$ [11], and $\Psi(X)$ preserves the maximum encoding of all the local features on each visual word. Thus, Eq.2 can be further transformed into

$$\sum_{v=1}^{|V|} \max_{x_j} \mathbf{w}^T \phi(x_j) + b, \tag{3}$$

wherein $\sum_{v=1}^{|V|} \max_{x_j} \mathbf{w}^T \phi(x_j)$ operates on each word separately to find a set of features x_j with the maximum scores $\mathbf{w}^T \phi(x_j)$. According to the studies on saliency [6], the features maximizing Eq.3 are salient for objects, and they are denoted as X_S, which satisfies $X_S \subset X$. Fig.2 illustrates this feedback, in which the salient features (yellow dots) mainly locate on the object.

Based on X_S, the second step goes from *low-level* to *location* and finds the best detection window B^*. Assume P object parts are used in DPM, the saliency distribution $S_B(\mathrm{X}_S)$ for each window B is constructed as follows:

$$\left(\frac{K_p}{H_p \times W_p}, \frac{K_p}{K}, \frac{(\mu_x)_p}{W}, \frac{(\mu_y)_p}{H}\right), \forall p = 0, 1, ..., P, \tag{4}$$

in which H and W are the height and width of the image, and K is the number of the salient features with the highest scores ($\mathbf{w}^T \phi(x_j)$) in the image. Similarly, K_p is the number in each window ($p = 0$) or part ($p = 1, ..., P$), H_p and W_p are the corresponding height and width, while $(\mu_x)_p$ and $(\mu_y)_p$ are the location expectations of these salient features. Therefore, the saliency distribution describes the density and location of saliency in each detection window, and it is constructed for all candidate windows in $\mathrm{BB}(i)$. Then, a linear SVM is trained with $S_B(\mathrm{X}_S)$ to update their confidence, and the one with the highest confidence is considered as the best window B^*. Fig.2 illustrates this feedback, in which the detection window (red rectangle) is rectified correctly.

*2) **Where Feedback**:* Based on the fixed B^*, the where feedback optimizes $\mathbf{w}$ and b . The basic idea is to obtain object structure for object matching. According to Fig.1(b), the where

feedback has two main steps: *location* to *low-level* and *low-level* to *category*.

The foreground has to be determined at first. Based on the optimized window B^*, we denote the image region in B^* as the foreground and the features in the foreground as foreground features X_F, which satisfies $X_F \subset X$. Fig.2 illustrates these features (yellow area), which are densely distributed in the foreground.

Based on X_F, the second step optimizes $\mathbf{w}$ and b. In DPM, each detection window is associated with some object parts, as the green rectangles shown in Fig.2. The basic idea is to consider these parts as object structure and construct object representation based on them. Assume B^* has P object parts, then X_F is organized into P subsets $X_F^p,\ \forall p = 1, ..., P$. The object representation $\Psi_O(X_F)$ based on X_F is given as follows:

$$\Psi_O(X_F) = \left[\Psi(X_F), \Psi\left(X_F^1\right), ..., \Psi\left(X_F^P\right)\right], \quad (5)$$

in which $\Psi(X_F)$ and $\Psi\left(X_F{}^p\right)$ are the representation on the detection window and object parts respectively. Fig.2 illustrates this step, in which the object matching can be adaptive to objects because the object structure can be robust to object variations.

Many studies show that context can also improve discrimination [2], thus we use the spatial pyramid matching (SPM) [4] based on X. Similarly, P_S rigid partitions are used in SPM, then X is organized into P_S subsets $X^p,\ \forall p = 1, ..., P_S$. We give the context representation $\Psi_C(X)$ based on X as follows:

$$\Psi_C(X) = \left[\Psi\left(X_S^1\right), ..., \Psi\left(X_S^{P_S}\right)\right]. \quad (6)$$

Then, we combine the object representation $\Psi_O(X_F)$ and the context representation $\Psi_C(X)$ to construct the final image representation $\Psi(X)$:

$$\Psi(X) = \left[\Psi_O(X_F), \Psi_C(X)\right]. \quad (7)$$

Finally, based on Eq.1, $\mathbf{w}$ and b are updated for a new iteration. Fig.2 shows this feedback, in which the object and context are combined to enhance classification.

III. Experiments and Results

A. Detailed Settings

Datasets and Evaluation: We evaluate the feedback model on the PASCAL VOC 2007 dataset. In training, the liblinear SVM is trained, and the penalty coefficient is determined by cross-validation. Besides, the Average Precision (AP) and Mean AP (mAP) are reported.

Object Classification: Firstly, SIFT features are densely extracted by every 4 pixels under 3 scales. Then, based on the vocabulary size of $32k$ and SPM with 1×1, 2×2 and 3×1, local-constrained linear coding (LLC) [3] is combined with maximum pooling for BoW representation.

Object Detection: The HoG features are densely extracted at first, then all the root and part filters are applied to score each window. Finally, candidate windows $\mathrm{BB}(i)$ are obtained. Specifically, the number of the object parts (P) and mixtures are set to be 8 and 6, and the voc-release 5.0 code is used.

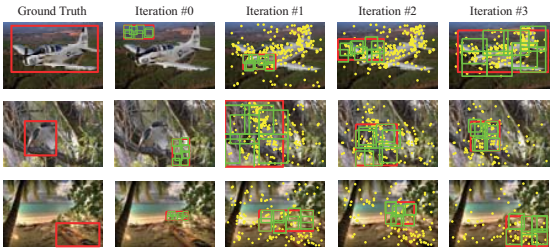

Fig. 3. Some examples of the iterative feedback. *Line 1*: large object in normal background. *Line 2*: small object in complex background. *Line 3*: occluded object in complex background.

B. Main Results

1) ***Iterative Performance****:* Fig.4 shows the classification and detection performance in each iteration. It is observed that the feedback model can improve both tasks consistently, and the improvement is quite large at the beginning, while it becomes narrower as the iteration increases, e.g., more than 1.5% at iteration $\#1$ and decreases to 0.1% at $\#4$. The reason may be that the feedback can largely rectify detection windows and matching at the beginning, and the subsequent iterations will gradually optimize this rectification. Some examples of the iterative feedback in different conditions are shown in Fig.3, in which the detection windows and object matching are rectified for better closeness and adaptiveness.

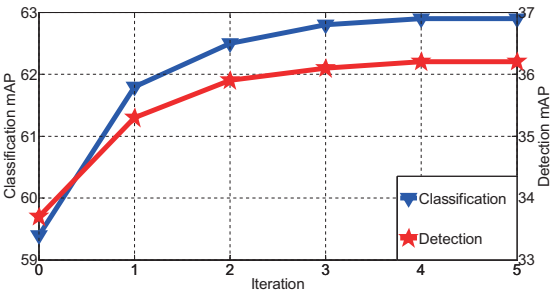

Fig. 4. The iterative performance of object classification and detection by the feedback model on the PASCAL VOC 2007 dataset.

2) ***Detection Evaluation****:* Table.I shows the detection performance of the feedback model and some related methods. The feedback model achieves the mAP of 36.2, which is the best among all the methods and also obtains the best AP on six object categories such as boat, car and person. Particularly, compared with the methods using multiple features [8], [15], the feedback model with the single SIFT features can be competitive. Finally, Fig.5 shows some examples of the detection windows rectified by the feedback model (yellow rectangles) based on DPM (red rectangles), and the feedback model can rectify detection windows to cover most part of the objects around the true locations.

3) ***Classification Evaluation****:* Table.II shows the comparison between the feedback model and some related methods. The feedback model achieves the highest mAP of 62.9, and obtains the best AP on 10 categories. Similar to detection, the feedback model with the single SIFT features can be comparable to the methods with multiple features, which shows the effectiveness of the feedback model and the potential

TABLE I. THE COMPARISON OF DETECTION PERFORMANCE BETWEEN THE FEEDBACK MODEL AND RELATED STUDIES ON PASCAL VOC 2007.

Methods	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	***mAP***
Layout [12]	28.8	56.2	3.2	14.2	29.4	38.7	48.7	12.4	16.0	17.7	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	*27.1*
INRIA-2009 [7]	35.1	45.6	10.9	12.0	23.2	42.1	50.9	19.0	18.0	**31.5**	17.2	17.6	49.6	43.1	21.0	**18.9**	**27.3**	24.7	29.	39.7	*28.9*
Hierarchy [13]	29.4	55.8	9.4	14.3	28.6	44.0	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	*29.6*
MKL [14]	**37.6**	47.8	**15.3**	15.3	21.9	20.7	50.6	**30.0**	17.3	33.0	22.5	**21.5**	51.2	45.5	23.3	12.4	23.9	28.5	45.3	48.5	*32.1*
DPM [5]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	*33.7*
HOG-LBP [15]	36.7	59.8	11.8	17.5	26.3	49.8	58.2	24.0	22.9	27.0	24.3	15.2	58.2	49.2	44.6	13.5	21.4	34.9	47.5	42.3	*34.3*
Sparse Codes [16]	32.2	58.3	11.5	16.3	**30.6**	49.9	54.8	23.5	21.5	27.7	**34.0**	13.7	58.1	51.6	39.9	12.4	23.5	34.4	47.4	45.2	*34.3*
And-or Tree [17]	35.3	60.2	11.0	16.6	29.5	53.0	57.1	23.0	22.9	27.7	28.6	13.1	58.9	49.9	41.4	16.0	22.4	37.2	48.5	42.4	*34.7*
Shared Structure [18]	32.5	60.1	11.1	16.0	31.0	50.9	59.0	26.1	21.2	26.5	25.4	16.4	61.7	48.3	42.2	16.1	28.2	30.1	44.6	46.3	*34.7*
Color Attribute [19]	34.5	61.1	11.5	19.0	22.2	46.5	58.9	24.7	21.7	25.1	27.1	13.0	59.7	51.6	44.0	19.2	24.4	33.1	48.4	**49.7**	*34.8*
DPM+Context [5]	36.6	**62.2**	12.1	17.6	28.7	**54.6**	60.4	25.5	21.1	25.6	26.6	14.6	60.9	50.7	44.7	14.3	21.5	**38.2**	**49.3**	43.6	*35.4*
Feedback Model	35.1	60.9	14.8	**20.6**	29.4	51.4	**60.5**	26.9	**23.0**	29.2	27.1	17.1	**61.7**	**51.7**	**46.2**	16.1	22.2	37.1	48.2	44.8	***36.2***

TABLE II. THE COMPARISON OF CLASSIFICATION PERFORMANCE BETWEEN THE FEEDBACK MODEL AND RELATED STUDIES ON PASCAL VOC 2007.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	***mAP***
Object+Context [20]	**80.2**	61.0	49.8	69.6	21.0	66.8	80.7	51.1	51.4	35.9	**62.0**	38.6	69.0	61.4	84.6	28.7	53.5	**61.9**	**81.7**	59.5	*58.4*
2007 Winner [21]	77.5	63.6	56.1	**71.9**	33.1	60.6	78.0	58.8	53.5	42.6	54.9	45.8	77.5	64.0	85.9	**36.3**	44.7	50.6	79.2	53.2	*59.4*
LLC+SPM [4]	73.7	65.7	49.9	68.7	28.1	66.2	78.4	60.4	55.9	49.4	52.6	45.5	77.4	68.0	84.3	29.1	46.8	56.3	77.0	53.7	*59.4*
LLC+OCP [2]	74.7	70.1	52.8	69.0	34.2	67.8	81.3	62.0	56.7	49.9	54.3	47.1	79.2	69.0	85.4	30.1	48.7	58.5	77.4	59.5	*61.4*
BoF+HOG [22]	75.0	68.3	**58.2**	69.5	33.3	**68.9**	80.0	**65.8**	55.9	50.9	60.6	**50.4**	77.6	70.6	86.2	31.6	49.6	56.9	78.9	55.5	*62.2*
Feedback Model	76.9	**71.0**	54.7	70.4	**35.1**	68.7	**82.3**	64.3	**57.8**	**51.7**	57.5	48.9	**80.1**	**70.7**	**86.4**	31.6	**50.2**	59.4	79.4	**60.3**	***62.9***

for further enhancement. Fig.3 shows some examples of the object matching by the feedback model. Though objects vary a lot in size, location and orientation, object structure can be adaptive, which validates the effectiveness of the "where" feedback in improving adaptiveness for classification.

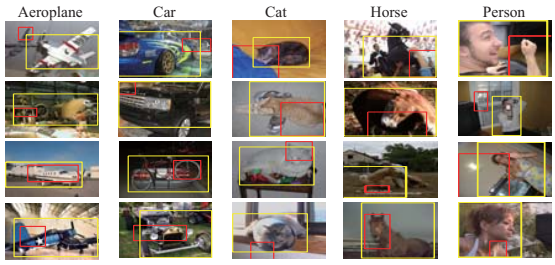

Fig. 5. Some rectified detection windows by the feedback model (yellow rectangles) based on the DPM (red rectangles).

IV. CONCLUSION

In this paper, we have proposed a robust object recognition framework by designing a computational model of the visual pathway feedback between classification and detection. In the feedback model, we discover feature saliency in the what feedback to enhance detection and exploit the object structure in the where feedback to enhance classification. Experiments on the challenging PASCAL VOC 2007 dataset demonstrate that the what and where feedback can effectively rectify detection windows and give robust object matching. As a result, better closeness and adaptiveness are achieved for robust object recognition, and encouraging improvements have been obtained. In the future, we will extend the feedback model to the multi-label problem.

ACKNOWLEDGEMENT

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209 and Grant No. 61175007), the National Key Technology R&D Program (Grant No. 2012BAH07B01).

REFERENCES

[1] J. Zhang, X. Zhao, Y. Huang, K. Huang, and T. Tan, "Semantic windows mining in sliding window based object detection," in *ICPR*, 2012.
[2] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *ECCV*, 2012.
[3] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.
[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.
[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *TPAMI*, vol. 32, no. 9, 2010.
[6] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A bayesian inference theory of attention," *Vision Research*, vol. 50, no. 22, 2010.
[7] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *ICCV*, 2009.
[8] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *CVPR*, 2011.
[9] Y. Zhang and T. Chen, "Weakly supervised object recognition and localization with invariant high order features," in *BMVC*, 2010.
[10] E. Barenholtz and M. J. Tarr, "Reconsidering the role of structure in vision," *The Psychology of Learning and Motivation*, vol. 47, 2007.
[11] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *CVPR*, 2010.
[12] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *ICCV*, 2009.
[13] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman, "Latent hierarchical structural learning for object detection," in *CVPR*, 2010.
[14] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *ICCV*, 2009.
[15] J. Zhang, K. Huang, Y. Yu, and T. Tan, "Boosted local structured hog-lbp for object localization," in *CVPR*, 2011.
[16] X. Ren and D. Ramanan, "Histograms of sparse codes for object detection," in *CVPR*, 2013.
[17] X. Song, T. Wu, Y. Jia, and S.-C. Zhu, "Discriminatively trained and-or tree models for object detection," in *CVPR*, 2013.
[18] X. Wang, L. Lin, L. Huang, and S. Yan, "Incorporating structural alternatives and sharing into hierarchy for multiclass object recognition and detection," in *ICCV*, 2013.
[19] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color attributes for object detection," in *CVPR*, 2012.
[20] J. Uijlin, A. Smeulders, and R. Scha, "What is the spatial extent of an object," in *CVPR*, 2009.
[21] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer, "Learning representations for visual object class recognition," in *ICCV*, 2007.
[22] T. Kobayashi, "Bof meets hog: Feature extraction based on histograms of oriented p.d.f gradients for image classification," in *CVPR*, 2013.