Wang C, Huang KQ. VFM: Visual feedback model for robust object recognition. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(2): 325–339 Mar. 2015. DOI 10.1007/s11390-015-1526-1

VFM: Visual Feedback Model for Robust Object Recognition

Chong Wang (王 冲), *Student Member, IEEE*, and Kai-Qi Huang* (黄凯奇), *Senior Member, IEEE*

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

E-mail: chongwang.nlpr@gmail.com; kqhuang@nlpr.ia.ac.cn

Received December 19, 2014; revised February 2, 2015.

Abstract Object recognition, which consists of classification and detection, has two important attributes for robustness: 1) closeness: detection windows should be as close to object locations as possible, and 2) adaptiveness: object matching should be adaptive to object variations within an object class. It is difficult to satisfy both attributes using traditional methods which consider classification and detection separately; thus recent studies propose to combine them based on confidence contextualization and foreground modeling. However, these combinations neglect feature saliency and object structure, and biological evidence suggests that the feature saliency and object structure can be important in guiding the recognition from low level to high level. In fact, object recognition originates in the mechanism of "what" and "where" pathways in human visual systems. More importantly, these pathways have feedback to each other and exchange useful information, which may improve closeness and adaptiveness. Inspired by the visual feedback, we propose a robust object recognition framework by designing a computational visual feedback model (VFM) between classification and detection. 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 match object structure for better adaptiveness. Experimental results show that the "what" and "where" feedback is effective to improve closeness and adaptiveness for object recognition, and encouraging improvements are obtained on the challenging PASCAL VOC 2007 dataset.

Keywords object recognition, object classification, object detection, visual feedback

1 Introduction

Object recognition is a fundamental problem in computer vision. It has been widely used in many vision applications such as image retrieval, scene understanding, and visual surveillance. Object recognition is easy to be confused with classification and detection. Usually, classification only gives object category, while detection only gives object location, such as the tasks in PASCAL VOC[1] and ImageNet[2]. For object recognition, based on the definition in Wikipedia①, its task is to identify and find objects in an image or video, which includes both classification and detection. Therefore, in this paper, we consider that object recognition has two basic tasks, classification and detection, and they identify the object category and find the object location respectively.

In the past decade, many studies on object recognition have been proposed, such as the bag-of-words model[3-7] and convolutional neural network (CNN)[8-12] in classification, and the deformable part model[13-16] and region-CNN[17-21] in detection. However, due to large object variations and cluttered backgrounds, it is quite challenging to achieve robust object recognition. Empirical studies propose two important attributes for robustness[22-23]: 1) closeness: detection windows should be as close to object locations as possible[22], and 2) adaptiveness: object matching should be adaptive to large variations within an object class[23]. As an illustration, Fig.1 shows some examples

Regular Paper

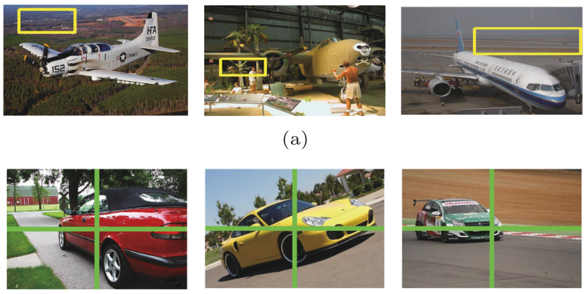
Special Section on Object Recognition

This work was supported by the National Basic Research 973 Program of China under Grant No. 2012CB316302, the National Natural Science Foundation of China under Grant Nos. 61322209 and 61175007, the National Key Technology Research and Development Program of China under Grant No. 2012BAH07B01.

* Corresponding Author

① http://en.wikipedia.org/wiki/Outline_of_object_recognition, Jan. 2015.

that do not satisfy these two attributes. In Fig.1(a), some detection windows do not cover any target objects, but they have a high confidence for detection; in Fig.1(b), due to large object variations in location, size, and orientation, rigid partitions cannot achieve robust matching.


(a)

(b)

Fig.1. Illustration of the recognition that does not satisfy closeness or adaptiveness. (a) Detection windows that do not satisfy closeness: some detection windows are not close to objects or do not cover any part of the objects. (b) Object matching that does not satisfy adaptiveness: the rigid matching cannot adapt to variations in location, size, and orientation.

In recent years, most studies consider classification and detection separately[4-5,7,13,24], and they are difficult to satisfy both closeness and adaptiveness. Based on some biological evidence and empirical studies which show the dependence between the two tasks[25-27], researchers enhanced the robustness of recognition by combining classification and detection using two primary methods. The first one is confidence contextualization, which concentrates on closeness and rectifies detection windows by taking the confidence (score) of classification as the context of detection[26-29]. Harzallah *et al.*[28] scored 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[26-27,29-30], Song *et al.*[29] proposed to use the confidence of all categories. The second one is foreground modeling, which focuses on adaptiveness and exploits possible foregrounds to model object matching for classification[23,31-33]. Russakovsky *et al.*[23] divided the foreground into rigid object partitions for matching, and Chen *et al.*[32] used the confidence of detection to segment objects by considering the spatial semantics of images. On some challenging datasets such as PASCAL VOC[1], these two primary methods have improved the robustness of object recognition.

However, the above methods have two limitations regarding to closeness and adaptiveness. For the closeness, the confidence contextualization considers the classification confidence (score) as context, but if the score is not correct, the detection windows will deviate from the object locations. As a result, the confidence contextualization can only provide limited context for post-processing. It neglects feature saliency[25,32,34], and biological evidence suggests that the saliency is important in guiding the recognition process from low level to high level[25,35-36]. Another limitation is that for the adaptiveness, the foreground modeling only considers rigid partitions of objects, but large object variations in size and pose make it difficult for these partitions to achieve robust object matching. Psychological studies show that object structure can preserve the deformation invariance of the objects within an object class to give robust object matching[37-38], but the foreground modeling does not take it into consideration.

Recently, researchers have sought a solution of robust recognition from the mechanism of human visual systems[25,39-41], which include "what" and "where" pathways, as shown in Fig.2(a). These pathways transfer visual information from low-level cortical areas to high-level ones, and they eventually yield object category and location. Therefore, the "what" and "where" pathways have the similar functionality to classification and detection respectively, and the robustness of object recognition may originate in the mechanism of these pathways. More importantly, Fig.2(a) shows that these two pathways have feedback to each other at the low-level areas, i.e., the "what" and "where" feedback from the high-level areas to the low-level ones. These two feedbacks carry useful information, e.g., the "what" feedback passes back the category information which provides feature saliency to improve closeness, and the "where" feedback passes back the object information which provides object structure to improve adaptiveness. Therefore, this feedback mechanism provides a suitable way of improving the robustness of object recognition.

In this paper, we propose robust object recognition by designing a computational visual feedback model (VFM) between classification and detection. Particularly, the bag-of-words (BoW)[3] and the deformable part model (DPM)[13] are used for classification and detection respectively. The feedback model is given in Fig.2(b), in which the *low level* represents local features, and the *mid-level* denotes the image representation after feature pooling in BoW[42]. Specifically, the "what" feedback first passes feature saliency back to local features, which construct saliency distribution to

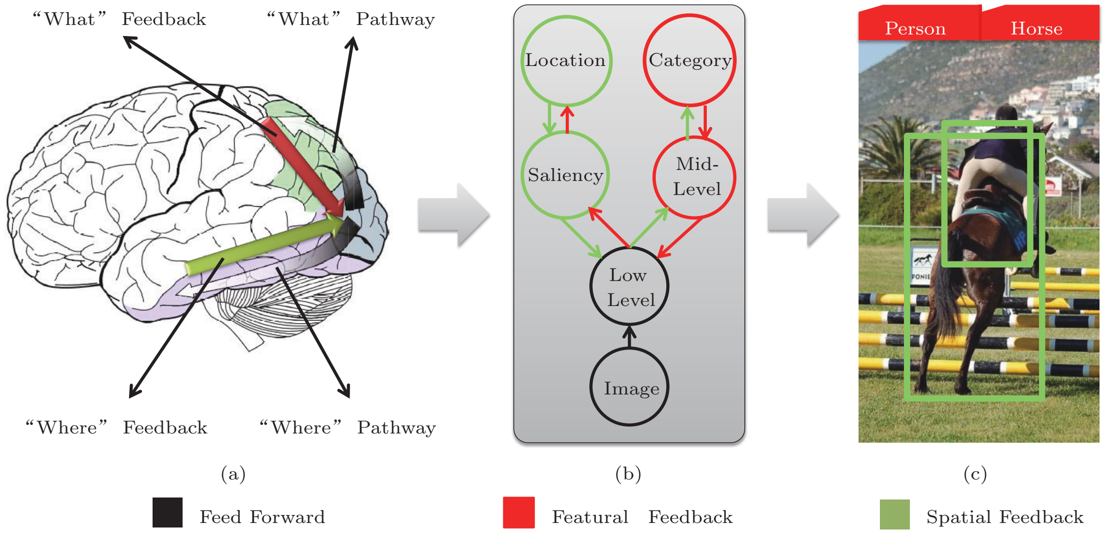


Fig.2. Illustration of the object recognition with the feedback of visual pathways. (a) Feedback mechanism in the human visual system. (b) Proposed computational model of the feedback mechanism in this paper. (c) Output of object recognition: object category and location.

rectify detection windows for better closeness. Then, the "where" feedback passes foreground information back to local features, which are combined with object parts to model the matching of object structure for better adaptiveness. Finally, these two steps are processed iteratively to achieve robust object recognition. The visual feedback model (VFM) is evaluated on the challenging and popular PASCAL VOC 2007 dataset. Experimental results show that the "what" and "where" feedback is effective to improve closeness and adaptiveness, and this robust recognition framework boosts simultaneously classification and detection by a large margin.

There are three contributions in this paper.

- We exploit a biologically inspired feedback mechanism in human visual systems to achieve robust object recognition.
- We propose a visual feedback model (VFM) between classification and detection by designing a computational way of the feedback mechanism.
- We accomplish simultaneous classification and detection to imitate human visual systems, and an initial exploration is taken in this paper.

The rest of the paper is organized as follows. In Section 2, we first review the studies in improving the robustness of object recognition. Then, Section 3 elaborates the visual feedback model for robust object recognition and Section 4 provides the detailed evaluation of the model. Finally, the paper is summarized with some concluding remarks in Section 5.

2 Related Work

In the past decade, many studies have considered robust recognition by combining classification and detection[3,8,13,17,43-47]. In some typical work of classification, the biological inspired model was first proposed based on biological evidence[43,48]. Inspired by the success of the bag-of-words model in document analysis[3], many dictionary learning[49-51] and feature encoding methods[4-7,42,51-54] have been proposed to enhance classification. Recently, with the great progress of theoretical achievements and parallel computing[44], deep neural networks further enhance classification with deep hierarchical structures[8-12,44-45]. Compared to classification, some models were also proposed in detection. To describe object shape, the rigid template matching was first proposed based on sliding window based searching[46]. Inspired by the importance of object structure[37-38,55-57], researchers proposed the popular deformable part model[13], which is a milestone in detection and triggered many popular models such as star models[14,58], tree models[59-60], grammar models[16] and graph matching based models[47,61-62]. Recently, based on the powerful representation of the convolutional neural networks, the region-CNN (RCNN) based methods have shown the impressive improvement on the detection tasks[17-21]. Besides, it has also become the state-of-the-art method in the popular ImageNet competition.

Though these methods have achieved promising performance in classification and detection, the separate consideration of these two tasks makes it difficult to satisfy both closeness in detection and adaptiveness in classification for robust object recognition. Based on some empirical studies on context[26-27,30,63] and biological evidence on the mechanism of object recognition in human visual system[25,64-66], researchers found that classification and detection are closely linked to each other. Therefore, many studies have been proposed to combine classification and detection together. These combinations can be categorized into two primary methods as follows.

The first primary method is confidence contextualization, which focuses on closeness in detection and rectifies detection windows by taking the confidence (score) of classification as the context of detection[28-30]. Based on the probabilistic model, Harzallah *et al.*[28] first combined the confidence of both tasks on each individual object category, and promising improvements were obtained. Given the fact that other categories provide co-occurrence information which is beneficial to infer the existence of other categories[26-27], Song *et al.*[29] proposed a context-SVM to combine the confidence of other categories as object co-occurrence. They have obtained the state-of-the-art performance on challenging datasets in both classification and detection. To further enhance the power of the co-occurrence context, Chen *et al.*[30] proposed to model the multi-order contextual co-occurrence and further improvements have been achieved. However, one limitation is that the confidence contextualization considers only the confidence. If the confidence (score) is not correct, the detection windows will deviate from the correct object locations. It neglects the feature saliency[25,32,34], which can be important in guiding the recognition process from low level to high level[25,35-36].

The second primary method is foreground modeling, which concentrates on adaptiveness in classification and models object matching by using possible foregrounds obtained in detection[23,31-33,67-69]. To determine the foreground region, Nguyen *et al.* applied a classifier trained on candidate detection windows, and then the window with the maximum score was considered as the foreground[33]. Based on the evidence that accurate foreground-background segmentation benefits object recognition[70], Chai *et al.*[67] used the co-segmentation to discover the possible foreground based on different image levels, and Chen *et al.*[32] further exploited the confidence of object detection to segment the discriminative foreground region. These studies mainly model the object matching in the foreground region by the bag-of-words representation, while they do not consider the spatial arrangement, which is important in matching objects. Based on this fact, Crandall *et al.*[68] proposed to learn the appearance and the spatial relation of local image parts simultaneously, and Zhang *et al.*[31] enhanced the spatial relation by considering the higher-order arrangement of local features in the foreground. To better exploit background information as useful context, Russakovsky *et al.*[23] considered the foreground-background representation based on the bag-of-words model[3], and Pandey *et al.*[71] used the deformable part model[13] to capture the component of the background scenes. Though these methods yield promising improvements, one limitation is that the rigid partitions of objects make it difficult to overcome large object variations. The foreground modeling neglects the object structure, which can be critical in preserving the deformation invariance of the objects within an object class to give robust object matching[37-38].

As introduced above, the previous two primary methods consider neither the feature saliency nor the object structure, which are important in discovering feature saliency for better closeness and preserving deformation invariance for better adaptiveness. To incorporate feature saliency and object structure for robust recognition, we use the "what" and "where" pathway feedback inspired by studies of human visual systems. The system consists of the "what" and "where" pathways, which have the same functionality to classification and detection[65-66]. More importantly, these two pathways have feedback to each other, named the "what" and the "where" feedback in Fig.2. The feedback of one task carries useful information to the other one; thus this feedback mechanism provides a suitable way to satisfy both closeness and adaptiveness. Inspired by this feedback, in this paper, we propose the visual feedback model (VFM) by designing a computation way of the feedback mechanism for robust object recognition.

3 Visual Feedback Model

In this section, we elaborate the visual feedback model (VFM) for object recognition. Firstly, we give the formulation of the model. Secondly, the algorithm of the feedback is provided in detail. Finally, we give the iterative procedure for the feedback between classification and detection.

3.1 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 denoting the image-level category. We formulate the visual feedback model as an optimization problem[23,33]:

$$\begin{aligned}&\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum\nolimits_{i=1}^{N} \xi_i\\ &s.t.\ y_i \max_{B\in BB(i)} (\boldsymbol{w}^{\mathrm{T}}\Psi_{B(\boldsymbol{w},b)}\left(I_i\right) + b) \geqslant 1 - \xi_i,\\ &\xi_i \geqslant 0,\ \forall i,\end{aligned} \tag{1}$$

wherein $\boldsymbol{w}$ is the weight vector and b is the bias term. $\Psi_{B(\boldsymbol{w},b)}\left(I_i\right)$ is the image representation of I_i given the detection window B, which belongs to a set of candidate windows $BB(i)$ generated after detection[13]. Particularly, $B(\boldsymbol{w}, b)$ denotes that the window B is determined by the feedback of classification.

The optimization of (1) 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 optimization problem. Firstly, $\boldsymbol{w}$ and b in classification are fixed to optimize B in detection, and this is the "what" feedback. Secondly, the window B in detection is fixed to optimize $\boldsymbol{w}$ and b in classification, which is the "where" feedback. Finally, these two feedbacks are processed iteratively to find the solution.

3.2 Algorithm

In this part, we give a computational way of finding the "what" and "where" feedback, and show how they can be used to improve closeness and adaptiveness. For classification and detection, the bag-of-words (BoW) and the deformable part model (DPM) are used respectively. As discussed in [51], the low level in Fig.2(b) denotes local features, and we denote $X = \{x_j | j = 1, ..., |X|\}$ as a set of $|X|$ local features, e.g., SIFT[72]. For the mid-level in the "what" pathway, it represents the image representation after feature pooling in BoW[51]. Let V be a visual vocabulary with $|V|$ visual words and $\phi\left(x_j\right)$ be the encoding of x_j on V. Then $\Psi\left(X\right)$ is the image representation by pooling $\phi\left(x_j\right)$ on V. Besides, for the saliency in the "where" pathway, it is the saliency distribution defined in this paper, and we will give the definition later. Finally, object category is represented by $\boldsymbol{w}$ and b in classification, and object location is represented by B in the set $BB(i)$ in detection. We begin the optimization with classification, and thus the initial $\boldsymbol{w}$ and b are known.

3.2.1 "What" Feedback

Based on the fixed $\boldsymbol{w}$ and b in classification, the "what" feedback optimizes B in detection. The idea is to exploit feature saliency to rectify detection windows. According to Fig.2(b), the "what" feedback transfers information from category to location. It has two main steps: category to low level and low level to location.

The first step is to go from category to low level and obtain feature saliency. The maximization term in the constraint of (1) can be transformed into

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

in which X_S is a subset of X, and we denote X_S as a set of salient local features locating in the object region B. We use the maximum pooling to construct $\Psi\left(X_S\right)$[4,42,51], which preserves the maximum encoding of all the local features on each visual word. Thus, (2) can be further transformed into

$$\sum\nolimits_{v=1}^{|V|} \max_{x_j} \boldsymbol{w}^{\mathrm{T}}\phi(x_j) + b, \tag{3}$$

wherein $\sum\nolimits_{v=1}^{|V|} \max\limits_{x_j}\ \boldsymbol{w}^{\mathrm{T}}\phi(x_j)$ operates on each word separately to find a set of features x_j with the maximum score $\boldsymbol{w}^{\mathrm{T}}\phi(x_j)$. According to the previous studies on saliency[25,34], the features (x_j) maximizing (3) are salient for objects. Therefore, the features x_j constitute the feature set X_S. Fig.3 illustrates X_S, wherein most local features are on the target object.

Based on X_S, the second step goes from low level to location and finds the best detection window B^*. Physiological evidence shows that X_S has an object-oriented saliency distribution[25], and we consider B^* as the window which best represents the distribution. To describe the distribution, we exploit the density and location of the saliency features, and we use both the detection window and its corresponding object parts in DPM[13] to accurately describe the distribution. Assume P object parts are used in DPM, and then the saliency distribution $S_B\left(X_S\right)$ 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 the width of the image, and K is the number of the salient features with the high scores $(\boldsymbol{w}^{\mathrm{T}}\phi(x_j))$ in the image. Similarly, K_p is the number of the salient features in each window $(p = 0)$ or part $(p = 1, ..., P)$, and H_p and W_p are

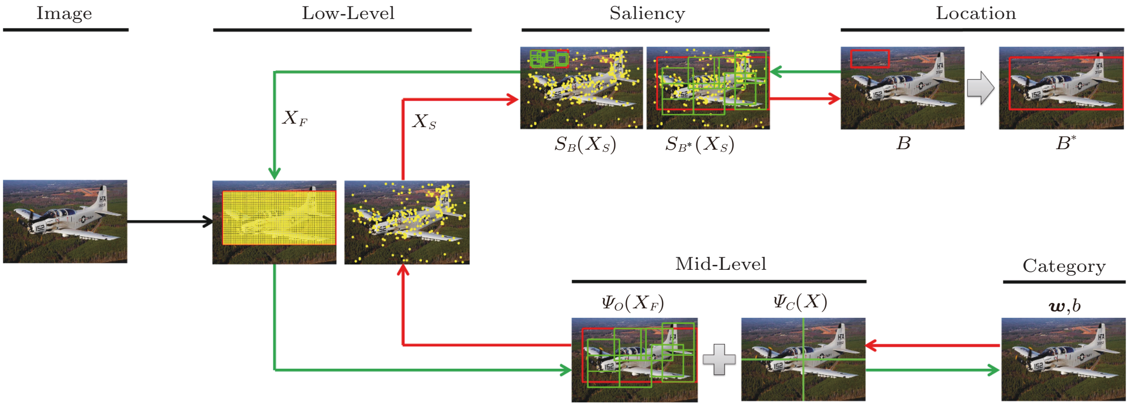


Fig.3. Illustration of the computational model of the "what" and "where" feedback. The red and the green arrows represent the "what" and the "where" feedback respectively, and the image is from the PASCAL VOC 2007 dataset.

the corresponding height and width. $K_p/(H_p \times W_p)$ and K_p/K measure the density distribution of X_S in each bounding box region. Compared to the density, $(\mu_x)_p/W$ and $(\mu_y)_p/H$ aim to describe the location distribution of the salient features, and they denote the mean location of X_S in each bounding region or part in both x and y axis. Therefore, the saliency distribution in (4) models the density and location of saliency in each detection window and object part, and it is constructed for all detection windows in $BB(i)$. Finally, a linear SVM classifier is trained with $S_B(X_S)$ to update the confidence of each window, and the one with the highest confidence is considered as the best window B^*. Fig.3 illustrates this window rectification, in which the detection window (red rectangle) is rectified correctly.

3.2.2 "Where" Feedback

Based on the optimized B^* in detection, the "where" feedback optimizes $\boldsymbol{w}$ and b in classification. The idea is to use the object parts in foreground as object structure to give robust object matching. According to Fig.2(b), the "where" feedback transfers information from location to category. It has two main steps: location to low level and low level to category.

The first step is to go from location to low level to determine the foreground region. Based on the optimized detection window B^*, which represents the most probable location of the target object, we denote the image region in B^* as the foreground and the local features in it as foreground features X_F, which satisfies $X_F \subset X$. Fig.3 illustrates these foreground features (yellow area), which are densely distributed in the foreground region.

Based on X_F, the second step is to optimize $\boldsymbol{w}$ and b. To obtain the object structure from the foreground, the DPM model provides a convenient way. In DPM, each detection window is associated with some object parts, as the green rectangles shown in Fig.3. The idea is to organize these object parts as object structure and construct an object representation based on it. Assume the optimized window B^* has P parts, then the foreground features X_F can be divided into P subsets X_F^p, $\forall p = 1, ..., P$. The object representation $\Psi_O\left(X_F\right)$ based on X_F is given as follows:

$$\Psi_O\left(X_F\right) = (\Psi(X_F), \Psi(X_F^1), ..., \Psi(X_F^P)),$$

in which $\Psi(X_F)$ and $\Psi(X_F^p)$ are the object representation in the detection window and object parts respectively. Fig.3 illustrates this step, in which the object matching can be adaptive to object variations because the object structure is preserved under the deformations in the image.

Many studies show that the surrounding context of objects also contributes a lot to classification[23,26-27]. To construct context representation, we use the popular spatial pyramid matching (SPM)[24] based on the feature set X. Similarly, P_S rigid partitions are used in SPM, and then X can be divided into P_S subsets X^p, $\forall p = 1, ..., P_S$. We give the context representation $\Psi_C\left(X\right)$ based on X as follows:

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

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

$$\Psi\left(X\right) = (\Psi_O(X_F), \Psi_C(X)).$$

Finally, based on (1), $\boldsymbol{w}$ and b are updated for a new iteration. Fig.3 shows this step, in which the object and the context are combined to enhance classification.

3.2.3 Unification

As a summary, the iterative treatment of the "what" and the "where" feedback in the visual feedback model (VFM) is given below. For initialization, we denote $\boldsymbol{w}(0)$, $b(0)$ and $\Psi(0)$ as the initial weight vector, bias term and image representation in classification, and denote $B^*(0)$ as the initial detection window in detection.

At the t-th iteration, $\boldsymbol{w}(t-1)$, $b(t-1)$ and $\Psi(t-1)$ are first used to construct the salient features $X_S(t)$, from which the saliency distribution $S_B(t)$ is constructed and used to update $B^*(t-1)$ to $B^*(t)$. Then, the foreground features $X_F(t)$ from $B^*(t)$ are obtained and used to construct the image representation $\Psi(t)$. Finally, $\Psi(t)$ is trained to update $\boldsymbol{w}(t-1)$ and $b(t-1)$ to $\boldsymbol{w}(t)$ and $b(t)$ for a new iteration. The iterative process terminates when the detection window B^* is unchanged or the maximum iteration number T is reached.

4 Experimental Evaluation

In this part, we give the detailed experimental evaluation of the proposed visual feedback model (VFM). The experimental setup is as follows.

Datasets and Evaluation. We use the popular PASCAL VOC 2007 dataset for evaluation[1]. The PASCAL VOC 2007 dataset is challenging because of its large object variations on illumination, size, deformation, occlusion, etc. The dataset contains 20 object categories, such as aeroplane and bicycle, as shown in Table 1. There are 9 963 images in total in this dataset, of which 5 011 are used for training/validation and the rest 4 952 for testing. For the evaluation of classification and detection, the average precision (AP) of each category and the mean average precision (mAP) of all the categories are reported. In detection, a detection window is considered as the correct detection if it has an overlap of more than 0.5 with the ground truth bounding box.

Object Classification. In classification, we use the common settings of the bag-of-words (BoW) model[5]. In detail, we use the VLFeat toolbox[76] to densely extract SIFT features[72] by every 4 pixels under 3 scales: 16 × 16, 24 × 24 and 32 × 32. Then, to construct a visual vocabulary with the size of 32 768, we adopt the Approximate Nearest Neighbor (ANN) algorithm, which is efficient in clustering large vocabulary. Finally, based on the partitions of the spatial pyramid matching (SPM) with 1×1, 2×2 and 3×1, local-constrained linear

Table 1. Detection Performance of VFM and Some Previous Methods on PASCAL VOC 2007

Object Categories	Layout[56]	INRIA-2009[28]	Hierarchy[15]	DPM[13]	HOG-LBP[73]	Shared Structure[74]	Color Attribute[75]	DPM+Context[13]	VFM
Plane	28.8	35.1	29.4	33.2	36.7	32.5	34.5	36.6	35.1
Bicycle	56.2	45.6	55.8	60.3	59.8	60.1	61.1	**62.2**	60.9
Bird	3.2	10.9	9.4	10.2	11.8	11.1	11.5	12.1	14.8
Boat	14.2	12.0	14.3	16.1	17.5	16.0	19.0	17.6	**20.6**
Bottle	29.4	23.2	28.6	27.3	26.3	31.0	22.2	28.7	29.4
Bus	38.7	42.1	44.0	54.3	49.8	50.9	46.5	**54.6**	51.4
Car	48.7	50.9	51.3	58.2	58.2	59.0	58.9	60.4	**60.5**
Cat	12.4	19.0	21.3	23.0	24.0	26.1	24.7	25.5	26.9
Chair	16.0	18.0	20.0	20.0	22.9	21.2	21.7	21.1	**23.0**
Cow	17.7	**31.5**	19.3	24.1	27.0	26.5	25.1	25.6	29.2
Table	24.0	17.2	25.2	26.7	24.3	25.4	27.1	26.6	27.1
Dog	11.7	17.6	12.5	12.7	15.2	16.4	13.0	14.6	17.1
Horse	45.0	49.6	50.4	58.1	58.2	**61.7**	59.7	60.9	**61.7**
Motor	39.4	43.1	38.4	48.2	49.2	48.3	51.6	50.7	**51.7**
Person	35.5	21.0	36.6	43.2	44.6	42.2	44.0	44.7	**46.2**
Plant	15.2	**18.9**	15.1	12.0	13.5	16.1	19.2	14.3	16.1
Sheep	16.1	**27.3**	19.7	21.1	21.4	28.2	24.4	21.5	22.2
Sofa	20.1	24.7	25.1	36.1	34.9	30.1	33.1	**38.2**	37.1
Train	34.2	29.0	36.8	46.0	47.5	44.6	48.4	**49.3**	48.2
TV	35.4	39.7	39.3	43.5	42.3	46.3	**49.7**	43.6	44.8
mAP	*27.1*	*28.9*	*29.6*	*33.7*	*34.3*	*34.7*	*34.8*	*35.4*	***36.2***

coding (LLC)[5] is combined with maximum pooling[4] to construct the final BoW image representation.

Object Detection. We use the same settings as for the deformable part model[13]. The variant histogram of gradients (HoG) features are densely extracted at first, and then all the root and part filters are applied to score each detection window. Finally, candidate detection windows $BB(i)$ are obtained after non-maximum suppression (NMS) with the overlap threshold of 0.5. Besides, the number of the object parts (P) and mixtures is set to be 8 and 6 respectively. The voc-released 5.0 code is used[77]②.

Training. There are three SVM classifiers in the visual feedback model: 1) the latent SVM in detection, which is used to train the detection model[13]; 2) one liblinear SVM in classification, which is used to train the classification model of the BoW image representation; 3) another liblinear SVM in classification, which is used to train the saliency distribution of all the detection windows. For the first one, we use the same settings to the DPM and the source code 5.0[13]. For the other two, we select the best SVM penalty term by cross-validation on the validation set.

4.1 Iterative Performance

Fig.4 shows the classification and detection performance in each iteration on the PASCAL VOC 2007 test set. It is observed that the visual feedback model (VFM) improves both tasks consistently, e.g., from iteration #1 to iteration #4, the improvement is at least 0.1% and 0.2% for detection and classification respectively. We also see that the improvement is quite large at the beginning, but it becomes smaller as the iteration increases, e.g., the improvement of detection is more than 1.5% at iteration #1 and decreases to 0.1% at #4. The reason of this may be that the initial classification and detection results can be easily enhanced by the "where" and the "what" feedback, and the VFM model will rectify most recognition as the iteration increases, and thus the performance will gradually achieve a stable value. Fig.5 shows some examples rectified by the feedback model in each iteration. These samples have different image conditions, such as large object variations in background, size, and occlusion. However, no matter in which condition, the feedback model can push the detection windows closer to the true object locations and finally achieve the correct localization.

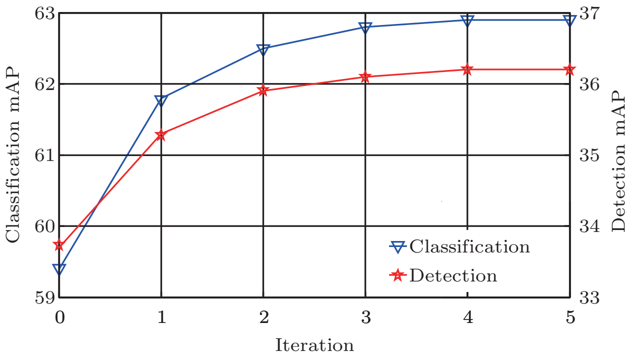
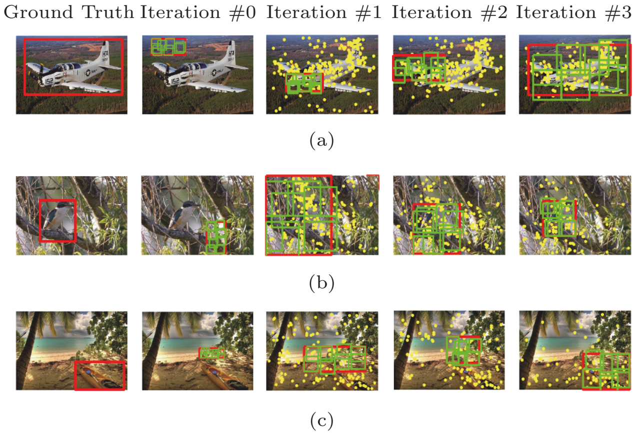


Fig.4. Iterative performance of object classification and detection by the feedback model on the PASCAL VOC 2007 dataset.



(a)

(b)

(c)

Fig.5. Some examples of the iterative feedback in different conditions. (a) Large object in normal background. (b) Small object in complex background. (c) Occluded object in complex background.

To obtain a consistent improvement in the visual feedback model, we should guarantee that each iteration produces an improvement. In our experiments, we use the objective value of the SVM classifier for validation. Fig.6 gives the average objective value of the two linear SVM classifiers for all the categories in both "what" and "where" feedback. These two classifiers are the one for the saliency distribution of $S_B\left(X_S\right)$ and the one for the BoW image representation of $\Psi\left(X\right)$. We use the dual solver for the liblinear SVM, and thus the higher objective value leads to a better solution. It is observed that the objective value of both classifiers increases consistently in each iteration. Similar to the case in Fig.4, the value gradually increases to a stable state as the iteration continues. These results demonstrate that the "what" and the "where" feedbacks are effective to enhance object detection and classification consistently.

② Girshick R B, Felzenszwalb P F, McAllester D A. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/rbg/latent-release5/, Jan. 2015.

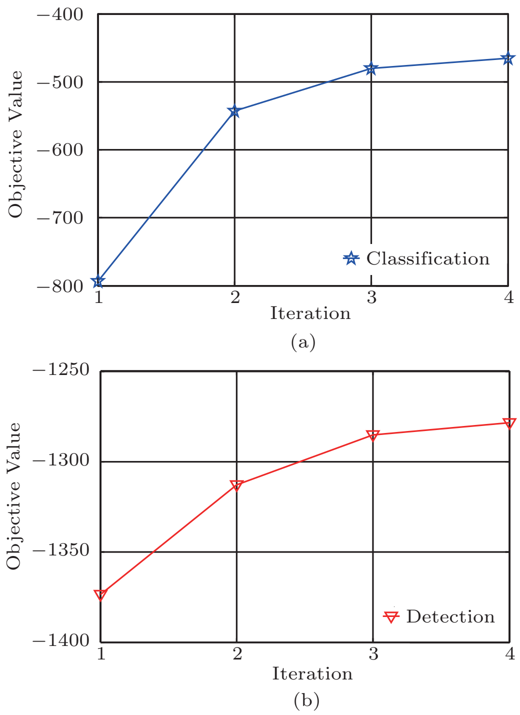


Fig.6. Average objective value of the liblinear SVMs for all object categories in (a) classification and (b) detection.

4.2 Detection Evaluation

Based on the iterative feedback, we evaluate to what degree this iterative procedure enhances the robustness of object detection. The detection result of the iteration #4 is used for comparison, and the salient features with the top 1 000 highest scores are selected to construct the saliency distribution. Table 1 shows the detection performance of the visual feedback model (VFM) and some previous methods. According to Table 1, the visual feedback model achieves the mean average precision (mAP) of 36.2, which is the highest among all the methods and is a considerable improvement over others, e.g., the improvement is 2.5% and 0.8% on DPM and DPM+Context respectively. Besides, the VFM also obtains the best AP on six object categories such as boat, car, motor, and person. In comparison with the methods using multiple low-level features[29,73], the visual feedback model with the single SIFT features is competitive, e.g., the mAP of VFM is 1.9% higher than that of HOG-LBP[73]. These results demonstrate that the "what" feedback is effective to enhance object detection.

To validate the contribution of the "what" feedback to the closeness in detection, we evaluate the overlap between the detection windows and the ground truth bounding box. Fig.7 shows the average best overlap (ABO), and ABO means the average best overlap between the detection windows and the ground truth bounding box[17]. It is popularly used to measure the quality of detection windows in some previous studies[17]. The definition of ABO is given as follows:

$$ABO = \frac{1}{|G^c|} \sum_{g_i^c \in G^c} \max_{d \in BB(i)} Overlap\,(g_i^c, d),$$

in which g_i^c is the ground truth bounding box in the i-th image for object category c, while G^c is the set of ground truth bounding boxes in category c, and $|G^c|$ is the number of the ground truth bounding boxes in category c. Besides, d denotes the detection window which has the best overlap with g_i^c in the i-th image. The overlap is measured as follows:

$$Overlap\,(g_i^c, d) = \frac{area(g_i^c) \cap area(d)}{area(g_i^c) \cup area(d)}.$$

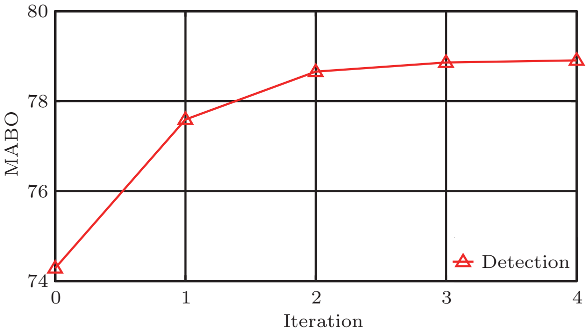


Fig.7. Mean average best overlap (MABO) of the visual feedback model in each iteration on the PASCAL VOC 2007 testing set.

It is observed that as the iteration increases, the ABO becomes larger, e.g., the ABO increases from 74 in iteration #1 to 79 in iteration #4, which validates that saliency distribution in the "what" feedback can effectively push the detection windows closer to the true object locations. Fig.8 shows this improvement for some examples of the rectified detection windows, in which the red rectangles are the detection windows of the baseline DPM 5.0, while the yellow ones denote the windows rectified by the visual feedback model based on DPM. We see that in most samples, the DPM detections only cover a part of the objects, while the VFM model rectifies these wrong detections to cover most part of the objects around the true locations.

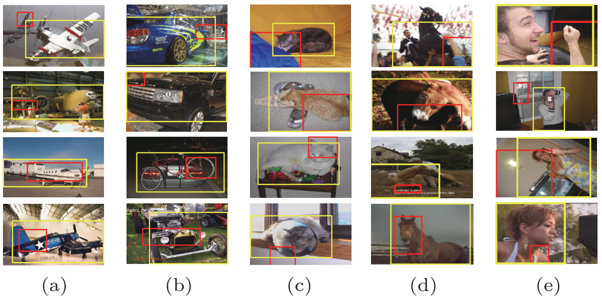


Fig.8. Some examples of the detection windows rectified by the visual feedback model (yellow rectangles) based on the DPM model (red rectangles).

4.3 Classification Evaluation

Based on the iterative feedback, we evaluate to what degree this iterative procedure can enhance the robustness in object classification. Table 2 shows the classification performance of the visual feedback model (VFM) and some previous methods on the PASCAL VOC 2007 test set. It is observed that the visual feedback model achieves the mAP of 62.9%, which is the highest among the single feature based methods and obtains promising improvements over others, e.g., 3.5% and 1.5% higher than LLC+SPM and LLC+OCP respectively. The proposed VFM obtains the best AP on 10 object categories, and the improvements on most of these categories are considerable, e.g., the improvement is about 1% on cow, bicycle, bottle, car, and chair. Similar to the comparison in detection, the visual feedback model with the single SIFT features is comparable to the methods with multiple features[32], e.g., the VFM is 0.7% and 3.5% higher than the 2007 Winner and BoF+HOG respectively, which shows the effectiveness of the VFM and the potential for further improvement. Fig.9 visualizes this improvement on the adaptiveness of object matching in object classification. Though objects vary a lot in size, location, and orientation, the

Table 2. Classification Performance Between VFM and Some Previous Methods on PASCAL VOC 2007

Method	Object+Context[78]	2007 Winner[79]	LLC+SPM[24]	LLC+OCP[23]	BoF+HOG[77]	VFM
Aero	**80.2**	77.5	73.7	74.7	75.0	76.9
Bicycle	61.0	63.6	65.7	70.1	68.3	**71.0**
Bird	49.8	56.1	49.9	52.8	**58.2**	54.7
Boat	69.6	**71.9**	68.7	69.0	69.5	70.4
Bottle	21.0	33.1	28.1	34.2	33.3	**35.1**
Bus	66.8	60.6	66.2	67.8	**68.9**	68.7
Car	80.7	78.0	78.4	81.3	80.0	82.3
Cat	51.1	58.8	60.4	62.0	**65.8**	64.3
Chair	51.4	53.5	55.9	56.7	55.9	**57.8**
Cow	35.9	42.6	49.4	49.9	50.9	**51.7**
Table	**62.0**	54.9	52.6	54.3	60.6	57.5
Dog	38.6	45.8	45.5	47.1	**50.4**	48.9
Horse	69.0	77.5	77.4	79.2	77.6	**80.1**
Motor	61.4	64.0	68.0	69.0	70.6	**70.7**
Person	84.6	85.9	84.3	85.4	86.2	**86.4**
Plant	28.7	**36.3**	29.1	30.1	31.6	31.6
Sheep	53.5	44.7	46.8	48.7	49.6	**50.2**
Sofa	**61.9**	50.6	56.3	58.5	56.9	59.4
Train	**81.7**	79.2	77.0	77.4	78.9	79.4
TV	59.5	53.2	53.7	59.5	55.5	**60.3**
mAP	*58.4*	*59.4*	*59.4*	*61.4*	*62.2*	***62.9***

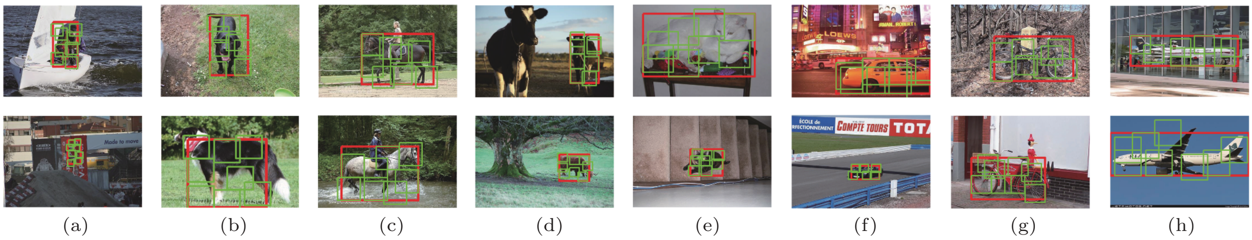


Fig.9. Some examples of the object matching based on object structure obtained by the visual feedback model. Red rectangles are the final detection windows, and the green ones denote the corresponding object parts. (a) Person. (b) Dog. (c) Horse. (d) Cow. (e) Cat. (f) Car. (g) Bicycle. (h) Aeroplane.

object structure can be adapted to different image conditions, which validates the importance of object structure and the effectiveness of the "where" feedback in improving adaptiveness for classification.

4.4 Parameter Selection

Finally, we provide some practical guidelines of the parameter selection in the VFM model. The selection includes three important factors: vocabulary size, object partition, and SVM penalty coefficients.

4.4.1 Vocabulary Size

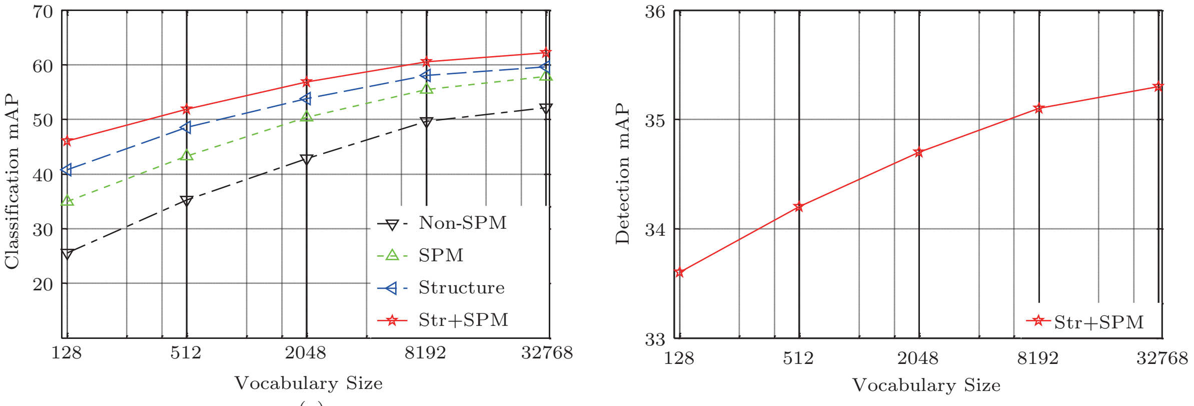
In the visual feedback model, the vocabulary size $|V|$ is an important factor, which will directly influence the number of the salient features obtained in the "what" feedback. Fig.10 shows the mAP of classification and detection under the vocabulary size of $128 \sim 32\,768$, in which Str+SPM represents the visual feedback model. In our implementation, we only use the salient features with the top 1 000 highest scores. It is observed that in both classification and detection, larger vocabulary size leads to higher performance, e.g., the highest mAP occurs at the size of 32k. The reason for this may be that larger vocabulary size can generate more salient features, which can represent the object saliency more accurately. However, one problem is that not all features are salient for objects. Due to the noise in cluttered scenes, some salient features may not locate on objects, while we observe that the features with higher score tend to locate on objects. Therefore, we use the salient features with the top 1 000 highest scores to construct the saliency distribution.

4.4.2 Object Partition

In this part, we evaluate different object partition methods to give the best object matching in classification. In the evaluation, we use four partition strategies: 1) SPM uses the typical spatial pyramid matching (SPM) for object matching; 2) Non-SPM matches objects without SPM and only based on the whole image; 3) Structure only uses the object parts for matching, while 4) Str+SPM combines both object parts and SPM for matching. Fig.10(a) shows the classification performance of these four methods based on the five vocabulary sizes, wherein Str+SPM is the proposed model. Firstly, we see that the highest performance of all the partition methods is obtained at the largest vocabulary size 32 k, which demonstrates the selection of vocabulary size in Subsection 4.4.1. Secondly, no matter which vocabulary size, these methods always have the same rank, i.e., the strategy with structure is better than the one with SPM. This rank implies that object structures are more robust than the SPM in object matching, and the structures are more beneficial to the context representation. Besides, by further considering SPM based on the structure, performance can be further improved, which demonstrates that context is useful for object classification. Therefore, in this paper, we adopt the combination of object structure and context to improve the discrimination of object representation.

4.4.3 Penalty Coefficient

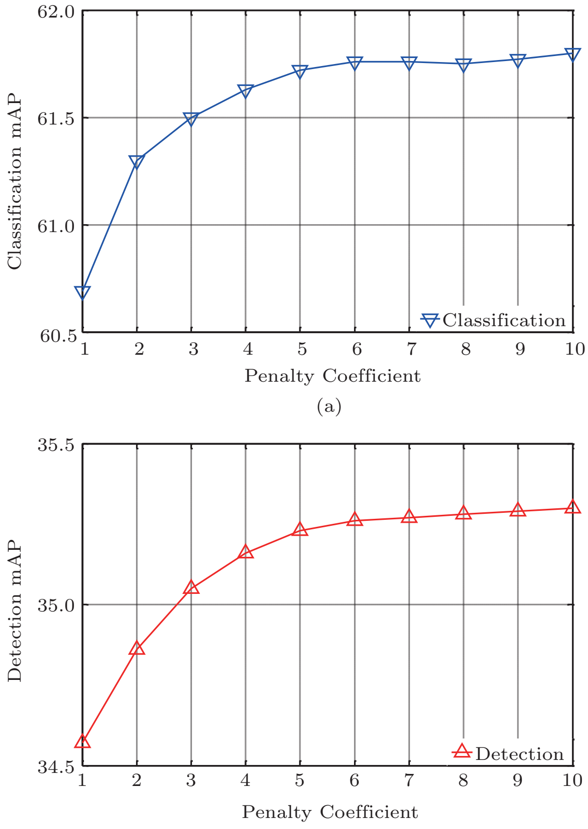
In the visual feedback model, the liblinear SVMs trained on the image representation $\Psi(X)$ and the saliency distribution $S_B\,(X_S)$ are important to guaran-



(a) (b)

Fig.10. Parameter selection of the vocabulary size and object partition in the visual feedback model for object (a) classification and (b) detection.

tee the effectiveness of each iteration, and the penalty coefficients C (in (1)) and C_S (for $S_B\left(X_S\right)$) are two important parameters in the classifier. In our implementation, we constrain their values to the range [1, 10], and use a 5-fold cross-validation on the validation set to select the best value. Fig.11 shows the classification and the detection performance based on the different C and C_S in the first iteration. It is observed that $C=10$ and $C_S=10$ yield the best performance. Furthermore, some values larger than 10 are also tested, while the performance remains stable or decreases. Based on these results, these two parameters are fixed to be 10 in the rest iterations and all the experiments.



(a)

(b)

Fig.11. Influence of the SVM penalty coefficients in the visual feedback model by the image representation $\Psi\left(X\right)$ and the saliency distribution $S_B\left(X_S\right)$. (a) Classification. (b) Detection.

5 Conclusions

In this paper, we proposed a visual feedback model (VFM) for classification and detection to achieve robust object recognition. Firstly, feature saliency is obtained from the "what" feedback to rectify detection windows to improve the closeness in detection. Secondly, object structure is obtained from the "where" feedback for object matching to enhance the adaptiveness in classification. Finally, we proposed an iterative procedure of the "what" and the "where" feedback to achieve robust object recognition. 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 visual feedback model to the multi-label problem and the weakly supervised detection problem.

Acknowledgement

Thank Steve Maybank for the revision.

References

[1] Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 2010, 88(2): 303-338.

[2] Deng J, Dong W, Socher R, Li L J, Li K, Li F F. ImageNET: A large-scale hierarchical image database. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2009, pp.248-255.

[3] Csurka G, Dance C R , Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In *Proc. European Conference on Computer Vision Workshop*, May 2004, pp.145-168.

[4] Yang J, Yu K, Gong Y, Huang T. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2009, pp.1794-1801.

[5] Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y. Locality-constrained linear coding for image classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2010, pp.3360-3367.

[6] Zhou X, Yu K, Zhang T, Huang T. Image classification using super-vector coding of local image descriptors. In *Proc. the 11th European Conference on Computer Vision*, September 2010, pp.141-154.

[7] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification. In *Proc. the 11th European Conference on Computer Vision*, September 2010, pp.143-156.

[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNET classification with deep convolutional neural networks. In *Proc. the 26th Annual Conf. Neural Information Processing Systems*, December 2012, pp.1106-1114.

[9] Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv:1405.3531*, 2014.

[10] Lin M, Chen Q, Yan S. Network in network. *arXiv:1312.4400*, 2014.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[12] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In *Proc. the 13th European Conference on Computer Vision*, September 2014, pp.818-833.

[13] Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(9): 1627-1645.

[14] Wang X, Bai X, Ma T, Liu W, Latecki L. Fan shape model for object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp.151-158.

[15] Zhu L, Chen Y, Yuille A, Freeman W. Latent hierarchical structural learning for object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2010, pp.1062-1069.

[16] Girshick R B, Felzenszwalb P F, McAllester D A. Object detection with grammar models. In *Proc. the 25th NIPS*, December 2011, pp.442-450.

[17] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2014, pp.580-587.

[18] Hoffman J, Guadarrama S, Tzeng E, Hu R, Donahue J, Girshick R, Darrell T, Saenko K. LSDA: Large scale detection through adaptation. In *Proc. NIPS*, December 2014, pp.3536-3544.

[19] Zhang N, Donahue J, Girshick R, Darrell T. Part-based R-CNNs for fine-grained category detection. In *Proc. the 13th European Conference on Computer Vision*, September 2014, pp.834-849.

[20] Gupta S, Girshick R, Arbeláez P, Malik J. Learning rich features from RGB-D images for object detection and segmentation. In *Proc. the 13th European Conference on Computer Vision*, September 2014, pp.345-360.

[21] Hariharan B, Arbeláez P, Girshick R, Malik J. Simultaneous detection and segmentation. In *Proc. the 13th European Conference on Computer Vision*, September 2014, pp.297-312.

[22] Zhang J, Zhao X, Huang Y, Huang K, Tan T. Semantic windows mining in sliding window based object detection. In *Proc. the 21st International Conference on Pattern Recognition*, November 2012, pp.3264-3267.

[23] Russakovsky O, Lin Y, Yu K, Li F F. Object-centric spatial pooling for image classification. In *Proc. the 12th European Conference on Computer Vision*, Oct. 2012, pp.1-15.

[24] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2006, pp.2169-2178.

[25] Chikkerur S, Serre T, Tan C, Poggio T. What and where: A Bayesian inference theory of attention. *Vision Research*, 2010, 50(22): 2233-2247.

[26] Galleguillos C, Belongie S. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 2010, 114(6): 712-722.

[27] Divvala S K, Hoiem D, Hays J H, Efros A A, Hebert M. An empirical study of context in object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2009, pp.1271-1278.

[28] Harzallah H, Jurie F, Schmid C. Combining efficient object localization and image classification. In *Proc. the 12th International Conference on Computer Vision*, Sept. 29–Oct. 2, 2009, pp.237-244.

[29] Song Z, Chen Q, Huang Z, Hua Y, Yan S. Contextualizing object detection and classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2011, pp.1585-1592.

[30] Chen G, Ding Y, Xiao J, Han T X. Detection evolution with multi-order contextual co-occurrence. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2013, pp.1798-1805.

[31] Zhang Y, Chen T. Weakly supervised object recognition and localization with invariant high order features. In *Proc. the British Machine Vision Conference*, Aug. 31–Sept. 3, 2010, pp.47:1-47:11.

[32] Chen Q, Song Z, Hua Y, Huang Z, Yan S. Hierarchical matching with side information for image classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2013, pp.3426-3433.

[33] Nguyen M H, Torresani L, de la Torre F, Rother C. Weakly supervised discriminative localization and classification: A joint learning process. In *Proc. International Conference on Computer Vision*, September 2009, pp.1925-1932.

[34] Huang Y, Huang K, Yu Y, Tan T. Salient coding for image classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2011, pp.1753-1760.

[35] Rybak I A, Gusakova V I, Golovan A V, Podladchikova L N, Shevtsova N A. A model of attention-guided visual perception and recognition. *Vision Research*, 1998, 38(15/16): 2387-2400.

[36] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(11): 1254-1259.

[37] Barenholtz E, Tarr M J. Reconsidering the role of structure in vision. *The Psychology of Learning and Motivation*, 2006, 47:157-180.

[38] Biederman I. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 1987, 94(2):115-147.

[39] Huang K, Wang Q, Wu Z. Natural color image enhancement and evaluation algorithm based on human visual system. *Computer Vision and Image Understanding*, 2006, 103(1): 52–63.

[40] Huang K, Wu Z, Wang Q. Image enhancement based on the statistics of visual representation. *Image and Vision Computing*, 2005, 23(1): 51–57.

[41] Huang K, Wu Z, Fung G S K, Chan F H Y. Color image denoising with wavelet thresholding based on human visual system model. *Signal Processing: Image Communication*, 2005, 20(2): 115–127.

[42] Boureau Y, Ponce J, LeCun Y. A theoretical analysis of feature pooling in visual recognition. In *Proc. the 27th International Conference on Machine Learning*, June 2010, pp.111-118.

[43] Serre T, Wolf L, Poggio T. Object recognition with features inspired by visual cortex. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2005, pp.994-1000.

[44] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, 313(5786): 504-507.

[45] LeCun Y, Kavukvuoglu K, Farabet C. Convolutional networks and applications in vision. In *Proc. IEEE International Symposium on Circuits and Systems*, May 30-June 2, 2010, pp.253-256.

[46] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 2005, pp.886-893.

[47] Wohlhart P, Donoser M, Roth P M, Bischof H. Detecting partially occluded objects with an implicit shape model random field. In *Proc. the 11th Asian Conference on Computer Vision*, November 2012, pp.302-315.

[48] Bogacz R, Usher M, Zhang J, McClelland J L. Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of The Royal Society of London, Series B, Biological Sciences*, 2007, 362(1485): 1655-1670.

[49] Yang J, Yu K, Huang T. Supervised translation invariant sparse coding. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2010, pp.3517-3524.

[50] Jurie F , Triggs B. Creating efficient codebooks for visual recognition. In *Proc. the 10th International Conference on Computer Vision*, Oct. 2005, pp.604-610.

[51] Boureau Y L, Bach F, LeCun Y, Ponce J. Learning mid-level features for recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2010, pp.2559-2566.

[52] Van Gemert J C, Veenman C J, Smeulders A W M, Geusebroek J M. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(7): 1271-1283.

[53] Jegou H, Perronnin F, Douze M, Sanchez J, Perez P, Schmid C. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(9): 1704-1716.

[54] Chatfield K, Lempitsky V, Vedaldi A, Zisserman A. The devil is in the details: An evaluation of recent feature encoding methods. In *Proc. the 22nd British Machine Vision Conference*, Aug. 29-Sept. 22, 2011, pp.76:1-76:12.

[55] Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005, 61(1): 55-79.

[56] Desai C, Ramanan D, Fowlkes C C. Discriminative models for multi-class object layout. *International Journal of Computer Vision*, 2011, 95(1): 1-12.

[57] Vedaldi A, Gulshan V, Varma M, Zisserman A. Multiple kernels for object detection. In *Proc. the 12th IEEE International Conference on Computer Vision*, Sept. 29-Oct. 2, 2009, pp.606-613.

[58] Pepik B, Stark M, Gehler P, Schiele B. Teaching 3D geometry to deformable part models. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp.3362-3369.

[59] Yang Y, Ramanan D. Articulated pose estimation using flexible mixtures of parts. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2011, pp.1385-1392.

[60] Zhu X, Ramanan D. Face detection pose estimation landmark localization in the wild. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp.2879–2886.

[61] Duchenne O, Joulin A, Ponce J. A graph-matching kernel for object categorization. In *Proc. IEEE International Conference on Computer Vision*, November 2011, pp.1792-1799.

[62] Song X, Wu T, Jia Y, Zhu S. Discriminatively trained and-or tree models for object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2013, pp.3278-3285.

[63] Carbonetto P, de Freitas N, Barnard K. A statistical model for general contextual object recognition. In *Proc. the 8th European Conference on Computer Vision*, May 2004, pp.350-362.

[64] Kosslyn S M, Flynn R A, Amsterdam J B, Wang G. Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition*, 1990, 34(3): 203-277.

[65] Mishkin M, Ungerleider L G, Macko K A. Object vision and spatial vision: Two cortial pathways. *Trends in Neurosciences*, 1983, 6: 414-417.

[66] Ungerleider L G, Mishkin M. Two Cortical Visual Systems. Cambridge, MA: MIT Press, 1982.

[67] Chai Y, Lempitsky V, Zisserman A. BiCoS: A bi-level co-segmentation method for image classification. In *Proc. IEEE International Conference on Computer Vision*, November 2011, pp.2579-2586.

[68] Crandall D J, Huttenlocher D P. Weakly supervised learning of part-based spatial models for visual object recognition. In *Proc. the 9th European Conference on Computer Vision*, May 2006, pp.16-29.

[69] Ren X, Ramanan D. Histograms of sparse codes for object detection. In *Proc. Computer Vision and Pattern Recognition*, June 2013, pp.3246-3253.

[70] Malisiewicz T, Efros A A. Improving spatial support for objects via multiple segmentations. In *Proc. the British Machine Vision Conference*, September 2007, pp.55:1-55:10.

[71] Pandey M, Lazebnik S. Scene recognition and weakly supervised object localization with deformable part-based models. In *Proc. IEEE International Conference on Computer Vision*, November 2011, pp.1307-1314.

[72] Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2): 91-110.

[73] Zhang J, Huang K, Yu Y, Tan T. Boosted local structured HOG-LBp for object localization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2011, pp.1393-1400.

[74] Wang X, Lin L, Huang L, Yan S. Incorporating structural alternatives and sharing into hierarchy for multiclass object recognition and detection. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, June 2013, pp.3334-3341.

[75] Shahba I, Khan F, Anwer R M, van de Weijer J, Bagdanov A D, Vanrell M, Lopez A M. Color attributes for object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp.3306-3313.

[76] Vedaldi A, Fulkerson B. VLFeat: An open and portable library of computer vision algorithms. In *Proc. International Conference on Multimedia*, Oct. 2010, pp.1469-1472.

[77] Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(9): 1627-1645.

[78] Uijlings J R R, Smeulders A W M, Scha R J H. What is the spatial extent of an object. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2009, pp.770-777.

[79] Marszasek M, Schmid C, Harzallah H, van de Weijer J. Learning representations for visual object class recognition. In *Proc. Visual Recognition Challenge Workshop for International Conference on Computer Vision*, October 2007.

[80] Kobayashi T. BFO meets HOG: Feature extraction based on histograms of oriented p.d.f gradients for image classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2013, pp.747-754.

Chong Wang received his B.S. degree from Beijing University of Posts and Telecommunications, and now he is a Ph.D. student in the National Laboratory of Pattern Recognition (NLPR), Institute of Automation (IA), Chinese Academy of Science, Beijing. His current research interests include pattern recognition, computer vision, and machine learning. He has published several papers on top conferences and journals like European Conference on Computer Vision and IEEE Transactions on Image Processing. In 2010 and 2011, he participated in the PASCAL VOC challenge and won prizes in both years. In 2014, he won the championship of the ImageNet 2014 classification challenge with additional data.

Kai-Qi Huang received his B.S. and M.S. degrees in electronic engineering from Nanjing University of Science Technology, China, and Ph.D. degree in signal and information processing from Southeast University, Nanjing, in 2003. He has worked in the National Laboratory of Pattern Recognition (NLPR), Institute of Automation (IA), Chinese Academy of Science (CAS), Beijing, and now he is a professor in NLPR. His current research interests include visual surveillance, digital image processing, pattern recognition, biological based vision, and so on. He has published over 80 papers in international journals and conferences such as IEEE TIPAMI, T-IP, T-SMC-B, TCSVT, Pattern Recognition (PR), Computer Vision and Image Understanding (CVIU), ECCV, CVPR, ICIP, ICPR. He received the Best Student Paper award from ACPR 2010, the winner prizes of the detection task in both PASCAL VOC 2010 and PASCAL VOC 2011, the winner prize of the classification task with additional data in ILSVRC 2014, and the honorable mention prize of the classification task in PASCAL VOC 2011. He is a senior member of IEEE and was the deputy general secretary of IEEE Beijing Section (2006∼2008).