A Combing Top-Down and Bottom-Up Discriminative Dictionaries Learning for Non-specific Object Detection

Yurui XIE (a), Nonmember, Qingbo WU (b), Student Member, Bing LUO (b), Chao HUANG (c), Nonmembers

SUMMARY In this letter, we exploit a new framework for detecting the non-specific object via combing the top-down and bottom-up cues. Specifically, a novel supervised discriminative dictionaries learning method is proposed to learn the coupled dictionaries for the object and non-object feature spaces in terms of the top-down cue. Different from previous dictionary learning methods, the new data reconstruction residual terms of coupled feature spaces, the sparsity penalty measures on the representations and an inconsistent regularizer for the learned dictionaries are all incorporated in a unitized objective function. Then we derive an iterative algorithm to alternatively optimize all the variables efficiently. Considering the bottom-up cue, the proposed discriminative dictionaries learning is then integrated with an unsupervised dictionary learning to capture the objectness windows in an image. Experimental results show that the non-specific object detection problem can be effectively solved by the proposed dictionary learning framework and outperforms some established methods.

key words: dictionary learning, object detection

1. Introduction

The aim of object detection is to determine the spatial position of object in an image. Although object detection has achieved great progress in recent years, most state-of-the-art methods [1]–[3] mainly focus on the specific object detection problem. However, the specific object detection method has two limitations: First, it can only be applicable to the trained object class, which is insufficient to model a new object category. Second, the previous methods for modeling a given object class require a large set of annotated data. Therefore, it is not always accessible in practice. In order to overcome the above limitations, many literatures seek to discover the non-specific objects within an image. Generally, these methods can be roughly classified into the segment-based [4]–[6] and the saliency-based [7]–[10]. The first method commonly utilizes single or multiple low-level segmentation to partition an image into the coherent regions. These regions help to find the spatial support for objects. The second method aims to automatically discover objects with the help of visual attention. There are also some recent approaches [11]–[13] that both adopt the segment-based and saliency-based cues, which aim to construct the effective objectness measure to exploit the non-specific object in an image.

Inspired by the sparse model [14] dictionary learning algorithm, it is currently a representative method to approximate the input signal as a linear combinations of a small number of atoms based on an over-complete dictionary. In this letter, we propose a framework that combines a novel discriminative coupled dictionaries learning and an unsupervised dictionary learning to address the problem of non-specific object detection. Different from most of the previous works, our proposed approach tends to focus on designing a discriminative learning model, which has the ability to incorporate the high-level knowledge for the objectness measure.

Figure 1 illustrates a flowchart of the proposed non-specific object detection method. Specifically, given an input image, we sample all the potential object positions and scales in the image. Then we score every potential object window incorporated with the top-down and bottom-up cues. As can be seen the objectness windows in Fig. 1, the red bounding boxes denote these potential object windows with high confidence score. Meanwhile, these windows with low score are shown in yellow.

2. Discriminative Coupled Dictionaries Learning Model

2.1 Discriminative Coupled Dictionaries Learning

Most existing dictionary learning methods mainly aim to learn an over-complete dictionary in a single feature space. Therefore, they are not suitable for the problem of cou-
pled object and non-object feature spaces. To further exploit the top-down cue in the dictionary learning process, we propose a novel supervised discriminative dictionaries learning method for the coupled feature spaces, where the learned dictionaries have the ability to reflect the discriminative power and inconsistence between the two spaces.

Formally, the joint dictionaries learning model for the coupled feature spaces can be expressed as the following energy minimization formulation:

\[
\min_{A_f, A_b} \left\| Y_f - D_f A_f \right\|_F^2 + \mu \left\| Y_b - D_b A_b \right\|_F^2 + \alpha \| A_f \|_1 + \beta \| A_b \|_1 + \gamma \| D_f^T D_b \|_F^2
\]

where \( Y_f \in \mathbb{R}^{n \times n_f} \) and \( Y_b \in \mathbb{R}^{n \times n_b} \) are the extracted features of training samples for the object and non-object feature spaces respectively, and \( d \) is the dimension of feature. \( D_f = \{d_f^1, d_f^2, \ldots, d_f^K\} \in \mathbb{R}^{n_f \times K} \) and \( D_b = \{d_b^1, d_b^2, \ldots, d_b^K\} \in \mathbb{R}^{n_b \times K} \) denote the coupled object and non-object dictionaries to be learned in the two feature spaces respectively. The \( d_f^l \in \mathbb{R}^{n_f \times 1} \) and \( d_b^l \in \mathbb{R}^{n_b \times 1} \) respectively represent the atom in dictionary \( D_f \) and \( D_b \). \( K \) is the size of the dictionary. Here, \( \| \cdot \|_F \) denotes the Frobenius norm. The \( \alpha, \beta, \gamma \) are the scalars to balance the terms in the objective function.

The first two terms \( \| Y_f - D_f A_f \|_F^2 \) and \( \| Y_b - D_b A_b \|_F^2 \) represent the reconstruction residuals by the coupled object and non-object feature dictionaries, which enforce the trained dictionaries to fit the corresponding sample data effectively. The terms \( \| A_f \|_1 \) and \( \| A_b \|_1 \) denote the sparsity imposed on representation coefficients. Here we utilize the \( l_1 \)-norm instead of \( l_0 \)-norm sparsity to make the objective function convex. Additionally, an inconsistent regularizer \( \| D_f^T D_b \|_F^2 \) for the coupled dictionaries is also added to the objective function. It can further penalize the learned dictionaries to be more discriminative in the coupled feature spaces.

2.2 Optimization Procedure

In our coupled dictionaries learning model, the energy function minimization problem (1) can be divided into two sub-problems. Then we design an iterative algorithm to alternatively optimize all the variables. 1) The first optimization sub-problem. Fixing the \( D_f \) and \( D_b \), updating \( A_f, A_b \). Thus it is essentially a \( l_1 \)-regularized least squares problem. This minimization can be achieved by some standard convex optimization techniques. 2) The second optimization sub-problem. Fixing the \( A_f, A_b \), we update \( D_f \) and \( D_b \) respectively. For this energy function minimization problem, we derive an efficient optimization processes to update each atom of the dictionaries.

**Step 1**: Updating \( A_f \) and \( A_b \). To update \( A_f, A_b \), the \( D_f \) and \( D_b \) are fixed. Then we solve the following two optimization objective functions accordingly.

\[
\min_{A_f} \| Y_f - D_f A_f \|_F^2 + \alpha \| A_f \|_1
\]

\[
\min_{A_b} \| Y_b - D_b A_b \|_F^2 + \beta \| A_b \|_1
\]

In this work, we use the algorithm [15] to solve the above optimization problems.

**Step 2**: Updating \( D_f \) and \( D_b \). For updating \( D_f \), the \( A_f \) and \( D_b \) are fixed. Then we obtain the following form:

\[
\min_{D_f} \| Y_f - D_f A_f \|_F^2 + \gamma \| D_f^T D_b \|_F^2
\]

The coefficient matrix \( A_f \) can be written as \( A_f = [a_f^1; a_f^2; \ldots; a_f^K] \in \mathbb{R}^{K \times n_f} \), where \( a_f^i \in \mathbb{R}^{1 \times n_f}, i = 1, \ldots, K \) is the \( i \)-th row vector of \( A_f \). Then we propose to update each atom of \( D_f \) one by one in iteration. Specifically, when updating \( d_f^i \), all the other atoms of \( D_f \) are fixed. Therefore, Eq. (4) is converted into the below formulation:

\[
J_{d_f} = \arg \min_{d_f^i} \| Y_f - \sum_l d_f^l a_f^l - d_f^i a_f^i \|_F^2 + \gamma \| d_f^i \|_2^2
\]

Let \( \overline{Y}_f = Y_f - \sum_l d_f^l a_f^l \), thus Eq. (5) can be reformulated as

\[
J_{d_f} = \arg \min_{d_f^i} \| \overline{Y}_f - d_f^i a_f^i \|_F^2 + \gamma \| d_f^i \|_2^2 + \eta_d \| D_f D_f^T d_f^i \|_2^2
\]

Let the first derivative of \( J_{d_f} \) w.r.t. be zero, then the optimum solution \( d_f^i \) is obtained below:

\[
d_f^i = \left( a_f^i a_f^i I + \gamma D_b D_b^T \right)^{-1} \overline{Y}_f a_f^i
\]

As an atom of dictionary, the updated atom \( d_f^i \) is further normalized by the \( L_2 \) norm as \( \overline{d_f^i} = d_f^i / \| d_f^i \|_2 \).

For updating \( D_b \), we fix the \( A_b \) and \( D_f \). Thus the optimization formulation can be defined as

\[
\min_{D_b} \| Y_b - D_b A_b \|_F^2 + \gamma \| D_f^T D_b \|_F^2
\]

Similarly, the \( t \)-th atom of \( D_b \) is updated as below:

\[
d_b^t = \left( a_b^t a_b^t I + \frac{\gamma}{\mu} D_f D_f^T \right)^{-1} \overline{Y}_b a_b^t
\]

where \( \overline{Y}_b = Y_b - \sum_t a_b^t a_b^t \). Then the updated atom is unitized by the \( L_2 \)-norm, i.e. \( \overline{d_b^t} = d_b^t / \| d_b^t \|_2 \).

According to the above optimization procedures, we can update all the atoms of the coupled dictionaries in each iteration. Therefore, the whole dictionary \( D_f \) and \( D_b \) are updated accordingly. It is straightforward that the value of the energy function decreases in each iteration, hence the algorithm converges.

3. Unsupervised Dictionary Learning

In order to incorporate the bottom-up cue, an unsupervised dictionary learning is also adopted in our dictionary learning
framework. Here, we assume that any object should have the visual inconsistency between the current region and its surrounding region. Then we propose to use a mid-level representation feature achieved by the local dictionary rather than the raw visual feature to describe the appearance difference.

In particular, the input image is first divided into many non-overlapped patches. Then all the features of patches within the image are fed into an unsupervised dictionary learning model. Formally, this local dictionary for a given image is trained by the following optimization formulation.

\[
\min_{D} \| D R - X \|^2_F \quad \text{s.t.} \quad \| r \|_0 \leq T_0
\]

where \( X = [x_1, x_2, \ldots, x_n] \) is the extracted raw features of all patches, \( x_i \) is the raw feature of \( i \)-th patch and \( n \) means the number of patches in the image. \( D \) is a local dictionary to be learned. The representation features for all patches are denoted as \( R = [r_1, r_2, \ldots, r_n] \), where \( r_i \) is the \( i \)-th representation feature associated with the \( i \)-th patch. \( T_0 \) is the sparsity parameter. This problem describes each given feature \( x_i \) as the sparsest representation \( r_i \) on an unknown dictionary \( D \).

It seeks to jointly find the optimum representation features and the dictionary. In this work, we adopt the K-SVD algorithm [14] to solve Eq. (10) due to its fast convergence. At last, these obtained representation features of all patches by the learned local dictionary are utilized to score the current potential object window.

4. Top-Down and Bottom-Up Cues Integration

In this letter, we adopt a weighted linear combination strategy to score these potential object windows. Though it is straightforward, the performance is verified effectively in the experiments. This formulation is defined as follows:

\[
S = \rho S_{\text{global}} + (1 - \rho) S_{\text{local}}
\]

where \( S_{\text{global}} \) is the global scores for all the potential object windows in terms of the top-down cue. \( S_{\text{local}} \) is the local scores, which encodes the bottom-up cue. The total scores for all windows are denoted as \( S \). More specifically, each global score for a potential object window is defined as:

\[
\| D_a a_f - x \|^2_F / \| D_f a_f - x \|^2_F, \quad \text{where} \quad x \text{ is the feature of current window}, \quad a_f \text{ is the feature of current window, and } \quad a_b \text{ can be obtained by solving the below optimization formulations accordingly:}
\]

\[
\min_{a_f} \| D_f a_f - x \|^2_F + \lambda \| a_f \|_1 \text{ and } \min_{a_b} \| D_b a_b - x \|^2_F + \lambda \| a_b \|_1. \quad \text{The above global score is measured by the reconstruction residuals of object and non-object space dictionaries. As for the local score, we suppose that any object should have the appearance differences from its surrounding region. The surrounding region is defined by the extended rectangular ring with respect to the window region in all directions. Specifically, the extended range of each window boundary is set to 20 pixel width in the experiment. If one of the extended window boundary exceeds the image boundary, this corresponding window boundary stops extending on the image boundary. Therefore, the local score of a window is evaluated by:}
\]

\[
D_{\chi^2}(\sum_{j=1}^{N} r_j, \sum_{j=1}^{M} r_j), \quad \text{where} \quad D_{\chi^2} \text{ denotes the } \chi^2 \text{ distance function, } N \text{ and } M \text{ are the number of patches in the current window and its surrounding region respectively. The } r_i \text{ and } r_j \text{ denote the representation features of the corresponding patches.}
\]

Finally, the obtained global scores \( S_{\text{global}} \) and local scores \( S_{\text{local}} \) for all the potential object windows are normalized and integrated into the total scores. \( \rho \) is a weighting parameter in (0, 1). In our experiment, the parameter \( \rho \) is set as 0.6 empirically. For the \( S_{\text{global}} \) and \( S_{\text{local}} \), the higher value the better. Figure 1 demonstrates the overview of our non-specific object detection approach.

5. Experiments

In this section, we evaluate the performance of our non-specific object detection method on the PASCAL VOC 08 dataset [16]. It is one of the challenging datasets to verify object detector. We use the test set of the VOC 08 detection dataset to report the results of different methods. Specifically, the performance is evaluated by the recall ratio, which is the standard measurement for object detection. The recall ratio is computed by the percentage of ground-truth objects been detected correctly from all the object locations. To be considered a correct detection by PASCAL challenge criterion, the area of overlap between the predicted bounding box and ground truth region must exceed 0.5.

For our discriminative coupled dictionaries learning, the training data are collected from the VOC 08 training set. More specifically, all the ground truth object windows are taken as the positive examples. As these negative examples, we first sample approximately 5000 windows that uniformly distributed within each image, then these generated windows that have an overlap of 0.2 to 0.4 with the positive examples are collected. The PHOG [17], LBP [18] and GIST [19] features are extracted from each window and further concatenated to form the final feature description. The regularization parameters are empirically set as \( \mu = 0.3 \), \( \alpha = \beta = 0.1 \) and \( \gamma = 0.5 \). We fix the size of the coupled dictionaries to 800 respectively and the initial dictionary is obtained by an identity matrix. For the unsupervised dictionary learning, the dictionary is initialized by the training data itself and the number of atoms in dictionary is fixed to 300. We set the sparsity parameter \( T_0 \) to 30 empirically in the experiment. Additionally, the patch size is set as 16 × 16 and we use the RGB color histogram features of all patches within an image to train the local dictionary.

We design two baselines by considering different components of our ‘Full’ dictionary learning framework. The first one is the ‘Global’ model using Eq. (1) only. The second one is the ‘Local’ model with Eq. (10). We also compare our approach and two variants with several representative methods, including the segmentation-based [4] and the saliency-based [7], [8]. For method [4], we obtain 50 objectness windows within an image by running segmentation method [20] in terms of multi-scale strategy. For all the other methods, the evaluations are achieved with on top 50 highest scoring windows per image. The detection results on
In this letter, we propose a new framework of combining a novel supervised discriminative coupled dictionaries learning and an unsupervised dictionary learning to tackle the challenging non-specific object detection problem. The main property of the proposed dictionary learning framework is that it has the ability to effectively exploit the top-down and bottom-up cues into the object detection. The experimental results demonstrate that the proposed method for non-specific object detection can achieve higher performance than some established methods.

6. Conclusions

In this letter, we propose a new framework of combining a novel supervised discriminative coupled dictionaries learning and an unsupervised dictionary learning to tackle the challenging non-specific object detection problem. The main property of the proposed dictionary learning framework is that it has the ability to effectively exploit the top-down and bottom-up cues into the object detection. The experimental results demonstrate that the proposed method for non-specific object detection can achieve higher performance than some established methods.