Object-Aware Dictionary Learning with Deep Features

Yurui Xie[1], Fatih Porikli[1,2], Xuming He[2]

Australian National University[1], NICTA[2]

Abstract. Visual dictionary learning has the capacity to determine sparse representations of input images in a data-driven manner using over-complete bases. Sparsity allows robustness to distractors and resistance against over-fitting, two valuable attributes of a competent classification solution. Its data-driven nature is comparable to deep convolutional neural networks, which elegantly blend global and local information through progressively more specific filter layers with increasingly extending receptive fields. One shortcoming of dictionary learning is that it does not explicitly select and focus on important regions, instead it either generate responses on uniform grid of patches or entire image. To address this, we present an object-aware dictionary learning framework that systematically incorporates region proposals and deep features in order to improve the discriminative power of the combined classifier. Rather than extracting a dictionary from all fixed sized image windows, our methods concentrates on a small set of object candidates, which enables consolidation of semantic information. We formulate this as an optimization problem on a new objective function and propose an iterative solver. Our results on benchmark datasets demonstrate the effectiveness of our method, which is shown to be superior to the state-of-the-art dictionary learning and deep learning based image classification approaches.

1 Introduction

Dictionary learning (DL) has attracted considerable amount of attentions in the past few years. The goal of DL is to learn an over-complete collection of atoms by a data-driven manner. The main property of learned dictionary is that it can approximate the input signal as a linear combination of a small number of atoms. Recently, the dictionary learning approaches have widely applied to various problems of computer vision area, such as image denoising [1, 2], image restoration [3], image synthesis [4, 5], visual tracking [6] and image classification [7–9].

The original intention of DL methods is to reconstruct the input data faithfully by a learned over-complete dictionary. Therefore, they are not appropriate for the visual recognition task. To overcome this problem, many literatures [7–19] aim to design the discriminative dictionary learning approaches for enhancing the representative ability of feature. For the visual recognition problem, our observation is that the local semantic information of image often provides the

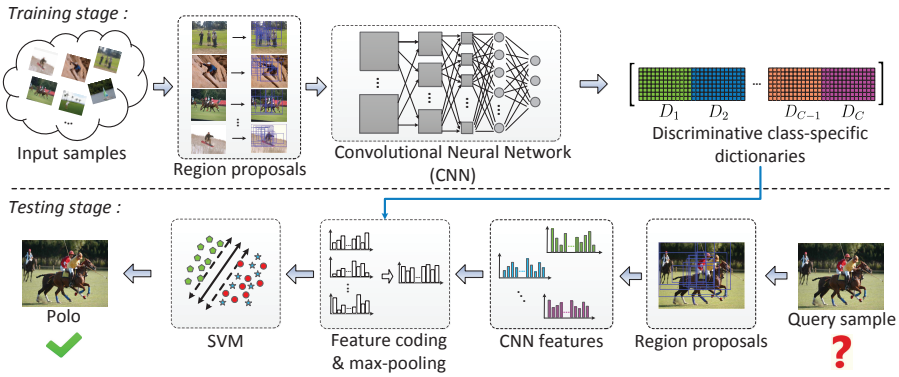

Fig. 1. Overall framework of our object-aware dictionary learning method.

important visual cues to improve the discrimination of feature representation. Thus it is beneficial for the image representation method that incorporates with the local image information. However, one main drawback of existing visual dictionary learning is that it is unable to select and focus on the important image regions explicitly. Instead, these methods only generate responses on the regular patches or the entire image. As a result, useful discriminative semantic information within image regions cannot be explored substantially in dictionary learning process.

Recently, the Convolutional Neural Network (CNN) has been shown to be successful in numerous visual recognition problems. One main advantage of the CNN is that it allows integrating the global context and local cues through multiple filter layers with increasingly extending receptive fields thanks to the pooling operations.

Inspired by this property of the CNN, we propose an Object-Aware Dictionary Learning (OADL) framework to address the above shortcoming of dictionary learning. To this end, we incorporate the deep features generated by the CNN into a region proposal framework to discover underlying local semantic information in the image. We design a new object aware objective for dictionary learning and then feed the deep features of region proposals to extract multiple discriminative class-specific dictionaries.

Unlike conventional dictionary learning approaches that extract a dictionary from all the fixed sized image patches or entire image, our method concentrates on a small set of object candidates. Since the local semantic information of image often provides important visual cues for recognition task, we concentrate on semantically meaningful image regions. To this end, we extract the region proposals. This facilitates the feature representation to consolidate the semantic information and suppress the distraction due to the background. In the final recognition stage, the learned a set of discriminative class-specific dictionaries are used to encode the deep features of all object candidates within image and

generates a global image representation by max-pooling. As for the proposed dictionary learning objective, we derive an efficient optimization algorithm to solve its variables alternatively. Figure 1 shows the framework of this new OADL method.

The remainder of this paper is organized as follows: The related works are briefly reviewed in Section 2. Section 3 presents the proposed object-aware dictionary learning framework that integrates with the region proposals and deep features systematically to improve the discriminative power of feature representation. The optimization algorithm of the OADL is also described in this section. Experimental results are given in Section 4.

2 Related Work

In order to enhance the representation power of image feature, many works aim to learn a discriminative dictionary for different visual recognition tasks. Existing dictionary learning approaches can be grouped as unsupervised and supervised methods. The goal of unsupervised dictionary learning is to compose an over-complete dictionary by minimizing the reconstruction error. A typical example for unsupervised dictionary learning would be the KSVD algorithm [15], which iteratively applies SVD to fit the atoms of a single dictionary to reconstruction error. To reduce the time complexity, Lee et al. [20] cast the standard sparse representation to the least squares problem.

To obtain a feature representation with more discriminative power, supervised dictionary learning incorporates additional classification objective to reconstruction using the labeled data. The existing supervised dictionary learning can be further grouped into two categories. Methods in the first category aim to make the representation coefficients discriminative by learning a single dictionary across all classes. The common characteristic of recent methods [11–13, 16, 17] is to combine a classification error term into the standard sparse dictionary learning formulation. Similarly, Jiang et al. [9] incorporate both a label consistent constraint and a linear classification cost into the KSVD objective for enhancing the representative power of features. Methods in the second category [7, 8, 10, 14, 18, 19] learn a set of class-specific sub-dictionaries, then these multiple sub-dictionaries are concatenated together to form a structured dictionary for feature representation. Specifically, Mairal et al. [18] integrate a softmax discriminative function with the KSVD model. Ramirez et al. [14] impose an incoherent constrain in the standard dictionary learning model, which encourages the learned class-specific dictionaries to be as independent as possible. Yang et al. [7] incorporate the Fisher Discriminant criterion into the dictionary learning for further improving the discriminative capability of class-specific dictionaries. Zhou et al. [8] propose to learn multiple class-specific dictionaries and a shared dictionary for the groups of classes that have the visually similar patterns. Gao et al. [19] also propose to train the class-specific dictionaries and a shared dictionary for addressing the fine-grained recognition problem. In addition, Gu et

al. [10] learn a structured synthesis dictionary and a structured analysis dictionary simultaneously for enhancing the representation power of feature.

For the feature generation task, the convolutional neural networks [21–23] provide powerful solutions. One advantage of CNN based methods is that they allow fusing the global and local information through gradually more specific filter layers with increasing receptive fields.

Recently, the region proposals approaches [24–27] provide an effective option to generate the object candidates from image. These methods utilize objectness measures derived from different visual cues. Compared with the traditional image interest points and sliding windows, region proposals are capable of detecting regions with higher semantic content.

3 Proposed Method

We first introduce the three components of our object-aware dictionary learning framework. Furthermore, the optimization algorithm is presented for solving all the variables in our OADL objective function.

3.1 Region Proposal Generation

In order to explore the local semantic information within image, we propose to take advantage of convolutional neural network (CNN) features and further integrate deep features with the region proposals systematically in our method. Compared with the fixed size of image patch, the region proposal is a mid-level element of image. Thus, it enables important region information in image to be collected for the recognition task.

We use the EdgeBox [27] algorithm to generate a set of initial region proposals within image. Then the non-maximum suppression (NMS) is adopted to refine these region windows, where the overlap rate of NMS is set to 0.8 IoU. Afterward, the deep feature is utilized to describe each region proposal in image. Finally, all the CNN features of region proposals from the training samples are fed into our OADL model to obtain these discriminative class-specific dictionaries.

3.2 Object-Aware Dictionary Learning (OADL)

Let $X = [X_1, X_2, ..., X_C]$ be a set of training data with C classes, where $X_i \in \mathbf{R}^{d \times N_i}, i = 1, 2, ..., C$ denotes the training samples corresponding to class i, d is the dimension of feature and N_i denotes the number of samples from class i. The goal of the OADL objective is to learn a structured dictionary $D = [D_1, D_2, ..., D_C, D_{C+1}] \in \mathbf{R}^{d \times K}$, which is used to transform the CNN feature into a discriminative feature space. $K = \sum_{i=1}^{C+1} K_i$ is the number of visual atoms in dictionary D, where K_i denotes the number of visual atoms in class-specific dictionary D_i. Since the background information of different classes may have the similar visual patterns, we further incorporate a background dictionary D_{C+1}

to separate the shareable visual patterns in the OADL model. We formulate the OADL model for C classes as:

$$\begin{aligned}\min_{A_i,Z_i,D_i,D_{C+1},w_i,b_i} \sum_{i=1}^{C}\{\sum_{n=1}^{N_i}[\| X_i^n - D_{\in i}Z_i^n \|_F^2 + \| X_i^n - DA_i^n \|_F^2 \\ + \alpha \| A_i^n - Z_i^n \|_F^2 + J(w_i, b_i, A_i^n, y_i) + \lambda_1 \| A_i^n \|_1 \\ + \lambda_2 \| Z_i^n \|_1] + \beta \sum_{j=1,j\neq i}^{C+1} \| D_i^T D_j \|_F^2\}\end{aligned} \tag{1}$$

where $X_i^n,\ n = 1, 2, ..., N_i$ denotes all the feature data of n-th image from class i, dictionary $D_{\in i}$ is defined as $[O_{d\times\Sigma_{q=1}^{i-1}K_q}, D_i, O_{d\times\Sigma_{q=i+1}^{C+1}K_q}] \in \mathbf{R}^{d\times K}$, O is the zero matrix. In other word, dictionary $D_{\in i}$ is a sub-dictionary associated with class i. Z_i^n is the representation coefficients of X_i^n on dictionary $D_{\in i}$. $D = [D_1, D_2, ..., D_C, D_{C+1}] \in \mathbf{R}^{d\times K}$ is a structured dictionary that concatenates all the class-specific dictionaries $D_i, i = 1, 2, ..., C$ and the additional background dictionary D_{C+1} together. A_i^n is the representation coefficients of X_i^n on the structured dictionary D. $J(\cdot)$ is defined as $J(w_i, b_i, A_i^n, y_i) = \| w_i \|^2 + R(w_i, b_i, A_i^n, y_i)$, where $R(w_i, b_i, A_i^n, y_i) = \eta\sum_{j=1}^{P_i^n}[\max(0, y_i \cdot w_i^T A_i^{n,j} + b_i - 1)]^2$ is the quadratic hinge loss due to the differentiable property [28], η is a constant, P_i^n denotes the number of region proposals in the n-th image from class i, y_i is the label of sample corresponding to class i, and $A_i^{n,j}$ denotes the representation feature of j-th region proposal within the n-th image from class i. The A_i, Z_i denote the representation coefficients of X_i on dictionaries D and $D_{\in i}$, respectively. $\| A_i^n \|_1$, $\| Z_i^n \|_1$ are the sparsity constrains imposed on the representation coefficients A_i^n and Z_i^n. $\alpha, \beta, \lambda_1, \lambda_2$ are the weighting parameters to balance the different terms in the objective function.

Discriminative reconstruction terms: The first two terms of Eq. (1) are the reconstruction residual terms. These two terms ensure the input data from class i not only to be represented using the class-specific dictionary $D_{\in i}$, but also be reconstructed by the structured dictionary D. Due to this property, the learned class-specific dictionaries have both the constructive and discriminative capabilities.

Coefficients consistency constraint: The third term $\| A_i^n - Z_i^n \|_F^2$ aims to make the representation coefficients have the consistent characteristic. In this energy term, A_i^n is the representation coefficients of X_i^n using the structured dictionary D, and Z_i^n denotes the representation coefficients of X_i^n on the class-specific dictionary $D_{\in i}$. This penalty term encourages the consistency between the representation coefficients A_i^n and Z_i^n. Therefore, the non-zero entries of A_i^n only appear on the indexes of visual atoms associated with class-specific dictionary D_i. In other words, it indicates that the structured dictionary D tends to represent the samples X_i^n of class i by choosing the visual atoms in dictionary D_i. Due to this consistency property, the discriminative power of feature representation can be strengthened.

Classification error term: The fourth term $J(w_i, b_i, A_i^n, y_i)$ of Eq. (1) is a loss function to measure the classification error. In our method, we incorporate a SVM formulation into our dictionary learning objective. w_i, b_i are the parameters of SVM classifier, A_i^n is the representation coefficients of feature sample X_i^n on dictionary D. The minimization of this term is to guide the dictionary learning process, which is beneficial to derive the discriminative feature representation.

Dictionary incoherent constraint: The seventh energy term, as in [14], is designed to ensure the learned class-specific dictionaries to be as independent as possible. Besides, we also impose the incoherent constraint between all the class-specific dictionaries $D_i, i = 1, 2, ..., C$ and the additional background dictionary D_{C+1}, which is used to separate the shared visual patterns and the class-specific visual patterns for all classes. Due to the incoherent property, the discriminative power of encoded feature can be enhanced.

3.3 Construction of Image-Level Feature

We first describe the proposed feature representation strategy using a group sparsity constraint. Moreover, we introduce the construction of image-level feature for final recognition task.

Given the learned discriminative structured dictionary $D = [D_1, D_2, ..., D_C, D_{C+1}]$, we propose to encode the deep feature of object proposal with the l_1/l_2-norm group sparsity constrain. Mathematically, the feature coding step is solved by the following l_1/l_2-norm regularized least squares problem.

$$\min_{B_i^n} \parallel X_i^n - D_{/C+1} B_i^n \parallel_2^2 + \rho \sum_{m=1}^{C} \parallel B_{i,m}^n \parallel_2 \tag{2}$$

where the dictionary $D_{/C+1}$ denotes the structured dictionary D when removing the visual atoms associated with the background class $C + 1$. Instead of using the overall dictionary D for feature representation, the shared visual patterns of all classes corresponding to potential background information can be separated by the dictionary $D_{/C+1}$. B_i^n is the representation coefficients of X_i^n on dictionary $D_{/C+1}$. In the feature coding step, we divide the representation coefficients into C non-overlapping groups, where $B_{i,m}^n$, $m = 1, 2, ...C$ denotes the m-th group of representation coefficients B_i^n. The entry indexes of $B_{i,m}^n$ is associated to the class-specific dictionary D_m. This feature representation strategy with l_1/l_2-norm sparsity encourages the dictionary $D_{/C+1}$ to represent feature sample by selecting the groups of visual atoms corresponding to the class-specific dictionaries. Therefore, the discriminative power of feature representation can be promoted effectively. ρ is a weighting parameter to balance the reconstruction term and the sparsity constrain in the objective function.

Figure 2 depicts the proposed feature representation property. In our method, we use the SLEP tool [29] to solve the minimization problem of Eq. (2). Once all the feature representations of region proposals within an image are computed, we then use them to construct the image-level feature by max-pooling for recognition task.

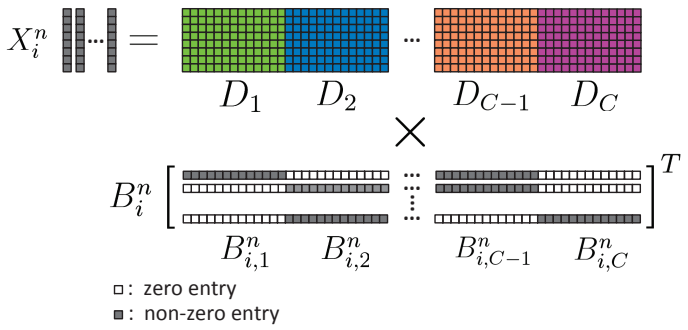

Fig. 2. Visual interpretation of the proposed feature coding strategy.

3.4 Optimization Algorithm

To solve the OADL objective, we derive an iterative optimization algorithm to compute all the variables in Eq. (1) alternatively. The detailed optimization procedures can be divided into the following five sub-problems: (1) updating variable A_i with fixed variables Z_i, D_i, D_{C+1} and w_i, b_i; (2) computing Z_i by fixing A_i, D_i, D_{C+1} and w_i, b_i; (3) updating dictionary D_i when fixing A_i, Z_i, D_{C+1} and w_i, b_i; (4) updating dictionary D_{C+1} with fixed A_i, Z_i, D_i, w_i, b_i. (5) updating w_i, b_i while fixing variables A_i, Z_i, D_i, D_{C+1}.

Updating A_i^n: With fixing the representation coefficients Z_i^n, dictionaries D_i, D_{C+1} and classifier parameters w_i, b_i, we can reduce the objective function of Eq. (1) with respect to A_i^n into the following optimization formulation.

$$\begin{aligned} \min_{A_i^n} \parallel X_i^n - DA_i^n \parallel_F^2 + \alpha \parallel A_i^n - Z_i^n \parallel_F^2 \\ + \ R(w_i, b_i, A_i^n, y_i) + \ \lambda_1 \parallel A_i^n \parallel_1 \end{aligned} \tag{3}$$

For updating the representation coefficients A_i^n, we propose to update each representation feature of region proposal within image one by one. The representation coefficients A_i^n can be rewritten as $A_i^n = [A_i^{n,1}, A_i^{n,2}, ..., A_i^{n,P_i^n}] \in \mathbf{R}^{K \times P_i^n}$, where $A_i^{n,j} \in \mathbf{R}^{K \times 1}$, $j = 1, 2, ..., P_i^n$ denotes the representation feature of j-th region proposal in the n-th image from class i, P_i^n is the number of region proposals in the n-th image from class i. To update the representation feature of each proposal, we first transfer the image label to these region proposals associated with this image. Then the classification error cost of $A_i^{n,j}$ is computed using a linear SVM with parameters w_i, b_i. If the predicted label of $A_i^{n,j}$ is consistent with the groundtruth, the classification error cost is set to zero. Otherwise, we use $\parallel b_i - 1 + y_i \cdot w_i^T A_i^{n,j} \parallel_F^2$ to approximate the quadratic hinge loss. Finally, the minimization problem of Eq. (1) with respect to each representation feature $A_i^{n,j}$ can be converted into the standard sparse coding formulation with l_1-norm.

Updating Z_i^n: Suppose that the variables A_i^n, D_i, D_{C+1} and w_i, b_i are fixed,

we can compute the representation coefficients Z_i^n as the following form:

$$\min_{Z_i^n} \parallel X_i^n - D_{\in i} Z_i^n \parallel_F^2 + \alpha \parallel A_i^n - Z_i^n \parallel_F^2 + \lambda_2 \parallel Z_i^n \parallel_1 \tag{4}$$

the above equation can be rewritten as

$$\min_{Z_i^n} \left\| \begin{pmatrix} X_i^n \\ \sqrt{\alpha} A_i^n \end{pmatrix} - \begin{pmatrix} D_{\in i} \\ \sqrt{\alpha} I \end{pmatrix} Z_i^n \right\|_F^2 + \ \lambda_2 \parallel Z_i^n \parallel_1 \tag{5}$$

where $I \in \mathbf{R}^{K \times K}$ denotes an identity matrix, K is the number of visual atoms in structured dictionary D. Let $\widetilde{X_i^n} = (X_i^n, \sqrt{\alpha} A_i^n)^T$, $\widetilde{D_{\in i}} = (D_{\in i}, \sqrt{\alpha} I)^T$, the minimization formulation of Eq. (5) is converted to a standard sparse coding problem. In our method, we use SPAMS solver [30] to achieve the optimal variable.

Updating D_i: When the representation coefficients A_i, Z_i, background dictionary D_{C+1} and classifier parameters w_i, b_i are fixed, each class-specific dictionary D_i can be updated. More specifically, we update these class-specific dictionaries class by class. While we compute the dictionary D_i, all the other dictionaries are fixed. Removing the terms that are independent of class-specific dictionary D_i, the optimization objective function (1) with respect to D_i is reduced to the following from:

$$\begin{aligned} \min_{D_i} \parallel X_i - D_{\in i} Z_i \parallel_F^2 + \parallel X_i - D A_i \parallel_F^2 \\ + \ \beta \sum_{j=1, j \neq i}^{C+1} \parallel D_i^T D_j \parallel_F^2 \end{aligned} \tag{6}$$

As for the above minimization function, we propose to compute each visual atom of dictionary $D_i = [d_i^1, d_i^2, ..., d_i^{K_i}] \in \mathbf{R}^{d \times K_i}$ one by one. When we compute the t-th visual atom d_i^t, the other visual atoms of D_i are fixed. We rewrite the representation coefficients Z_i and A_i as $Z_i = [z_i^1; z_i^2; ...; z_i^K] \in \mathbf{R}^{K \times N_i}$, $A_i = [a_i^1; a_i^2; ...; a_i^K] \in \mathbf{R}^{K \times N_i}$, where $z_i^t \in \mathbf{R}^{1 \times N_i}$, $a_i^t \in \mathbf{R}^{1 \times N_i}$, $t = 1, 2, ..., K$ denote the t-th row vector of Z_i and A_i, respectively. To update visual atom d_i^t, we let the first derivative of d_i^t equal to zero. Therefore, the t-th visual atom in dictionary D_i is computed as the closed-form

$$d_i^t = \ (\parallel z_i^t \parallel_2^2 I + \parallel a_i^t \parallel_2^2 I + \beta H_1 H_1^T)^{-1} \cdot (Y_1 \cdot z_i^{t^T} + Y_2 \cdot a_i^{t^T}) \tag{7}$$

where $Y_1 = X_i - \sum_{u=1, u \neq t}^{K_i} d_i^u z_i^u$, $Y_2 = X_i - \sum_{u=1, u \neq t}^{K_i} d_i^u a_i^u - \sum_{h=1, h \neq i}^{C+1} D_h A_i^h$ and $H_1 = [D_1, D_2, ..., D_{i-1}, \mathbf{O}_{d \times K_i}, D_{i+1} ..., D_C, D_{C+1}]$. The A_i^h denotes the submatrix of representation coefficients A_i corresponding to the indexes of h-th class, $\mathbf{O}$ is a zero matrix associated with the indexes of class-specific dictionary D_i.

As an visual atom in dictionary, the atom d_i^t is further normalized by the l_2-norm, *i.e.* $\hat{d_i^t} = d_i^t / \parallel d_i^t \parallel_2$. Therefore, we can compute all the visual atoms of class-specific dictionary D_i accordingly.

Updating D_{C+1}: In order to compute the background dictionary D_{C+1}, the

other variables A_i, Z_i, D_i, w_i, b_i are fixed. When removing the independent terms with respect to D_{C+1}, the minimization formulation of Eq. (1) is converted into the following optimization problem.

$$\min_{D_{C+1}} \parallel X_i - DA_i \parallel_F^2 + \beta \sum_{j=1, j\neq C+1}^{C+1} \parallel D_{C+1}^T D_j \parallel_F^2 \tag{8}$$

In our method, we propose to update each visual atom of background dictionary D_{C+1} one by one. When the t-th visual atom is updated, the rest visual atoms in D_{C+1} are fixed. Thus, we can compute the t-th visual atom d_{c+1}^t by the closed-form solution.

$$d_{c+1}^t = (\parallel a_i^t \parallel_2^2 I + \beta H_2 H_2^T)^{-1} \cdot (Y_3 \cdot a_i^{t^T}) \tag{9}$$

where $Y_3 = X_i - \sum_{u=1,u\neq t}^{K_{C+1}} d_{c+1}^u a_i^u - \sum_{h=1,h\neq C+1}^{C+1} D_h A_i^h$ and $H_2 = [D_1, D_2, ..., D_C, \mathbf{O}_{d\times K_{C+1}}]$. The a_i^t, $t = 1, 2, ...K$ denotes the t-th row vector of representation coefficients A_i, and A_i^h is the sub-matrix of A_i corresponding to the indexes of h-th class, $\mathbf{O}$ denotes a zero matrix associated with background dictionary D_{C+1}. The updated dictionary atom is then normalized by the l_2-norm. Once all the visual atoms in D_{C+1} is computed, the background dictionary D_{C+1} is updated.

Updating w_i, b_i: To update the classifier parameters w_i, b_i, the other variables are fixed. In our method, we cast the SVM classifier learning problem with C classes into the C one-vs-all SVM sub-problems. More specifically, we first transfer the image label to these region proposals corresponding to this image. Then all the feature representations of region proposals across all classes are used to train multiple SVM classifiers. In our OADL, a linear SVM solver [28] is adopted to learn the parameters of SVM classifiers.

Initialization: As for our OADL objective, we need to initialize the variables $\{D_i, Z_i, A_i,\ i = 1, 2, ..., C\}$ and D_{C+1}. For class-specific dictionary D_i, it is initialized by the K-SVD [15] algorithm using all the region proposals of images from class i. We also adopt the K-SVD algorithm to initialize the background dictionary D_{C+1} using the region proposals of training samples across all classes. The representation coefficients Z_i and A_i are initialized by solving the sparse coding problem with $l_{2,1}$-norm: $\min_{Z_i} \parallel X_i - D_{\in i} Z_i \parallel_F^2 + \rho_1 \parallel Z_i \parallel_{2,1}$, $\min_{A_i} \parallel X_i - DA_i \parallel_F^2 + \rho_2 \parallel A_i \parallel_{2,1}$, respectively. ρ_1, ρ_2 are the scale parameters to balance the different energy terms. The proposed optimization procedures of the OADL objective are summarized in Algorithm 1.

4 Experiments

In this section, we verify the effectiveness of our method with other competing methods on the UIUC8 Sport [31] and Graz-02 [32] public datasets. The goal of our OADL method is to learn a feature subspace for improving the discriminative power of feature representation. In the experiments, we adopt two

Algorithm 1 Object-Aware Dictionary Learning

Require: training samples $X = [X_1, X_2, ..., X_C]$, the number of visual atoms $K_i,\ i = 1, 2, ..., C, C+1$ for each class-specific dictionary and an additional background dictionary, parameters α, β, λ_1 and λ_2.
Ensure: class-specific dictionary $D_i,\ i = 1, 2, ..., C$, background dictionary D_{C+1}
initialize $D_i, Z_i, A_i,\ i = 1, 2, ..., C$, and D_{C+1}
while not convergence and the maximum number of iterations is not reached **do**
  for $i = 1 \longrightarrow C$ **do**
    update representation coefficients A_i^n by solving Eq. (3);
    update representation coefficients Z_i^n using Eq. (5);
    for $t = 1 \longrightarrow K_i$ **do**
      update the t-th visual atom d_i^t of class-specific dictionary D_i by Eq. (7);
    end for
    for $t = 1 \longrightarrow K_{C+1}$ **do**
      update the t-th visual atom d_{c+1}^t in background dictionary D_{C+1} by Eq. (9);
    end for
    compute the parameters w_i, b_i of SVM classifier ;
  end for
end while

convolutional neural networks: VGG-F [22] and VGG-VeryDeep19 [23] models to generate the CNN features and further evaluate our method. The OADL(F) and OADL(VD19) denote our method integrated with the different deep features for brevity. As for the OADL, the dimension of deep feature is reduced to 1000 by PCA in the experiments. The weighting parameters $\alpha, \beta, \lambda_1, \lambda_2$ of the OADL model are empirically set as 0.01, constant η is set to 0.2. In the recognition stage, the scale parameter ρ for the regularization term of group sparsity is set to 0.5. Finally, the obtained global image representation is fed into a linear SVM classifier for predicting the label of image.

4.1 UIUC8 Sport Dataset

We first evaluate our method and several competing approaches on the UIUC8 Sport [31] event recognition dataset. This dataset have eight sport classes and 1792 images in total, including rowing, badminton, polo, bocce, snowboarding, croquet, sailing and rockclimbing. The number of images from each class varies from 137 to 250. Several example images of this dataset are shown in Fig. 3.

Following the common experimental setting on this dataset [33], we randomly choose 70 images from each class as the training samples and randomly select 60 images from the rest images as the testing samples in the experiment. As for our OADL model, we learn the class-specific dictionary with 200 visual atoms for each class. The number of visual atoms in background dictionary is also set to 200. We show the confusion matrices of our method on the UIUC8 Sport dataset in Fig. 4. More specifically, the confusion matrix of OADL incorporating with the deep feature by VGG-F model is demonstrated in Fig. 4(a).

Fig. 3. Sample images from different classes on the UIUC8 Sport dataset.

Table 1. Performance comparisons between our method and the state-of-the-art approaches on the UIUC8 Sport dataset.

Method	Accuracy (%)	**Method**	Accuracy (%)
KSPM [34]	79.98	RSP [42]	79.6
ScSPM [28]	82.74	LPR [43]	86.25
LLC [35]	81.77	LSC [33]	82.79
K-SVD [15]	82.21	LScSPM [44]	85.31
SPMSM [36]	83.0	Fusion [45]	94.8
LRSC [37]	88.17	DSFL+DdCAF [46]	96.78
VLAD [38]	79.16	VGG-F [22]	94.5
VC+VQ [39]	88.4	VGG-VD19 [23]	95.45
OB [40]	77.88	OADL(F)	96.9
ISPR [41]	89.5	OADL(VD19)	**98.09**

Figure 4(b) shows the obtained confusion matrix by our OADL with the deep feature of VGG-VeryDeep19 model on the UIUC8 Sport dataset. Moreover, we evaluate our method with several competing approaches on this dataset, such as KSPM [34], ScSPM [28], LLC [35], KSVD [15], SPMSM [36], LRSC [37], VLAD [38], VC+VQ [39], OB [40], ISPR [41], RSP [42], LPR [43], LSC [33], LScSPM [44], Fusion [45], DSFL+DdCAF [46] and the two deep features by VGG-F [22] and VGG-VD19 [23] models. The recognition results of different methods are summarized in Table 1. It is noticed that our OADL model outperforms the state-of-the-art methods, including the recent dictionary learning and two powerful deep learning based image classification approaches. Finally, our method achieves the highest performance on the UIUC8 Sport dataset. In addition, we can observe that the OADL(VD19) gains the better recognition accuracy than the OADL(F). It indicates the discrimination of feature representation by our OADL method can be further enhanced with the increase of depth in convolutional network.

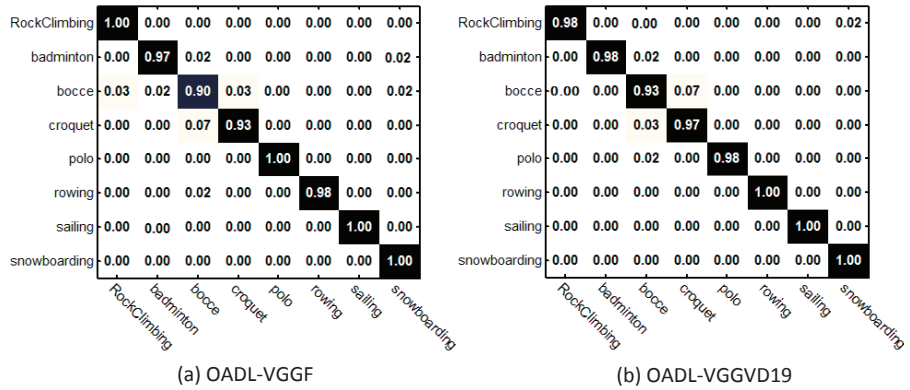

Fig. 4. Confusion matrices on the UIUC8 Sport dataset for our method. (a) Confusion matrix with the deep feature of VGG-F model. (b) Confusion matrix with the deep feature of VGG-VeryDeep19 model.

Table 2. Recognition results of different methods on the Graz-02 dataset.

Method	Bike	Car	People	Total
[47]	89.5	80.2	85.2	84.9
[48]	91.2	87.5	85.3	88.0
[49]	-	-	-	82.2
VGG-F [22]	94.44	96.05	85.71	92.48
VGG-VD19 [23]	96.91	**97.74**	89.29	94.98
OADL(F)	98.15	97.18	88.57	94.99
OADL(VD19)	**98.77**	**97.74**	**91.43**	**96.24**

4.2 Graz-02 Dataset

The Graz-02 dataset contains 1096 images with three classes, including bike, car and people. It is also a challenge object recognition dataset because the objects from each class have the large intra-class differences in location, scale and viewpoint, as shown in Fig. 5. The effectiveness of our method is also tested on this dataset following the standard evaluation setting [32]. In detail, the class-specific dictionary with 400 visual atoms is learned for each class. For the background dictionary in OADL model, the number of visual atoms is also set to 400 in the experiment. Furthermore, we compare the OADL method with several competing approaches [47–49] and the two CNN features by VGG-F [22] and VeryDeep-19 [23] models on this dataset. The recognition results of different methods are summarized in Table 2. As can been seen, our OADL method is superior to the deep features and other competing approaches on the Graz-02 dataset. It is also observed that the discriminative power of feature generated by our OADL method can be promoted effectively with the increasing depth of convolutional network.

Fig. 5. Sample images of different classes from the Graz-02 dataset.

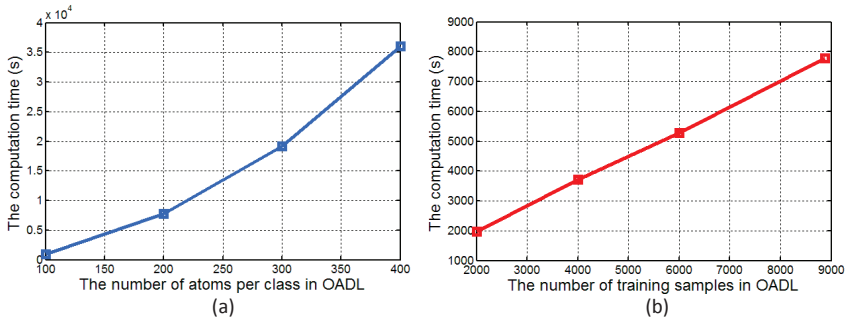

Fig. 6. Computation time analysis on the Graz-02 dataset. (a) Training time of OADL as a function of the number of visual atoms per class. (b) Running time of OADL with the increase of training samples across all classes.

In addition, we give the computation time of our method on the Graz-02 dataset in Fig. 6. Specifically, the number of training samples across all classes is first fixed to 8874 in the experiments, then we vary the number of visual atoms per class with $[100, 200, 300, 400]$. The running time of OADL during each iteration as a function of the number of visual atoms is shown in Fig. 6(a). We can see that the computation time of OADL increases with the number of visual atoms per class gradually. With fixed the number of visual atoms per class to 200, we change the number of training samples from all classes in the range $[8874, 6000, 4000, 2000]$. Figure 6(b) demonstrates that the runtime of OADL increases with the growth of training samples. All experiments are performed using a single CPU Intel Core at 3.0GHz.

5 Conclusion

Visual dictionary learning provides a data-driven manner to represent image data as a linear combination of a few atoms from an over-complete dictionary.

However, a critical problem of existing dictionary learning is that it does not select and concentrate on the important image regions explicitly. Thus, discriminative semantic information within image regions cannot be exploited effectively during dictionary learning procedure. Currently, the convolutional neural network (CNN) has the capacity to combine the global and local information within image by means of designed specific filter layers with the increasingly receptive fields. Motivated by the advantage of deep feature, we proposed an object-aware dictionary learning framework that integrates the deep features and region proposals systematically to overcome this problem. Instead of extracting a dictionary from all the fixed size of image patches or entire image, our method focuses on the small amounts of object candidates, which ensure the local semantic information can be encoded into the feature representation of image. We treat this as an optimization problem and derive an iterative algorithm to solve it. The experimental results on the public benchmark datasets demonstrate that our method outperforms the state-of-the-art dictionary learning and deep learning based image classification approaches.

References

1. Elad, M., Aharon, M.: Image denoising via learned dictionaries and sparse representation. In: CVPR. (2006)
2. Fu, Y., Lam, A., Sato, I., Sato, Y.: Adaptive spatial-spectral dictionary learning for hyperspectral image denoising. In: ICCV. (2015)
3. Bao, C., Cai, J.F., Ji, H.: Fast sparsity-based orthogonal dictionary learning for image restoration. In: ICCV. (2013)
4. Wang, S., Zhang, L., Liang, Y., Pan, Q.: Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: CVPR. (2012)
5. Huang, D.A., Wang, Y.C.F.: Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In: ICCV. (2013)
6. Wang, N., Wang, J., Yeung, D.: Online robust non-negative dictionary learning for visual tracking. In: ICCV. (2013)
7. Yang, M., Zhang, D., Feng, X., Zhang, D.: Fisher discrimination dictionary learning for sparse representation. In: ICCV. (2011)
8. Zhou, N., Shen, Y., Peng, J., Fan, J.: Learning inter-related visual dictionary for object recognition. In: CVPR. (2012)
9. Jiang, Z., Lin, Z., Davis, L.: Label consistent k-svd: Learning a discriminative dictionary for recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013) 2651–2664
10. Gu, S., Zhang, L., Zuo, W., Feng, X.: Projective dictionary pair learning for pattern classification. In: NIPS. (2014)
11. Cai, S., Zuo, W., Zhang, L., Feng, X., Wang, P.: Support vector guided dictionary learning. In: ECCV. (2014)
12. Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., Bach, F.R.: Supervised dictionary learning. In Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., eds.: NIPS. (2009)
13. Yang, J., Yu, K., Huang, T.: Supervised translation-invariant sparse coding. In: CVPR. (2010)

14. Ramirez, I., Sprechmann, P., Sapiro, G.: Classification and clustering via dictionary learning with structured incoherence and shared features. In: CVPR. (2010)
15. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54** (2006) 4311–4322
16. Zhang, Q., Li, B.: Discriminative k-svd for dictionary learning in face recognition. In: CVPR. (2010)
17. Yang, L., Jin, R., Sukthankar, R., Jurie, F.: Unifying discriminative visual codebook generation with classifier training for object category recognition. In: CVPR. (2008)
18. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Discriminative learned dictionaries for local image analysis. In: CVPR. (2008)
19. Gao, S., Tsang, I.H., Ma, Y.: Learning category-specific dictionary and shared dictionary for fine-grained image categorization. IEEE Transactions on Image Processing **23** (2014) 623–634
20. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: NIPS. (2007)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
22. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: British Machine Vision Conference. (2014)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2014)
24. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR. (2010)
25. Manen, S., Guillaumin, M., Gool, L.V.: Prime object proposals with randomized prim's algorithm. In: ICCV. (2013)
26. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.H.S.: BING: Binarized normed gradients for objectness estimation at 300fps. In: CVPR. (2014)
27. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV. (2014)
28. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR. (2009)
29. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections. Arizona State University. (2009)
30. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11** (2010) 19–60
31. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: ICCV. (2007)
32. Marszałek, M., Schmid, C.: Accurate object localization with shape masks. In: IEEE Conference on Computer Vision & Pattern Recognition. (2007)
33. Liu, L., Wang, L., Liu, X.: In defense of soft-assignment coding. In: ICCV. (2011)
34. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
35. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR. (2010)
36. Kwitt, R., Vasconcelos, N., Rasiwasia, N.: Scene recognition on the semantic manifold. In: ECCV. Volume 7575. (2012)
37. Zhang, T., Ghanem, B., Liu, S., Xu, C., Ahuja, N.: Low-rank sparse coding for image classification. In: ICCV. (2013)

38. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010)
39. Li, Q., Wu, J., Tu, Z.: Harvesting mid-level visual concepts from large-scale internet images. In: CVPR. (2013) 851–858
40. Li, L.J., Su, H., Lim, Y., Li, F.F.: Objects as attributes for scene classification. In: ECCV. (2010)
41. Lin, D., Lu, C., Liao, R., Jia, J.: Learning important spatial pooling regions for scene classification. In: CVPR. (2014)
42. Jiang, Y., Yuan, J., Yu, G.: Randomized spatial partition for scene recognition. In: ECCV. (2012)
43. Sadeghi, F., Tappen, M.F.: Latent pyramidal regions for recognizing scenes. In: ECCV. (2012)
44. Gao, S., Tsang, I.W.H., Chia, L.T., Zhao, P.: Local features are not lonely: Laplacian sparse coding for image classification. In: CVPR. (2010)
45. Koskela, M., Laaksonen, J.: Convolutional network features for scene recognition. In: Proceedings of the 22Nd ACM International Conference on Multimedia. (2014)
46. Zuo, Z., Wang, G., Shuai, B., Zhao, L., Yang, Q., Jiang, X.: Learning discriminative and shareable features for scene classification. In: ECCV. (2014)
47. Tuytelaars, T.: Vector quantizing feature space with a regular lattice. In: ICCV. (2007)
48. Krapac, J., Verbeek, J., Jurie, F.: Learning Tree-structured Descriptor Quantizers for Image Categorization. In: BMVC. (2011)
49. Hong, Y., Li, Q., Jiang, J., Tu, Z.: Learning a mixture of sparse distance metrics for classification and dimensionality reduction. In: ICCV. (2011)