Multi-Level Discriminative Dictionary Learning towards Hierarchical Visual Categorization

Li Shen[1], Shuhui Wang[2], Gang Sun[1,3], Shuqiang Jiang[2], and Qingming Huang[1,2]

[1]University of Chinese Academy of Sciences, Beijing, China
[2]Key Lab of Intell. Info. Process. (CAS), Inst. of Comput. Tech., CAS, Beijing, China
[3]State Key Lab. of Computer Science, Inst. of Software, CAS, Beijing, China
{lshen, shwang, sqjiang, qmhuang}@jdl.ac.cn, sung@ios.ac.cn

Abstract

For the task of visual categorization, the learning model is expected to be endowed with discriminative visual feature representation and flexibilities in processing many categories. Many existing approaches are designed based on a flat category structure, or rely on a set of pre-computed visual features, hence may not be appreciated for dealing with large numbers of categories. In this paper, we propose a novel dictionary learning method by taking advantage of hierarchical category correlation. For each internode of the hierarchical category structure, a discriminative dictionary and a set of classification models are learnt for visual categorization, and the dictionaries in different layers are learnt to exploit the discriminative visual properties of different granularity. Moreover, the dictionaries in lower levels also inherit the dictionary of ancestor nodes, so that categories in lower levels are described with multi-scale visual information using our dictionary learning approach. Experiments on ImageNet object data subset and SUN397 scene dataset demonstrate that our approach achieves promising performance on data with large numbers of classes compared with some state-of-the-art methods, and is more efficient in processing large numbers of categories.

1. Introduction

Visual categorization serves as a challenging issue in computer vision and machine learning research. The prospect of classifying large numbers of categories in real world scenario has drawn much attention. For visual categorization, the importance of features has been addressed in a mass of work [14, 15, 16]. Among a variety of visual features, the Bag-of-Words representation built on local descriptors (*e.g.*, SIFT) has achieved great success in visual categorization [9, 24]. The dictionary applied to quantize the local descriptors is often obtained by some clustering method, such as k-means.

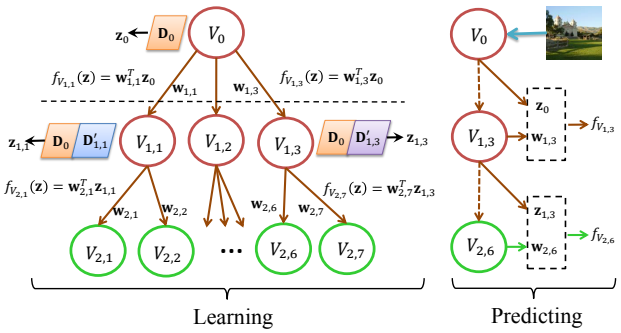

Figure 1. An example framework in this paper. In learning stage, D_0 denotes the dictionary used by first level nodes ($V_{1,1}$,$V_{1,2}$ and $V_{1,3}$). $D_{1,1}$ consists of the inherited part D_0 and specific part $D'_{1,1}$. The corresponding representation $z_{1,1}$ is used for classification model learning among the child nodes ($V_{2,1}$ and $V_{2,2}$). In predicting stage, the input image goes through the tree by selecting the node with maximal response.

In recent work [28, 18], sparse coding based dictionary learning has reported more promising results. The dictionary is optimized via minimizing reconstruction error. Local coding [26, 31] which puts the local geometry constraint on data points achieves impressive results. Furthermore, it is shown in [17, 2] that the dictionaries via supervised learning are beneficial for performance improvement by encoding more discriminative information into the representations. However, when the number of categories is large, these methods suffer from considerable time overhead during the supervised dictionary learning and predicting stages.

Much existing work has shown that using object hierarchy to guide the model learning for classification can

bring in improvements in both efficiency and accuracy [20, 19, 35, 11]. The categories are usually organized in the form of tree-structured hierarchy [6] and are treated as the leaf nodes in the bottom of the tree. Each internode corresponds to one hyper-category that is composed of a group of categories with semantic relevance or visual similarity, so that the structure reflects the hierarchical correlation among categories.

In this paper, we observe some key statements on the benefit of the object hierarchy for classification. Firstly, due to diversified inter-correlation among different layers, the sibling nodes in higher levels are less related than the ones in lower levels, thus the discrimination between nodes in higher levels is easier. Conversely, it is much more difficult to distinguish a node from its adjacent ones in lower levels [6, 8]. For example, it is easy to distinguish *tree* and *dog* rather than *poplar* and *willow*. Secondly, The features of different spatial granularity can be spotted from natural images. Simple features extracted from relatively small regions are useful to classify less related concepts. On the other hand, the features extracted from larger regions describe more class-specific patterns [16, 10]. Finally, lower level categories are supposed to possess the general properties from the higher level categories and additional class-specific details. In other words, the classification may rely on different feature representations at different layers, even features chosen by different internodes at the same layer would likely be different. For example, the feature that best discriminates *rose* from *lily* is not effective for discriminating *flower* from *animal*, as well as *panda* from *monkey*.

Based on the above discussion, we propose a Multi-Level Discriminative Dictionary Learning (ML-DDL) approach for hierarchical categorization. The learnt visual dictionary set and feature representations make better use of the category correlation encoded by the hierarchy. As the features exacted from larger receptive fields encode more complex and specific patterns, the dictionaries learnt in lower layers are designed to encode the descriptors at larger scale. Given the structure, we learn one discriminative dictionary and a set of discriminative models for each hyper-category (internode). Besides, our learnt dictionaries in lower levels consist of additional part inherited from ancestor nodes, so that categories in lower levels are described with multi-scale visual information.

The framework is illustrated in Fig. 1. For internal node $V_{1,1}$, the corresponding dictionary $D_{1,1}$ consists of two parts D_0 and $D^{'}_{1,1}$. D_0 represents the dictionary inherited from node V_0. The specific dictionary $D^{'}_{1,1}$ and the class models of nodes $V_{2,1}$ and $V_{2,2}$ are learnt in a discriminative formulation simultaneously. On predicting stage, an image is labeled sequentially by choosing the node which outputs the largest response among its siblings until it reaches a leaf node. When it is going through one tree path from the root to a leaf node, only two $(L-1)$[1] visual features have to be computed. The proposed framework is similar with the branch-and-bound strategy based on the tree structure.

In this paper, the main contributions can be summarized as follows:

- We propose a local coding based Multi-Level Discriminative Dictionary Learning method for hierarchical categorization. The dictionaries learnt at different layers encode different scale information. Moreover, each dictionary consists of the general part inherited from upper layers and the specific part learnt from its child nodes. Compared with unsupervised dictionary [28] or class-specific dictionary [17, 2], our learnt dictionaries capture the information of multi-level visual (hyper-)categories in a more effective way.
- The time complexity of training and prediction can be significantly reduced compared with supervised dictionary learning such as [17]. On training stage, the number of dictionaries is equal with the number of internodes in the tree, which is far less than the number of categories. On prediction stage, a test image only needs to go through one tree path with $L-1$ dictionaries.

The rest of the paper is organized as follows. Section 2 introduces related work, and Section 3 presents our approach and model solution. Experimental results and analysis will be provided in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

Current dictionary learning approaches can be categorized into two main types: unsupervised and supervised dictionary learning. In the field of unsupervised dictionary learning, the dictionary is optimized only based on the reconstruction error of the signals [18]. Yang *et al*. [28] propose to learn a unique unsupervised dictionary by sparse coding for classification and achieve impressive results. Yu *et al*. [30] develop a hierarchical sparse coding method which models the spatial neighborhood dependency of local patches. Besides the sparsity, other constraints help to incorporate more information. Jenatton *et al*. [12] employ a tree-structured sparsity to encode the dependencies between dictionary atoms. Local coding [31, 26] integrates sparse property with local geometry, which ensures that similar signals tend to have similar codes. However, these methods do not take the discriminative information into account.

The dictionaries learnt in a supervised way have stronger discriminative power. For example, label information is fed in Fisher discrimination criterion or logistic regression model for dictionary learning [17, 29]. A shared dictionary or multiple class-specific dictionaries can be obtained by su-

[1] L denotes the number of levels

pervised learning. Boureau *et al*. [2] adapt supervised dictionary learning to local features, which achieves better discriminative power than features extracted by [28]. Zhou *et al*. [34] propose to learn multiple specific dictionaries and a global dictionary shared by all categories. Meanwhile, when the number of categories is large, the globally shared strategy is not suitable for dealing with the complex category correlation.

When the number of categories is large, the advantage of hierarchical structure over flat structure emerges in terms of efficiency and accuracy [22, 33, 3]. Much work exploits semantic or visual similarity requirements in the hierarchy for classification. Some studies follow the regularization-type routine by imposing statistical *similarity* or *dissimilarity* constraints between the nodes in the hierarchy. The *similarity* constraint assumes that sibling nodes should share similar statistical property [25, 22, 5, 1]. This constraint makes adjacent nodes belonging to a hyper-node much closer. On the other hand, the *dissimilarity* constraint encourages that the classifiers in different layers tend to use the different features [33]. Another solution of hierarchical learning is empirical-loss-type, *e.g*., the tree-induced loss [3, 32] regarding the hierarchical structure. Besides, some work leverages the hierarchical structure to capture the contextual information, such as object co-ocurrence and spatial layout [4]. These models are mostly based on the hand-crafted features which limits the model capacity.

3. Approach

3.1. Local Coding-based Dictionary Learning

Low-level visual descriptors (*e.g*., SIFT) contain rich information and are usually integrated to higher-level representation. The integration stages are composed of transformation (coding) [21, 31, 26] and pooling operation [2]. As one of the coding strategies, local coding aims to learn a set of anchor points (dictionary) to encode signals incorporating with locality constraint, which ensures that similar signals tend to have similar codes. As the supervised learning approach has been shown beneficial to dictionary learning, in this paper we propose to introduce a supervised formulation for local coding.

Consider training samples $X = [x_1, ..., x_n]$, and $\hat{x}_{i,p} \in \mathbb{R}^m$ denotes the p-th local descriptor belonging to sample x_i. Given a dictionary D_b, $\hat{x}_{i,p}$ can be reconstructed by:

$$\begin{aligned}\alpha_{i,p}(\hat{x}_{i,p}, D_b) \;=\; & \arg\min \frac{1}{2}\left\|\hat{x}_{i,p} - D_b\alpha_{i,p}\right\|_2^2 \\ & + \mu \sum_j \left|\alpha_{i,p}^j\right| \cdot \left\|d_j - \hat{x}_{i,p}\right\|_2^2 \quad (1)\end{aligned}$$

where $D_b \triangleq \left\{D_b \in \mathbb{R}^{m\times K_b},\, s.t. \forall j \in \{1, ..., K_b\}, \|d_j\|_2 \leqslant 1\right\}$, D_b consists of K_b atoms in the dictionary, and d_j denotes the j-th atom.

The codes of the descriptors are pooled together for image representation. Among the pooling strategies, spatial pooling makes use of the spatial layout of the local features and has brought promising result for image classification [28]. Due to the fact that max pooling is not differentiable, average pooling is more appropriate to task-driven dictionary learning [2]. We use *average spatial pooling* during training dictionary. The pooling result of the local descriptors belonging to the sample x_i is denoted by z_i.

Let $Z \in \mathbb{R}^{K_b^* \times n}$ denote the image representation set, and $Y \in \mathbb{R}^n$ be the set of labels corresponding to sample set X. We aim to learn a discriminative model in which classifier matrix W is trained based on image representations Z. For classifying U categories, we employ a multinomial logistic regression model. Given a sample x_i, the probability that x_i belongs to category u can be written as:

$$P(y_i = u|x_i) = \frac{\exp(w_u^\top z_i)}{\sum_{s=1}^U \exp(w_s^\top z_i)} \quad (2)$$

where w_u is the classifier for class u. And the loss function can be formulated by cross-entropy error:

$$loss = -\sum_{i=1}^n \sum_{u=1}^U I(y_i = u) \log \frac{\exp(w_u^\top z_i)}{\sum_{s=1}^U \exp(w_s^\top z_i)} \quad (3)$$

where $I(*)$ denotes the indicator function.

To prevent D from being arbitrarily large, we constrain its columns $\{d_1, d_2, ..., d_K\}$ to be bounded by unit l_2 norm. To jointly learn the dictionary and classification model, the model can be formulated as following:

$$\min_{W,D} \frac{\lambda}{2}\|W\|_F^2 + loss\big(W, y, z(x, D)\big) \quad (4)$$

where $D \triangleq \left\{D \in \mathbb{R}^{m\times K},\; s.t.\; \forall j \in \{1, ..., K\}, \|d_j\|_2 \leqslant 1\right\}$. The $loss$ function is defined in Eq. (3).

3.2. Multi-Scale Dictionary Learning

Given the tree structure, our goal is to learn a set of discriminative dictionaries for discrimination among the sibling nodes. As the general patterns (such as edges) can be extracted from the hyper-category in higher levels, they are useful for classifying general concepts. For example, regular horizontal and vertical lines can be found in *"building"*, while arcs are frequently observed in *"fruit"*. Instead, a mass of class-specific patterns can be discovered from specific categories, such as shapes and object parts. These patterns encode larger receptive fields and thus contain more specific information. Based on the above analysis, we propose to learn the dictionaries in different layers to encode the descriptors of different scales.

Let T denote the set of leaf nodes (categories) in the tree, $\bar{T}$ denote the set of internodes which represent the hyper-categories, and $T^+ = T \cup \bar{T}$ denote the set containing all

the nodes in the tree. We assume the tree has L levels and $l \in \{0, 1, ..., L-1\}$ is the level number. For each node $t \in \bar{T}$, $C(t)$ is identified as the set of child nodes of t. For classification, we aim to learn a function $F : X \to Y$ determined by the following process:

$$F = \left\{ \begin{array}{l} v = root \\ while \;\; v \notin T \\ \quad v = \arg\max_{u \in C(v)} w_u^\top x \\ end \end{array} \right\} \tag{5}$$

which means a sample should be labeled as the category with the maximal response compared with the sibling ones through the tree path until reaching a leaf node.

Under this formulation, each internal node t corresponds to a multiclass classification problem on its child nodes. We need to learn a single dictionary D_t shared by its child nodes. For the child node v of node t, we define a response function $f(*)$. Given the sample x_i, its response in node v can be written as:

$$f_v(x_i) \triangleq f_v(x_i, D_t) = w_v^\top z(x_i, D_t) \tag{6}$$

For the hierarchical model, the tree loss can be formulated by extending Eq. (3) as:

$$loss = -\sum_{i=1}^{n} \sum_{t \in \bar{T}} \sum_{v \in C(t)} I\big(v \in y_i^+\big) \log \frac{\exp\big(f_v(x_i)\big)}{\sum_{u \in C(t)} \exp\big(f_u(x_i)\big)} \tag{7}$$

where y_i^+ represents the label set of sample x_i in the tree. The dictionaries in different layers are learnt to discover the valuable properties with different scales.

3.3. Multi-Level Dictionary Learning for Hierarchical Representation

Given the hierarchical structure, we learn a set of discriminative dictionaries to encode the descriptors of different scales. The dictionary learnt on an internode consists of the specific properties for discriminating the children nodes, and these properties are supposed to be embodied in their children nodes. Therefore, the dictionaries in the higher levels can be regarded as the sharing properties for the groups of correlated categories in lower levels, and they can be inherited by the child nodes through the tree path.

For example, considering the node $V_{1,1}$ as shown in Fig. 1, the corresponding $D_{1,1}$ is expressed as $D_{1,1} = [D_0, D_{1,1}^{'}]$. D_0 denotes the inherited dictionary from V_0, and $D_{1,1}^{'}$ denotes the specific dictionary learnt in node $V_{1,1}$. Suppose the sample i belongs to node $V_{1,1}$, and the descriptor x_i consists of two parts as $x_i = [x_i^0, x_i^1]$. x_i^0 denotes the descriptor set with smaller scale, and x_i^1 denotes the one with larger scale. Then, for the child node $V_{2,1}$, the response function of sample i (Eq. (6)) will be rewritten as:

$$\begin{aligned} f_{V_{2,1}}(x_i, D_{1,1}) &= w_{2,1}^\top z(x_i, D_{1,1}) \\ &= w_{2,1}^\top [z(x_i^0, D_0), z(x_i^1, D_{1,1}^{'})] \end{aligned} \tag{8}$$

where $z(x_i, D_{1,1})$ is the image representation with multi-scale information in current layer .

The dictionary and classifier learning based on hierarchical structure can be revised from Eq. (4) as:

$$\min_{W, D^+} \frac{\lambda}{2} \|W\|_F^2 + loss(W, D^+, X, Y) \tag{9}$$

where D^+ represents the set of dictionaries in the tree and W denotes the classifier matrix embedded in the structure. The $loss$ function is given by Eq. (7).

In this algorithm, the information propagates via multi-level dictionaries in a top-down fashion. For the nodes in lower layers, the learnt dictionaries are desired to encode more specific information. On the other hand, the inherited dictionaries from ancestors consist of more general information, based on which the response z should be used to minimize the classification loss. We will demonstrate this scheme is useful for performance improvement in experiments.

3.4. Model Learning

Although the overall learning problem on the whole tree is very complicated, it can be decomposed into a set of sub-problems. From top layer to the bottom layer, the learning tasks are processed sequentially. However, the tasks corresponding to the nodes at the same layer can be processed independently. For an internode t, the learning process of dictionary and class models is done iteratively, which consists of two steps: 1) Coding: by fixing the dictionary D_t, we compute the coefficients and generate the features z_t of the samples. 2) Dictionary and class models updating: based on the features computed by previous dictionary, we optimize the class models and dictionary simultaneously. Particularly, the *specific* part of the dictionary needs to be updated rather than the *inherited* part.

In fact, the inherited dictionary has been optimal in the higher layers, it should be inherited without any update by the classification models of the descendant nodes in lower levels. With regard to class models, the classifier learning is a traditional classification task. The dictionary updating is a loss minimization problem through the learnt features z. As the loss function is differentiable with respect to dictionary and class model parameters, the gradient of specific dictionary D_t of internode t and class model of its child node v

can be computed as following:

$$\begin{cases} \dfrac{\partial loss}{\partial w_v} = -\sum\limits_{i\in A(t)}\left[\sum\limits_{v\in C(t)} I(v\in y_i^+)z_i^t - \dfrac{\exp(w_v^\top z_i^t)z_i^t}{\sum\limits_{u\in C(t)}\exp(w_u^\top z_i^t)}\right] \\ \dfrac{\partial loss}{\partial D_t} = \sum\limits_{i\in A(t)}\left\{\dfrac{1}{|P(i)|}\sum\limits_{p\in P(i)}\left[2\mu\Big(\hat{x}_i^p\,\gamma_{i,p}^\top - D_t\odot(\mathbf{1}\cdot\gamma_{i,p}^\top)\Big)\right.\right. \\ \qquad\left.\left. -D_t\beta_{i,p}\alpha_{i,p}^\top + \Big(\hat{x}_i^p - D_t\alpha_{i,p}\Big)\beta_{i,p}^\top\right]\right\} \end{cases} \tag{10}$$

where:

$$\beta_{i,p}^{\Lambda} = \left(D_t^{\Lambda\,\top} D_t^{\Lambda}\right)^{-1}\nabla_{z_i^{\Lambda}} loss, \quad \beta_{i,p}^{\bar{\Lambda}} = 0$$

$$\nabla_{z_i^t} loss = -\sum_{v\in C(t)} I(v\in y_i^+)w_v + \frac{\sum\limits_{u\in C(t)}\exp(w_u^\top z_i^t)w_u}{\sum\limits_{u\in C(t)}\exp(w_u^\top z_i^t)}$$

$$\gamma_{i,p} = \beta_{i,p}\odot \mathrm{sgn}(\alpha_{i,p})$$

where $\odot$ denotes the element-wise multiplication. For the internal node t, $A(t)$ represents the set of samples belonging to the node t. Λ denotes the indices of α with non-zero value, and $\bar{\Lambda}$ denotes the complement set of Λ. We employ stochastic gradient descent algorithm and mini-batch strategy for model optimization. After learning the dictionary and discriminative class models, we can directly use the dictionary to calculate the visual feature z, and use w for classification. As the max pooling is helpful to enhance the performance, we then test the dictionary with max spatial pooling [2].

The optimization proceeds in a top-down fashion. The procedure is shown in Algorithm 1. For an internode t in layer l, the training image set X_l is given [2], and the classifier matrix W_t is initialized to zeros. For dictionary D_t, $\bar{D}_t$ denotes the dictionaries inherited from ancestor nodes and the specific part $D_t^{'}$ is initialized by unsupervised dictionary $initD_l$. For each iteration, only the specific part $D_t^{'}$ is updated, and the W_t is the weight of the representation which is generated via the whole dictionary D_t. Due to the fact that inherited dictionary is not updated, the corresponding representations can be saved and directly used for classification in lower levels.

Algorithm 1: Multi-Level Discriminative Dictionary Learning

Input :
Data: $X, T^+, Y^+, initD$
Parameters: L, λ, $nBatch$, $nIter$, η_0, ρ
Output:
Dictionaries: D
for $l = 0$ **to** $L-2$ **do**
  foreach $t \in T_l^+$ **do**
    Initialize: $\{x,y\}_t \leftarrow \{x \in X_l,\ s.t.\ t \in y^+\}$, $W_t \leftarrow \emptyset,\ D_t^{'} \leftarrow initD_l,\ D_t^0 \leftarrow [\bar{D}_t, D_t^{'}]$
    for $k = 1$ **to** $nIter$ **do**
      1. **Randomly select** $A_k \subset \{x,y\}_t$, where $|A_k| = nBatch$
      2. **Local coding and average spatial pooling** Eq. (1)
      3. **Update**
      $\eta_k = \eta_0 \cdot \rho/(\rho + k)$
      $W_t \leftarrow \Pi(W_t - \eta_k(\nabla_W Loss + \lambda W))$
      $D_t^{'} \leftarrow \Pi(D_t^{'} - \eta_k \nabla_D Loss)$
    end
  end
end

4. Experiments

In this section, we evaluate our dictionary learning approach on two databases: SUN397 [27] and ImageNet subset [7]. The SUN database is a large scale database for scene recognition. We use all the 397 categories to evaluate our method. The data has been split in 10 partitions given by [27] and each one has 50 training images and 50 testing images per class. We train the model in all the partitions and obtain the result by averaging the performance. Besides that, we use the hierarchy provided by the SUN397 dataset as the category structure, which is a 4 level (include the root level) structure established based on semantic relation. Since some nodes in the hierarchy connect to more than one parent nodes, we change the original structure by choosing one parent node for them.

ImageNet [7] is a large scale dataset where the visual categories are organized based on the WordNet structure. We randomly select 185 categories which covers wide domains of semantics, such as *sports, animal, clothes and shops*. We generate a 4 level tree structure based on visual similarity [23]. For each category, we randomly choose 100 images for training and 100 for testing. We use 128-dimensional Dense-SIFT from patches with three scales 16, 32 and 48 as the input descriptors. Considering the time complexity, we choose the median dictionary sizes (256, 256, 512) for the specific part of the dictionaries in different layers in our method. Thus, the dictionary sizes in different layers of the hierarchy are 256, 512 and 1024 in our experiments. The dictionary size in other method is 1024. The regularization term λ is 1.0 in the experiments. In order to further evaluate our method, we evaluate our method on ImageNet1K used

[2]The samples consist of multiple scales if t is not the root node, details can be found in Section 3.3.

	Bi-ScSPM	H-ScSPM	Bi-TDDL	ML-DDL
Training	10.5	6.5	>700	29
Testing	5.5	5.5	>600	6.5
Training	8	6	>320	27
Testing	5	5	>270	5.5

Table 1. Time cost (in hours) in training and testing the model in SUN397 (the first two rows) and ImageNet185 (the last two rows).

in ILSVRC10, containing 1000 categories from the ImageNet. We select 100 images of each category for training, considering the time cost. A 5 level tree structure is generated based on semantic distance of the dataset. The learnt specific dictionary sizes in different layers are all set as 512 in our experiments.

For our method, it consists of three basis components: hierarchical structure, multinomial logistic classification, dictionary learning. In order to evaluate our method (ML-DDL), we compare it with the following state-of-the-art methods:

1. Binary SVM + ScSPM [28] (Bi-ScSPM). Based on the task-independent dictionary learnt by Sparse Coding, this method trains the linear SVM with one-vs-all strategy.

2. StructSVM + ScSPM (H-ScSPM). Based on the task-independent dictionary learnt by Sparse Coding, we use the similar empirical loss in [33] and train the SVM with the hierarchical structure by SVM-struct package [13].

3. Binary SVM + Task-driven dictionary learning (Bi-TDDL) [2]. The method trains the SVM with one-vs-all strategy, and learn a dictionary for each category.

4.1. Efficiency

The time cost of training and testing is critical for large scale problems. We compare the time complexity with the above-mentioned baseline models in following aspects: model training time and model testing time, as shown in Table 1. With respect to the training time, we accumulate the time of three steps: dictionary learning, mid-level feature computation and model learning. For the unsupervised dictionary learning methods (Bi-ScSPM and H-ScSPM), the time cost in dictionary learning is trivial compared with the other two steps. The mid-level features for each training sample only need to be computed once. The training time difference between Bi-ScSPM and H-ScSPM is brought by using different discriminative models. For the supervised methods, it is shown that our method is much faster than Bi-TDDL with the help of hierarchy. The number of dictionaries we need to learn is equal with the number of internodes and $L-1$ features are computed for each image. However, the number of dictionaries that Bi-TDDL needs to learn is equal to the number of categories (much larger than the number of internodes) as well as the number of features for each image.

	Bi-ScSPM	H-ScSPM	Bi-TDDL	ML-DDL
SUN397	23.4%	20.8%	24.0%	23.1%
ImageNet185	26.5%	25.1%	27.2%	28.6%

Table 2. Recognition rate compared among the four methods.

	Flat Error			Hierarchical Error		
Algorithm	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
Bi-ScSPM	0.862	0.826	0.803	9.978	8.660	7.784
ML-DDL	0.871	0.830	0.801	7.914	7.725	7.596

Table 3. Classification results (flat and hierarchical errors) of the methods in ImageNet1K.

Considering the testing time cost, only one feature is generated for each test sample in Bi-ScSPM and H-ScSPM. Thus, the testing time of Bi-ScSPM is almost equal with H-ScSPM. For the supervised dictionary learning, the time cost lies on the number of dictionaries. In our method, only $L-1$ features are computed for each test image. Moreover, the specific dictionaries in different layers are used to encode the single-scale descriptors, so the time cost does not increase obviously compared with the unsupervised methods. However, the descriptors need to be computed through all the dictionaries in Bi-TDDL method, so the time complexity increases drastically.

4.2. Performance of Multi-Level Categorization

Table 2 summarizes the performance achieved by different methods. ML-DDL obtains the best result in ImageNet185 which benefits from the favorable clustering of categories in hierarchy. Using the visual similarity for generating the hierarchy is more suitable to find the sharing structure for feature learning. For SUN397, the three methods except H-ScSPM have better results compared with the baseline of SIFT using k-means quantization and kernel (21.5% reported in [27]). However, our method does not achieve outstanding performance, probably because there exists gap from semantic relation to visual similarity despite of relatively consistency. Even so, our method also outperforms H-ScSPM with the help of hierarchical dictionary learning.

For experiments on ImageNet1K dataset, we adopt two types of error used in ILSVRC10 to measure the methods. Specifically, due to the limit in grouping of leaf nodes, some groups have few categories (at least three). So the tested algorithm will produce a list of three categories, and performance is measured using the top-n error rate, as shown in Table 3. We can clearly see the promising results of ML-DDL on hierarchical error, especially on the top-1 error rate, since our proposed method fully optimizes both the multi-level dictionaries and the discriminative models towards the

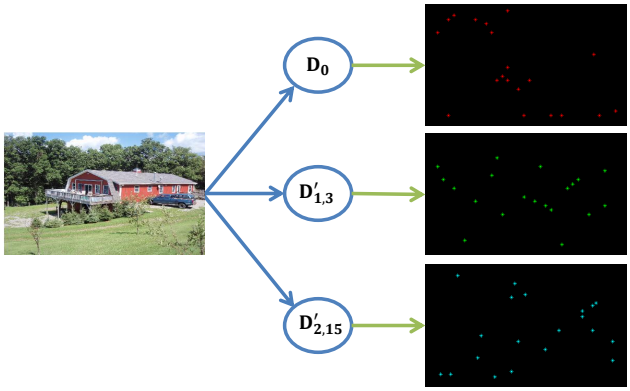

Figure 2. The property of the learnt dictionaries in different layers.

	SUN397			ImageNet185		
Algorithm	level 1	level 2	level 3	level 1	level 2	level 3
H-ScSPM	81.7%	48.9%	20.8%	81.0%	52.5%	25.1%
ML-DDL^0	83.4%	49.4%	21.2%	81.9%	53.2%	26.0%
ML-DDL	83.4%	51.0%	23.1%	81.9%	53.9%	28.6%

Table 4. Performance in different levels of hierarchical structure in SUN397 and ImageNet185.

semantic hierarchy. Our method also achieves good results on flat error.

4.3. Result Analysis and Discussion

To investigate the relation among the dictionaries in different layers, we use another strategy (named as ML-DDL^0) for dictionary learning in the hierarchical structure. In ML-DDL^0, the dictionaries in lower levels do not inherit the dictionary from ancestor nodes, in other words, the dictionaries in different layers only have the specific parts which are learnt by discrimination models. Compared with H-ScSPM, the performance in different levels of the hierarchy is shown in Table 4.

It is shown that the accuracy of the model decreases when going down the hierarchy, especially when reaching the leaf nodes. Besides misclassification in current layer, the errors which the samples misclassified in higher levels are propagated through the path. On the other hand, the nodes in lower levels are so visually similar that they are much harder to be distinguished compared with nodes in higher levels. However, the comparison between ML-DDL^0 and H-ScSPM shows that the problem could be relieved by the benefit from dictionary learning.

Furthermore, the effect of the dictionary inheritance has also been revealed by the performance difference between ML-DDL and ML-DDL^0. We can see that the accuracy has been improved in the lower layers, especially at the leaf nodes. This implies that the properties captured from the ancestor nodes are of great importance for child nodes. Different from the sharing model based on pre-computed features (sibling child nodes inherit the common information from ancestor nodes) [22], these properties can be selected and weighted via class models in child nodes, thus have more flexibilities. Due to dictionary inheritance in the hierarchy, the image representations in lower levels integrate all the useful information of multiple scales, which are beneficial to promote model capacity. Conversely, we test the Bi-ScSPM method by using the patches with multi-scale and single-scale as the input, and observe slight improvement (less than 0.5%). This observation is consistent with the analysis in [28]. Compared with directly pooling over multiple patch scales, our approach makes better use of the information of multiple scales.

Given a test image, its local descriptors are encoded based on the specific dictionary in each layer, and the reconstruction coefficients regarded as the response of the dictionary are pooled to represent the image. The dictionaries learnt in each layer consist of atoms which are biased to capture different information. Thus, the distribution of the response is divergent on different layers, as shown in Fig. 2. The points with different colors denote the locations having large values of the response using the learnt dictionaries in different levels. It shows that the dictionaries in different layers consist of specific atoms, thus the different part of visual information are highlighted at different layers.

Compared with binary dictionary learning with flat structure (Bi-TDDL), the recognition accuracy of a lot of categories has been improved by ML-DDL. Some categories have large improvement beneficial of the hierarchy. For example, in SUN397 dataset, {wine cellar / bottle storage +42.3%}, {iceberg +40.4%}, {kitchen +39.6%}, and in ImageNet {baboon +30.6%}, {butternut squash +28.2%}, {toy shop +22.1%}). Meanwhile, accuracy in some categories decreases, such as {sky -49.3%}, {stadium / baseball -47.9%}, {discotheque -46.5%} in SUN397, and {yurt -25.3%}, {freight car -24.1%}, {totem pole -20.9%} in ImageNet subset. These concepts have distinguished appearance with others, and grouping these concepts with other categories will lead to inappropriate model sharing structure. Besides, the misclassification in higher levels spreads through the path and finally incurs the misclassification of the child nodes. Therefore, the visual coherence of the hierarchy is directly related to the result, and finding a better tree structure is very important for visual classification.

5. Conclusion

In this paper, we present a hierarchical dictionary learning approach. The dictionaries in different layers are learnt to capture the discriminative information of different scales. Besides, the inheritance of dictionaries in the hierarchical structure enables the categories in lower levels to exploit the features of multiple scales. The experimental results on

two challenging databases show that our model efficiently deal with the classification task with large numbers of categories.

As our approach relies on a given category hierarchy, the information sharing mechanism among (hyper-)categories are predefined by some other processing such as [7]. Therefore, it may not be the optimal solution for effective information transfer. In future work, we would like to study how to automatically establish the hierarchical structure which best optimizes the multi-level classification performance.

Acknowledgements

This research is supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61070108, 61272326. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

[1] E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *CVPR*, 2008.

[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[3] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Incremental algorithms for hierarchical classification. *JMLR*, 7:31–54, 2006.

[4] M. Choi, A. Torralba, and A. Willsky. A tree-based context model for object recognition. *TPAMI*, 34(2):240–252, 2012.

[5] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *ICML*, 2004.

[6] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8] T. Deselaers and V. Ferrari. Visual and semantic similarity in ImageNet. In *CVPR*, 2011.

[9] L. Fei-Fei and P. Perona. A bayesian heirarchical model for learning natural scene categories. In *CVPR*, 2005.

[10] S. Fidler, M. Boben, and A. Leonardis. Similarity-based cross-layered hierarchical representation for object categorization. In *CVPR*, 2008.

[11] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *CVPR*, 2008.

[12] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, 2010.

[13] T. Joachims, T. Finley, and C. Yu. Cutting plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[15] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.

[16] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun. ACM*, 54:95–103, 2011.

[17] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *TPAMI*, 34(4):791–804, 2012.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *JMLR*, pages 19–60, 2010.

[19] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, 2007.

[20] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[21] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37:3311–3325, 1997.

[22] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011.

[23] L. Shen, S. Jiang, S. Wang, and Q. Huang. Learning-to-share based on finding groups for large scale image classification. In *PCM*, 2011.

[24] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *ICCV*, 2005.

[25] A. Torralba, K. Murphy, and W. Freeman. Sharing visual features for multiclass and multiview object detection. *TPAMI*, 29(5):854–869, 2007.

[26] J. Wang, J. Yang, K. Yu, and F. Lv. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[27] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[28] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[29] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, 2011.

[30] K. Yu, Y. Lin, and J. Lafferty. Learning image representations from the pixel level via hierarchical sparse coding. In *CVPR*, 2011.

[31] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.

[32] B. Zhao, L. Fei-Fei, and E. Xing. Large-scale category structure aware aware categorization. In *NIPS*. 2011.

[33] D. Zhou, L. Xiao, and M. Wu. Hierarchical classification via orthogonal transfer. In *ICML*, 2011.

[34] N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *CVPR*, 2012.

[35] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, 2007.