Computer Vision and Image Understanding 136 (2015) 23–31

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Learning representative and discriminative image representation by deep appearance and spatial coding

Bingyuan Liu, Jing Liu *, Hanqing Lu

National Laboratory of Pattern Recognition, CASIA, Beijing 100190, China

ARTICLE INFO

Article history:
Received 15 April 2014
Accepted 4 March 2015

Keywords:
Image classification
Deep learning
Structured sparsity

ABSTRACT

How to build a suitable image representation remains a critical problem in computer vision. Traditional Bag-of-Feature (BoF) based models build image representation by the pipeline of local feature extraction, feature coding and spatial pooling. However, three major shortcomings hinder the performance, *i.e.*, the limitation of hand-designed features, the discrimination loss in local appearance coding and the lack of spatial information. To overcome the above limitations, in this paper, we propose a generalized BoF-based framework, which is hierarchically learned by exploring recently developed deep learning methods. First, with raw images as input, we densely extract local patches and learn local features by stacked Independent Subspace Analysis network. The learned features are then transformed to appearance codes by sparse Restricted Boltzmann Machines. Second, we perform spatial max-pooling on a set of over-complete spatial regions, which is generated by covering various spatial distributions, to incorporate more flexible spatial information. Third, a structured sparse Auto-encoder is proposed to explore the region representations into the image-level signature. To learn the proposed hierarchy, we layerwise pre-train the network in unsupervised manner, followed by supervised fine-tuning with image labels. Extensive experiments on different benchmarks, *i.e.*, UIUC-Sports, Caltech-101, Caltech-256, Scene-15 and MIT Indoor-67, demonstrate the effectiveness of our proposed model.

1. Introduction

The task of recognizing semantic category of an image remains one of the most challenging problems in computer vision. How to build suitable image representations is the most critical. In previous decades, Bag-of-Feature (BoF) [8] based models have achieved impressive success for image representations. Usually, the models employ carefully hand-designed features, *e.g.* SIFT [38], HOG [9] and LBP [1], along with visual dictionary learning for local feature coding, and then obtain the image-level signatures by spatial concatenation of the local codes, and the powerful SVM are utilized to perform the classification task at last. However, three major problems hinder the performance of such a pipeline, *i.e.*, the limitation of hand-designed descriptors, the information loss brought by feature coding and the lack of spatial information. How to alleviate the above problems in the BoF-based framework to enhance the representative and discriminative abilities of image features is a challenging but vital task, and also becomes our focus in this paper.

There are several improvements proposed to address the above problems. Yang et al. [53] and Wang et al. [51] reduce the information loss by minimizing the reconstruction error along with effective priors such as sparse regularizer or locally-constrained linearity constraints. However, these methods are performed in a purely unsupervised way without any high-level guidance. Another inherent drawback is the lack of spatial information as the BoF-based representation describes an image as an orderless collection of local features. To incorporate the spatial information, one popular extension, *i.e.* Spatial Pyramid Matching (SPM) [29], is effective. It requires to partition each image into a fixed sequence of increasingly finer uniform grids ($1 \times 1, 2 \times 2, 4 \times 4$) and then concatenates the BoF features in each grids forming an image representation. Obviously, this simple partition and concatenation scheme can not reflect various spatial distributions in different categories of images. It is demonstrated recently that deep feature learning models, inspired by the hierarchical nature of human vision cortex, are effective to learn high-level image features [20]. The deep architectures are also effective to reduce the information loss by integrating unsupervised pre-training and supervised fine-tuning [17], and may generalize to different situations.

Motivated by the traditional image prior knowledge and recently developed deep feature learning, this paper proposes a

* Corresponding author. Fax: +86 10 82544594.
E-mail address: jliu@nlpr.ia.ac.cn (J. Liu).

http://dx.doi.org/10.1016/j.cviu.2015.03.006
1077-3142/

novel deep appearance and spatial coding architecture. The whole network is built based on Restricted Boltzmann Machines (RBM) and Auto-encoder (AE), which take advantage of unsupervised learning and supervised learning to explore the latent generative and discriminative properties. As shown in Fig. 1, it is a hierarchical architecture consisting of three modules: appearance coding, over-complete spatial max-pooling and spatial coding. With an image as input, our model first extracts dense local patches to learn local features by the stacked Independent Subspace Analysis (SISA) network, which is demonstrated effective to learn robust features [30]. Then the learned features are encoded into high-dimensional appearance codes by a sparse RBM (SRBM) layer. To incorporate more flexible spatial layout information, we adopt the ideas of over-completeness and structured sparsity. A over-complete spatial partition set including various spatial distributions is created in a flexible scheme and then max-pooling is carried out within each region. The resulting region features are concatenated as input to the next spatial coding module. In the spatial coding, we hypothesize that only a few dimensions of the mid-level region representations are effective. That is, partial spatial partitions are suitable to describe images. A structured sparse Auto-encoder (SSAE) approach is adopted to sparsely select the useful dimensions of the concatenated features as the image-level signature. An additional AE layer is further added to improve the performance. To learn the proposed deep model, we apply layer-by-layer unsupervised training and then fine-tune the parameters with image labels to enhance the discrimination. Finally, the output image representations are employed to train a one-versus-others SVM classifier to perform classification. We evaluate our model on widely used image benchmarks (*i.e.* UIUC-sports, Caltech-101, Caltech-256, Scene-15 and MIT Indoor-67). The extensive experiments demonstrate the effectiveness of our method in comparison with baselines and related work.

The rest of this paper is organized as follows. Section 2 reviews the related work of traditional image representation models and feature learning In Section 3, we elaborate our proposed model by introducing the three modules in details. The experimental evaluations and conclusions are given in Section 4 and Section 5 respectively.

2. Related work

Over the past few years, many researches have been conducted based on the BoF-based framework to address the existing problems and improve the performance.

To overcome the information loss in feature coding phase, some tried to learn discriminative visual codebooks [42,27]. Co-occurrence information of visual words was also considered in a generative framework [3]. In [16], the idea of visual word ambiguity was proposed to soft assign each local descriptor to multiple visual words in the learned codebook. As sparse coding was proven effective in feature representation, Yang et al. [53] utilized it to encode the local features into high-dimensional sparse codes. This method unitedly learned the codebook and searched the sparse weights for each local feature. Inspired by this, Wang et al. [51] further proposed to use locality constraint to guide the sparse coding learning process, and obtained better performance with less computation cost. Some other work [54]17 also tried to jointly learn the codebooks and appearance codes. Zhou et al. [61] learned a global Gaussian Mixture Model to randomly distribute each feature into one Gaussian component and then formed a supervector by the normalized means of the feature distribution. However, most methods were performed in a purely unsupervised way without any high-level guidance and independent of the local features extraction.

Many subsequent researches have been done to incorporate spatial information, as the traditional histogram-like representation discard the spatial relationship. One direction is to incorporate the local spatial layout in image, *i.e.* the relative or pairwise positions of local features. Savarese et al. [46] explored the combination of correlograms and visual words to represent spatially neighboring image regions. Liu et al. [36] proposed an efficient feature selection method based on boosting to mine high-order spatial features, while [40] proposed to jointly cluster feature space to build a compact local pairwise codebook and capture correlation between local descriptors. The spatial orders of local features were further considered in [41]. Since images often have spatial preferences, another direction is to incorporate global spatial layout property, *i.e.*, the absolute positions in image. Lazebnik et al. [29] pioneered this direction and proposed the SPM model. In SPM, the image was divided into uniform grids at different scales (*e.g.* $1\times1, 2\times2, 4\times4$), and the features are concatenated over all cells. It was also demonstrated that the combinations of SPM with sparse coding [53], locality-constrained coding [51], super vector [60] and fisher vector [43] models are very effective. However, the simple spatial partitions chosen in the ad-hoc manner without any optimization are too simple for complex nature situations. To solve this problem, Harada et al. [19] proposed to form the image feature as a weighted sum of semi-local features over all pyramid levels and the weights were automatically selected to maximize a discriminative power. To design better spatial partition, Sharma et al. [48] defined a space of grids, where each grid is obtained by a series of recursive axis aligned splits of cells and learned via a maximum margin formulation. Jia et al. [25]

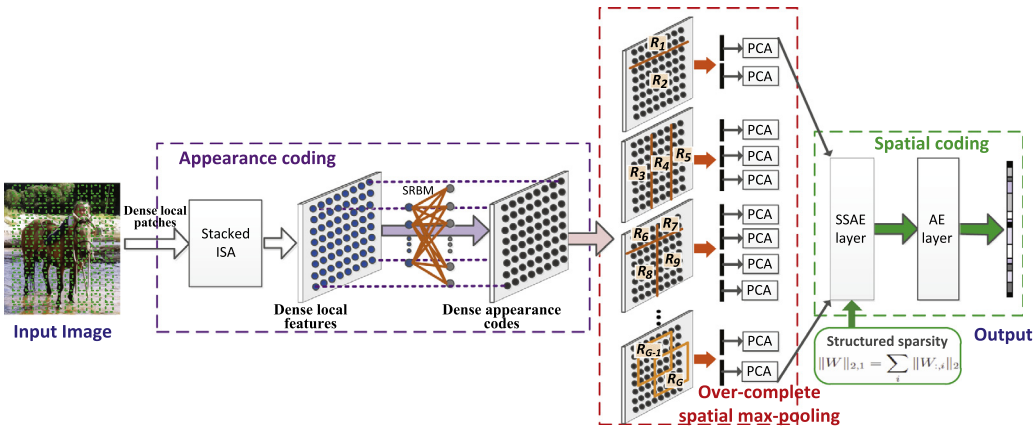

Fig. 1. Architecture of the proposed model.

formulated the problem in a multi-class fashion with structured sparse regularizer for feature selection.

It is recently considered that the performances of traditional methods are fundamentally limited by the hand-crafted local descriptors (*e.g.* SIFT, HOG and LBP), because these features are task-dependent and difficult to detect more complex structures beyond edges. A growing amount of researches have focused on automatically feature learning [28,31]. The models are usually built in a hierarchical framework by stacking shallow generative models with greedy layerwise scheme. One class of feature learning algorithms is based on the encoder-decoder architecture (*e.g.* Auto-encoder) [22]. The input is fed to the encoder which produces a feature vector and the decoder module then reconstructs the input from the feature vector with the reconstruction error measured. Deep Belief Networks (DBN) [21] built multiple layers of directed sigmoid belief nets with the top layer as a Restricted Boltzmann Machines. Lee et al. [31] extended DBN with convolution operation for the purpose of extracting latent features from raw image pixels. Yu et al. [55] proposed a hierarchical sparse coding model to learn image representations from local patches. Different from these models, we apply a stacked Independent Subspace Analysis [30] to learn features from raw image pixels, which is able to learn robust local features and performs well when combined with other feature learning modules. Benefit from the deep architecture, our deep appearance and spatial coding network further solves the information loss problem.

3. The proposed model

The proposed model for image representation is a hierarchical architecture, including three modules: appearance coding, over-complete spatial max-pooling and spatial coding. We will present the three modules as follows.

3.1. Appearance coding

Given an image, we densely extract local patches and learn local features via stacked convolutional ISA network and then encode them into high-dimensional vectors by a SRBM layer. In this subsection, we revisit the ISA and SRBM models respectively, and present how to utilize them to perform appearance coding.

3.1.1. Independent Subspace Analysis

Independent Subspace Analysis (ISA) [23] is an unsupervised learning algorithm for feature representation. In this paper, we explore it to learn features from image patches as an alternative to hand-designed descriptors. An ISA network is a two-layer structure with square and square-root nonlinear operations. The parameters of the network are the weights F and V in the two layers respectively. In particular, F is learned to represent the subspace structure of the units in the first layer, while V is fixed to pool over a small neighborhood of adjacent first layer units. With x^t as input, the output of each unit in the second layer is defined as:

$$p_i(x) = \sqrt{\sum_{k=1}^{m} V_{ik}\left(\sum_{j=1}^{n} f_{kj} x_j^t\right)^2} \tag{1}$$

Here, $F \in R^{m\times n}$ represents the weights between the input vector to the first layer, and $V \in R^{m\times k}$ is the weights of the pooling units in the second layer. Then the learning objective is defined as:

$$\min_F \sum_{t=1}^{N}\left(\sum_{i=1}^{m} p_i(x^t) + \mu\|FF^T x^t - x^t\|^2\right) \tag{2}$$

where N denotes the number of training instances. The first term in the objective encourages the sparsity of the learned features, while the second term ensures the important information is preserved. Besides sparsity, the learned features of ISA also have the property of invariance because of the pooling operation.

With raw image patches as input, we build a stacked ISA (SISA) to learn local patch features as shown in Fig. 2. We first train an ISA layer on small input patches and then let the learned network convolve with a larger image patch. The concatenated responses of the convolution are taken as input to learn the second ISA layer. PCA is employed as a prepossessing step to whiten the data and reduce the dimensions between the two stages.

3.1.2. Sparse Restricted Boltzmann Machine

The sparse Restricted Boltzmann Machine (SRBM) model is employed to obtain robust high-dimensional codes with the visible nodes corresponding to the dimensions of the learned local feature. It is a particular type of bi-partite undirected graphical with a two-layer structure, defining a joint probability distribution over a hidden layer $h \in \{0,1\}^{N_h}$ and a visible layer $p \in \{0,1\}^{N_p}$:

$$P(p,h) = \frac{1}{Z} exp(p^T U h + p^T b_p + h^T b_h) \tag{3}$$

where Z is the partition function, $U \in R^{N_p\times N_h}$ represents the undirected weights and $b_p \in R^{N_p}, b_h \in R^{N_h}$ are the bias terms. With an additional sparsity penalty incorporated, the overall cost function is:

$$E(p,h) = -logP(p,h) + \lambda\sum_{j=1}^{N_h}\left(\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}\right) \tag{4}$$

where ρ is a constant sparsity parameter, typically a small value close to zero and $\hat{\rho}_j$ is the average activation of hidden unit $j : \hat{\rho}_j = \frac{1}{N}\sum_{i=1}^{N} h_j(p^t)$. The second term in Eq. (4) is actually the Kullback–Leibler divergence between a Bernoulli random variable with mean ρ and a Bernoulli random variable with mean $\hat{\rho}$, encouraging the hidden units' activations to be sparse.

3.1.3. Appearance coding model learning

To learn the appearance coding model, we employ the greedily layerwise scheme. At first, we learn the SISA by optimizing Eq. (1) for each layer. As the gradient of this objective function is trackable, we apply the batch projected gradient descent algorithm while ensuring the orthogonal constraint by projection with symmetric pronominalization [23]. In particular, it is required to project F to the constraint set by computing $(FF^T)^{-\frac{1}{2}}F$ in the projected gradient descent. It is noted that the inverse square root of the matrix usually needs solving an eigenvector problem, which requires cubic time. The convolution and stacking ideas address this problem by slowly expanding the receptive fields via convolution. We also resort to PCA for whitening and reducing dimension to make the learning step much less expensive.

With the parameters of SISA layers obtained, we train the SRBM layer by minimizing Eq. (4). Since it is expensive to compute the gradient of the log-likelihood term, we adopt the contrastive divergence learning algorithm which gives an efficient approximation to the gradient of the log-likelihood [21]. Based on this algorithm, in each iteration we apply one step of contrastive divergence update rule, followed by another gradient descent step by the gradient of the regularization term. In our implementation the features of the two ISA stages are combined as input to the SRBM layer, which shows better performance.

3.2. Over-complete spatial max-pooling

For aggregating the local appearance features to image-level representation, the pooling operation is usually needed. We

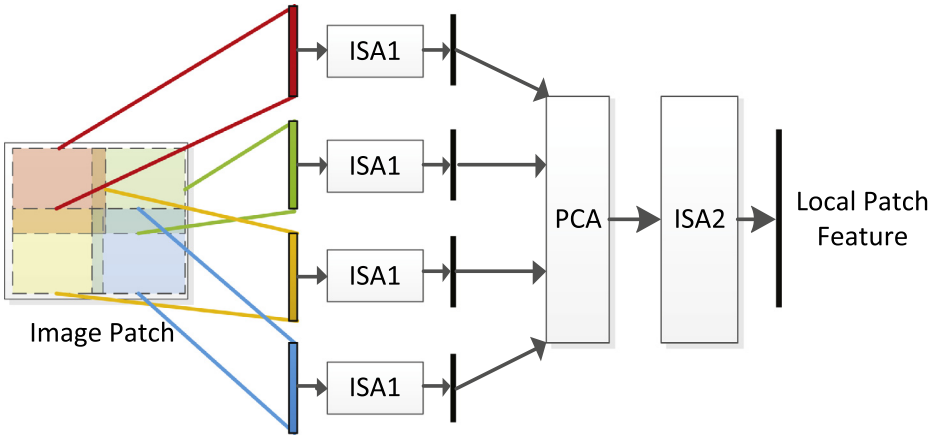

Fig. 2. Architecture of stacked convolutional ISA network for learning local features.

propose to build a set of randomly generated spatial regions to perform the pooling operation, for the purpose of incorporating more general and adaptive spatial information. As described in Fig. 3, we first apply uniform horizontal and vertical grids to divide the image into rectangular grids (the dotted grids). These grids are considered as candidate grids to generate a certain kind of spatial partition. Then a type of spatial partition is created by randomly choosing a subset of the candidate grids. By covering all the possible combinations of the grids, the spatial partition is able to present various spatial layout information. In addition, we generate some randomly sampled grids to provide more flexible spatial information. All the partitioned regions are collected as the over-complete spatial region set. Max-pooling operation is then performed on the local appearance codes within each partitioned region. We use r_i to denote the representation of the i-th region and G to denote the number of the regions. All the region vectors are concatenated as input to the next spatial coding layers. As the appearance codes and spatial regions are highly over-complete, we apply PCA on each region vector to reduce the dimension.

3.3. Spatial coding

With the region vectors $r = [r_1 r_2 \dots r_G]$ as input, spatial coding is performed to fuse them into global image representations. As this feature is very high-dimensional and redundant, we propose a structure sparse Auto-encoder (SSAE) to explore the semantically meaningful dimensions. Given r as input, our model transforms it into latent vector z via the encoder, which is defined as $z = f(Wr + b)$ (f is the sigmoid function). On the contrary, the decoder maps the latent representation back into the input space, producing a reconstruction $\tilde{r} = f(W'z + b')$.

It is noted that only a few dimensions among the region feature are effective, indicating structured semantic prior to compress the feature. Recent analysis and application of the mixed norm regularization [2,7,47] show that under certain conditions the coefficient vector W enjoys the structured sparse property, encouraging content-based structured feature selection in high-dimensional feature space. Therefore, we adopt the idea of structured sparsity and introduce the objective for the SSAE:

$$\min_{W,b,W',b'} \sum_{t=1}^{N} \|r^t - \tilde{r}^t\|^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|W'\|_2 \tag{5}$$

Here N denotes the number of training instances. This objective is to minimize the square reconstruction loss with two regularizers. As the decoder weights W' may not need any structural properties, we employ l_2 norm on it to prevent over-fitting. The regularizer on W is the $l_{2,1}$ norm:

$$\|W\|_{2,1} = \sum_i \|W_{:,i}\|_2 \tag{6}$$

where $W_{:,i}$ denotes the i-th column of W. This regularizer introduces structured sparsity by encouraging the encoder matrix W to be column-wise sparse, and lets the encoder select the effective dimensions and discard the redundant information. To optimize Eq. (5), we adopt the efficient algorithm proposed in [24,39] to solve the non-smooth penalty function. The dual of the proximal problem associated with the norm can be reformulated as a quadratic min-cost flow problem, which is able to be efficiently computed in polynomial time. In our model we further train a AE layer after the SSAE layer to enhance the performance.

3.4. Supervised fine-tuning of the network

After bottom-up layer-by-layer unsupervised learning, we fine-tune the network parameters of the appearance and spatial coding respectively through supervised learning.

To fine-tune the appearance coding, we associate each obtained appearance code to the image label and improve the discrimination of the latent feature with respect to the association. The appearance coding is essentially a three-layer neural network, where the set of learned parameters (F, V, U, b_p, b_h) map the input to the latent feature h. For the image dataset with L categories, we add an auxiliary classifier layer with the parameters of weights

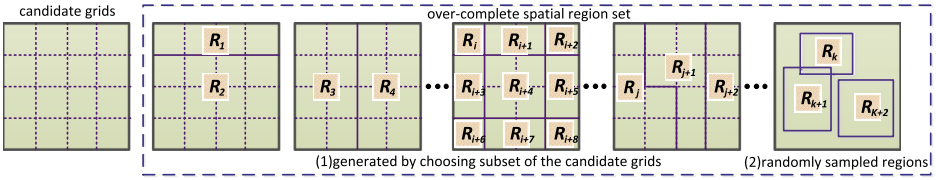

Fig. 3. The generation of over-complete spatial region set.

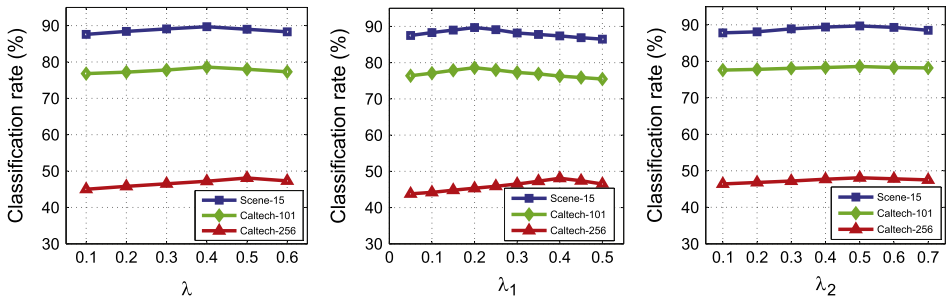

Fig. 4. Performance changes with different λ (left), λ_1 (middle) and λ_2 (right).

Table 1
Performance (%) comparison of different feature coding methods on three datasets.

Algorithms	UIUC-Sports	Scene-15	Caltech-101
SIFT+Hard [29]	80.9 ± 0.8	81.1 ± 0.3	64.6 ± 0.8
SIFT+SC [53]	82.7 ± 1.5	80.3 ± 0.9	73.2 ± 0.5
SIFT+LLC [51]	82.1 ± 0.60	80.9 ± 0.5	73.4
HOG+SC	80.5 ± 0.5	77.9 ± 0.4	71.5 ± 0.5
SIFT+SRBM	83.1 ± 0.3	82.1 ± 0.2	73.1 ± 0.3
SISA+Hard	82.8 ± 0.1	81.7 ± 0.4	68.2 ± 0.2
SISA+SC	84.2 ± 0.3	83.7 ± 0.4	72.2 ± 0.2
SISA+SAE	83.9 ± 0.3	84.1 ± 0.2	72.9 ± 0.5
SISA+SRBM	84.5 ± 0.2	84.8 ± 0.1	74.3 ± 0.4
Supervised SISA+SRBM	85.8 ± 0.5	86.1 ± 0.5	76.2 ± 0.5

$M \in R^{N_h \times N_L}$ and biases $d \in R^{N_L}$, transforming $h \in R^{N_h}$ to the one-hot coded class labels $y \in R^{N_L}$ via feed-forward soft-max activations:

$$\hat{y}_l = \frac{e^{\sum_{i=1}^{N_h} m_{il} h_i + d_l}}{\sum_{j=1}^{N_L} e^{\sum_{i=1}^{N_h} m_{ij} h_i + d_j}} \tag{7}$$

In the fine-tuning phase, we initialize the parameters of appearance coding by values learned in the unsupervised phase and introduce supervision with the error back propagation algorithm. Soft-max loss is employed to measure the error E between the hypothesized class $\hat{y}$ and the ground truth y:

$$E = -\sum_{l=1}^{N_L} 1\{y = i\} \log \hat{y}_l \tag{8}$$

Then the network is updated by the gradient descent with respect to the parameters across the layers. After fine-tuning, the auxiliary classifier layer is discarded since they are no longer needed in the test process.

After fine-tuning the appearance coding layers, we fine-tune the spatial coding layers in the same scheme by associating the output to the image label and learning an additional soft-max layer. The parameters (W, b) of the two spatial coding layers are also updated by the back propagated gradient from the soft-max loss. with the trained hierarchy, image representations are obtained by direct feed-forward matrix operation. To carry out the recognition task, we turn to train a one-versus-others SVM classifier for each class.

4. Experiments

In the experiments, we widely evaluate the proposed architecture on five public image benchmarks: UIUC-Sports, Caltech-101, Caltech-256, Scene-15 and MIT Indoor-67. Since the proposed framework is complex and hybrid, we first evaluate the effect of each part in the network with an in-depth analysis. Then we report the classification performance on the five datasets compared with the famous hand-designed features(*i.e.* SIFT [38] and HOG [9]), some related feature coding methods (*i.e.* KSPM [29], ScSPM [53], LLCSPM [51], SSRBM [17]), some work considering spatial information [19,48,25] and other state-of-the-arts results.

4.1. Evaluations of each part in the model and the parameters

The proposed model includes three parts: appearance coding, over-complete spatial max-pooling and spatial coding. For the SISA in the appearance coding, the inputs to the two layers are of size 16×16 and 20×20 respectively, where local patches are densely sampled with spacing of 4 pixels. According to our

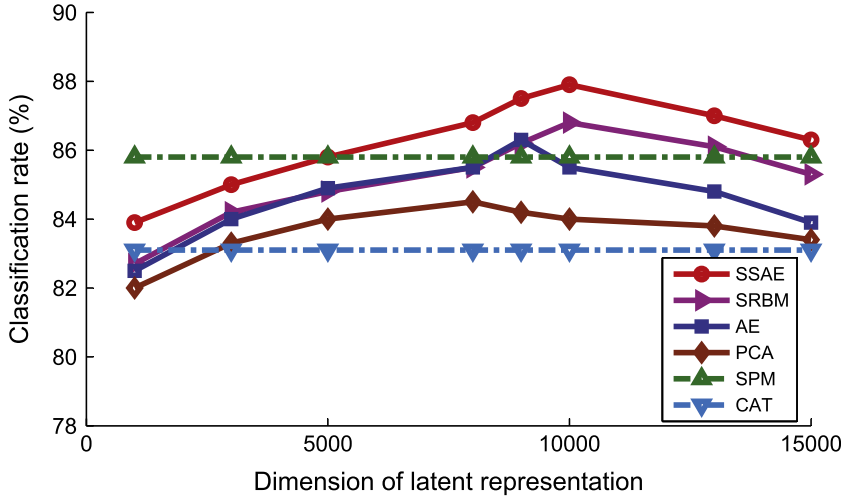

Fig. 5. Performance comparisons of different spatial coding method on UIUC-Sports.

Table 2
Classification rate (%) comparison on UIUC-Sports.

Algorithms	Classification rate
Li et al. [34]	77.9
Dixit et al. [10]	84.4
Liu et al. [37]	84.6 ± 1.5
Gao et al. [15] (LScSPM)	85.7 ± 1.3
Perronnin et al. [43] (FV)	88.6 ± 1.2
Our (unsupervised)	88.7 ± 0.4
Our (supervised)	89.8 ± 0.2

implementation, the number of output units is set to 120 in the first layer and 200 in the second. The μ in Eq. (2) is fixed to 10. The dimensions of hidden units in the SRBM is set to 1024 as a fair comparison [53] and the target sparsity is empirically set to 10%. In the over-complete spatial max-pooling section, the over-complete spatial region set contains 50 kinds of regions and the dimension of each region vector is reduced to 300 by PCA. To train the final SVM classifiers, we employ the one-versus-others linear SVM provided by LIBLINEAR toolbox [11] for its advantages in speed and good performance. We report the results by repeating the experimental process 5 times with different randomly selected training and testing images.

In our method, the most important parameters are λ in Eq. (4) and λ_1, λ_2 in Eq. (5). Fig. 4 shows the performance changes with varied λ, λ_1 and λ_2 on the three datasets. It is shown that the performances are not sensitive to the value of λ_2 and we may fix it to 0.5 on all the conditions. In the experiments, we find that the performances are good when small $\lambda(0.4)$ and $\lambda_1(0.2)$ are used for UIUC-Sports, Scene-15, Caltech-101 datasets, and relatively large $\lambda(0.5)$ and $\lambda_1(0.4)$ are used for Caltech-256 and MIT Indoor-67.

We evaluate the effect of our appearance coding by analyzing the relative contributions of the ISA and SRBM, as displayed in Table 1. We compare our method to some related work with only difference in the local feature extraction and coding scheme. The final image representations are all complied by SPM [29]. 'SISA+SRBM' denotes our appearance model, and the best performance is achieved by 'Supervised SISA+SRBM', which is the result after supervised fine-tuning. This indicates that the fine-tuning process is effective improve the discrimination. It is shown that our method outperforms the hand-designed features, as well as traditional feature coding method, *i.e.*, hard assignment (Hard) [29], sparse coding (SC) [53] and locality-constrained linear coding (LLC) [51]. Note that 'SISA+SC' represents the method of applying sparse coding on the SISA features and the performance is similar to our model. We also compare with the method sparse Auto-encoder (SAE), and the performances of SAE and SRBM are similar as they basically do similar feature transformation. The 'SISA+SRBM' consistently outperforms 'SISA+SAE' by about 1 percentage in the three datasets.

Fig. 5 experimentally demonstrates the effect of the spatial coding model. In this figure, the local feature extraction method is fixed as 'supervised SISA+SRBM', and we compare the proposed SSAE with AE, SRBM, PCA, SPM and the method of directly concatenating all the region vectors (CAT). The performance of 'CAT' is weak because of the high redundancy. By applying the dimension reduction techniques, the classification rate is obviously improved as some redundant information is discarded. SSAE beats all the others as the incorporated structured sparse prior may be more semantically reasonable. It is also shown that the performances changes by varying the number of latent units. The best results are achieved with the number of hidden units set to about 10,000, which is more compact than SPM.

Compared to traditional framework like [53,51], the proposed model needs more time to train because of the high dimension. However, the test process is much more efficient as only forward matrix multiplication operation is needed. On the desktop with a 8-core 3.40 GHz CPU, it quires about 8 h to train the entire model. In the test process, only 0.9 s per image is required to obtain the final representation, while the time in ScSPM is about 4 s.

4.2. Results on UIUC-Sports

The UIUC-Sports dataset is collected by Li and Li [33], containing 1792 images of eight sport categories. The number of images in each class ranges from 137 to 250. Following the standard setup [53], we randomly select 70 from each class for training and test on the rest. Table 2 gives our results compared with related work on this dataset. The best result is achieved by 'Our(supervised)', where the contribution of supervised fine-tuning phase is about 1 percentage. Liu et al. [37] propose the soft assignment coding method and Gao et al. [15] (LScSPM) use laplacian sparse coding to encode the local features. Our model outperforms the two methods by 5.2 and 4.1 percentage respectively. Perronnin et al. [43] employs fisher vectors to encode the local features. Compared with this method, the accuracy is similar but our obtained representation is much more compact, which speeds up the classification.

4.3. Results on Caltech-101 and Caltech-256

The Caltech-101[12] dataset contains 9144 images totally from 102 different categories, including 101 object categories and 1 additional background category, with high shape variability. As an extension, the Caltech-256 [18] dataset totally holds 29,780 images in 256 object categories and is much more challenging as it possesses much higher intra-class variability and object location variability. In the both datasets, the number of images in each class varies from 31 to 800. In the experiments, we follow the common setup on the two dataets. For Caltech-101, we train on 15 and 30 images per category and testing on the rest, while for Caltech-256 we evaluate our model on 15, 30, 45 and 60 training images per class respectively.

Table 3 gives the detailed performance comparisons on the Caltech-101 dataset. Macrofeatures [4] is a type of mid-level representation based on SIFT. HSC [55] and DN [57] are two models employing deep architecture to learn image features. Our model outperforms these methods because of the incorporated various spatial information. Boureau et al. [5] improves SPM by restricting pooling to codes that are nearby in descriptor space. Our model behaves slightly better than it, even a denser local feature sampling scheme is adopted in [5]. It is noted that our best result also outperforms Jia et al. [25], which also utilizes the idea of over-complete spatial partition but they perform feature selection in a multi-class discriminative learning fashion. To the best of our

Table 3
Classification rate (%) comparison on Caltech-101.

Algorithms	15 training	30 training
KSPM [29]	56.40	64.6 ± 0.8
Macrofeatures [4]	–	75.7 ± 1.1
ScSPM [53]	67.0 ± 0.45	73.2 ± 0.5
HSC [55]	–	74.0
DN [57]	–	71.1 ± 1.0
Jia et al. [25]	–	75.3 ± 0.70
Boureau et al. [5]	–	77.1 ± 0.70
Chatfield et al. [6] (IFV)	–	77.8 ± 0.6
Our (unsupervised)	67.9 ± 0.30	76.9 ± 0.4
Our (supervised)	70.4 ± 0.2	78.6 ± 0.2
Todorovic et al. [50]	73.0	83.0
CNN-SVM [56] (no extra data)	22.8 ± 1.5	46.5 ± 1.7
CNN-SVM [56] (with extra ImageNet)	83.8 ± 0.5	86.5 ± 0.5

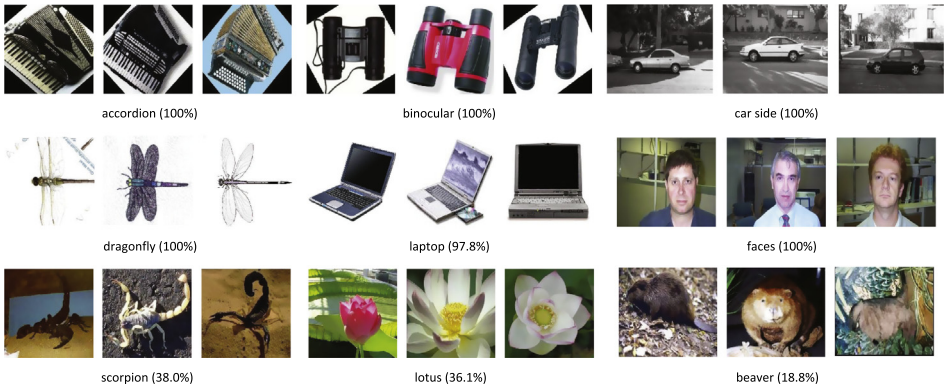

Fig. 6. Examples images from classes with highest and lowest classification accuracy from the Calteche-101 dataset.

Table 4
Classification rate (%) comparison on Caltech-256.

Algorithms	15 train	30 train	45 train	60 train
KSPM [29]	28.3	34.1	–	–
ScSPM [53]	27.7 ± 0.5	34.0 ± 0.4	37.5 ± 0.6	40.1 ± 0.9
LLCSPM [51]	34.4	41.2	45.3	47.7
GLP [14]	35.8	43.2	47.3	–
Our (Unsupervised)	36.6 ± 0.5	46.9 ± 0.3	47.9 ± 0.1	50.8 ± 0.4
Our (Supervised)	37.4 ± 0.6	48.1 ± 0.2	49.2 ± 0.2	52.0 ± 0.4
CNN [56](no extra data)	9.0 ± 1.4	22.5 ± 0.7	31.2 ± 0.5	38.8 ± 1.4
CNN [56](with ImageNet)	65.7 ± 0.2	70.6 ± 0.2	72.7 ± 0.4	74.2 ± 0.3

knowledge, the performance of the proposed model outperforms all published results for a single descriptor type and fair experimental setups, although some better results have been reported. Todorovic et al. employ the segmentation tree for subcategory learning, while CNN-SVM et al. [56] train a large Convolutional Neural Network (CNN) using a large scaled auxiliary ImageNet dataset. In Fig. 6, we show some example classes with the highest and lowest classification accuracy from the Caltech-101 dataset with 30 training images per class. Our model performs well on categories with little clutter (*e.g.* laptop and faces) or strong spatial layout priors (*e.g.* dragon fly and car side), but less successful on classes with large intra-class variation and highly diversity (*e.g.* beaver and lotus).

The performance comparison results on the Caltech-256 dataset are shown in Table 4. On the much more challenging dataset, our model also consistently leads the performance on all the conditions and outperforms the baseline ScSPM by more than 10%. GLP [14] is a method of using discriminatively learned pooling operation to aggregate local features and our model also behaves better than it as we explore more kinds of spatial distributions and integrate the regions beyond the simple concatenation scheme. Zeiler and Fergus [56] achieves the accuracy of 74.2% with 60 training images per class, as they pre-train a very large scaled CNN with extra ImageNet dataset.

Table 5
Classification rate (%) comparison on Scene-15.

Algorithms	Classification rate
Lazebnik et al. [29] (SPM)	81.1 ± 0.3
Zhou et al. [61]	84.1 ± 0.5
Boureau et al. [4]	85.6 ± 0.3
Zhou et al. [59]	85.2
Goh et al. [17] (ssRBM)	86.0 ± 0.5
Feng et al. [14] (GLP)	83.2
Sharma et al. [48] (DS)	80.10 ± 0.6
Sharma et al. [49] (DSS)	84.6 ± 0.7
Our (Unsupervised)	88.5 ± 0.3
Our (Supervised)	89.7 ± 0.2

4.4. Results on Scene-15

We further evaluate the effect of our model on a standard scene classification benchmark, *i.e.* Scene-15 dataset. This dataset is compiled by several researchers [13,29], including 15 different classes of scene categories (*e.g.* kitchen, coast, highway) with each class containing 200 to 400 images. Following the common setup, 100 images per class are randomly selected for training with the rest for testing. Table 5 gives the detailed comparison results. Our model also achieves better than related work on this scene benchmark, demonstrating the generalization to handle various kinds of images. Zhou et al. [61] and Goh et al. [17] (ssRBM) are two methods for improving the feature coding. In [61] a Gaussian Mixture Model is trained to form a supervector by the normalized means of the feature distribution, while [17] uses a sparse and selective regularized RBM. The advantage of our model is that we employ high-level information to guide the feature coding. Sharma et al. [48] (DS) and [49] (DSS) are two recent methods considering better spatial information coding. The reason that our model behaves better may be the employment of deep architecture to select the most effective dimensions from the over-complete spatial regions through unsupervised and supervised learning.

Fig. 7 gives the confusion table between the 15 scene categories. It is seen that our method performs well in most categories. Confusions mainly occur between the outdoor building classes (*e.g.* industrial and tallbuilding), some natural scenes (*e.g.* open-country and coast) and also some indoor classes (*e.g.* living room and bedroom), as these categories are too similar to distinguish.

4.5. Results on MIT Indoor-67

The MIT Indoor-67 dataset [44] is the currently largest indoor scene recognition dataset, consisting of 15,620 images in 67 categories. The similarity of the objects present in different indoor scenes makes the dataset especially difficult. Following the protocol of [44], we use the same training and test split where each category has about 80 training images and 20 test images.

Table 6 gives the detailed performance comparisons. Our method consistently outperforms related methods like SPM [51], and achieves the best performance with single representation on the difficult dataset. This again proves the effectiveness of the proposed method. Xie et al. [52] (OPM) is an improvement of SPM, which uses the 3D orientations to form the pyramid and produce the pooling regions. Lin et al. [35] (ISPR) is another work

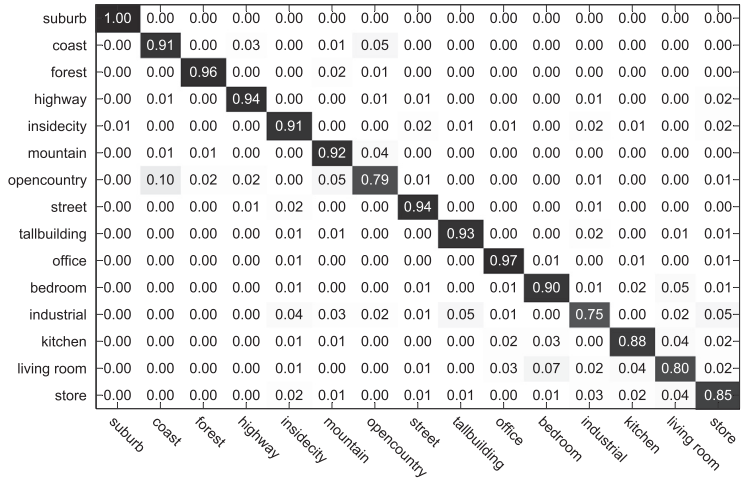

Fig. 7. Confusion table for the Scene-15 dataset. Average classification rates for individual classes are listed along the diagonal. The entry in the *i*-th row and *j*-th column is the percentage of images from class *i* that are misidentified as class *j*.

Table 6
Classification rate (%) comparison on MIT Indoor-67.

Algorithms	Accuracy
Quattoni et al. [44]	26
li et al. [32]	37.6
Juneja et al. [26] (BoP)	46.1
Wang et al. [51] (SPM)	54.62
Xie et al. [52] (OPM)	51.45
Lin et al. [35] (ISPR)	50.1
IFV+SPM [52]	61.2
Our (Unsupervised)	60.3
Our (Supervised)	62.9
Hybrid-Parts+GIST-color+SPM [58]	47.2
ISPR+IFV+SPM [35]	68.5
CNNaug-SVM [45]	69.0

considering more spatial information by jointly learning the appearance and important spatial pooling regions. It is noted that our performance beats the two methods by more than 10 percentage. On this dataset, more state-of-the-art results are obtained by combining multiple image representations and part detector model, like 'ISPR+IFV+SPM' [35]. In 'CNNaug-SVM' [45], they use extra large scaled ImageNet dataset to train a large CNN network.

5. Conclusion

In this paper, we introduce a novel image representation, which is a hierarchical architecture including appearance coding, over-complete spatial max-pooling and spatial coding. Specifically, we apply stacked Independently Subspace Analysis combined with sparse Restricted Boltzmann Machines to learn local appearance codes from raw pixels. To incorporate more flexible spatial layout information, we create an over-complete spatial partition set and perform max-pooling within each region. Then we proposed a structured sparse Auto-encoder to encode the mid-level region representations into the global image signature. The training of the hierarchy consists of unsupervised layer-by-layer pre-training and parameters fine-tuning with image labels. The experimental results on several public databases outperform baselines and other related work, demonstrating the effectiveness of the proposed method. In the future work, we will further extend the network to deal with more large-scaled dataset, *i.e.* ImageNet.

Acknowledgments

This work was supported by 973 Program (2012CB316304) and National Natural Science Foundation of China (61272329, 61472422, 61332016 and 61273034).

References

[1] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 2037–2041.
[2] S. Bengio, F. Pereira, Y. Singer, D. Strelow, Group sparse coding, in: NIPS, 2009, pp. 676–684.
[3] O. Boiman, E. Shechtman, M. Irani, In defense of nearest-neighbor based image classification, in: CVPR, 2008, pp. 1–8.
[4] Y.L. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, in: CVPR, 2010, pp. 2559–2566.
[5] Y.L. Boureau, N.L. Roux, F. Bach, J. Ponce, Y. LeCun, Ask the locals: multi-way local pooling for image recognition, in: ICCV 2011, IEEE, 2011, pp. 2651–2658.
[6] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: BMVC, 2011.
[7] X. Chen, X.T. Yuan, Q. Chen, S. Yan, T.S. Chua, Multi-label visual classification with label exclusive context, in: ICCV, 2011, pp. 834–841.
[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: ECCV 2004 Workshop on statistical learning in computer vision, 2004.
[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, 2005, pp. 886–893.
[10] M. Dixit, N. Rasiwasia, N. Vasconcelos, Adapted gaussian models for image classification, in: CVPR, 2011, pp. 937–943.
[11] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[12] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, Comput. Vis. Image Underst. 106 (2007) 59–70.
[13] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: CVPR, 2011, pp. 524–531.
[14] J. Feng, B. Ni, Q. Tian, S. Yan, Geometric lp-norm feature pooling for image classification, in: CVPR, 2011, pp. 2697–2704.
[15] S. Gao, I.W.H. Tsang, L.T. Chia, P. Zhao, Local features are not lonely – laplacian sparse coding for image classification, in: CVPR, 2010, pp. 3555–3561.
[16] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, J.M. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1271–1283.
[17] H. Goh, N. Thome, M. Cord, J. Lim, Unsupervised and supervised visual codes with restricted boltzmann machines, in: ECCV, 2012, pp. 298–311.
[18] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset, Technical Report 7694, California Institute of Technology, 2007.
[19] T. Harada, Y. Ushiku, Y. Yamashita, Y. Kuniyoshi, Discriminative spatial pyramid, in: CVPR, 2011, pp. 1617–1624.
[20] G. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[21] G.E. Hinton, S. Osindero, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[22] F.J. Huang, Y. lan Boureau, Y. Lecun, Unsupervised learning of invariant feature hierarchies with applications to object recognition, in: CVPR, 2007, pp. 1–8.
[23] A. Hyvrinen, J. Hurri, P.O. Hoyer, Natural Image Statistics, Springer, 2009.

[24] R. Jenatton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for sparse hierarchical dictionary learning, in: ICML, 2010, pp. 2297–2334.
[25] Y. Jia, C. Huang, T. Darrell, Beyond spatial pyramids: receptive field learning for pooled image features, in: CVPR, 2012, pp. 3370–3377.
[26] M. Juneja, A. Vedaldi, C.V. Jawahar, A. Zisserman, Blocks that shout: distinctive parts for scene classification, in: CVPR, 2013, pp. 923–930.
[27] F. Jurie, B. Triggs, Creating efficient codebooks for visual recognition, in: ICCV, 2005, pp. 604–610.
[28] K. Kavukcuoglu, P. Sermanet, Y.L. Boureau, K. Gregor, M. Mathieu, Y. LeCun, Learning convolutional feature hierarchies for visual recognition, in: NIPS, 2010, pp. 1090–1098.
[29] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: CVPR, 2006, pp. 2169–2178.
[30] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: CVPR, 2011, pp. 3361–3368.
[31] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: ICML, 2009, pp. 609–616.
[32] L. Li, H. Su, E. Xing, L. Fei-Fei, Object bank: a high-level image representation for scene classification and semantic feature sparsification, NIPS 2010 (2010) 1378–1386.
[33] L.J. Li, F.F. Li, What, where and who? classifying events by scene and object recognition, in: ICCV, 2007, pp. 1–8.
[34] L.J. Li, H. Su, Y. Lim, F.F. Li, Objects as attributes for scene classification, in: ECCV 2010 Workshops on Parts and Attributes, 2010, pp. 57–69.
[35] D. Lin, C. Lu, R. Liao, J. Jia, Learning important spatial pooling regions for scene classification, in: CVPR, 2014, pp. 3726–3733.
[36] D. Liu, G. Hua, P. Viola, T. Chen, Integrated feature selection and higher-order spatial feature extraction for object categorization, in: CVPR, 2008.
[37] L. Liu, L. Wang, X. Liu, In defense of soft-assignment coding, in: ICCV, 2011, pp. 2486–2493.
[38] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[39] J. Mairal, R. Jenatton, G. Obozinski, F. Bach, Network flow algorithms for structured sparsity, in: NIPS, 2010, pp. 1558–1566.
[40] N. Morioka, S. Satoh, Building compact local pairwise codebook with joint feature space clustering, in: ECCV, 2010, pp. 692–705.
[41] N. Morioka, S. Satoh, Learning directional local pairwise bases with sparse coding, in: BMVC, 2010.
[42] F. Perronnin, C. Dance, G. Csurka, M. Bressan, Adapted vocabularies for generic visual categorization, in: ECCV, 2006, pp. 464–475.
[43] F. Perronnin, C.R. Dance, Fisher kernels on visual vocabularies for image categorization, in: CVPR, 2007.
[44] A. Quattoni, A. Torralba, Recognizing indoor scenes, in: CVPR, 2006, pp. 413–420.
[45] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, in: CVPR 2014 Workshops, 2014.
[46] S. Savarese, J. Winn, A. Criminisi, Discriminative object class models of appearance and shape by correlations, in: CVPR, 2006, pp. 2033–2040.
[47] M. Schmidt, K. Murphy, G. Fung, R. Rosales, Structure learning in random fields for heart motion abnormality detection, in: CVPR, 2008, pp. 1–8.
[48] G. Sharma, F. Jurie, Learning discriminative spatial representation for image classification, in: BMVC, 2011, pp. 6.1–6.11.
[49] G. Sharma, F. Jurie, C. Schmid, Discriminative spatial saliency for image classification, in: CVPR, 2012, pp. 3506–3513.
[50] S. Todorovic, N. Ahuja, Learning subcategory relevances for category recognition, in: CVPR, 2008.
[51] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: CVPR, 2010, pp. 3360–3367.
[52] L. Xie, J. Wang, B. Guo, B. Zhang, Q. Tian, Orientational pyramid matching for recognizing indoor scenes, in: CVPR, 2014, pp. 3734–3741.
[53] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: CVPR, 2009, pp. 1794–1801.
[54] J. Yang, K. Yu, T.S. Huang, Supervised translation-invariant sparse coding., in: CVPR, 2010, pp. 3517–3524.
[55] K. Yu, Y. Lin, J. Lafferty, Learning image representations from the pixel level via hierarchical sparse coding, in: CVPR, 2011, pp. 1713–1720.
[56] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: ECCV, 2014, pp. 818–833.
[57] M.D. Zeiler, G.W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: ICCV, 2011, pp. 2018–2025.
[58] Y. Zheng, Y.G. Jiang, X. Xue, Learning hybrid part filters for scene recognition, in: ECCV, 2012, pp. 172–185.
[59] X. Zhou, N. Cui, Z. Li, F. Liang, T. Huang, Hierarchical gaussianization for image classification, in: ICCV, 2009, pp. 1971–1977.
[60] X. Zhou, K. Yu, T. Zhang, T.S. Huang, Image classification using super-vector coding of local image descriptors, in: ECCV, 2010, pp. 141–154.
[61] X. Zhou, X. Zhuang, H. Tang, M. Hasegawa-Johnson, T.S. Huang, A novel gaussianized vector representation for natural scene categorization, in: ICPR, 2008, pp. 1–4.