CSPS: An Adaptive Pooling Method for Image Classification

Jinzhuo Wang, Wenmin Wang, Ronggang Wang, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*—This paper proposes an adaptive approach to learn class-specific pooling shapes (CSPS) for image classification. Prevalent methods for spatial pooling are often conducted on predefined grids of images, which is an ad-hoc method and, thus, lacks generalization power across different categories. In contrast, our CSPS is designed in a data-driven fashion by generating plenty of candidates and selecting the optimal subset for each class. Specifically, we establish an overcomplete spatial shape set that preserves as many geometric patterns as possible. Then, the class-specific subset is selected by training a linear classifier with structured sparsity constraints and color distribution cues. To address the high computational cost and the risk of overfitting due to the overcomplete scheme, the image representations for CSPS are first compressed according to dictionary sensitivity and shape importance. These representations are finally fed to SVMs for the classification task. We demonstrate that CSPS can learn compact yet discriminative geometric information for different classes that carries more semantic meaning than other methods. Experimental results on four datasets demonstrate the benefits of the proposed method compared with other pooling schemes and illustrate its effectiveness on both object and scene images.**

***Index Terms*—Class-specific pooling shapes (CSPS), dictionary sensitivity, image classification, multi-shape matching kernel, representation compression.**

I. Introduction

IMAGE classification has been a challenging task in multimedia analysis field for decades. Most successful approaches for this task are based on the bag-of-features (BoF) model [1], which has shown promising results on many popular data sets. The standard BoF model starts by identifying local image patches, using either normalized raw pixel values or handcrafted features such as SIFT [2] or HoG [3]. Such low-level features are then encoded into an over-complete representation using various algorithms such as K-means or sparse coding [4]. Finally, the histogram of the summed feature codes for the entire image is regarded as a signature for classification. However, spatial layout information is completely neglected in such methods because each feature code in an image is treated equally. To overcome this drawback, [5] pioneered the direction of exploiting spatial layout properties and proposed spatial pyramid (SP) to embed spatial information of local features. In practice, this method first partitions an image into a fixed sequence of increasingly finer uniform grids (such as 1×1, 2×2, 4×4) and then concatenates the BoF representation in each grid with a certain pooling scheme to achieve the final representation. The idea of spatial pooling dates back to an analysis of complex cells in the mammalian visual cortex [6], which identifies mid-level image features that are invariant to small spatial shifting. The spatial invariance property also reflects the concept of locally order-less images [7], which suggests that low-level features can be grouped spatially to provide information about the overall semantics. Most recent research on spatial pooling aims to find a good pooling operator, which can be considered as a function that produces informative statistics based on local features in a specific spatial area. For example, average and max pooling strategies have been found in various algorithms, respectively, and systematic comparisons between such pooling strategies have been presented and discussed in [8], [9]. The SP-based representation guides most approaches of image classification and benefits many state-of-the-art systems [10]–[12].

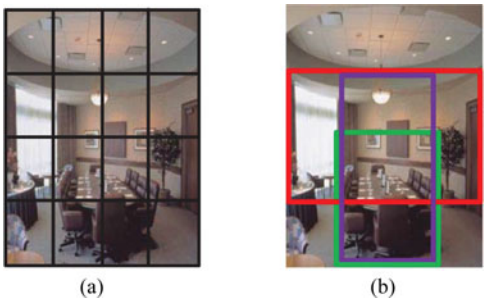


Fig. 1. Comparison of the standard SP and the CSPS learned by our method. (a) Three-level SP. (b) CSPS.

However, one obvious limitation in standard SP is the uniform feature pooling style, which pre-defines the grids for an image and uses all the spatial shapes equally. We argue that this uniform scheme lacks the capability to capture adaptive spatial information. For instance, an image belonging to "meeting-room" class in MIT-67 dataset [13] is coped with a three-level standard SP and the proposed class-specific pooling shape (CSPS) method shown in Fig. 1(a) and (b), respectively. It is obvious that the spatial pooling shapes learned by CSPS separate the target and background properly, providing more reasonable and semantical spatial information. In such cases which are common in natural images, the hand-crafted and uniform pooling shapes in

Manuscript received November 6, 2015; revised January 15, 2016; accepted March 14, 2016. Date of publication March 18, 2016; date of current version May 13, 2016. This work was supported by the ShenZhen Peacock Plan. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Balakrishnan Prabhakaran. *(Corresponding author: Wenmin Wang.)*

J. Wang is with the School of Electronics Engineering and Computer Science, Peking University Shenzhen Graduate School, Shenzhen 518055, China (e-mail: cr7or9@163.com).

W. Wang is with the Department of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China (e-mail: wangwm@ece.pku.edu.cn).

R. Wang is with the School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China (e-mail: rgwang@pkusz.edu.cn).

W. Gao is with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: wgao@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2544099

standard SP can not focus effectively on the regions of interest and thus lack generalization power across different classes on large datasets.

In this paper, we present a data-driven approach that adaptively learns CSPS to overcome the aforementioned limitation. Our idea is motivated by the observation that images in the same class often share common spatial layout properties, i.e., the background and target tend to follow similar spatial distributions. In practice, we first adopt the concept of over-complete set and establish a set of spatial shapes containing as many shapes as possible. This scheme helps us collect more flexible patterns of spatial distribution. Instead of using pre-defined spatial shapes as in standard SP, we train a linear classifier with structured sparsity constraint and color distribution cues to select the optimal subset for each class. In particular, the sparsity term encourages the classifier to extract a small but essential subset, thus avoiding redundancy. The color term makes the selected shapes more semantically reasonable by following color distribution inference. We expect to learn certain pooling shapes for each class that capture the most semantically useful geometric properties. On the other hand, although the over-complete process helps preserve more candidates, it also suffers from increased computational cost and the risks of over-fitting. To solve this issue, we compress the representations over the CSPS according to the shape importance and dictionary sensitivity. These are finally fed to an SVM with a multi-shape matching kernel to perform the image classification task. The main contributions of this paper can be summarized as follows.

1) We address the limitation of standard SP pooling and propose a novel framework that uses CSPS for image classification. CSPS can naturally generate more geometric patterns and learn adaptive yet semantically important spatial information for each class through sparsity constraints and color distribution cues.
2) We introduce two methods to compress the image representations over CSPS that are based on dictionary sensitivity and shape importance. The latter can be further used in SVMs with a proposed multi-shape matching kernel that preserves the spatial impact of different spatial shapes.
3) We validate the effectiveness of the proposed approach through extensive experiments. The results demonstrate the benefits of CSPS compared with standard spatial pooling techniques and other state-of-the-art methods on both object and scene images.

The remainder of the paper is organized as follows. Section II reviews related works concerning spatial information exploration for image classification. Section III presents our image classification framework including the approach for adaptively learning CSPS, representation compression and the multi-shape matching kernel. Extensive experiments on four data sets along with analysis and comparison are provided in Section IV. We conclude the paper in Section V.

II. Related Work

The studies on image classification in the literature can be divided into two categories: one that relies on the BoF framework and hand-crafted descriptors to achieve image-level representations while the another that uses deep neural networks (DNNs) to learn powerful features from raw pixels. In this section, we briefly review typical DNN models, particular the recent successful convolutional neural networks (CNNs), and then focus on the BoF framework and spatial pooling techniques.

DNN is a family of learning models that can offer good representations of data using a structure stacked into multiple layers, in which each layer represents a different degree of abstraction of data features. One typical model is CNNs, which saw heavy use in the 1990s in the literature such as LeNet [14], but then fell out of fashion with the rise of SVMs. In 2012, AlexNet [15] rekindled the interest of the computer vision community in CNNs, by showing substantially higher image classification accuracy on the imagenet large scale visual recognition challenge [16]. This success resulted from training a large CNN on 1.2 million labeled images, together with a few twists on LeNet such as rectified linear units and "dropout" regularization. Since then, many attempts have been made to design more complex models including GoogLeNet [17], ZF-net [18], VGG-net [19] and so on. These models have proven powerful but require a great deal of training data. However, recent work [20] shows that BoF can yield competitive or superior performance on smaller data sets than DNN models. Thus, the purpose of this paper is to further improve the BoF framework with a focus on spatial pooling.

The BoF model originated from text processing [21] and was initially introduced to image analysis in [22]. Specifically, each image in this model is represented as an order-less histogram of local features or their codewords (these codewords are often acquired from a carefully constructed dictionary using some unsupervised clustering techniques such as k-means clustering [23] and sparse coding [10]). However, in this model, spatial layout information is completely discarded. To overcome this limitation, various extensions have been proposed from two directions: the properties of local spatial layout and global spatial layout. Here, we review related work from these two perspectives.

Local spatial layout information mainly explores the relative positions or pairwise positions of local features. Researchers in [24] used the combination of correlograms and visual words to jointly incorporate appearance and shape information. Compact spatial modeling without loss of discrimination is achieved through the adaptive vector quantized correlograms, which the authors call "correlatons." In [25], an efficient feature selection method based on boosting was introduced to mine high-order spatial features. Selected lower order features are employed to avoid exhaustive computation. Qi *et al.* [26] believe that one single model has difficulties in representing various spatial context in different images and proposed constructing a prototype set of kernelized spatial-context for image classification. Relative features in [27] were learned on a reference basis and the authors proposed an adaptive pooling technique to assemble the learned multiple relative features. They also achieved good performance on image classification tasks. Recently, [28] presented a novel localized visual feature named PixNet that embeds relative spatial information by learning different image parts while preserving a compact representation. Moreover, [29] presented a novel feature selection method based on a class-specific codebook, which was shown to learn distinctive local features.

On the other hand, global spatial layout information leverages the absolute positions in images, an idea that is identical to our focus in this paper. Based on the pioneering work [5] where the original SP was proposed, [10] and [30] showed that incorporating advanced feature coding strategies can improve the classification performance. Moreover, combinations with super vector [31] and fisher vector [32] have been demonstrated to be effective in obtaining a good image representation. More recently, several advanced image classification systems have been based on SP, but involving different parameters including the number of pyramid levels and the structure of the grids. For instance, [5] and [10] uses up to four pyramid levels with uniform grids of 1 × 1, 2 × 2, 4 × 4 and 8 × 8, whereas the winner of the Pascal VOC 2007 competition [33] (and many others such as [31]) used three pyramid levels with grids of 1 × 1, 2 × 2, and 3 × 1. However, these SP parameters are still chosen in an ad-hoc manner and few works report systematic construction of the representation.

Although the extensions of standard SP have been largely explored, less attention has been paid to pursuing adaptive spatial pooling shapes in a learning procedure. In this paper, we address this issue using a data-driven approach to adaptively learn the optimal pooling shapes for each class of images. The work most related to ours is [11] which also adopted the idea of using an over-complete set and formulated the problem in a multi-class fashion to learn discriminative spatial shapes for the whole data set. However, because different categories often have different distribution of spatial properties, we attempt to assemble local features through adaptive pooling shapes rather than fixed spatial patterns, and learn the corresponding CSPS. Moreover, we also try to leverage color distribution information which is often used for region of interest (ROI) detection and segmentation in our learning procedure to select more semantically reasonable spatial information by following color cues.

III. Approach

In this section, we present our image classification framework with a focus on learning CSPS and discuss the process from establishing an over-complete spatial shape set to learning CSPS with sparsity and color constraints. Robust yet compact image representations over CSPS are then compressed with dictionary sensitivity and shape importance. We finally feed the representation over CSPS fed to SVMs using a proposed multi-shape matching kernel for classification.

A. Over-Complete Spatial Shape Set

We first establish an over-complete spatial shape set which preserves candidates with more spatial distribution patterns. Instead of only using certain uniform squares in standard SP as in Fig. 2(a), we use all the rectangular shapes, including as many geometric properties of the local features as possible. Specifically, let a and b represent the number of horizontal and vertical lines to separate an image, we collect a total of $\mathcal{R} = \binom{a+1}{2} \times \binom{b+1}{2}$ rectangles as in Fig. 2(b) and the over-complete spatial shape set is denoted by $\mathcal{S} = \{s_1, s_2, \ldots, s_{\mathcal{R}}\}$.

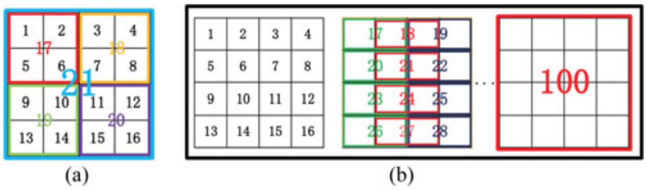


Fig. 2. Toy example of pooling shapes generated by standard SP 2 (a) and the proposed CSPS 2 (b) on a 4 × 4 grid. Standard SP yields $1 \times 1 + 2 \times 2 + 4 \times 4 = 21$ rectangle grids, whereas ours can produce $\binom{4+1}{2} \times \binom{4+1}{2} = 100$ candidates.

Note that the over-complete scheme makes it possible to obtain flexible shapes such as circles and polygons, which can capture more adaptive and semantically meaningful geometric properties for particular visual recognition tasks. For simplicity of comparison with the standard SP scheme, we apply only the increasing horizontal and vertical lines to form rectangular shapes in our implementation.

B. CSPS Learning (CSPSL)

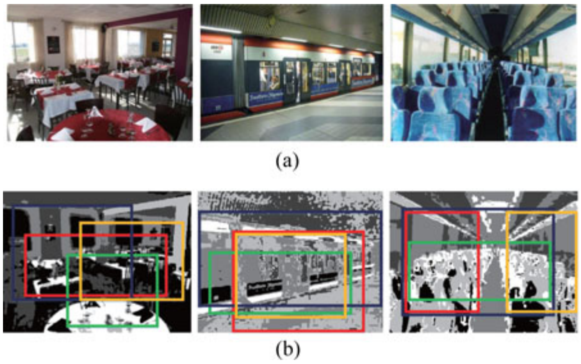
Because $\mathcal{S}$ is over-complete with much redundancy, we attempt to select the optimal subset for each class because images in the same class often share common spatial layout distributions. In other words, we want to achieve $\mathrm{S_L} = \{\mathcal{S}^1, \mathcal{S}^2, \ldots, \mathcal{S}^t\}$ for the image set containing t classes where $\mathcal{S}^i$ denotes a certain subset of $\mathcal{S}$ for class i.

To this end, given a set of images $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$ we first the extract local features and employ a feature coding algorithm to obtain a dictionary $\mathcal{D} = \{d_1, d_2, \ldots, d_t\}$. These two steps are addressed a lot in the literature and we use different strategies according to different datasets in practice. Next, spatial pooling of the feature codes is conducted on each shape of $\mathcal{S}$. In this manner, the ith image can be represented by concatenating the pooled feature codes as $\mathbf{x}_i = \{\mathbf{x}_i^{s_1}, \mathbf{x}_i^{s_2}, \ldots, \mathbf{x}_i^{s_{\mathcal{R}}}\}$ and the image set can be represented as $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Then, a linear classifier is trained using the one-versus-all scheme to select the optimal subset for each class, leading to the following optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b, y_n) + \lambda \mathrm{Reg}(\mathbf{w}) \tag{1}$$

where vector $\mathbf{w}$ and scalar b are the parameters to be estimated; $\mathbf{x}_n$ is the feature vector of the nth sample; $y_n \in \{-1, +1\}$ is the label of the nth sample, indicating class i and the "rest-of-the-world;" $\mathcal{L}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b, y_n)$ is a certain non-negative convex loss function to punish a certain set of $\{\mathbf{w}, b\}$, $\mathrm{Reg}(\mathbf{w})$ is a regularizer term and $\lambda \in \mathbb{R}$ is the regularization coefficient. In practice, we choose the binomial negative log likelihood as the loss function, as shown below

$$\mathcal{L}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b, y_n) = \ln(1 + \exp(-y_n(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n + b)). \tag{2}$$


(a)

(b)

Fig. 3. Examples of raw images and the corresponding learned CSPS with color distribution cues in L*a*b space on MIT-67 dataset. (a) Raw images. (b) Learned shapes with color segmentation.

The regularization term $\mathrm{Reg}(\mathbf{w})$ in (1) selects the subset for each class containing the most representative and discriminative spatial shapes. Thus we employ two regularization terms; $\mathrm{Reg}(\mathbf{w})$ can then be reformulated as

$$\mathrm{Reg}(\mathbf{w}) = \mathrm{Reg}_s(\mathbf{w}) + \mathrm{Reg}_c(\mathbf{w}) \tag{3}$$

where $\mathrm{Reg}_s(\mathbf{w})$ and $\mathrm{Reg}_c(\mathbf{w})$ are the sparsity constraint term and color distribution constraint term, respectively. These two regularization terms are described in the following.

Color distribution cues. To leverage color distribution information as learning cues, we apply color segmentation to assign a certain color channel to each pixel in advance. In practice, we simply employ the K-means algorithm to cluster an image with k colors by converting the RGB color space to the L*a*b space, denoted by $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$. Because the learned shapes capture dominant channels in color space based on the observation that target objects tend to follow different color distribution from the background, we define the color regularization term in (3) as

$$\mathrm{Reg}_c(\mathbf{w}) = \sum_{i=1}^{\mathcal{R}} \max_j \left\{ \left(\frac{\mathrm{N}(c,i)}{\mathrm{P}(j)} \right)^{\frac{\mathrm{N}(c,i)}{\mathrm{N}(c)}} \right\}, j = \{1, 2, \ldots, k\} \tag{4}$$

where $N(c_i, j)$ denotes the number of the ith color in s_j, $N(c_i)$ is the number of pixels of the ith color and $P(j)$ is the number of pixels in the s_j region. The base term indicates the proportion of a color in each shape and the exponent term stands for the proportion of that color in a specific channel. Using this regularization term, the classifier tends to select semantically reasonable shapes by following color distribution inferences. Some examples of learned shapes are shown in Fig. 3.

Structured sparsity regularization. Although significant efforts have been conducted on the design of sparse regularization terms such as the squared Frobenius norm and $\ell_{1,\infty}$ norm [11], recent analysis [34] shows that mixed-norm regularization enjoys the group sparsity propert under certain conditions, encouraging content-based structured feature selection in high-dimensional feature space. Following the instructions in [34], we adopt the idea of structured sparsity and define the sparse regularization term as a ℓ_2/ℓ_1 norm regularizer

Algorithm 1: Class-Specific Pooling Shape Learning

```
Input: $n$ images $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ belonging to $t$
       classes
Output: CSPS $\mathrm{S_L} = \{\mathcal{S}^1, \mathcal{S}^2, \ldots, \mathcal{S}^t\}$
for $i = 1, \ldots, t$ do
    Set the label of class-$i$ images $+1$ and the others $-1$
    Set the feature set $\mathcal{F}$ empty
    for $ii = 1, \ldots, \eta$ do
        while $\{\mathbf{w}, b\}_{\mathrm{ii}}$ not stable do
            Select a subset $\widehat{\mathcal{F}}$ with the largest score
             using (6)
            $\mathcal{F} \leftarrow \mathcal{F} \bigcup \widehat{\mathcal{F}}$
            Solve Eq. (1) using $\widehat{\mathcal{F}}$
        $\{\mathbf{w}, b\}_{\mathrm{ii}}^{\mathrm{new}} \leftarrow \mathrm{argmax}\, \mathcal{L}(\{\mathbf{w}, b\}_{\mathrm{ii}}^{\mathrm{old}})$ using (2)
        end while
    end for
    $\mathcal{S}^i \leftarrow \{\mathbf{w}_1, \ldots, \mathbf{w}_\eta\}$
end for
return $\mathrm{S_L} = \{\mathcal{S}^1, \mathcal{S}^2, \ldots, \mathcal{S}^t\}$
```

$$\mathrm{Reg}_s(\mathbf{w}) = \|\mathbf{w}\|_{2,1} = \sum_{i=1}^{\mathcal{R}} \|\mathbf{w}_{s_i}\|_2 \tag{5}$$

where $\mathbf{w}_i$ is the ith group of parameters corresponding to s_i. This regularizer motivates dimensions in the same group to be jointly zero. Thus the optimization procedure tends to select a much smaller but more discriminative subset. Beyond the usual ℓ_1 norm regularization, sparsity is now imposed on spatial shape level rather than merely at the feature level. To solve the joint learning with mixed norm regularization problem, we employ the primal-dual algorithm proposed in [35].

Fast learning. Although the over-complete scheme provides flexible spatial shapes with more geometric patterns, optimizing (1) is still a computationally challenging task despite its convexity, because it has a high dimensional search space. In practice, we adopt the greedy approach proposed in [11] by starting with an empty set of selected features and incrementally adding features to the set. Specifically in each iteration, for the feature i that has not been selected, we compute the score of the ℓ_2 norm of the gradient of (2) as follows:

$$\mathrm{score}(i) = \left\| \frac{\partial \mathcal{L}(\mathbf{w}, b)}{\partial \mathbf{w}_i} \right\|_{\mathbf{Fro}}^2. \tag{6}$$

We then select the feature with the largest score and add it to the feature set. The selection procedure can be controlled by a threshold to limit the size of feature set. In practice, we follow the suggestion of [11] and set the active set size to 100 (this choice is discussed in Section IV-C). Algorithm 1 summarizes the overall procedure of our CSPSL method.

C. Representation Compression

Another essential issue of concern is the high dimensional representations over CSPS, which are even larger than those in standard SP due to our over-complete grids. We address this problem by finding a more compact representation that maintains or even improves their original counterparts. To this end, several works address the problem by searching compact vocabulary construction [36]–[38]. In particular, [37] and [39] provide guidelines for the compression of vocabularies using agglomerative information bottleneck (AIB) theory. In the following, we briefly review the AIB compression algorithm and present our corresponding compression method.

AIB compression. AIB characterizes the discriminative power of the dictionary $\mathcal{D}$ as mutual information $I(d, c)$ from the visual word d and the category c as

$$I(d, c) = \sum_{d \in \mathcal{D}} \sum_{c=1}^{\mathcal{C}} P(d, c) \log \frac{P(d, c)}{P(d)P(c)} \tag{7}$$

where the joint probability $P(d, c)$ is estimated from training images by simply counting the number of occurrences of each visual word $d \in \mathcal{D}$ in each category $c \in \{1, \ldots, \mathcal{C}\}$. AIB iteratively compresses the dictionary $\mathcal{D}$ by merging the two visual words d_i and d_j that cause the smallest decrease D_{ij} in the mutual information $I(d, c)$, an operation can be regarded as the discriminative power. Let $[x]_{\text{ij}}$ denote two visual words corresponding to the dictionary after merge; then, D_{ij} is

$$D_{\text{ij}} = I(d, c) - I([x]_{\text{ij}}, c). \tag{8}$$

The information $I(d, c)$ is monotonically reduced after each merge. Merging is often conducted iteratively until the desired number of visual words are obtained.

Dictionary sensitivity. The use of non-optimized dictionaries for building CSPS results in its huge dimensionality. Our aim is to optimize the dictionaries in a manner that maintains or improves the original CSPS performance. To this end, we investigated the use of the AIB algorithm as proposed in [31]. Specifically, we propose to learn the most compact vocabulary $\overline{\mathcal{D}}$ that best suits each CSPS by eliminating occurrences of the least informative features from each specific shape in our CSPS. For instance, for an $a \times b$ over-complete shape set, we first eliminate the smallest shapes but those bigger than 1×1 (i.e., 1×2) the least informative visual word from all its $\mathcal{R}$ spatial occurrences. Subsequently, we eliminate them from 1×3, 1×4 and 2×2, etc. To this end, the probability of a spatial visual word $P(d_i)$ at shape i is then computed as follows:

$$P(d_i) = \begin{cases} \text{occ}(d_i|c), & \text{if } i = 1 \\ \sum\limits_j P(d_j), & \text{else} \end{cases} \tag{9}$$

where j indicates a specific region inside of shape i, and $\text{occ}(d_i|c)$ is the number of occurrences of each spatial visual word $d_i \in \mathcal{D}$ in each category c. Finally, we use the information content criteria to measure the discriminative power of the spatial vocabulary $\overline{\mathcal{D}}$ and converge at a certain threshold

Fig. 4. Samples from PASCAL VOC 2007 dataset.

$$I(\overline{\mathcal{D}}, \mathcal{C}) = \sum_i \sum_t p(d_i, c_t) \log \frac{p(d_i, c_t)}{p(d_i)p(c_t).} \tag{10}$$

Shape importance. Except the dictionary compression, we expect the most valuable shapes to be preserved. To measure the importance of each shape, we apply a leave-one-out paradigm; then the importance value of a particular shape j is defined as the increase of training error after neglecting the shape dimensions

$$I_j = \frac{\text{Error}_j - \text{Error}_0}{\text{Error}_0} \tag{11}$$

where Error_0 denotes the training error over all the training data. The largest I_j indicates that the neglected dimensions of the jth shape are more important and discriminative. In practice, we select the top $\eta\, I_j$ for each class. Beyond the target of representation compression, I_j can also be used in the subsequent multi-shape matching kernel as a weight term.

D. Multi-Shape Matching Kernel

At this point, we have learned the top η CSPS for each class. We then employ SVMs for classification. Notice that standard SP treats each pooling shape equally in the matching kernel, neglecting the differences between spatial shapes. We attempt to weight the shapes because important regions should be given more attention. Specifically, we use the shape importance value I in (11) as a weight to indicate the importance of different shapes and define the multi-shape matching kernel as the weighted sum of the separate shape kernels

$$\mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{m=1}^{\eta} I_m \cdot K(\mathbf{x}_1^{s_m}, \mathbf{x}_2^{s_m}) \tag{12}$$

where the kernel K can be any kernel function under Mercer's theorem. With this multi-shape matching kernel, a one-versus-others classifier is prepared for the classification task.

IV. Experiments

In this section, we report the experimental results on both object and scene data sets for the image classification tasks. Our experiments are conducted using a single core of an eight-core Intel Core i7-3770K CPU running at 3.50 GHz with 16.0 GB RAM. We evaluate the performance of our CSPS scheme mainly by comparing with traditional SP-based schemes and other competitive approaches that exploit geometric information to obtain image-level representations. Because most methods rely on different processing procedures using low-level descriptors according to different data sets, we describe our implementation details and report the quantitative results separately.

TABLE I
CATEGORY-WISE ACCURACY (%) WITH DIFFERENT POOLING STRATEGIES ON PASCAL VOC 2007 DATASET

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	AP
MP [30]	68.2	57.7	39.9	61.6	24.0	57.4	73.4	53.5	49.7	36.9	42.3	39.6	73.4	62.2	79.4	23.8	42.7	48.4	68.0	47.7	52.5
MWLP [40]	68.4	56.9	41.1	62.9	23.8	58.8	73.9	53.4	50.1	37.2	41.7	40.4	74.3	62.1	79.5	24.1	42.4	49.3	68.8	48.8	52.8
OCP [41]	**74.2**	63.1	45.1	65.9	29.5	**64.7**	**79.2**	61.4	51.0	**45.0**	54.8	45.4	76.3	67.1	84.4	21.8	44.3	**48.8**	70.7	51.7	57.2
SP [42]	70.5	58.6	42.9	61.6	28.3	59.4	74.8	54.8	51.4	39.4	44.3	41.1	74.9	65.5	81.8	27.6	43.9	48.9	69.9	49.8	54.5
CSPS	73.1	**66.7**	**49.4**	**72.2**	**36.5**	62.3	75.2	**66.4**	**53.0**	41.0	**55.8**	**48.0**	**76.8**	**67.4**	**86.2**	**29.5**	**46.2**	44.3	**73.1**	**52.0**	**57.8**

TABLE II
AVERAGE PRECISION (%) WITH DIFFERENT SETTINGS OF λ AND k

k \ λ	0.2	0.4	0.6	0.8	1.0	1.2
2	42.9	46.3	45.2	47.5	48.2	46.8
4	44.5	47.1	45.5	50.4	49.1	49.5
6	45.6	45.0	49.7	51.9	49.3	49.9
8	48.4	49.6	50.8	55.2	53.5	49.3
10	51.1	50.2	52.1	55.4	54.4	52.1
12	53.5	52.1	52.3	**57.2**	55.2	53.3
14	53.7	50.5	54.8	57.0	55.1	54.4

Fig. 5. Some samples (CD, duck, and iPod from top to bottom) from Caltech-256 dataset.

A. *Object Classification*

PASCAL VOC 2007 [43] is a challenging dataset composed of images from 20 object categories gathered from Flickr and characterized by a high variability of viewing angle, object size, illumination, pose and appearance. We use the standard protocol which consists of training and validating on the "train" and "val" sets, respectively, and then testing on the test set. The classification performance is evaluated using the average precision (AP) measure, which is the standard metric used by the PASCAL challenge [43]. Some sample images are shown in Fig. 4.

We report the benefits of our strategy in Table I by comparing CSPS with related adaptive pooling methods. For a fair comparison, we run the code for low-level feature extraction and feature coding provided by [30]. In particular, we set the over-complete shape sets to 5×5 and choose the corresponding parameters in a cross-validate fashion. Note that [30] is a standard max pooling strategy that uses pre-defined grids as discussed in Section III, whereas [42] learns a supervised pooling method by separating the image and feature domains. A more recent from [40] uses multiple methods for local pooling. In addition, [41] leverages a separate representation of background and foreground by first inferring the object locations.

There are two important parameters in our method. One is λ as shown in (1), which controlls the regularization term for CSPSL, and the other is k from (4), which affects the color distribution cues in (1). We show the results in Table II as a function of the choice of λ and k. In most situations, for a given number of λ, the classification ability improves as k increases. On the other hand, when k is fixed, the optimal choice is often at $\lambda = 0.8$. Overall, the best performance is obtained when $k = 12$ and $\lambda = 0.8$.

As shown in Table I, our proposed CSPS outperforms the other approaches in most cases. An interesting contribution is proposed in [41] where the authors report a very similar performance (57.2%) by adopting a pooling strategy that assumes *a priori* knowledge of the location of the object of interest. However, in our method, we do not rely on such a priori knowledge because the learning procedure in our proposed scheme selects several essential candidates that naturally capture semantically meaningful shapes, and these shapes lead us to focus on the salient regions of images. Finally, note that the results reported in [41] (59.3%) were higher than the ones we obtained running their code. This might be due to different low-level features and possibly to post-processing of the resulting image features because we conduct the stages of feature extraction and feature coding following the instructions of [30].

Caltech-256 [44] is another popular object data set that contains 29 780 images in 256 object categories. The number of images in each category varies from 31 to 800. Some samples of this data set are shown in Fig. 5. Following the common experimental setup for this data set, we validate our method on $\{5, 10, \ldots, 60\}$ training images for each class, respectively, and test on the rest. To cross-validate the parameters, we use half the training data for training, the other half for validation and then we retrain with the optimal parameters on the full training data. We repeat each experiment ten times. We measure the classification accuracy for each class and report the average as well as the standard deviation. In Fig. 6, we compare a popular SP baseline [10] with our CSPS (using only dense SIFT descriptors) as a function of the number of training samples. We can observe that CSPS consistently outperforms the standard SP. This is a favour of the advantage of our adaptive pooling method.

We also report our results in Table III and compare them with the results from several state-of-the-art approaches. We consider both the case where we use only SIFT descriptors and the case where we use SIFT descriptors along with local color statistic descriptors (which prove to be effective on this dataset [20]) with a simple weighted linear combination. We

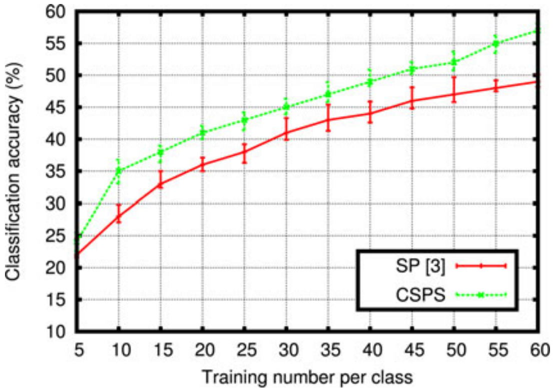


Fig. 6. Comparison of SP and CSPS on Caltech-256 dataset (both using only use SIFT descriptors). We report the mean and three times the average deviation.

TABLE III
PERFORMANCE (%) COMPARISON OF TWO IMPLEMENTATIONS OF CSPS WITH OTHER APPROACHES ON CALTECH-256 DATASET

Method	15tr.	30tr.	45tr.	60tr.
Griffin *et al.* [44]	-	34.1 (0.2)	-	-
Boiman *et al.* [46]	-	42.7 (-)	-	-
Bo and Sminchisescu [49]	23.2 (0.6)	30.5 (0.4)	34.4 (0.4)	37.6 (0.5)
Gehler and Nowozin [47]	27.7 (0.5)	34.0 (0.4)	37.5 (0.6)	40.1 (0.9)
VanGemert *et al.* [48]	34.2 (-)	45.8 (-)	-	-
Wang *et al.* [30]	34.4 (-)	41.2 (-)	45.3 (-)	47.7 (-)
Boureau *et al.* [8]	-	41.7 (0.8)	-	-
Feng *et al.* [50]	35.8 (-)	43.3 (-)	47.3 (-)	-
Kulkarni and li [51]	39.4	45.8 (-)	49.3 (-)	51.4 (-)
Bergamo and Torresani [45]	39.5 (-)	45.8 (-)	-	-
Bo *et al.* [52]	40.5 (0.4)	48.0 (0.2)	**51.9 (0.2)**	**55.2 (0.3)**
CSPS + SIFT	39.2 (0.4)	46.1 (0.2)	50.8 (0.2)	55.1 (0.4)
CSPS + CI	**42.2 (0.3)**	**49.1 (0.3)**	51.8 (0.4)	53.6 (0.2)

now provide more details about the different techniques. The baseline [10] is a reimplementation of the original SP [5]. Several systems are based on the combination of multiple channels that correspond to many different features, including those from [45]–[48]. Other works consider a single type of descriptors, typically SIFT descriptors [2]. The work in [49] makes use of the efficient match kernel framework, which embeds patches in a higher-dimensional space in a non-linear fashion. In [30], the authors consider different variants of sparse coding and [8], [40], [50] design different spatial pooling strategies. The method described in [51] extracts on the order of a million patches per image by computing SIFT descriptors from several affine transforms of the original image and uses sparse coding in combination with Adaboost. Among the advanced methods, [52] reports the best results on this dataset. It uses a deep architecture that stacks three layers, each one consisting of three steps: coding, pooling and contrast normalization. Note that this deep architecture also makes use of color information. Our CSPS which combines the SIFT and color information in the descriptors in [52], outperforms that method with 15 and 30 training images.

Fig. 7. Samples of each class from Scene-15 dataset.

TABLE IV
CATEGORY-WISE ACCURACY (%) COMPARISON ON SCENE-15 DATASET

Class	ScSPM [10]	SP+RSC [56]	CSPS + LLC	CSPS+RSC
suburb	85.29 ± 1.42	89.55 ± 1.23	89.79 ± 0.95	**93.55 ± 1.32**
coast	90.53 ± 1.51	**93.03 ± 1.47**	92.15 ± 0.61	90.03 ± 1.31
forest	84.85 ± 0.91	97.67 ± 1.55	89.12 ± 1.30	**91.67 ± 1.87**
highway	86.25 ± 2.67	88.85 ± 2.18	**90.12 ± 1.34**	88.85 ± 2.18
insidecity	88.94 ± 1.16	89.50 ± 1.10	92.04 ± 1.43	**94.50 ± 1.10**
mountain	84.67 ± 2.70	85.67 ± 2.35	87.50 ± 2.96	**87.61 ± 2.05**
opencountry	74.19 ± 3.33	83.37 ± 0.50	86.03 ± 1.55	**89.37 ± 0.72**
street	84.63 ± 2.29	93.91 ± 2.07	92.79 ± 3.13	**95.91 ± 1.31**
tallbuilding	93.57 ± 0.35	**98.52 ± 0.28**	94.05 ± 0.33	96.52 ± 0.28
office	86.96 ± 2.25	86.45 ± 1.29	87.83 ± 2.84	**88.45 ± 1.29**
bedroom	67.24 ± 5.57	84.21 ± 2.54	88.35 ± 1.03	**92.21 ± 2.14**
industrial	56.40 ± 2.00	57.34 ± 3.07	76.25 ± 2.67	**79.24 ± 1.07**
kitchen	66.36 ± 3.44	69.83 ± 3.78	76.55 ± 2.54	**82.83 ± 1.55**
living room	62.43 ± 2.92	65.69 ± 2.38	78.02 ± 2.55	**83.69 ± 2.38**
store	69.77 ± 2.70	72.47 ± 1.96	**84.53 ± 2.50**	83.32 ± 1.05

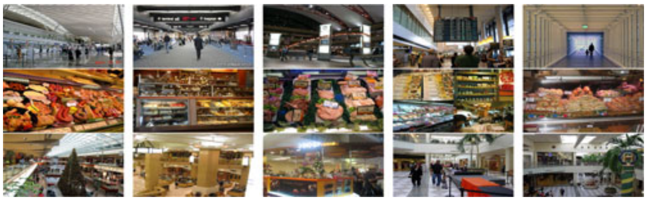


Fig. 8. Samples of three categories (top to bottom: airport, deli, and mall) from MIT-67 dataset.

B. Scene Classification

Scene-15 [5] dataset was built gradually.The initial eight classes were collected by [53], and then five categories were added by [54]. Finally, two additional categories were introduced by [5]. Scene-15 dataset has 15 categories, including suburb, coast, forest, highway, insidecity, mountain, opencountry, street, tallbuilding, office, bedroom, industrial, kitchen, living room and store. It contains 4482 gray-scale images in total. The image resolution is approximately 250 × 300, and there are 210 to 410 images per category. Some samples are shown in Fig. 7. In our experiments, we resize the images to have a minimum dimension of 256 pixels (while maintaining the aspect ratio). The gray histogram features are extracted from each image region [55]. Following the common settings on this dataset [10], we randomly select 100 images per class as training data and use the remaining images as test data.

As for the implementation details for this dataset, we use a single SIFT descriptor, by densely extracting local patches of 16 × 16 pixels computed over a grid with a spacing of

TABLE V
CATEGORY-WISE ACCURACY (%) FOR CSPS, RBoW [58], DPM [57], AND GIST-COLOR [63] ON MIT-67 DATASET

Category	CSPS	RBoW	DPM	GC	Category	CSPS	RBoW	DPM	GC	Category	CSPS	RBoW	DPM	GC	Category	CSPS	RBoW	DPM	GC
greenhouse	**88**	75	65	55	c. store	**64**	44	33	11	hos. room	**44**	40	5	15	lobby	25	10	**30**	30
buffet	**87**	65	75	50	c. room	**63**	56	22	28	subway	**44**	33	38	38	artstudio	**23**	15	5	10
i. subway	**85**	81	62	10	nursery	**63**	55	60	50	grocerystore	**44**	33	19	43	lab. wet	22	**27**	5	9
elevator	**85**	62	52	67	corridor	**62**	43	57	48	fastfoodrest	**43**	24	12	18	j. shop	**22**	0	5	5
clositer	83	80	**90**	80	garage	**61**	44	56	28	winecellar	42	29	14	**43**	deli	**20**	0	5	16
c.inside	**82**	74	63	74	auditorium	**55**	44	11	22	meeting r.	39	41	**75**	45	toystore	**19**	14	9	14
casino	**81**	47	32	32	tv studio	**54**	22	6	33	bathroom	39	33	**50**	33	o. room	18	5	5	**26**
inside bus	**81**	78	43	48	poolingside	**53**	10	0	55	d. office	37	**48**	24	33	shoeshop	18	**21**	16	11
bowling	79	**85**	35	45	stairscase	**53**	35	35	35	bakery	36	5	11	**37**	museum	18	**39**	13	4
pantry	**78**	55	75	40	r. kitchen	**53**	26	4	17	air. inside	**36**	20	5	5	gameroom	18	**45**	40	10
florist	78	**84**	79	63	library	**52**	50	0	35	ch. room	**35**	28	6	17	mall	18	**25**	25	20
stu. music	**75**	37	32	42	movie t.	52	**55**	45	25	living r.	**35**	5	20	10	warehouse	17	19	**24**	24
classroom	**75**	72	67	39	gym	**51**	28	22	11	bedroom	**34**	19	5	0	waiting r.	16	**24**	5	14
closet	**73**	61	44	50	kitchen	**51**	38	29	43	laundromat	32	36	**45**	18	bar	14	**44**	11	11
k. garden	**70**	40	15	25	videostore	**50**	41	18	18	prisoncell	30	**45**	40	35	office	**14**	10	10	10
concert h.	**69**	65	65	60	bookstore	**48**	30	45	20	locker r.	**28**	19	19	5	restaurant	**13**	10	5	0
trainstation	**68**	45	35	55	din. room	46	28	28	**50**	hairsalon	28	19	43	**29**					

TABLE VI
CLASSIFICATION ACCURACY (%) ON MIT-67 DATASET

HOG	22.8	DPM+GIST-Color	39.0
GIST-Grayscale [53]	22.0	DPM +SPM	40.5
MM-Scene [53]	28.0	DPM + GIST-Color + SPM	43.1
SP [5]	34.4	RBoW [58]	37.9
GIST-Color + SPM	38.5	SPMSM [58]	44.0
ROI+GIST [13]	26.5	CSPS + Object Bank	38.9
DPM [57]	30.4	CSPS + ROI-GIST	41.2
CENTRIST [59]	36.9	CSPS + GIST-Grayscale	46.9
Object Bank [60]	37.6	CSPS + GIST-Color	**52.5**

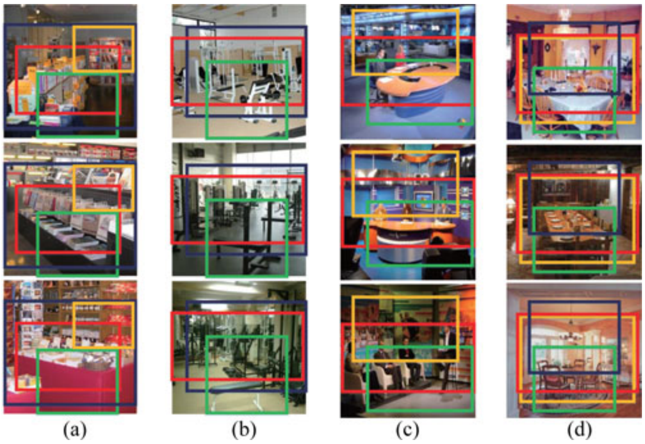


Fig. 9. Samples learned by CSPS on MIT-67 dataset. (a) Bookstore. (b) Gym. (c) TV studio. (d) Dining room.

8 pixels. We set the dictionary size to $1,024$ and the color cluster parameter k in (4) to (5). We apply χ^2 kernel for the kernel K in (12). The tradeoff parameters to the sparsity regularization term and the SVM regularization term are chosen via five-fold cross validation on the training data. We compare our proposed pooling strategy CSPS with the original SP and the state-of-the-art method from [56], and list the specific quantitative results and comparisons in Table IV. We can observe that our CSPS achieves extremely high performance on scene classification and outperforms ScSPM by nearly 10%. This is probably because the scene images contain plentiful geometric properties across different categories, which benefits CSPS when capturing adaptive spatial information. However, we notice that SP + RSC [56] achieves better results in some classes than CSPS, which is perhaps due to the discriminative power of their feature codes on this data set. We then turn to employ their feature coding strategy. By incorporating robust sparse coding (RSC), our method can obtain additional improvement of about 5%.

MIT-67 [13] is a more challenging scene data set that includes $15\,620$ images of indoor scenes in 67 different categories. It consists of different types of stores (e.g., bakery, grocery) residential rooms (e.g., nursery room, bedroom), public spaces (e.g., inside bus, library, prison cell), leisure places (e.g., buffet, fast food, bar, movie theater) and working places (e.g., office, operating room, tv studio). Some samples are shown in Fig. 8. Following common settings on this dataset [13], we use 80 images per class for training and 20 images per class for testing. Train/test configuration files were provided by the authors of [13]. We resize the images to have a maximum dimension of 400 pixels (while maintaining the aspect ratio).

Many works [13], [57], [58] indicate that the popular SIFT descriptors can not yield satisfactory performance on this dataset. Thus, we evaluate the combination of our CSPS with four different feature extraction implementations, i.e., Object Bank, ROI-GIST, GIST-Grayscale and GIST-Color, and compare them with the well-known methods including HOG, GIST, GIST-color, SP, Object Bank, DPM, and several other recently published methods. We list a very specific category-wise performance comparison in Table V for further insight and discussion. The average classification accuracy is shown in Table VI. We can observe that our CSPS achieves the best performance on 51 categories out of the full 67 categories, compared using RBoW [58], DPM [57], and Gist-color [57] with standard SPM. Unlike DPM based methods, our method does not need to detect any specific

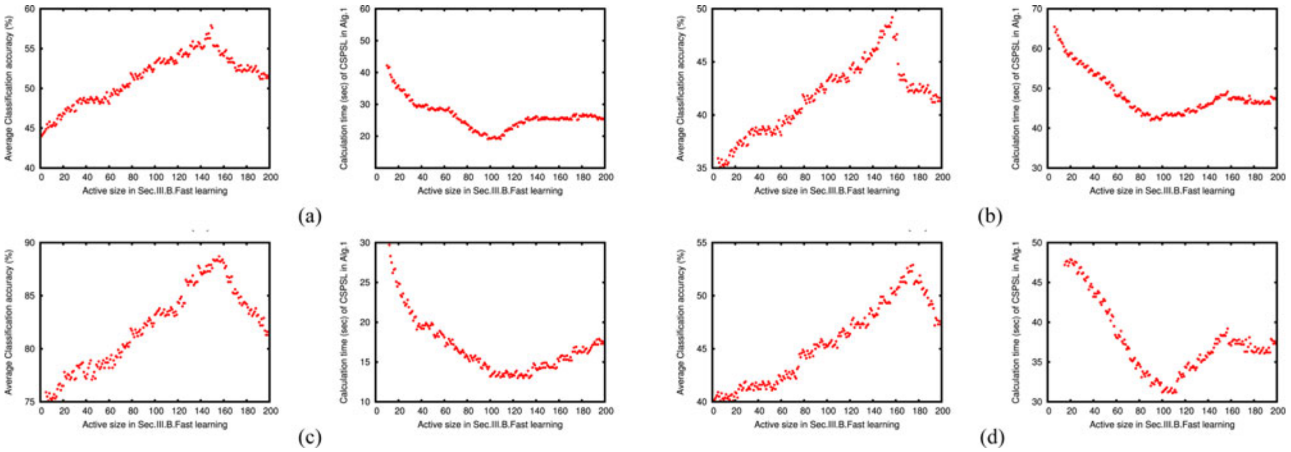


Fig. 10. Calculation time and classification ability with different active sizes of fast learning on four datasets. (a) PASCAL VOC 2007. (b) Caltech-256. (c) Scene-15. (d) MIT-67.

"objects" or "parts" explicitly, and its computational cost is extremely low. Finally, we show some typical learned CSPS results from MIT-67 data set in Fig. 9. As the figure show, CSPS captures the most dominant spatial information and the captured regions are compact yet discriminative.

C. Calculation Time

We evaluate the calculation time of the CSPSL procedure in this part. Because the class-specific selection from an over-complete basis is a computationally demanding task, our fast learning technique in Section III-C plays an important role. The key point during fast learning is the active size of the feature set when iteratively selecting the features with the largest score and updating the feature set. We report the calculation time and the classification performance in Fig. 10 with different active sizes of the feature set. The conclusions are as follows. In all four data sets, the fast learning procedure can improve the calculation time for CSPSL in Algorithm 1. When the active size is small, we obtain better performance as the active size increases. However, the classification ability begins to decrease when the active size is larger than 120. In all four datasets, accuracy peaks are obtained at approximately 120. One interesting finding is that in Scene-15 dataset, the best classification accuracy is obtained at 60. This result is perhaps because this data set contains fewer categories and simpler images than the other three data sets. As for the calculation time, in all data sets, fast learning can helps reduced the learning calculation time required for CSPS. Specifically, when the active size is fewer than 100, CSPSL tends to improve as the active size increases, but start to converge when the active size of the feature set is larger than 100. Taking both classification ability and calculation time into account, we suggest setting the active size to approximately 100.

V. Conclusion

In this paper, we propose a data-driven approach to adaptively learn CSPS for image classification. In contrast to the standard SP, which uses uniform spatial pooling shapes, our CSPS determines adaptive and semantical spatial patterns for feature pooling, which proves capable of capturing more class-specific geometric information. Our method outperforms standard SP and most other relevant works on four diverse datasets (PASCAL 2007, Caltech-256, Scene-15 and MIT-67). The experimental results show that it effectively captures valuable spatial information for both object and scene images. In the future, our study will concentrate primarily on the design of different over-complete shape sets and faster solutions for CSPSL.

References

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vis.*, 2004, vol. 1, pp. 1–22.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[4] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 801–808.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2169–2178.

[6] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

[7] J. J. Koenderink and A. J. Van Doorn, "The structure of locally orderless images," *Int. J. Comput. Vis.*, vol. 31, nos. 2/3, pp. 159–168, 1999.

[8] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 2559–2566.

[9] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 111–118.

[10] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1794–1801.

[11] Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3370–3377.

[12] J. Sánchez, F. Perronnin, and T. De Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recog. Lett.*, vol. 33, no. 16, pp. 2216–2223, 2012.

[13] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 413–420.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[16] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, "Imagenet Large Scale Visual Recognition Competition (ILSVRC2012)," 2012. [Online]. Available: http://www.image-net.org/challenges/LSVRC/2012/

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842.

[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556.

[20] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[21] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern Information Retrieval*, vol. 463. New York, NY, USA: ACM Press, 1999.

[22] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 2, pp. 1470–1477.

[23] P. Quelhas *et al.*, "Modeling scenes with local descriptors and latent aspects," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 1, pp. 883–890.

[24] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2033–2040.

[25] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[26] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Image classification with kernelized spatial-context," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 278–287, Jun. 2010.

[27] M. Shao *et al.*, "Learning relative features through adaptive pooling for image classification," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2014, pp. 1–6.

[28] N. Pourian and B. Manjunath, "Pixnet: A localized feature representation for classification and visual search," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 616–625, May 2015.

[29] U. L. Altintakan and A. Yazici, "Towards effective image classification using class-specific codebooks and distinctive local features," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 323–332, Mar. 2015.

[30] J. Wang *et al.*, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3360–3367.

[31] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 141–154.

[32] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[33] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer, "Learning object representations for visual object class recognition," in *Proc. Vis. Recog. Challange Workshop, Conjunction IEEE Int. Conf. Comput. Vis.*, 2007. [Online]. Available: http://lear.inrialpes.fr/pubs/2007/MSHV07

[34] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 82–89.

[35] R. Jenatton, J. Mairal, F. R. Bach, and G. R. Obozinski, "Proximal methods for sparse hierarchical dictionary learning," in *Proc. 27th Int. Conf. Mach. Learning*, 2010, pp. 487–494.

[36] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 2, pp. 1800–1807.

[37] B. Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 179–192.

[38] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 7, pp. 1294–1309, Jul. 2009.

[39] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *CoRR*, 2000. [Online]. Available: http://arxiv.org/abs/physics/0004057.

[40] Y.-L. Boureau, N. L. Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: multi-way local pooling for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2651–2658.

[41] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 1–15.

[42] S. R. Fanello, N. Noceti, C. Ciliberto, G. Metta, and F. Odone, "Ask the image: supervised pooling to preserve feature locality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 851–858.

[43] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[44] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. UCB/CSD-04-1366, California Inst. Technol., Pasadena, CA, USA, 2007.

[45] A. Bergamo and L. Torresani, "Meta-class features for large-scale object categorization on a budget," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3085–3092.

[46] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[47] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep.-Oct. 2009, pp. 221–228.

[48] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[49] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 135–143.

[50] J. Feng, B. Ni, Q. Tian, and S. Yan, "Geometric p-norm feature pooling for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 2609–2704.

[51] N. Kulkarni and B. Li, "Discriminative affine sparse codes for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1609–1616.

[52] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 660–667.

[53] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[54] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 2, pp. 524–531.

[55] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[56] C. Zhang *et al.*, "Image classification using spatial pyramid robust sparse coding," *Pattern Recog. Lett.*, vol. 34, no. 9, pp. 1046–1052, 2013.

[57] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. IEEE Int. Conf. Comput. Vis*, Nov. 2011, pp. 1307–1314.

[58] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2775–2782.

[59] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.

[60] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

Jinzhuo Wang received the B.S. degree in electronic engineering and computer science from Peking University, Beijing, China, in 2009, and is currently working toward the Ph.D. degree in electronic engineering and computer science at Peking University.

His research interests include computer vision and deep learning.

Wenmin Wang received the Ph.D. degree in computer architecture from the Harbin Institute of Technology, Harbin, China, in 1989.

After completing the Ph.D. degree, he was an Assistant Professor and an Associate Professor with the Harbin University of Science and Technology, Harbin, China, as well as the Harbin Institute of Technology. Since 1992, he gained about 18 years of oversea industrial experiences in Japan and America, in where served as a Staff Engineer, Chief Engineer, General Manager of Software Division, etc., for various companies. In 2009, he became a Professor with the School of Electronic and Computer Engineering, Peking University, Shenzhen, China. His current research interests include computer vision, multimedia retrieval, artificial intelligence, and machine learning.

Ronggang Wang (M'12) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006.

He was a Research Staff Member with Orange (France Telecom) Laboratories, Lannion, France, from 2006 to 2010. He is currently an Associate Professor with Peking University Shenzhen Graduate School, Shenzhen, China. He has authored or coauthored more than 50 papers in international journals and conferences. He holds more than 40 patents. His research interests includes video coding and processing.

Mr. Wang lead the MPEG Internet Video Coding (IVC) standard, has served as MPEG IVC AHG Co-Chair since 2012, and has served as the AVS Implementation Sub-Group Co-Chair since 2015.

Wen Gao (S'87–M'88–SM'05–F'09) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

He is a Professor of Computer Science with Peking University, Beijing, China. Before joining Peking University, he was a Professor of Computer Science with the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology of Chinese Academy of Sciences, Beijing, China. He has authored or coauthored five books and more than 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics.

Prof. Gao served on the Editorial Board for several journals, such as the IEEE TRANSACTIONS ON CIRUCITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, the *Eurasip Journal of Image Communications*, and the *Journal of Visual Communication and Image Representation*. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.