LEARNING CLASS-SPECIFIC POOLING SHAPES FOR IMAGE CLASSIFICATION

Jinzhuo Wang[*], *Wenmin Wang*[*], *Ronggang Wang*[*], *Wen Gao*[*,†]

[*]Digital Media R & D Center, Peking University Shenzhen Graduate School
[†]National Engineering Laboratory for Video Technology, Peking University
[*]wangjz@sz.pku.edu.cn, [*]wangwm@ece.pku.edu.cn, [*]rgwang@pkusz.edu.cn, [†]wgao@pku.edu.cn

ABSTRACT

Spatial pyramid (SP) representation is an extension of bag-of-feature model which embeds spatial layout information of local features by pooling feature codes over pre-defined spatial shapes. However, the uniform style of spatial pooling shapes used in standard SP is an ad-hoc manner without theoretical motivation, thus lacking the generalization power to adapt to different distribution of geometric properties across image classes. In this paper, we propose a data-driven approach to adaptively learn class-specific pooling shapes (CSPS). Specifically, we first establish an over-complete set of spatial shapes providing candidates with more flexible geometric patterns. Then the optimal subset for each class is selected by training a linear classifier with structured sparsity constraint and color distribution cues. To further enhance the robust of our model, the representations over CSPS are compressed according to the shape importance and finally fed to SVM with a multi-shape matching kernel for classification task. Experimental results on three challenging datasets (Caltech-256, Scene-15 and Indoor-67) demonstrate the effectiveness of the proposed method on both object and scene images.

Index Terms— Image classification, class-specific pooling shapes (CSPS), representation compression, multi-shape matching kernel

1. INTRODUCTION

Bag-of-feature (BoF) model [1] is one of the most powerful and popular framework for image classification. The standard BoF model starts with handcrafted features such as SIFT [2] extracted from either interesting points or densely sampled patches. Such raw features are then either quantized or coded to obtain a dictionary. Finally the histogram of the feature codes over the whole image is regarded as a signature to make classification. However, the spatial layout information in this fashion is completely neglected. To overcome this drawback, [3] pioneers the direction of exploiting spatial layout property and proposes spatial pyramid (SP) to embed spatial information of local features. In detail, it 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 feature in each grid with a certain pooling scheme to achieve the final representation. The SP-based representation guides most approaches of image classification and benefits many state-of-the-art systems [4–6].

(a) 3-level SP (b) CSPS

Fig. 1. Comparison of standard SP and class-specific pooling shapes (CSPS) learned by our method.

However, one obvious limitation in standard SP is the uniform feature pooling style, which uses all the spatial shapes and treats them equally, thus lacking the capability to capture adaptive spatial information. For instance, an image belonging to "meeting-room" class in Indoor-67 dataset [7] is coped with an 3-level standard SP and the proposed class-specific pooling shape (CSPS), as shown in Fig.1(a) and Fig.1(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 handcrafted and uniform pooling shapes in standard SP lack the power of generalization across different classes.

In this paper, we propose a data-driven approach to adaptively learn class-specific pooling shapes (CSPS). Our idea is motivated by the observation that images in the same class often share common spatial layout properties, i.e., background and target tend to follow similar spatial distribution. In practice, we first adopt the concept of over-complete set and establish a set of spatial shapes providing as many candidates as possible. This scheme helps us collect more flexible pattern-

This work was partly supported by the grant from Shenzhen municipal government for basic research on Information Technologies (No. JCYJ20130331144751105), and Shenzhen Peacock Plan.

s of spatial distribution. Instead of using pre-defined spatial shapes in standard SP, we train a linear classifier with structured sparsity constraint and color distribution cues to select optimal subset for each class. In particular, the sparsity term encourages the classifier to extract small but essential subset avoiding redundancy and the color term makes the selected shapes more semantically reasonable following color distribution inference. To limit the complexity, we compress the representations over CSPS according to the shape importance which are finally fed to SVM with a multi-shape matching kernel for image classification task.

The remainder of the paper is organized as follows. Section 2 reviews the related work of standard SP and its variations, and Section 3 proposes our image classification framework with focus on the approach to adaptively learn class-specific pooling shapes (CSPS). Experimental results with analysis and comparison are presented in Section 4, and we conclude the paper in Section 5.

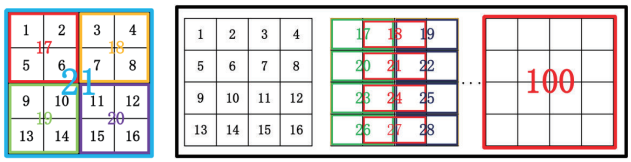
2. RELATED WORK

Most approaches for image classification are built upon the BoF model which regards an image as an order-less histogram of features, where the spatial layout property is completely discarded. To overcome this limitation, various extensions have been proposed from the following two directions, i.e., the property of local spatial layout and global spatial layout.

Local spatial layout information mainly explores the relative positions or pairwise positions of the local features. [8] uses the combination of correlograms and visual words to represent spatially neighboring image regions. In [9], an efficient feature selection method based on boosting is introduced to mine high-order spatial features. [10] learns relative features by reference basis (RB) and proposes an adaptive pooling technique to assemble the learned multiple relative features, achieving good performance.

On the other hand, global spatial layout information leverages the absolute positions in images, which is also our focus in this paper. Based on the pioneer work [3] where the original SP is proposed, [4] and [11] show that incorporating advanced feature coding strategies is able to improve the classification performance. Moreover, the combinations with super vector [12] and fisher vector [13] are demonstrated effective to obtain a good image representation.

More recently, several advanced image classification systems are based on SP, but involving different parameters including the number of pyramid levels and the structure of the grids at each level. For instance, [3] and [4] use up to 4 pyramid levels with uniform grids of 1×1, 2×2, 4×4 and 8×8, while the winner of Pascal VOC 2007 competition [14] followed by many others such as [12] use 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.



(a) (b)

Fig. 2. Toy example of pooling shapes by standard SP 2(a) and the proposed over-complete spatial shape set 2(b) on a 4×4 grid. Standard SP yields $1 \times 1 + 2 \times 2 + 4 \times 4 = 21$ grids while ours can produce $\binom{4+1}{2} \times \binom{4+1}{2} = 100$ candidates.

Although the extensions of standard SP are addressed a lot, rather little attention has been paid to achieve spatial pooling shapes in a learning procedure. In this paper, we address this issue by a data-driven approach to adaptively learn the optimal pooling shapes for each class of images. The most related work to ours is [5] which also adopts the idea of over-complete set and formulates the problem in a multi-class fashion to learn discriminative spatial shapes for the whole dataset. However, different classes often have different distribution of spatial properties, we thus attempt to learn class-specific pooling shapes (CSPS). Moreover, we try to leverage the color distribution information which is often used for region of interest (ROI) detection and segmentation in our learning procedure to select more semantically reasonable pooling shapes following color cues.

3. APPROACH

In this section, we detail our image classification framework with focus on learning class-specific pooling shapes (CSPS), from establishing an over-complete spatial shape set, to learning CSPS with sparsity and color constraints. Image representations over CSPS are then compressed and finally fed to SVM classifier with a proposed multi-shape matching kernel.

3.1. Over-complete spatial shape set

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

Note that the over-complete scheme makes it possible to obtain more flexible shapes such as circles and polygons, which can capture more adaptive and semantical geometric properties for particular recognition task. For the simplicity to compare with standard SP, we only apply the increasing

horizontal and vertical lines to form rectangular shapes in our implementation.

3.2. Learning class-specific pooling shapes (CSPS)

Since $\mathcal{S}$ is over-complete with a lot of redundancy, we attempt to select the optimal subset for each class due to the observation that images in the same class often share the common spatial layout distribution. 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 extract local features and employ feature coding algorithm to obtain a dictionary $\mathcal{D} = \{d_1, d_2, \ldots, d_t\}$. Afterwards spatial pooling of feature codes is conducted on each shape of $\mathcal{S}$. By this way, the i-th 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 with one-versus-all fashion to select 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 n-th sample, $y_n \in \{-1, +1\}$ is the label of the n-th sample indicating class i and "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

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

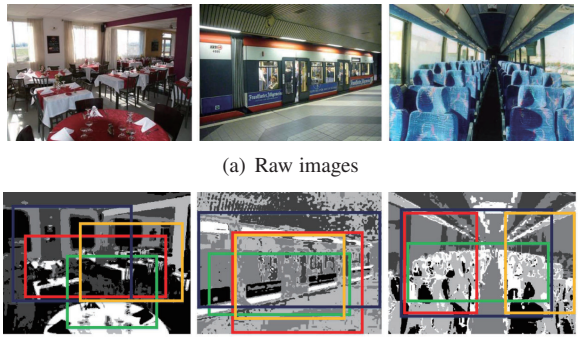
The regularization term in Eq.1 is expected to select the subset for each class containing the most representative and discriminative spatial shapes. Thus we employ two regularization terms and $\mathrm{Reg}(\mathbf{w})$ can then be reformulated as

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

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

3.2.1. *Color distribution cues*

To leverage the color information as learning cues, we apply color segmentation to assign a certain color channel to each pixel in advance. For fast implementation, we simply employ k-means algorithm to cluster an image with k colors by converting RGB color space to L*a*b space, denoted as $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$. Since the learned shapes are expected



(a) Raw images

(b) Learned shapes with color segmentation

Fig. 3. Examples of raw images and the corresponding CSPS with color distribution cues on Indoor-67 dataset.

to capture dominant channels in color space, we define the color regularization term in Eq.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 i-th color in s_j, $N(c_i)$ is the pixel number of the i-th color, $P(j)$ is the pixel number of s_j. 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. With this regularization term, the classifier tends to select semantically reasonable shapes following color distribution inference. Some examples of learned shapes are shown in Fig.3.

3.2.2. *Structured sparsity constraint*

While a lot of significant efforts have been placed on the design of sparse regularizer such as squared Frobenius norm and $\ell_{1,\infty}$ norm [5], recent analysis [15] shows that the mixed norm regularization under certain conditions enjoys the group sparsity property, encouraging the content-based structured feature selection in high-dimensional feature space. We adopt the idea of structured sparsity in [15] and define the sparse regularization term as a ℓ_2/ℓ_1 norm regularizer

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

where $\mathbf{w}_i$ is the i-th 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 regular ℓ_1 norm regularizer, the sparsity is now imposed on spatial shape level rather than merely on feature level. To solve the jointly learning with mixed norm regularizer, we USE the primal-dual algorithm proposed in [16].

3.3. Fast learning

Although the over-complete scheme provides flexible spatial shapes with more geometric patterns, jointly optimizing Eq.1 is a computationally challenging task due to its high dimensional searching space. We employ a greedy approach proposed in [5] 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 Eq.1

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

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 [5] and set the active set size as 100.

3.4. Representation compression

Another issue we concern is the high dimensional representations which are even larger than that in standard SP. To overcome this drawback, we compress the representations by only remaining the feature dimensions corresponding to the important shapes. To measure the importance of each shape, we apply a leave-one-out scheme and the importance value of a particular shape j is defined as the training error increase after neglecting the shape dimensions

$$I_j = \frac{\text{Error}_j - \text{Error}_0}{\text{Error}_0} \tag{7}$$

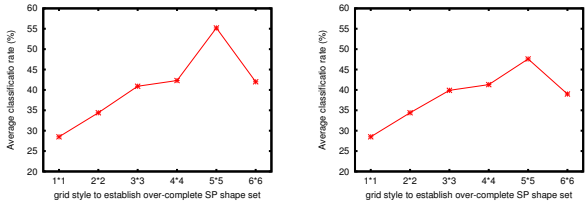
where Error_0 denotes the training error over all the training data. The largest I_j indicates that the neglected dimensions of the j-th shape are more important and discriminative. Beyond the target of representation compression, I_j can be also used in our multi-shape matching kernel as a weight term.

3.5. Multi-shape matching kernel

By now we have learned the CSPS for each class. We then turn to employ SVMs to make classification. Notice that standard SP treats each pooling shape equally in the matching kernel where the difference of each spatial shape is neglected. We attempt to weight the shapes due to the fact that the important region should be paid more attention. Specifically, we use the shape importance value $\mathbf{I}$ in Eq.7 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

$$\mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{m=1}^{\mathcal{R}} I_m \cdot K(\mathbf{x}_1^{s_m}, \mathbf{x}_2^{s_m}) \tag{8}$$

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



(a) Caltech-256 (b) Indoor-67

Fig. 4. Performance with different grid styles of our method on Caltech-256 dataset (60-train) and Indoor-67 dataset.

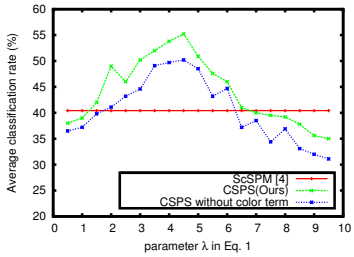


Fig. 5. Performance on Caltech-256 dataset (60-train) with varying tradeoff λ and color term $\text{Reg}_c(\mathbf{w})$ in Eq.1.

4. EXPERIMENTS AND RESULTS

In this section, we report the experimental results on three diverse datasets (Caltech-256, Scene-15 and Indoor-67). We compare our method (CSPS) mainly with KSPM [3] which is standard SP, its popular extensions ScSPM [4] and LCC [11], some related works exploiting spatial pooling shapes [5,6,17] and other relevant approaches [7, 18–25].

4.1. Experiment setup

Although several coding and pooling strategies can be used, we employ sparse coding and max pooling following the instructions of [4] for fair comparison. We use a single SIFT descriptor, by densely extracting local patches of 16×16 pixels computed over a grid with spacing of 8 pixels. For all the experiments, we fix the codebook size as $1,024$. The color cluster parameter k in Eq.4 is set 5. We apply χ^2 kernel for the kernel K in Eq.8. The trade-off parameters to the sparsity regularization term and the SVM regularization term are chosen via 5-fold cross validation on the training data. Other parameter settings are detailed in Sec.4.2. Following the common benchmarking procedures, we repeat the experimental process by 5 times with different randomly selected training and testing images to obtain reliable results.

4.2. Analysis

We provide analysis of our method with focus on the initialization of $\mathcal{S}$, the tradeoff parameter λ of the regularization

terms in Eq.1, and the effectiveness of the color regularizer $\mathrm{Reg}_c(\mathbf{w})$ in Eq.3. Recall that given an $a \times b$ grid over an image, the proposed $\mathcal{S}$ can yield $\mathcal{R} = \binom{a+1}{2} \times \binom{b+1}{2}$ candidates. Although the finer grid setting provides more spatial patterns, spatial shapes produced by too fine grids fail to capture object-oriented information. Experiments on Caltech-256 and Indoor-67 dataset indicate that 5×5 grid style which produces 225 rectangle shapes consistently achieves the best performance as shown in Fig.4(a) and 4(b). On the other hand, the free parameter λ in Eq.1 designed to control sparsity and color distribution of the solution is needed to determine. We conduct the corresponding experiment on Caltech-256 dataset. As shown in Fig.5, a typical curve demonstrates that the best performance is obtained by $\lambda = 4.5$. To validate the effect of color regularizer $\mathrm{Reg}_c(\mathbf{w})$, we test without $\mathrm{Reg}_c(\mathbf{w})$ and the performance lowers about average 2.3% (Fig.5), which indicates that color distribution cues provides complementary information to our model.

4.3. Results

4.3.1. Results on Caltech-256 dataset

Caltech-256 dataset [23] consists of images from 256 object classes containing images from 80 to 827 per class. The significance of this dataset is its large inter-class variability, as well as intra-class variability. The performance comparison results are shown in Fig.6. ScSPM [4] and LCC [11] are two popular extensions of standard SP which employ advanced feature coding strategy, while [19] and [18] are two published leading approaches on this dataset. It is indicated that our method consistently leads the performance and outperforms state-of-the-art by more than 3% averagely. However, the improvement of our method is limited when training number is 45, which is worthy to be noticed.

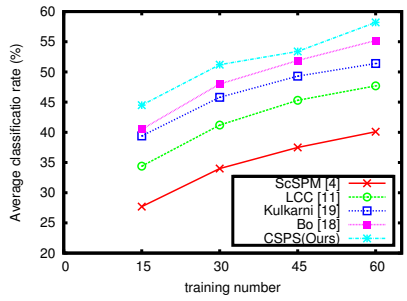


Fig. 6. Performance comparison with different training number on Caltech-256 dataset.

4.3.2. Results on Scene-15 dataset

We further evaluate the performance of our approach for scene images. Scene-15 is a popular scene dataset consisting of 15 natural scene categories, e.g., "building, "bedroom, with total $4,485$ images. We follow the setup in [6] that randomly selects 100 images from each class for training and the rest of the images for testing. Table 1 shows the detailed comparison for each class. We notice that SP + RSC [20] achieves better results in some classes than ours, which is perhaps due to the discriminative power of their feature codes on this dataset. By incorporating robust sparse coding (RSC), our method can further obtain about 2% improvement.

Table 1. Classification accuracy $(\%)$ on Scene-15 dataset.

Class	ScSPM [4]	SP+RSC [20]	CSPS (Ours)	CSPS+RSC
Bedroom	67.24 ± 5.57	84.21 ± 2.54	88.35 ± 1.03	$\mathbf{92.21 \pm 2.14}$
CALsuburb	85.29 ± 1.42	89.55 ± 1.23	89.79 ± 0.95	$\mathbf{93.55 \pm 1.32}$
Industrial	56.40 ± 2.00	57.34 ± 3.07	76.25 ± 2.67	$\mathbf{79.24 \pm 1.07}$
Kitchen	66.36 ± 3.44	69.83 ± 3.78	76.55 ± 2.54	$\mathbf{82.83 \pm 1.55}$
Livingroom	62.43 ± 2.92	65.69 ± 2.38	78.02 ± 2.55	$\mathbf{83.69 \pm 2.38}$
Coast	90.53 ± 1.51	$\mathbf{93.03 \pm 1.47}$	92.15 ± 0.61	90.03 ± 1.31
Forest	84.85 ± 0.91	97.67 ± 1.55	89.12 ± 1.30	$\mathbf{91.67 \pm 1.87}$
Highway	86.25 ± 2.67	88.85 ± 2.18	$\mathbf{90.12 \pm 1.34}$	88.85 ± 2.18
Insidecity	88.94 ± 1.16	89.50 ± 1.10	92.04 ± 1.43	$\mathbf{94.50 \pm 1.10}$
Mountain	84.67 ± 2.70	85.67 ± 2.35	87.50 ± 2.96	$\mathbf{87.61 \pm 2.05}$
Opencountry	74.19 ± 3.33	83.37 ± 0.50	86.03 ± 1.55	$\mathbf{89.37 \pm 0.72}$
Street	84.63 ± 2.29	93.91 ± 2.07	92.79 ± 3.13	$\mathbf{95.91 \pm 1.31}$
Tallbuilding	93.57 ± 0.35	$\mathbf{98.52 \pm 0.28}$	94.05 ± 0.33	96.52 ± 0.28
PARoffice	86.96 ± 2.25	86.45 ± 1.29	87.83 ± 2.84	$\mathbf{88.45 \pm 1.29}$
Store	69.77 ± 2.70	72.47 ± 1.96	$\mathbf{84.53 \pm 2.50}$	83.32 ± 1.05

Table 2. Classification accuracy $(\%)$ on Indoor-67 dataset.

ROI GIST [7]	26.5
SP + HOG [24]	29.8
SP + SIFT [24]	34.4
Scene DPM [17]	30.4
MM Scene [21]	28.0
Centrist [25]	36.9
Object Bank [22]	37.6
CSPS (Ours)	**47.6**

4.3.3. Results on Indoor-67 dataset

Indoor scene dataset [7] is another scene dataset characterized by 67 indoor classes with high intra-class variations. We use the same training and test split as in [17] where each class has 80 training and 20 test images. Fig.7 shows some examples of the learned CSPS for a few classes. As can be seen, the proposed method is able to learn adaptive spatial information. Instead of using popular scene-specific features such as GIST in this dataset, we only use single SIFT features to demonstrate the effectiveness of CSPS. The detailed comparison results are listed in Table 2 which shows that our method outperforms the SP-based approaches [7, 17] and other relevant works [21, 22, 25], yielding 47.6% performance which, to our knowledge, is the best on this dataset.

Note that [17, 25] report there is an improvement space by adding some scene-specific features on this dataset, which indicates that our method has the potential to be further improved when combining other features such as GIST and GIST-color [17] on this dataset.

(a) Bookstore (b) Gym (c) Tv-studio (d) Dining-room

Fig. 7. A few samples of class-specific pooling shapes (CSPS) learned by our method on Indoor-67 dataset.

5. CONCLUSION

In this paper, we propose a data-driven approach to adaptively learn class-specific pooling shapes (CSPS) for image classification. Different from standard SP using uniform spatial pooling shapes, our CSPS provides adaptive and semantical spatial patterns for feature pooling, which is able to capture more class-specific information. Our method outperforms standard SP and other methods on three diverse datasets (Caltech-256, Scene-15 and Indoor-67), the experimental results have shown its effect to capture valuable spatial information for both object and scene images.

6. REFERENCES

[1] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, 2004, vol. 1, pp. 1–2.

[2] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2169–2178.

[4] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[5] Yangqing Jia, Chang Huang, and Trevor Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3370–3377.

[6] Jorge Sánchez, Florent Perronnin, and TeóFilo De Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2216–2223, 2012.

[7] Ariadna Quattoni and Antonio Torralba, "Recognizing indoor scenes," 2009.

[8] Silvio Savarese, John Winn, and Antonio Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2033–2040.

[9] David Liu, Gang Hua, Paul Viola, and Tsuhan Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[10] Ming Shao, Sheng Li, Tongliang Liu, Dacheng Tao, Thomas S Huang, and Yun Fu, "Learning relative features through adaptive pooling for image classification," in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–6.

[11] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[12] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S Huang, "Image classification using super-vector coding of local image descriptors," in *Computer Vision–ECCV 2010*, pp. 141–154. Springer, 2010.

[13] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010*, pp. 143–156. Springer, 2010.

[14] Marcin Marszałek, Cordelia Schmid, Hedi Harzallah, Joost Van De Weijer, et al., "Learning object representations for visual object class recognition," in *Visual Recognition Challange workshop, in conjunction with ICCV*, 2007.

[15] Samy Bengio, Fernando Pereira, Yoram Singer, and Dennis Strelow, "Group sparse coding," in *Advances in Neural Information Processing Systems*, 2009, pp. 82–89.

[16] Rodolphe Jenatton, Julien Mairal, Francis R Bach, and Guillaume R Obozinski, "Proximal methods for sparse hierarchical dictionary learning," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 487–494.

[17] Megha Pandey and Svetlana Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1307–1314.

[18] Liefeng Bo, Xiaofeng Ren, and Dieter Fox, "Multipath sparse coding using hierarchical matching pursuit," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 660–667.

[19] Naveen Kulkarni and Baoxin Li, "Discriminative affine sparse codes for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1609–1616.

[20] Chunjie Zhang, Shuhui Wang, Qingming Huang, Jing Liu, Chao Liang, and Qi Tian, "Image classification using spatial pyramid robust sparse coding," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1046–1052, 2013.

[21] Jun Zhu, Li-Jia Li, Li Fei-Fei, and Eric P Xing, "Large margin learning of upstream scene understanding models," in *Advances in Neural Information Processing Systems*, 2010, pp. 2586–2594.

[22] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[23] Gregory Griffin, Alex Holub, and Pietro Perona, "Caltech-256 object category dataset," 2007.

[24] Saurabh Singh, Abhinav Gupta, and Alexei A Efros, "Unsupervised discovery of mid-level discriminative patches," in *Computer Vision–ECCV 2012*, pp. 73–86. Springer, 2012.

[25] Jianxin Wu and Jim M Rehg, "Centrist: A visual descriptor for scene categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1489–1501, 2011.