Generalized Regular Spatial Pooling for Image Classification

Lingxi Xie, Qi Tian and Bo Zhang

May 21, 2014

Abstract

This paper discusses spatial pooling, a basic and crucial problem in the Bag-of-Features (BoF) model. Conventional algorithms such as Spatial Pyramid Matching (SPM) [1] hierarchically divide the image into exclusive and regular regions for feature summarization, but we propose an extremely simple algorithm named Generalized Regular Spatial Pooling (GRSP), which allows the pooling bins in the same layer have relatively denser or sparser distributions. With the proposed algorithm, it is possible to enhance the "representation power" on each of the pooling layers. State-of-the-art classification accuracy is achieved on several challenging image classification datasets.

1 Introduction

Image classification has been a basic task in the computer vision community. It is an intrinsic challenge towards image understanding and implies a wide range of real-world applications. Today, one of the most popular methods is to represent images with long vectors, and use a generalized classifier such as SVM [2] for training and testing.

The Bag-of-Features (BoF) model [3] [4] is widely used for image representation. It is a statistics-based model which summarizes local features into an image-level feature. As the final stage of the BoF model, pooling is widely adopted to capture the spatial invariance of the image. Beyond the primary sum-pooling and max-pooling methods, efforts are made towards better image representation. Among the numerous spatial pooling methods, the most successful methods are probably Pyramid Matching (PM) [5] and Spatial Pyramid Matching (SPM) [1]. By dividing an image into several hierarchical regions for feature summarization, it is possible to capture richer semantic information in the individual parts of the image. Many efforts are also made [6] [7] to improve the spatial pooling methods. However, conventional algorithms often define the pooling bins as exclusive and regular grids on the image plane, which limits the flexibility of the model and makes it difficult to fit on larger-scale image collections. To overcome this shortcoming, we propose Generalized Regular Spatial

Pooling (GRSP), an extremely simple pooling method which allows the bins in the same layer have denser or sparser distributions, so that the "representation power" of spatial pooling could be adjusted and enhanced. Despite the simplicity, the proposed method is verified to achieve better image representation than the spatial pyramids, and produces the state-of-the-art classification accuracy on some challenging image classification datasets.

The remainder of this paper is organized as follows. First, we provide a brief overview of the BoF model in Section 2. Then Section 3 presents the Generalized Regular Spatial Pooling (GRSP) algorithm. After extensive experiments and discussions are given in Section 4, we draw our conclusions in Section 5.

2 The Bag-of-Features Model

The Bag-of-Features model starts from a raw image $\mathbf{I} = (a_{ij})_{W \times H}$, where a_{ij} is the **pixel** at position (i, j). For better local representation, a set of SIFT [8] descriptors is extracted: $\mathcal{D} = \{(\mathbf{d}_1, \mathbf{l}_1), (\mathbf{d}_2, \mathbf{l}_2), \ldots, (\mathbf{d}_M, \mathbf{l}_M)\}$, where $\mathbf{d}_m$ and $\mathbf{l}_m$ denote the **description vector** and the **spatial location** of the m-th descriptor, respectively. M is the number of descriptors.

For encoding the descriptors, a **codebook** $\mathbf{C}$ is trained using clustering methods. $\mathbf{C}$ is a $B \times D$ matrix consisting of B vectors with dimension D, each of which is called a **codeword**. Descriptors are then projected onto the space spanned by the codewords. Typical encoding methods include Locality-constrained Linear Coding (LLC) [9] and Improved Fisher Vector (IFV) [10]. The encoded vector $\mathbf{w}_m$ is named the corresponding **visual word** of **descriptor** $\mathbf{d}_m$. Let $\mathcal{W}$ be the set of visual words: $\mathcal{W} = \{(\mathbf{w}_1, \mathbf{l}_1), (\mathbf{w}_2, \mathbf{l}_2), \ldots, (\mathbf{w}_M, \mathbf{l}_M)\}$, and the visual words in $\mathcal{W}$ are then aggregated into a single representation vector $\mathbf{f}$ using max-pooling (for LLC encoding) or sum-pooling (for IFV encoding).

The global pooling algorithm ignores rich spatial information which could be very useful for image understanding. The state-of-the-art image classification systems [1] [11] often divide images into smaller regions for spatial context modeling. Explicitly, let $\mathcal{J} = \{1, 2, \ldots, M\}$ be the index set of the descriptors in $\mathcal{D}$. The spatial pooling algorithm defines S subsets of $\mathcal{J}$, denoted as $\{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_S\}$, and summarize the feature vectors in each subset individually, obtaining S individual pooled vectors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_S\}$. Finally, the pooled vectors are concatenated as a long vector $\mathbf{F}$ which is the output of the BoF model.

3 Generalized Regular Spatial Pooling

Different spatial pooling algorithms define different index subsets $\{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_S\}$. In this section, we illustrate the definition of index subsets in Spatial Pyramid Matching (SPM) [1] and the proposed Generalized Regular Spatial Pooling (GRSP) algorithm.

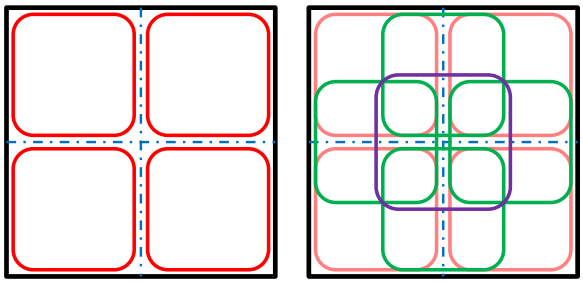

Figure 1: An example of original (left) and denser (right) spatial pooling in the 1st layer (pooling bin size is $\frac{W}{2} \times \frac{H}{2}$). We set $s_1 = 3$, so that each pooling bin shares half of its pixels with its neighboring bins.

3.1 Spatial Pyramid Matching

The Spatial Pyramid Matching (SPM) [1] algorithm, also known as Regular Spatial Pooling, defines the number of layers L for spatial matching, and divides the image region recursively into subregions for feature pooling.

Mathematically, let $\mathcal{P}$ be the set of pixels in image $\mathbf{I}$. We also define the (only one) pooling bin in the zeroth layer as $\mathcal{P}_{0,0} = \mathcal{P}$. For $l \geqslant 0$ and $0 \leqslant t < 4^l$, we divide the t-th pooling bin in the l-th layer as 4 bins in the $l+1$-th layer, *i.e.*, $\mathcal{P}_{l+1,4t} = \mathcal{P}_{l,t}^{\mathrm{UL}}$,$\mathcal{P}_{l+1,4t+1} = \mathcal{P}_{l,t}^{\mathrm{UR}}$, $\mathcal{P}_{l+1,4t+2} = \mathcal{P}_{l,t}^{\mathrm{LL}}$ and $\mathcal{P}_{l+1,4t+3} = \mathcal{P}_{l,t}^{\mathrm{LR}}$, where $\mathcal{P}_{l,t}^{\mathrm{UL}}$, $\mathcal{P}_{l,t}^{\mathrm{UR}}$, $\mathcal{P}_{l,t}^{\mathrm{LL}}$ and $\mathcal{P}_{l,t}^{\mathrm{LR}}$ denote the upper-left, upper-right, lower-left and lower-right corners of $\mathcal{P}_{l,t}$, respectively, after $\mathcal{P}_{l,t}$ is divided into 2×2 equal-sized subregions. One can easily see that there are $2^l \times 2^l$ pooling bins with size $\frac{W}{2^l} \times \frac{H}{2^l}$ in the l-th layer. We define the index sets straightforwardly using the pooling bins: $\mathcal{J}_{l,t} = \{m \mid 1 \leqslant m \leqslant M \wedge \mathbf{l}_m \in \mathcal{P}_{l,t}\}$. The number of index sets is equal to the number of pooling bins, $\sum_{l=0}^{L-1} \left(2^l\right)^2$, in the L-layer SPM model.

3.2 Generalized Regular Spatial Pooling

The Generalized Regular Spatial Pooling (GRSP) algorithm follows the basic rules of Spatial Pyramid Matching, but allows the bins within the same pooling layer have either denser or sparser distributions.

First let us still assume the bins in the l-th layer have size $\left(W/2^l\right) \times \left(H/2^l\right)$, *i.e.*, this is the same setting as in the regular spatial pooling algorithm. Then we define a sequence $(s_0 = 1, s_1, s_2, \ldots, s_{L-1})$, which means that there are $s_l \times s_l$ equal-sized pooling bins in the l-th layer. We put a pooling bin with size $\left(W/2^l\right) \times \left(H/2^l\right)$ at the upper-left corner of the image, move the bin along the axis of both sides of the image, from upper-left to lower-right corner, and make sure that the spatial stride in each move is the same. When $s_l = 2^l$, $l = 1, 2, \ldots, L-1$, the proposed method degenerates to Spatial Pyramid Matching, otherwise the pooling bins would become either denser ($s_l > 2^l$) or sparser

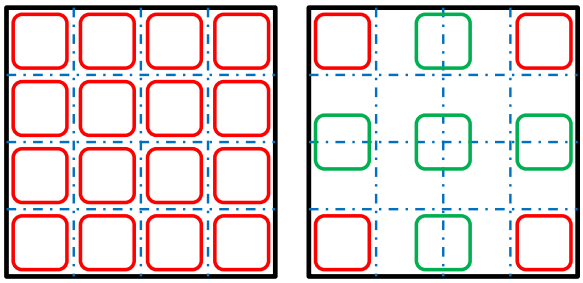

Figure 2: An example of original (left) and sparser (right) spatial pooling in the 2nd layer (pooling bin size is $\frac{W}{4} \times \frac{H}{4}$). We set $s_2 = 3$, so that some regions on the image plane are not occupied by any one of the pooling bins.

($s_l < 2^l$) on the image plane. Figure 1 illustrates the denser spatial pooling on the 1st-layer ($s_1 = 3$), and Figure 2 illustrates the sparser spatial pooling on the 2nd-layer ($s_2 = 3$).

With the pooling bins, we can obtain the index set in the same way as in regular pooling. The number of index sets in GRSP is $\sum_{l=0}^{L-1} s_l^2$. When denser pooling is performed on some layer, some local features could be summarized in more than one bins, while sparser pooling might ignore a fraction of the local features (not included in any bins).

3.3 Comparison to Previous Works

There are many works aimed at providing a better way of spatial pooling beyond Spatial Pyramid Matching [1]. In [6], the authors propose to compute smaller codebooks for feature encoding in the lower levels, while [12] suggests to combine Sparse Coding algorithms with Spatial Pyramids towards better image representation. Maybe the most relevant work to our idea is [7], in which a number of possible bins are extracted on the image plane, and it remains to select a small number of them which best capture the spatial saliency of the images. In [13], it is also suggested to group the local features with orientational pooling bins. In comparison with previous works, Generalized Regular Spatial Pooling (GRSP) is extremely simple: one need only few lines of codes to implement the GRSP algorithm, which only differs from SPM in the way of filtering local features according to their coordinates.

4 Experiments

In this section, we focus on the selection of parameters to improve the classification accuracy. Detailed discussions are also provided to explain the impact of different models and parameters. In the final part, we compare our classification

Table 1: Classification results of different parameters on four widely used image collections.

Case No.	Enco-ding	s_l 0th	s_l 1st	s_l 2nd	Feature Dims	Sport-8 Acc. (%)	Scene-15 Acc. (%)	Indoor-67 Acc. (%)	Caltech101 Acc. (%)
1	LLC	1×1	2×2	3×3	28K	87.28	81.34	43.21	73.24
2	LLC	1×1	2×2	4×4	42K	87.03	81.66	43.55	74.47
3	LLC	1×1	2×2	6×6	82K	86.73	81.76	44.63	75.96
4	LLC	1×1	2×2	8×8	138K	86.46	81.27	44.40	76.18
5	LLC	1×1	3×3	3×3	38K	87.60	81.89	43.22	75.43
6	LLC	1×1	3×3	4×4	52K	87.44	81.83	43.17	75.66
7	LLC	1×1	3×3	6×6	92K	87.09	81.90	45.15	76.70
8	LLC	1×1	3×3	8×8	148K	86.78	81.49	44.86	76.68
9	LLC	1×1	4×4	3×3	52K	87.58	81.48	44.04	75.61
10	LLC	1×1	4×4	4×4	66K	87.56	81.57	44.22	75.96
11	LLC	1×1	4×4	6×6	106K	87.18	81.67	45.07	76.55
12	LLC	1×1	4×4	8×8	162K	86.98	81.43	44.99	76.77
13	IFV	1×1	2×2	–	200K	90.82	87.54	61.22	80.73
14	IFV	1×1	3×3	–	400K	91.38	87.79	62.55	81.86
15	IFV	1×1	4×4	–	680K	91.16	87.75	62.57	82.04

accuracy on several widely used image collections with recently reported results.

4.1 Datasets and Basic Settings

We report the classification accuracy on four widely used image collections.

- **The UIUC Sport-8 dataset [14]** contains 8 sporting scenes and 1579 images. Images are divided into easy and medium difficulties according to their qualities.
- **The Scene-15 dataset [1]** contains 15 scenes and 4485 images. All the instances are grayscale images collected from outdoor environments. It is one of the most widely used dataset for scene understanding tasks.
- **The MIT Indoor-67 dataset [15]** contains 67 indoor scenes and 15620 images. It is a large and challenging dataset for indoor scene recognition.
- **The Caltech101 dataset [16]** contains 9144 images of 102 classes. There exists significant deformation among different objects from the same category.

The basic setting follows the recent proposed BoF models [9] [10]. Images are scaled, with the aspect ratios preserved, so that the larger axis is 600 pixels.

Table 2: Comparison of our classification results with previous works.

Algorithm	UIUC Sport-8	Scene-15	MIT Indoor-67	Caltech101
Yang *et.al.* [12]	–	80.4	–	73.2
Boureau *et.al.* [17]	–	84.3	–	75.7
Jia *et.al.* [7]	–	–	–	75.3
Xie *et.al.* [18]	88.17 ± 0.78	83.77 ± 0.69	46.38 ± 0.75	78.14 ± 0.80
Kobayashi *et.al.* [19]	90.42	85.63	58.91	–
Wang *et.al.* (LLC) [9]	87.10 ± 0.82	81.66 ± 0.36	43.55 ± 0.63	74.47 ± 0.91
Ours (LLC + GRSP)	$\mathbf{87.60} \pm 0.73$	$\mathbf{81.89} \pm 0.50$	$\mathbf{45.15} \pm 0.46$	$\mathbf{76.70} \pm 0.79$
Perronin *et.al.* (IFV) [10]	90.82 ± 0.92	87.54 ± 0.58	61.22 ± 0.65	80.73 ± 0.82
Ours (IFV + GRSP)	$\mathbf{91.38} \pm 0.86$	$\mathbf{87.79} \pm 0.59$	$\mathbf{62.55} \pm 0.45$	$\mathbf{81.86} \pm 0.94$

We use the VLFeat [20] library to extract dense RootSIFT [21] descriptors. The spatial stride and window size of dense sampling are 10 and 16 for all the datasets. The dimension of descriptors are reduced to 80 using PCA in the case of IFV encoding. We then cluster the descriptors with K-Means clustering ($K = 2048$) and Gaussian Mixture Model (GMM, $K = 256$), respectively, for the LLC [9] and IFV [10] encoding methods. The number of descriptors for clustering does not exceed 2 million. We use LLC and IFV algorithms to encode local descriptors, and the encoded vectors are normalized individually within each spatial pooling bin [22]. The number of layers for spatial pooling is 3 for LLC encoding, and 2 for IFV encoding. We will discuss the parameters of the Generalized Regular Spatial Pooling algorithm in the next section. We use LibLINEAR [23], a scalable SVM for evaluating the image representation. For each dataset, we select a fixed number of images per category for training the model, and test it on the remaining images to calculate the average classification accuracy over all the categories. The numbers of training images per category for the four dataset are 70, 100, 80 and 30, respectively. We repeat the random selection 10 times and report the averaged results.

4.2 Models and Parameters

In this section, we observe the impact of different parameters in the Generalized Regular Spatial Pooling algorithm. The results are summarized in Table 1.

First, we compare different settings used with LLC encoding [9]. One can see that, when we increase the number of bins from 2×2 to 3×3 on the 1st layer, the classification accuracy is usually improved (see case pairs $(1, 5)$, $(2, 6)$, $(3, 7)$, and $(4, 8)$). However, when the number is further increased from 3×3 to 4×4, we only observe limited accuracy gain or even accuracy drop (see case pairs $(5, 9)$, $(6, 10)$, $(7, 11)$, and $(8, 12)$). This suggests that denser spatial pooling bins do provide complementary information into image representation, but using too many bins could also introduce considerable redundance which

actually harms the classification accuracy. Similar discipline is also observed on different numbers of pooling bins on the 2nd layer (see case groups $(1, 2, 3, 4)$, $(5, 6, 7, 8)$, and $(9, 10, 11, 12)$). We benefit from the complementary information by increasing the number of pooling bins from 4×4 to 6×6, meanwhile suffer from the redundance introduced by too many (8×8) bins.

Similar discipline is also summarized from the results using IFV encoding [10]. To prevent the image-level features have too high dimensionality, we only use two layers of bins for spatial pooling. When the originally used 2×2 grid is replaced by a 3×3 grid, the classification accuracy is improved significantly, whereas the even denser 4×4 grid does not help much to provide complementary information in image representation.

It is also interesting to observe the relationship between the number of pooling bins and the number of categories in the dataset. In the UIUC Sport-8 dataset, there are only few categories, therefore too high-dimensional feature vectors might cause over-fitting. As the number of categories increases, the benefit of using more pooling bins becomes more and more significant. Taking the results using LLC encoding as the example. On the UIUC Sport-8 dataset, 3×3 grid on the 2nd layer produces the best classification accuracy, whereas on the Caltech101, recognition is more accurate when the pooling bins are denser (*e.g.*, 8×8 on the 2nd layer). This suggests that in the large-scale datasets such as Caltech256 [24], SUN-397 [25] or ImageNet [26], it is instructive to use more pooling bins or even more pooling layers for better image representation.

In conclusion, we use 1×1, 3×3 and 6×6 grids in the 3-layer pooling model with LLC encoding, except for the UIUC Sport-8 dataset in which 3×3 grid is used in the lowest (2nd) layer. It produces 92K-dimensional feature vectors (except for the UIUC Sport-8 dataset in which it is 38K), which is about twice as long as original SPM vectors (1×1, 2×2 and 4×4 grids, 42K dimensions). With IFV encoding, we use 1×1 and 3×3 grids in the 2-layer model on all the datasets, producing 400K-dimensional feature vectors which is of exactly twice length of original SPM vectors (1×1 and 2×2 grids, 200K dimensions).

4.3 Comparison with the State-of-the-Art

Here, we report the classification accuracy with some competitors on all the four datasets. To make the comparison fair, we only compare our algorithm to those only using grayscale SIFT descriptors. As the proposed algorithm is focused on spatial pooling, we do not compare the results with those complex encoding algorithms. The results are listed in Table 2. One can see that our algorithm achieves very competitive classification performance, which outperforms the recent published work [19] on all three scene recognition datasets.

5 Conclusions

In this paper, we aim at providing a better way of spatial context modeling towards image understanding. We propose Generalized Regular Spatial Pooling

(GRSP), which generalizes the Spatial Pyramid Matching (SPM) algorithm by allowing the pooling bins have either denser or sparser distributions on the image plane. Despite the simplicity of our model, it is verified very efficient at several challenging image classification tasks. In the future, we shall investigate the use of our model on the larger-scale datasets.

References

[1] Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. (2006)

[2] Vapnik, V.: Statistical Learning Theory. (1998)

[3] Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. International Conference on Computer Vision (2003)

[4] Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. Workshop on Statistical Learning in Computer Vision, ECCV (2004)

[5] Grauman, K., Darrell, T.: The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. International Conference on Computer Vision (2005)

[6] Liu, X., Wang, D., Li, J., Zhang, B.: The Feature and Spatial Covariant Kernel: Adding Implicit Spatial Constraints to Histogram. International Conference on Image and Video Retrieval (2007)

[7] Jia, Y., Huang, C., Darrell, T.: Beyond Spatial Pyramids: Receptive Field Learning for Pooled Image Features. Computer Vision and Pattern Recognition (2012)

[8] Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal on Computer Vision (2004)

[9] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-Constrained Linear Coding for Image Classification. Computer Vision and Pattern Recognition (2010)

[10] Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-scale Image Classification. European Conference on Computer Vision (2010)

[11] Xie, L., Tian, Q., Hong, R., Yan, S., Zhang, B.: Hierarchical Part Matching for Fine-Grained Visual Categorization. International Conference on Computer Vision (2013)

[12] Yang, J., Yu, K., Gong, Y., Huang, T.: Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. Computer Vision and Pattern Recognition (2009)

[13] Xie, L., Wang, J., Guo, B., Zhang, B., Tian, Q.: Orientational Pyramid Matching for Recognizing Indoor Scenes. Computer Vision and Pattern Recognition (2014)

[14] Li, L., Fei-Fei, L.: What, Where and Who? Classifying Events by Scene and Object Recognition. International Conference on Computer Vision (2007)

[15] Quattoni, A., Torralba, A.: Recognizing Indoor Scenes. Computer Vision and Pattern Recognition (2009)

[16] Fei-Fei, L., Fergus, R., Perona, P.: Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. Computer Vision and Image Understanding (2007)

[17] Boureau, Y., Bach, F., LeCun, Y., Ponce, J.: Learning Mid-Level Features for Recognition. Computer Vision and Pattern Recognition (2010)

[18] Xie, L., Tian, Q., Zhang, B.: Spatial Pooling of Heterogeneous Features for Image Applications. ACM Multimedia (2012)

[19] Kobayashi, T.: (BoF meets HOG: Feature Extraction based on Histograms of Oriented pdf Gradients for Image Classification)

[20] Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms. ACM Multimedia (2010)

[21] Arandjelovic, R., Zisserman, A.: Three Things Everyone Should Know to Improve Object Retrieval. Computer Vision and Pattern Recognition (2012)

[22] Xie, L., Tian, Q., Zhang, B.: Feature Normalization for Part-based Image Classification. International Conference on Image Processing (2013)

[23] Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research (2008)

[24] Griffin, G., Holub, A., Perona, P.: Caltech-256 Object Category Dataset. Technical Report (2007)

[25] Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun Database: Large-scale Scene Recognition from Abbey to Zoo. Computer Vision and Pattern Recognition (2010)

[26] Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: A Large-scale Hierarchical Image Database. (2009)