Spatial Pooling for Transformation Invariant Image Representation

Xia Li
Dept. of Computer Science
University of Texas
at San Antonio
TX, 78249, U.S.A
xial@cs.utsa.edu

Yan Song
Department of EEIS
University of Science and
Technology of China, Hefei,
230027, China
songy@ustc.edu.cn

Yijuan Lu
Dept. of Computer Science
Texas State University
San Marcos, TX, 78666,
U.S.A
yl12@txstate.edu

Qi Tian
Dept. of Computer Science
University of Texas
at San Antonio
TX, 78249, U.S.A
qtian@cs.utsa.edu

ABSTRACT

Spatial Pyramid Matching (SPM) [2] has been proposed to extend the Bag-of-Word (BoW) model for object classification. By reserving the finer level information, it makes image matching more accurate. However, for not well-aligned images, where the object is rotated, flipped or translated, SPM may lose its discrimination power. To tackle this problem, we propose novel spatial pooling layouts to address various transformations, and generate a more general image representation. To evaluate the effectiveness of the proposed approach, we conduct extensive experiments on three transformation emphasized datasets for object classification task. Experimental results demonstrate its superiority over the state-of-the-arts. Besides, the proposed image representation is compact and consistent with the BoW model, which makes it applicable to image retrieval task as well.

Categories and Subject Descriptors

I.4.8 [Scene Analysis]: Object recognition.

General Terms

Algorithms, Experimentation, and Verification.

Keywords

Image Representation, Spatial Transformation, and Object Classification.

1. INTRODUCTION

Bag-of-Word model (BoW) has shown its superiority over many conventional global features in image classification and retrieval systems [5, 7, 11-13, 19, 20]. However, it has been shown that quantization error may degrade the performance of image classification and retrieval. To address this issue, many methods have been proposed recently. They can be categorized into three groups: 1) improving the codebook quality and reducing quantization error [5, 7], 2) constructing a robust visual representation, *e.g.* visual phrase [11], and 3) utilizing the geometric information to remove false matched feature pairs, *e.g.* bundled feature [12], spatial coding [13], and spatial matching [14].

In this paper, we mainly focus on image matching method for object recognition. As BoW discards the spatial order of local features, the descriptive power of its image representation is limited. The previously proposed spatial pyramid matching (SPM) [2] provides a good solution for this problem. Typically, SPM partitions the image into regions on multiple resolution levels. As shown in Fig. 1, each column displays the 3-level SPM partition. The final image representation is the concatenation of the BoW representations of all the regions. This "subdivide and reorder" strategy achieves more accurate matching by reserving the spatial information. Although SPM can greatly improve the performance of image classification on some benchmark datasets, including Caltech101 [3] and Scenes15 [15], it cannot handle object transformations. Three object transformations are illustrated in Fig. 1. The guitars in image (a) and (b) are rotated different angles. The ducks in image (c) and (d) are flipped horizontally. The bottles in image (e) and (f) appears at different locations. We can see that the corresponding SPM regions of similar images fail to match in the three cases. Our experiments demonstrate that SPM performs even worse than BoW with translation variance. Spatial Bag-of-Visual Word (SBoW) [8] is proposed to extend SPM. The image retrieval experiments show the effectiveness of SBoW. However, its image representation consists of thousands of long histograms, which are too complex for real-time application. So labeled data and boosting algorithm are used to select the most useful histograms for a given dataset.

Different from SBoW, this paper aims to find out a compact image representation robust to rotation, flipping and translation variances. We propose three kinds of spatial layouts: spatial pyramid ring, reordered SPM, and relative SPM, which are evaluated by extensive experiments on different transformation datasets. Experimental results show that compared with SPM and SBoW, our approaches can handle these spatial variances more efficiently and effectively. Furthermore, the simplicity of the proposed layouts makes it a promising potential for real time applications.

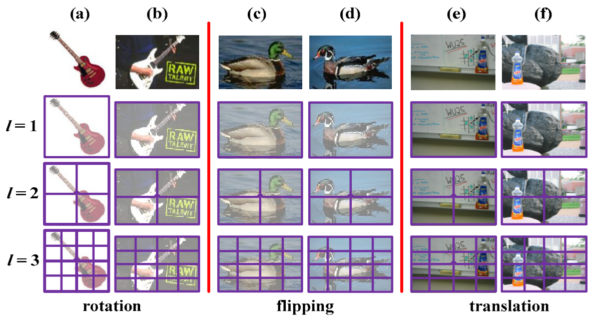

Figure 1. SPM division on rotated, flipped and translated objects.

*Area Chair: Nicu Sebe.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11...$10.00.

We will introduce our approaches in Section 2. In Section 3, the experiments and results are given. Finally, discussion and future work are concluded in Section 4.

2. THE PROPOSED APPROACH

Our objective is to construct a spatial transformation invariant image representation. We will first introduce our rotation- and flip- invariant method, and then the translation-invariant method will be introduced next.

2.1 Rotation and Flipping Invariance

An object may rotate any angle on the image plane, like the exemplars of electronic guitar in Fig. 1(a) and (b), or flip horizontally or vertically, like the duck images in Fig. 1(c) and (d). We can see that different spatial transformations may cause the same visual features fall into different SPM bins, which degrades the discrimination power of SPM. Note that in this section, we assume images are well-aligned after rotation or flipping.

2.1.1 Spatial Pyramid Ring (SPR)

To handle object rotation like image (a) and (b) in Fig. 1, SBoW calibrates sectors according to a certain structure pattern. However, to deal with any rotation angle in the meanwhile avoiding being too strict, the appropriate sector number needs to be predetermined, which is a challenging task. Furthermore, SBoW is not flipping free. As we can see from the flipping example in Fig. 1, when object structure is flipped, partial shifting method cannot find matching pairs, as the order of sectors is reversed.

In order to solve this problem, we propose a pooling method based on spatial pyramid ring, which can better handle both rotation and flipping transformations. This is inspired from the observation that visual words falling into a spatial ring area remain stable on rotated or flipped objects.

We first partition the image into concentric rings on the polar coordinate. Then we compute BoW histogram within each ring. Motivated from the "subdivide and reorder" strategy of SPM, we also build a spatial pyramid to represent the image. Assume we have L-level pyramid, on the l-th level; the image is divided into 2^{l-1} concentric rings from the image center. The step size between rings is set to D/2^{l-1}, where D denotes the half diagonal length of the image. Fig. 2 displays our spatial pyramid ring partitions on the horizontally flipped duck images. This method will yield $\sum_{l=1}^{L} 2^{l-1}$ histograms for one image based on the rings. All the histograms can be concatenated to form a long histogram for representing the entire image.

Fig. 2 illustrates that this simple ring layout is flipping invariant as the corresponding rings of the two images are still matched. It also works for rotations. This mainly benefits from the fact that the same parts of an object will remain within a certain ring area in spite of rotation or flipping. And the pyramid representation can be matched on different resolution levels of the rings structure.

2.2 Translation Invariance

In this section, we will address the translation transformation problem. As shown in Fig. 1(e) and (f), objects may appear anywhere in an image, which cannot be handled by the layout based on SPM. Hence, we further propose two schemes: Reordered SPM and relative SPM.

2.2.1 Reordered SPM

The reason that SPM fails under translation transformation is that SPM requires matching each region pairs correspondingly between any two images. This problem can be fixed to some extent by resetting the matching order of SPM cells.

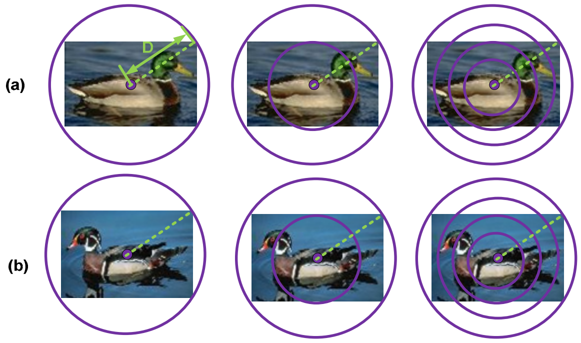

Figure 2. Illustration of SPR on image (a) and (b). From left to right are 1st, 2nd, and 3rd pyramid level.

First, we build SPM with L levels, which are $\sum_{l=1}^{L} 2^{2\times(l-1)}$ cells in total. Then on the l-th level, we sort the $2^{2\times(l-1)}$ bins for each visual word channel according to either ascending or descending order and form the re-ordered histogram for each visual word channel. For example, Fig. 3 contains three visual words (represented by yellow triangle, green star, and blue square) in the image (a) and (b), and it displays the SPM partition on the 2nd level. The right of the image shows the term frequency of each visual word in every region in descending order. Then the histograms of all visual word channels are connected to form the l-th level histogram representation, and finally representations of all the levels are concatenated together to form the entire image representation. With reordered SPM, objects appearing within different regionss will be matched because their histograms on each word channel have been aligned well.

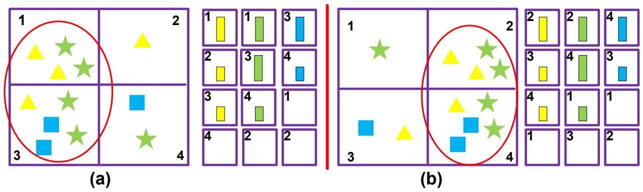

Figure 3. Illustration of Reordered SPM for translation invariance. Image (a) and (b) have the same object circled by a red ellipse. Images are divided into 2×2 grids. Their right parts show the reordered histograms from top to bottom for each visual word colored in yellow, green and blue separately.

2.2.2 Relative SPM

In the quantization and pooling step of image classification, some recent work has shown that Locality-constrained Linear Coding (LLC) [7] with max pooling [6] can yield good performance. LLC softly quantizes local features to their k nearest words. Max pooling then selects the nearest one feature for a certain visual word. Hence the selected feature can be regarded as salient points and we are able to derive the specific location where a visual word appears in an image.

We propose to utilize the geo information of salient points generated from the coarser level of SPM to generate the current level grids. For example, after max pooling on the l-th level, we can compute the center of the selected salient points within each region on this level, then each center is used as the origin point to divide its belonging region into another 2×2 grids for the (*l*+1)-th level. See the toy example in Fig. 4 for reference.

We call this method relative SPM as its grid division is relative to the distribution of salient points. This method benefits from the approximate tracking of the object position.

3. EXPERIMENTS

In this section, we will verify the proposed spatial pooling methods for object classification on three different spatial transformation datasets including flipping, rotation and translation variance datasets.

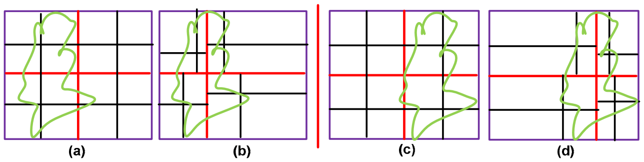

Figure 4. Illustration of Relative SPM. The same object is sketched in green curve. (a) and (c) are the original SPM partition. (b) and (d) show the partition of our approach. Red line is the 2nd level partition and black line is the 3rd level partition.

3.1 Experimental Settings

Image representation: The Bag-of-Word histogram is adopted. We use *K*-Means clustering to generate visual codebooks of size from 512 to 2048. After quantizing local feature descriptors into visual words, each image is represented by a visual word histogram. Here, Locality-constrained Linear Coding [7] is used for feature quantization, and max pooling [6] is used for histogram generation due to their sound performance on standard image classification benchmarks. To extract local feature descriptors, we adopt two popular methods: SIFT [1] detected from sparse regions and densely sampled SIFT. It has been shown that dense SIFT sampling gives better performance than sparse region of interest detectors, especially for object classification. This is because the so called "brute force" metric of dense sampling which covers all the information with overlap. However, the SIFT descriptors generated by dense sampling are not scale and orientation invariant. So we choose the appropriate one to avoid the feature level influence in different experiments. For dense sampling, we extract SIFT descriptors from 16×16 patches with step size 4. For sparse region of interest detection, we use the open source software VLFeat [9] with default parameters.

Model training: Three datasets are generated to test flipping, rotation, and translation variances respectively. We use LIBSVM [10] with linear kernel for object classification. For every object, we randomly select 30 images for training. The left data are used for testing. We run the experiments five times and calculate the average classification accuracy of all categories for each run.

Baseline system: We use BoW to denote the baseline image level classification. SPM denotes adding SPM on the baseline method, where max pooling is performed within each cell. Three-level SPM is adopted with 1+4+16=21 cells.

3.2 Flipping Invariance Experiment

Flipping Dataset: To test flipping invariant methods, we assume no other spatial variance exists, *e.g.* objects are center-aligned, and occupy the most part of an image with similar scale. We use Caltech101 [3] with 9144 images of 101 categories in this experiment and we randomly flip images horizontally or vertically. Flipping samples are shown in Fig. 5.

We use dense SIFT descriptors to cover the whole image. Although dense SIFT is not flipping invariant, It has been proposed in [18] to combine the metrics of both dense sampling and sparse detectors to make the extracted feature more stable and repeatable. Therefore, in our experiments, we make an assumption that there is a good way to address the feature level variance to avoid its influence on feature matching. With this assumption, we can focus on analyzing the spatial transformation. Hence, the descriptors from the non-flipped images are used for the flipped images instead, but the location information of descriptors are flipped.

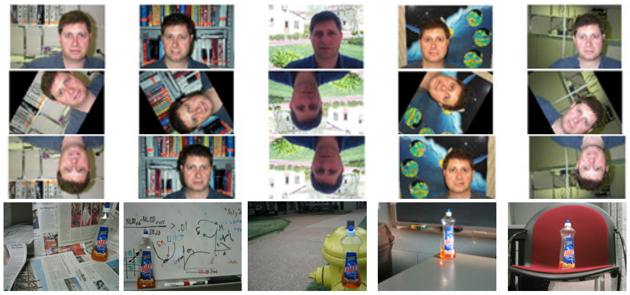

Figure 5. Image samples. The first three rows are from Caltech101: 1st row shows the original images; 2nd row shows images after random rotation; 3rd row shows images after randomly horizontal or vertical flipping. The last row shows sample images from the "AjaxOrange" category of SIVAL.

In the experiment, we combine BoW model with three different spatial coding algorithms: SPM [2], calibrated circular projection (CCP) in SBoW [8], and spatial pyramid ring (SPR) pooling. The codebook size is 2048 as it is widely used in previous work [6]. For CCP, the images are divided into 1, 4, and 16 sectors. In our spatial pyramid ring pooling, 1, 2, 4, and 8 rings are constructed from the 1st to 4th level, which yields 1+2+4+8=15 rings. We test the three algorithms on the flipping dataset and obtain the average classification accuracy shown in Table 1.

Table 1. Caltech101 classification accuracy comparison on flipping and rotation data

Method	Before transformation	After flipping	After rotation
BoW	52.17%	-	26.81%
SPM[2]	74.34%	57.94%	29.84%
CCP[8]		57.04%	29.23%
SPR		**66.61%**	**34.95%**
SPR + SPM		**70.19%**	**35.20%**

The accuracy of SPM drops to 57.94% from 74.34% after flipping. The performance of CCP is close to SPM. Our proposed SPR gives the best improvement of 14% over the image level classification, *i.e.*, BoW, and about 8.6% improvement over SPM and CCP. These results clearly demonstrate that spatial pyramid ring pooling is very effective against flipping variances.

3.3 Rotation Invariance Experiment

Rotation Dataset: We randomly rotate the images in Caltech101 with an angle between 0 and 360 counterclockwise. See Fig. 5 for sample images.

In this section, we want to show that spatial pyramid ring pooling also works well with rotation variance. As we mentioned above, dense SIFT is not rotation invariant. So in this experiment, we use the sparse region of interest detector to extract SIFT features.

Experimental results of the three spatial pooling methods are shown in the last column of Table 1. Both SPM and CCP improve the performance. However, spatial pyramid ring pooling method achieves the best performance. We further test combining SPM with SPR pooling by connecting their histograms together to form the final image representation. It outperforms all other methods on both flipping and rotation.

3.4 Translation Invariance Experiment

Translation Dataset: To test translation invariant methods, we use SIVAL benchmark [17]. This dataset contains 1, 500 images in 25 categories. An image contains a primary object located randomly and independently on the image with diverse backgrounds. Objects do not occupy the entire image, but have similar scale at most time. Since the biggest variance of this dataset is translation, we use it to evaluate the methods for handling translation variance only. Fig. 5 shows some sample images from this dataset.

Table 2. SIVAL classification accuracy

Codebook Size	512	1024	2048
BoW	51.03%	64.86%	76.22%
SPM[2]	26.39%	32.63%	39.57%
Relative SPM	34.31%	40.27%	46.15%
Reordered SPM	59.50%	69.94%	77.83%
Reordered Relative SPM	61.29%	71.53%	78.94%

We test BoW, SPM, Relative SPM, Reordered SPM, and Reordered Relative SPM on this dataset. Their classification accuracy with different codebook sizes are listed in Table 2. We can see SPM fails on this dataset. It is even worse than image level classification. The Relative SPM method is designed for adjusting the division of SPM to make it focus on the object. This method does improve the SPM performance by about 7%. The most effective method is Reordered SPM, which gains improvement over the image level classification. Combining the two methods together achieves the best performance on this dataset. This improvement is relatively less than what can be gained by using SPM on well-aligned image datasets as can be seen from Table 1. This means without object localization, it is hard to do fine level matching globally due to the large translation variance of objects and irrelatively diverse background noise. We don't compare our methods with SBoW on translation problem, as it requires a complex learning process to determine which line projections should be used.

4. CONCLUSIONS

In this paper, we propose several effective spatial pooling methods to handle various spatial transformations. The generated spatial transformation invariant image representation can improve the entire image matching very well. Experimental results show that Spatial Pyramid Ring (SPR) pooling works best for rotation and flipping variances. A refined version of SPM (Relative Reordered SPM) can improve the matching accuracy with large translation variance.

Furthermore, the proposed approaches are not limited to classification task. They also can be applied on object detection and image retrieval applications as well. That is because we invariantly embed the spatial layout information into classical BoW image representation, which ensures that all current techniques following the image representation step can be seamlessly applied.

In the future, we will continue to address other spatial transformations of objects, *e.g.* scale and view angle change. We also would like to fuse these methods together to design a unified representation, which can be tolerant with all kinds of layout variances.

5. ACKNOWLEDGEMENT

This work is supported in part to Dr. Qi Tian by NSF IIS 1052851, Faculty Research Awards by Google, FXPAL and NEC Laboratories of America, in part to Dr. Yan Song by Anhui Provincial Natural Science Foundation under grant 090412056, and in part to Dr. Yijuan Lu by Research Enhancement Program (REP) and start-up funding from the Texas State University respectively.

6. REFERENCES

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

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

[3] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," *CVPR Workshop on Generative-Model Based Vision*, 2004.

[4] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *In proc. CVPR*, 2005.

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

[6] Y. Boureau, F. Bach, Y. LeCun, J. Ponce, "Learning mid-level features for recognition," *In proc. CVPR*, 2010.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," *In proc. CVPR*, 2010.

[8] Y. Cao, C. Wang, Z. Li, L. Zhang, "Spatial-bag-of-features," *In proc. CVPR*, 2010.

[9] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," 2008, http://www.vlfeat.org/

[10] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/-cjlin/libsvm.

[11] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive Visual Words and Visual Phrases for Image Applications," *In ACM MM*, 2009.

[12] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," *In proc. CVPR*, 2009.

[13] W. Zhou, Y. Lu, H. Li, Y. Song, Q. Tian, "Spatial coding for large scale partial-duplicate web image search," *In ACM MM*, 2009.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," *In proc. CVPR*, 2007.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *IJCV*, 2001.

[16] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *IJCV*, 2009

[17] SIVAL: http://accio.cse.wustl.edu/sg-accio/SIVAL.html

[18] T. Tuytelaars, "Dense interest points," *In proc. CVPR*, 2010.

[19] M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, Y. Song, "Unified Video Annotation Via Multi-Graph Learning," *IEEE CSVT*, 2009.

[20] M. Wang, X. Hua, T. Mei, R. Hong, G. Qi, Y.Song and L. Dai, "Semi-Supervised Kernel Density Estimation for Video Annotation," *CVIU*, 2009.