Spatial Statistics for Spatial Pyramid Matching Based Image Recognition

Toshihiko Yamasaki*†‡ and Tsuhan Chen*
* Cornell University, Ithaca, United States
E-mail: ty273@cornell.edu
† The University of Tokyo, Tokyo, Japan
‡ JSPS Postdoctoral Fellow for Research Abroad, Japan

***Abstract*—This paper presents an image feature extraction algorithm that enhances the object classification accuracy in the spatial pyramid matching (SPM) framework. The proposed method considers the spatial statistics of the feature vectors by calculating the moment vectors. While the original SPM algorithm captures the spatial distribution of the image feature descriptors, the proposed algorithm describes how such spatial distribution is variant. The experiments are conducted using two state-of-the-art SPM-based methods for two commonly used datasets. The results demonstrates the validity of our proposed algorithm. The cases where the proposed algorithm works well are also investigated. In addition, it is demonstrated that the proposed feature and adding more layers improve the classification accuracy in different situations.**

I. Introduction

This paper proposes additional features for the SPM-based image recognition, which use the statistical information among the feature vectors in sub-regions. In conventional SPM, the input image is divided into $1\times1, 2\times2, 4\times4$ sub-regions. And the features are extracted from each sub-region. Although the spatial distributions of the local descriptors are somehow encapsulated in the spatial pyramid structure, the statistical information such as variance among the sub-blocks has not been used so far. We demonstrate such spatial statistics helps us to improve the classification accuracy. The proposed feature has different effects from adding more layers. Therefore, it is also possible to combine them to achieve better classification accuracy.

Experiments were conducted using sparse coding SPM (Sc-SPM) [1] and locality-constrained linear coding SPM (LLC-SPM) [2] for Caltech-101 [3] and Caltech-256 [4]. The results demonstrate that the proposed feature representation outperforms the conventional SPM-based approaches. In addition, the proposed method is better than simply adding one more layer of 3×3 when the intra-class variance is relatively small (i.e., the Caltech-101 dataset). And, it is comparable to the additional layer approach even when the intra-class variance is large (i.e., the Caltech-256 dataset). It is also demonstrated that the proposed features work very well in some cases although the averaged classification accuracy improvement seems small. The proposed feature can be incooperated into other algorithms as long as they extract features from multiple sub-regions.

The organization of this paper is as follows. Related works are summarized in Section II. In Section III, the proposed feature extraction algorithm based on the spatial statistics is described. The experimental results showing the validity of our algorithm is demonstrated in Section IV. An application example is presented in V. The concluding remarks are given in Section VI.

II. Related Work

Bag-of-features (BoF) representation [5], [6] in conjunction with spatial pyramid matching (SPM) [7] is a *de facto* standard approach for image classification and object recognition. In the BoF representation, local appearance descriptors are extracted using such as scale-invariant feature transform (SIFT) [8], histograms of oriented gradients (HoG) [9], etc. and they are encoded and pooled to form a feature vector. If the feature vector is generated only for the whole input image, information of the spatial layout of the descriptors is discarded. On the other hand, if the input image is partitioned into small pieces of sub-regions and the feature vector is generated from each of them, spatial-layout-aware feature vectors would be generated but at the same time they would be too sensitive to spatial variations. To overcome this problem, the SPM was proposed, which partitioned the image into increasingly fine-grid sub-regions and extracted histograms of local descriptors in each sub-region. The concept is shown on the left half in Fig. 1. In the SPM, the global feature was captured in the coarse-grid histograms and local spatial information was described in the finer-grid histograms.

After the success of the BoF model with the SPM, a lot of techniques have been proposed to enhance the image classification accuracy. For the BoF representation, instead of doing hard voting [5] using vector quantization (VQ), soft code assignment techniques were developed using Gaussian mixture model (GMM) [10], distance-in-feature-space-based soft code assignment [11], code word uncertainty (UNC) model [12], and so on. Yang *et al.* [1] demonstrated that the sparse coding (Sc) with max pooling outperformed conventional histogram-based feature generation (i.e., average pooling). Then, locality-constrained linear coding (LLC) [2] was proposed to ensure that descriptors that were closer in the feature space were assigned similar code words whereas the sparse coding did not guarantee such constraint.

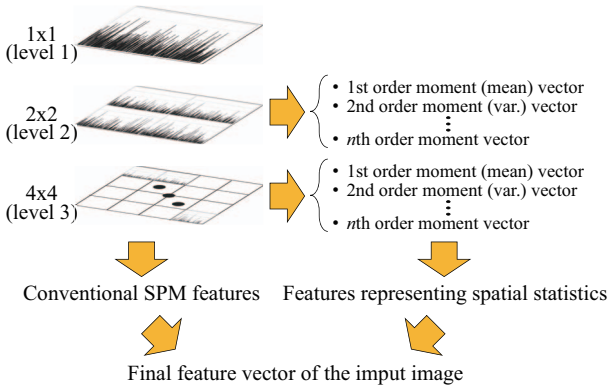


Fig. 1. Concept of the proposed feature extraction method.

Regarding the SPM, partitioning strategy and weight assignment to sub-regions are often discussed. For instance, the pyramid structure of $1\times1, 2\times2, 4\times4$ is employed in [1], [2], [7] while $1\times1, 2\times2, 1\times3$ is also popular [13], [14], [15], [16]. Lazebnik *et al.* [7] used a pyramid match kernel to calculate the weights for each level of spatial pyramid. In [17], pyramid kernel learning was proposed to obtain the optimal kernel fusing weights from multiple scales, locations, as well as codebooks instead of doing the brute-forth search as in [18]. Harada *et al.*, on the other hand, proposed a discriminative spatial pyramid representation, which formed the image feature as a weighted sum of semi-local features over all pyramid levels [19]. Besides, simply concatenating the feature vectors is also frequently used [1], [2].

Previous works that consider the spatial distribution of the descriptors more in detail [7], [17], [18], [19] use only the SPM structure and assign weights to the sub-regions. The difference of our proposed algorithm is the spatial variance of the features are statistically analyzed and used as additional features.

III. Feature extraction based on spatial statistics

A. Feature extraction algorithm

The concept of the proposed algorithm is illustrated in Fig. 1. In addition to the conventional SPM-based features, statistical information from the 1st order moment (mean) vector up to the nth order moment vector is extracted for each layer of the spatial pyramid. For instance, in the layer of level 2, the average (1st order moment) vector is obtained by averaging the $2\times2 = 4$ feature vectors. Note that the moments are calculated for each layer considering all the sub-regions in the layer, not in each sub-region. The dimension of the moment vectors is same as that of the feature vectors extracted from the sub-regions, e.g., the codebook size. The SPM-based features and the extracted moment vectors are then concatenated to form a final feature vector of the input image. Using the moments in the feature vectors is common in image retrieval [20], image classification [21], and so on. However, such spatial statistics has not been discussed in the SPM framework so far as far as the we know.

The proposed spatial statistics contributes in two ways. First, such statistics analyze how the extracted features (i.e., the response to the codebook) are spatially variant. Although the local spatial information is embedded in the conventional hierarchical spatial pyramid, how such information is variant among the others is not captured. The other contribution is its non-linear feature representation of the extracted SPM-based features. In most cases, a linear SVM is used for the classification because both the number and the dimension of the feature vectors are very large. By adding non-linear features and projecting the feature vectors to a higher dimensional space helps the SVM to define better hyper-plain in many cases. One might argue that such non-linear projection is not mathematically optimal. However, even when a non-linear SVM is employed, users have to conduct try-and-error-based kernel selection from Gaussian kernels, polynomial kernels, and so on along with optimal parameter setting.

The dimension of the generated feature vectors is

$$(1^2 + 2^2 + 4^2)B + 2mB$$

for the $1\times1, 2\times2, 4\times4$ configuration where m is the number of moments and B is the codebook size. The extra memory usage and computational cost are the penalty we have to pay.

B. The cases where the proposed algorithm works better

One might think that it is apparent that the performance would be improved if we use more features such as an additional layer of 3×3[1] or 8×8 sub-blocks to the original SPM. Hereafter, we call SPM $(1\times1, 2\times2, 4\times4)$, SPM $(1\times1, 2\times2, 4\times4, 8\times8)$, and SPM $(1\times1, 2\times2, 3\times3, 4\times4)$ as 3-layer SPM, 4-layer SPM, and additional-layer SPM, respectively. In fact, as demonstrated in Section IV, the averaged performance improvement of our proposed algorithm as compared to such approaches seems relatively small. However, there are some cases in which the proposed algorithm works much better than the original SPM and the SPM with additional layers. Here, we classify such cases into three categories.

1) Case 1: Objects are shifted from the center: Although objects seem to be placed in the center of the images, it is not always true. In some cases, the objects are simply shifted from the center and placed in the upper half or lower half of the images as shown in Fig. 2(a). In other cases, although the objects are aligned in the center, the the characteristic features may exist in different sub-regions. For instance, in Fig. 2(b), the black spots of the wild cats exists in the upper half in the left image but they are in the left half in the right image depending on the posture of the wild cat. The conceptual comparison of the generated feature vectors is shown in Fig. 2(c). In the original SPM method, the generated feature vectors are totally different because the the descriptors are in different sub-regions. On the other hand, the spatial statistics (i.e., the moment values of the sub-regions) is the almost the same.

2) Case 2: Background features are uniformly distributed over the image: Since the images usually contain not only

[1]though it is recommended that the l-th layer have $2^{l-1}\times2^{l-1}$ sub-blocks, it is also possible to add such a layer.

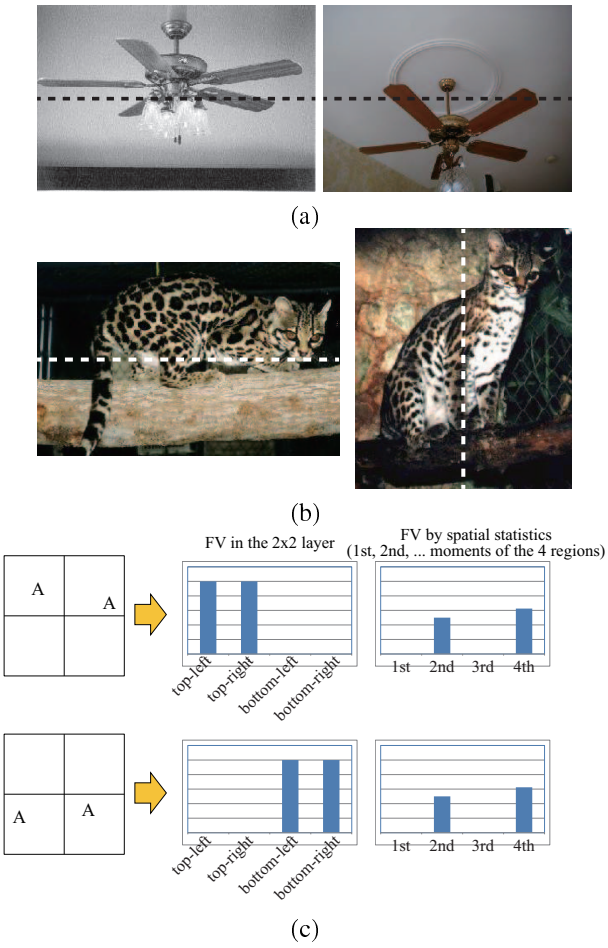


Fig. 2. Case 1: the proposed works better when the objects are shifted from the image center. (a) The objects are simply in different regions. (b) The objects are almost in the center of the image but the characteristic features are in different regions. (c) Generated features in such cases where "A" represents the code assigned to the descriptors.

the target objects to classify but also unrelated background, the descriptors extracted from the background region can become a noise to degrade the classification performance. When the background images have uniformly distributed texture as shown in Fig. 3(a), the moments for the codes in background would become close to zero as explained in Fig. 3(b) and yield no information while some meaningful information is extracted from the moments for the foreground objects.

3) Case 3: Image similarity/dissimilarity is enhanced: Adding another layer such as 3×3 sub-blocks tries to capture the distribution of the descriptors with different spatial coarseness. On the contrary, the proposed spatial statistics provides different aspect of of the spatial distribution of the descriptors. The similarity or dissimilarity of the images can be enhanced in some cases. This also means that if the intra-class variance is large, the proposed work would not contribute to the performance improvement.

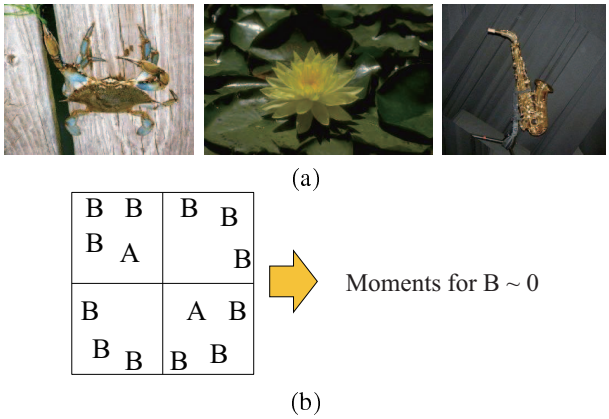


Fig. 3. Case 2: the proposed works better when the background texture is uniform. (a) Sample images. (b) Generated features in such cases where "A" and "B" represent the code assigned to the descriptors of the foreground objects and that to the background, respectively.

C. Combining multiple approaches

The additional-layer SPM, on the other hand, considers a finer level of details. It works better than the SPM + spatial statistics in a different situation: when the object in the image is well aligned. Since the conditions where the additional-layer SPM works better and those where the SPM + spatial statistics works better are different from each other, it is also possible to combine the multiple approaches. One possible approach is comparing the confidence values yielded from the classifiers as proposed in [22].

IV. Experimental Results

A. Experimental setup

In this paper, two state-of-the-art approaches are employed: ScSPM [1] and LLCSPM [2]. Note that the proposed algorithm is general and can be applied to any other approaches as long as they extract feature vectors from sub-regions such as the spatial pyramid framework. The proposed algorithm has been implemented based on the source codes available on the authors' project page[2] in order to purely focus on the difference between the original SPM and our proposed method. Our proposed work is also compared with the SPM-based method with an additional layer of 3×3 and 8×8 (only for Caltech-101 dataset due to the limitation of the computational resources). The parameter for the constraints violation (C) is optimized for each case.

Our descriptor extraction followed the one in [1]. Namely, the SIFT descriptors extracted from 16×16 pixel patches were densely sampled from each image on a grid with step size of 6 pixels. The images were all preprocessed into gray scale. Note that the local descriptor used in [2] was HoG [9] but in this paper SIFT is employed because the source code mentioned above also assumed SIFT. A simple linear SVM was used as the classifier.

[2]http://www.ifp.illinois.edu/ jyang29/resources.html

The order of the moment n was set up to 6. And all the possible combination of moments were tested. In the combination method, the accuracy for each class was investigated using the test data by an one-leave-out method and the most probable class was taken.

B. Overall Performance for Caltech-101 Dataset

The Caltech-101 dataset [3] contains 9,145 images in 101(+1 for background) classes including a variety of objects (animals, vehicles, flowers, etc), with large variance in shape. On the other hand, the position, orientation, and size of the objects are roughly aligned. The number of images per category was from 31 to 800. The images were resized to be no larger than 300×300 pixels with preserved aspect ratio. We mostly followed a common testing procedure as in: 30 samples were randomly picked up from each class and the rest of them were used for testing. This process was repeated 5 times and the performance was measured by using average accuracy over 102 classes (102 accuracy values were averaged). The codebook sizes were set as 1,024 for Sc and 2,048 for LLC, respectively, by following the settings in [1], [2].

The performance comparison between the SPM and the proposed method is shown in Table I. The classification accuracy is not from the papers [1], [2] but is obtained by running the authors' source codes. It is shown that the SPM + spatial statistics works better than the 3-layer SPM (by 1.8% for Sc and 2.2% for LLC) and the additional-layer SPM (by 0.8% for Sc and 0.7% for LLC). It is even better than the 4-layer SPM ($1\times1, 2\times2, 4\times4, 8\times8$) by 0.5-0.6%. In fact, the 4-layer SPM is only 0.2% better than the additional-layer SPM even though the feature vector dimensions are very different: $30B$ for the additional-layer SPM and $73B$ for the 4-layer SPM. This indicates that the proposed feature grabs another aspect of the spatial distribution of the descriptors and is different from simply adding extra SPM layers.

We applied Welch's t test to determine the statistical significance of the differences. In Sc, both additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$) and SPM + spatial statistics were better than the original SPM and the difference is significant ($p < 0.05$). The difference between the additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$) and the SPM + spatial statistics were marginally significant ($p < 0.1$). It can be observed that although the dimension of the feature vectors in our proposed work is slightly smaller than the additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$), we can achieve better classification performance. In LLC, the difference between the original SPM and the other two is statistically significant ($p < 0.01$). The difference between our proposed method and the additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$) is also statistically significant ($p < 0.05$).

Table II shows the performance comparison with recent related works. We can observe that the proposed work outperforms even some of the state-of-the-art works [19], [23], [24], [25], [26]. Although there are some works that yields better accuracy than ours such as [27], [28], [29], they also employ the SPM framework. Therefore, we believe that our algorithm can also be incooperated into such approaches.

TABLE II
CLASSIFICATION RATE (%) COMPARISON ON CALTECH-101. THE RESULTS WITH (*) WERE OBTAINED BY EXECUTING THEIR SOURCE CODE.

Algorithm	Accuracy
SPM [7]	64.40±0.80
Caltech-256 [4]	67.60
NBNN [30]	70.40
ML+CORR [31]	69.60
Kernel Codebook [12]	64.14±1.18
ScSPM [1]	73.2±0.54
*ScSPM [1]	*72.1
LLCSPM [2]	73.44
*LLCSPM [2]	*71.2
Code Relation [23]	74.25
D-SP [19]	67.21±0.67
LC-KSVD2 [24]	73.60
RLDA [25]	73.7±0.8
Hie Sc [26]	74.00
Proposed	
ScSPM + spatial statistics	73.9
LLCSPM + spatial statistics	73.4

C. Detailed Analysis for Caltech-101 Dataset

If only the average accuracy is considered, the performance improvement of the SPM + spatial statistics seems relatively small. However, the SPM + spatial statistics works much better than the 3-layer SPM and the additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$) in some classes. Actually, there are 12 classes where the proposed algorithm works more than 5% better than the other two approaches and there are three classes where it works more than 5% worse out of the 102 classes. Such classes and the performance comparison are summarized in Tables III and IV. And sample images are shown in Figs. 4 and 5. For instance, Fig. 4(a) and Fig. 4(l) correspond to the case 1 in Section III-B and some images in Fig. 4(b), Fig. 4(e), Fig. 4(f) and Fig. 4(i) correspond to the case 2. The objects in Fig. 4 have less intra-class variance in their appearance as compared to the other classes. For example, inline skates (Fig. 4(d)), metronome (Fig. 4(g)), and musical instruments (Fig. 4(c) and Fig. 4(j)) are non-rigid industrial products and the objects in the same class look very similar to each other. Even the wild animals such as platypus (Fig. 4(i)) and wild cat (Fig. 4(l)) have very small intra-class variances in their textures.

On the other hand, the images in the butterfly class in Fig. 5 are very different from each other in terms of texture although they are all butterflies. In the case of camera and tick classes, the intra-class variances do not seem very large, but they contain a few groups of clearly different appearances. For example, the images in Fig. 5(b) can be classified to two different sub-classes: compact camera and single-lens reflex (SLR) camera. If such different sub-classes are mixed in a single class, the performance of our proposed algorithm would be degraded.

In order to clarify the advantage and disadvantage of the SPM + spatial statistics, we conducted further experiments

TABLE I

COMPARISON BETWEEN CONVENTIONAL SPM AND PROPOSED METHOD USING THE CALTECH-101 DATASET. (*)THE RESULTS OF SCSPM AND LLCSPM WERE OBTAINED BY EXECUTING THE SOURCE CODES DOWNLOADED FROM THE AUTHORS' SITES, NOT FROM THEIR PAPERS.

	3-layer SPM ($1\times1, 2\times2, 4\times4$)	additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$)	4-layer SPM ($1\times1, 2\times2, 4\times4, 8\times8$)	SPM + spatial statistics
Sc [1]	*72.1	73.1	73.3	73.9 (with 2nd, 4th, 5th and 6th moments)
LLC [2]	*71.2	72.7	72.9	73.4 (with 1st, 3rd, and 6th moments)

TABLE V

PERFORMANCE IMPROVEMENT OF SC BASED ON WHICH MOMENTS TO USE.

Used moments	Accuracy
2,4,5,6	73.9
1,5,6	73.8
1,2,3	73.6
...	
2,5	71.7
2,3,4,5,6	71.6
1,2	71.5

using the camera class. We split the images into two subclasses: compact camera and SLR camera as shown in Fig. 6. Then, the three classifiers (3-layer SPM, additional-layer SPM and SPM + spatial statistics) are re-trained. Note that the task here is only 2-class classification. The results are demonstrated in Fig. 7. The figure shows the classification accuracy as a function of the number of training data for each class. It can be observed that the SPM + spatial statistics always works better than the other two and also the performance difference becomes more obvious when the training data are increased. For example, when the number of training data per class is 5, the accuracy is 78%, 79%, and 83%, respectively. If the number of training data is increased up to 15, the accuracy is 81%, 82%, and 88%, respectively. The results support our claim that similarity and dissimilarity of images are enhanced (case 3 in Section III-B).

The condition for the additional-layer SPM to work better or worse than the other two approach is simple. Since it considers the distribution of the descriptors more strongly, it works better when the objects are well aligned and vice versa. Some sample images are shown in Figs. 8 and 9. It is observed that the orientation and/or the size of the objects are different from each other in Fig. 9.

Table V shows how the image classification accuracy is affected by the spatial statistics. When the moments are not properly selected, the performance would become worse than the original SPM. Which moments contribute to the performance enhancement was not clear.

D. Caltech-256 Dataset

The Caltech-256 dataset holds 30,608 images in 256 categories with much higher intra-class variability and higher object location variability compared with Caltech-101. Each category contains at least 80 images. Same as in IV-B, all the images were resized so that they are no larger than 300×300 with aspect ratio being preserved. We trained a codebooks with 4,096 bases. The number of training data was 30 for each object class.

(a) (b)

Fig. 6. New definition of the two camera classes: (a) compact, (b) SLR.

TABLE VII

CLASSIFICATION RATE (%) COMPARISON ON CALTECH-256. THE RESULTS WITH (*) WERE OBTAINED BY EXECUTING THEIR SOURCE CODE.

Algorithm	Accuracy
Caltech-256 [4]	34.10
Kernel CB [12]	27.17
ScSPM, [1]	34.02
*ScSPM, [1]	34.0
LLCSPM [2]	41.19
*LLCSPM [2]	*35.8
D-SP (tr=15), [19]	30.24
LR-Sc+SPM (tr=15), [27]	35.31
Proposed	
ScSPM + spatial statistics	39.0
LLCSPM + spatial statistics	36.3

As shown in Table VI, the performance of our proposed method is better than the original 3-layer ScSPM and LLCSPM even with the Caltech-256 dataset. It is improved by 5% in the Sc case and 0.4% in the LLC case, respectively. However, the performance is equivalent to that of the additional-

TABLE III
CLASSES IN WHICH THE SPM + SPATIAL STATISTICS WORKS MORE THAN 5% BETTER THAN THE OTHER TWO APPROACHES.

	3-layer SPM ($1\times1, 2\times2, 4\times4$)	additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$)	SPM + spatial statistics
ceiling fan	64	64	71
crab	42	41	49
garfield	75	75	84
inline skate	60	80	88
lotus	48	48	53
mandolin	77	77	84
metronome	80	80	94
octopus	24	36	43
platypus	45	55	81
saxophone	66	66	76
schooner	68	68	74
wild cat	35	30	41

TABLE IV
CLASSES IN WHICH THE SPM + SPATIAL STATISTICS WORKS MORE THAN 5% WORSE THAN THE OTHER TWO APPROACHES.

	3-layer SPM ($1\times1, 2\times2, 4\times4$)	additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$)	SPM + spatial statistics
butterfly	60	59	53
camera	82	83	74
tick	84	86	79

TABLE VI
COMPARISON BETWEEN CONVENTIONAL SPM AND PROPOSED METHOD USING CALTECH-256 DATASET. (*)THE RESULTS OF SCSPM AND LLCSPM WERE OBTAINED BY EXECUTING THE SOURCE CODES DOWNLOADED FROM THE AUTHORS' SITES, NOT FROM THEIR PAPERS.

	3-layer SPM ($1\times1, 2\times2, 4\times4$)	additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$)	SPM + spatial statistics
Sc [1]	*34.0	39.1	39.0 (with 1st, 3rd, 4th, and 6th moments)
LLC [2]	*35.8	36.6	36.3 (with 1st, 2nd, and 5th moments)

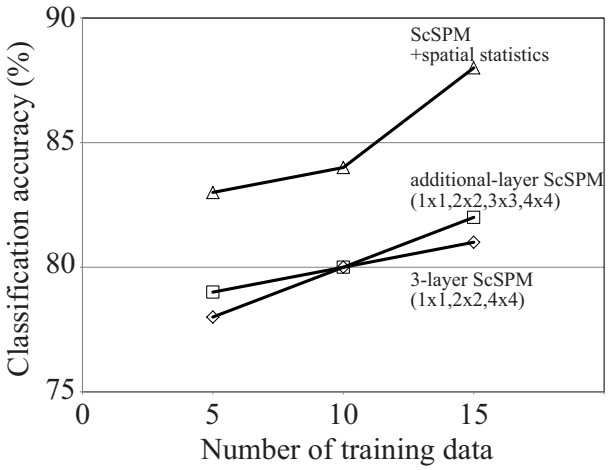


Fig. 7. Classification accuracy of compact cameras and SLR cameras as a function of the number of training data. Three algorithms are compared.

layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$). No statistical significance was observed between the two approaches. This is because the objects in the images are not aligned well and there are large intra-class variations. Not only that the appearance of the objects are different from each other, but also there are sometimes unrelated objects in the background, multiple objects in a single image, and so on. Such noise makes spatial statistics less informative than the Caltech-101 case. The 4-layer SPM ($1\times1, 2\times2, 4\times4, 8\times8$) was not evaluated for this dataset due to the limitation of computational resources.

The comparison with previous works is shown in Table VII. Same as the experiments in Caltech-101, there are some algorithms which perform better than our algorithm [28], [29]. We would like to emphasize that it is not fair to compare our work with such approaches because the algorithms are totally different. As shown in Table VI, the proposed algorithm outperforms the baseline approach (3-layer SPM).

E. Discussion

So far, we discussed our spatial statistics in the spatial pyramid framework because it is one of the *de facto* standards in object recognition. Some of the state-of-the-art works propose "post-SPM" spatial pooling. For example, in [32], [33], receptive field learning with overcomplete rectangular bins is used instead of simply dividing the input image into 2×2, 4×4, and so on. As mentioned in IV-A, the proposed algorithm is a general idea and we believe that it can also be incooperated into such sophisticated algorithms.

The extra cost for calculating the spatial statistics is negligibly small as compared to the SIFT feature description and code assignment. This is one of the advantages of the proposed work because most of the state-of-the-art algorithms requires a large amount of computation. The memory usage is also smaller than the additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$)

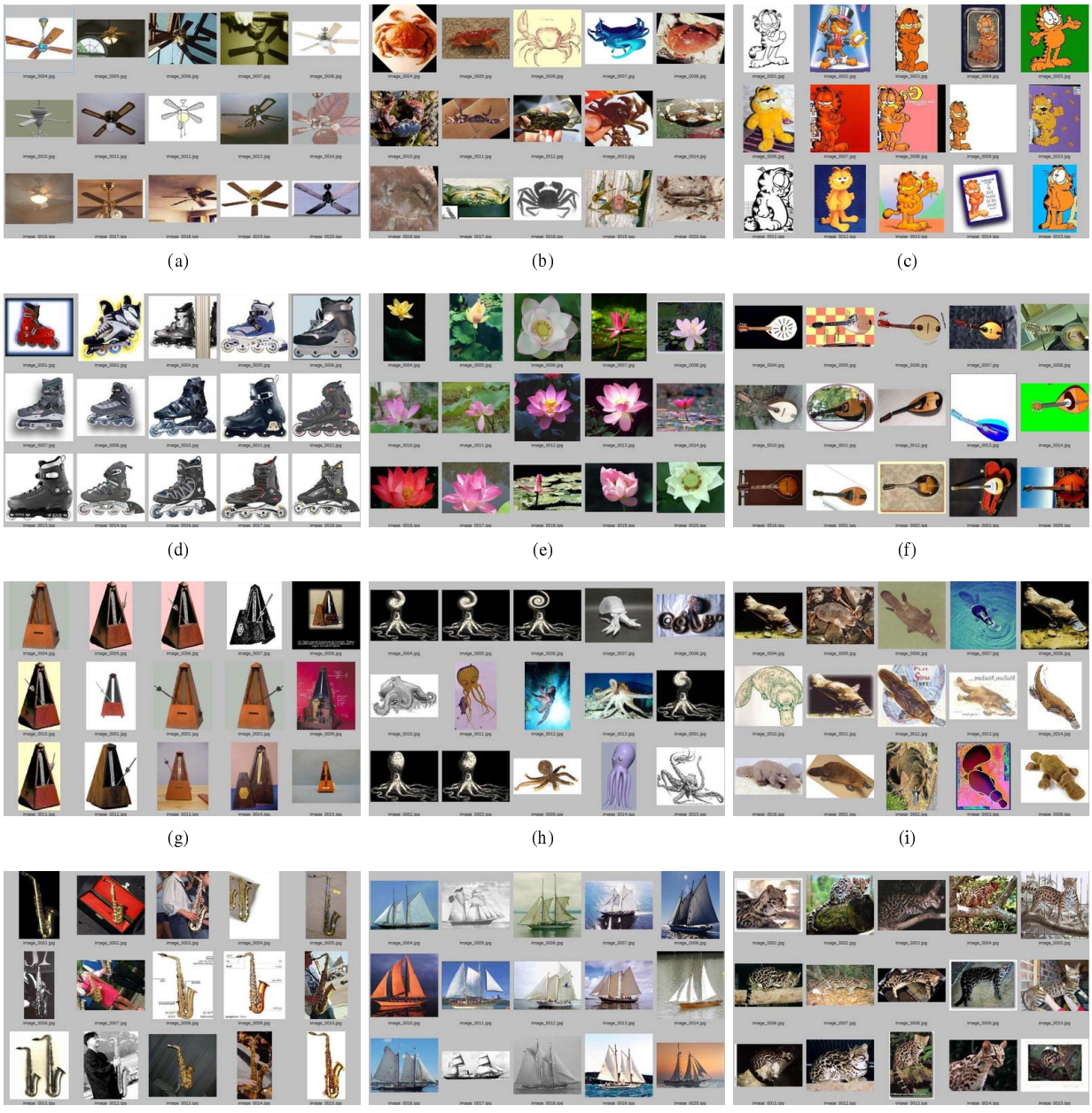


Fig. 4. Sample images of the classes where the SPM + spatial statistics work more than 5% better than the other two: (a) ceiling fan, (b) crab, (c) garfield, (d) inline skate, (e) lotus, (f) mandolin, (g) metronome, (h) octopus, (i) platypus, (j) saxophone, (k) schooner, (l) wild cat.

and the 4-layer SPM ($1\times1, 2\times2, 4\times4, 8\times8$) while achieving better results. This is a non-trivial matter when the size of the data becomes quite large.

As shown in the previous section, the proposed work does not work well when the intra-class variation is very large (though it is still better than the original SPM). From this point of view, the proposed feature representation would work well with such classifiers that can respond to specific exemplars [34].

V. Application

In this section, another application is demonstrated. Recently, an avatar CAPTCHA (completely automated public Turing test to tell computers and humans apart) system [35], which asks users to classify human faces and avatar faces, have been proposed. The object recognition/classification such as face classification is more challenging task than character recognition, but humans can intuitively solve the problem. In [35], 63% of the users successfully classified 12 human/avatar faces and more than 90% of the users showed positive response to the system. Sample images of their human/avatar faces are

(a) (b) (c)

Fig. 5. Sample images of the classes where the SPM + spatial statistics work more than 5% worse than the other two: (a) butterfly, (b) camera, (c) tick.

(a) (b) (c)

Fig. 8. Sample images of the classes where the additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$) work more than 5% better than the other two: (a) beaver, (b) crocodile head, (c) lamp.

(a) (b) (c)

Fig. 9. Sample images of the classes where the additional-layer SPM ($1\times1, 2\times2, 3\times3, 4\times4$) work more than 5% worse than the other two: (a) binocular, (b) hedgehog, (c) strawberry.

shown in Fig. 10.

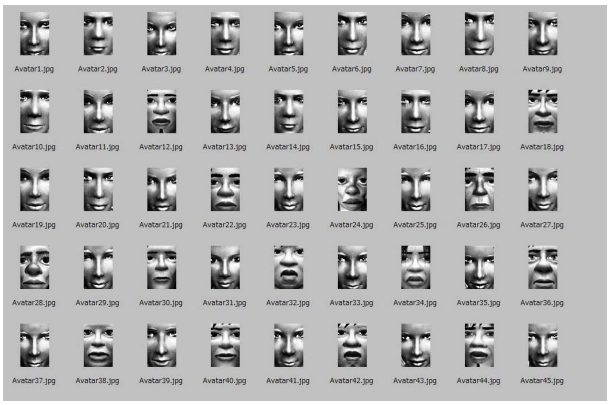
Conventional approaches consider how real images and CG images are "statistically" different. Therefore, they are mainly based on low-level features. On the other hand, since the texture of faces is different, object classification approaches can also be applied to human/avatar face classification. In particular, as discussed in III-B, the proposed algorithm enhances the similarity/dissimilarity of the input images.

The proof-of-concept experiments were conducted with 50 avatar face images and 50 human face images [35]. The avatar faces were collected from virtual communities such as Entropia Universe and Second Life. In our experiment, k images were randomly sampled from each class for training and the rest were used for testing. k was changed from 1 to 40. This procedure was repeated 100 times and the average accuracy was calculated. The proposed algorithm is compared with two different method: one is a simple SVM-based method using the raw pixel values and the other is ScSPM [1]. It is shown that the proposed algorithm rarely fails in distinguishing human/avatar faces. For more details, please refer to [36].

VI. Conclusions

The paper presented a feature representation algorithm using spatial statistics for SPM-based image classification. By calculating the moments of feature vectors in sub-regions in each level and concatenating them to the SPM-based feature vectors, spatial variances of the feature distribution were considered. Experimental results using Caltech-101 and Caltech-256 have demonstrated that the proposed feature extraction algorithm can improve the classification accuracy as compared to the original 3-layer SPM. The proposed algorithm was more effective than simply adding a layer for Caltech-101 and comparable to it for Caltech-256.

Although the averaged performance improvement over the additional-layer SPM seems limited, the proposed method worked more than 5% better in as many as 12 classes while



(a)

(b)

Fig. 10. Sample images: (a) avatar, (b) human.

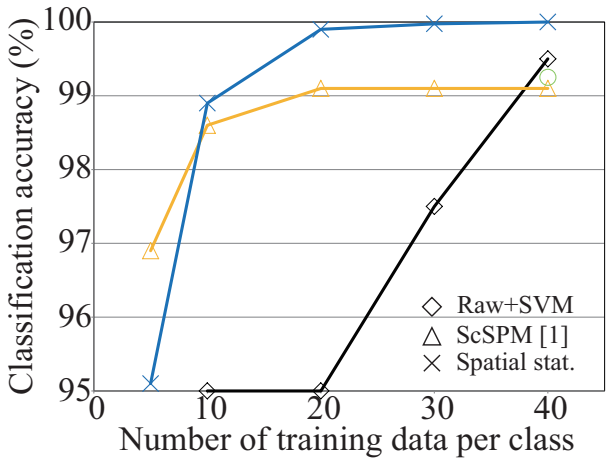


Fig. 11. Classification accuracy as a function of the number of training data per class.

it worked more than 5% worse only in three classes for the Caltech-101 dataset. We also clarified the cases where the proposed algorithm works better.

REFERENCES

[1] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

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

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

[4] Griffin, G. Holub, and P. AD. Perona. Caltech-256 object category dataset. In *Technical Report 7694, California Institute of Technology*, 2007.

[5] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

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

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[10] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. In *Technical report, University of Southampton*, 2005.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: improving particular object retrieval in large scale image databases. In *CVPR*, 2008.

[12] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, 2008.

[13] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. In *Visual Recog. Challange workshop*, 2007.

[14] F. Perronnin, J. S'anchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[15] J. Yang, K. Yu, and T. Huang. Efficient highly over-complete sparse coding using a mixture model. In *ECCV*, 2010.

[16] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.

[17] J. He, S.-F. Chang, and L. Xie. Fast kernel learning for spatial pyramid matching. In *CVPR*, 2008.

[18] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.

[19] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi. Discriminative spatial pyramid. In *CVPR*, 2011.

[20] M. Stricker and A. Dimai. Color indexing with weak spatial constraints. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, volume 2670, pages 29–39, 1996.

[21] K. Kitamura, T. Yamasaki, and K. Aizawa. Foodlog: Capture, analysis and retrieval of personal food images via web. In *Proc. Workshop on Multimedia for Cooking and Eating Activities*, pages 23–29, 2009.

[22] T. Yamasaki and T. Chen. Confidence-assisted classification result refinement for object recognition featuring topn-exemplar-svm. In *ICPR (in press)*, 2012.

[23] Y. Huang, K. Huang, C. Wang, and T. Tan. Exploring relations of visual codes for image classification. In *CVPR*, 2011.

[24] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR*, 2011.

[25] S. Karayev, M. Fritz, S. Fidler, and T. Darrell. A probabilistic model for recursive factorized image features. In *CVPR*, 2011.

[26] K. Yu, Y. Lin, and J. Lafferty. Learning image representations from the pixel level via hierarchical sparse coding. In *CVPR*, 2011.

[27] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma. Image classification by non-negative sparse coding, low-rank and sparse decomposition. In *CVPR*, 2011.

[28] J. Feng, B. Ni, Q. Tian, and S. Yan. Geometric l_p-norm feature pooling for image classification. In *CVPR*, 2011.

[29] N. Kulkarni and B. Li. Discriminative affine sparse codes for image classification. In *CVPR*, 2011.

[30] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[31] P. Jain, B. Kullis, and K. Grauman. Fast image search for learned metrics. In *CVPR*, 2008.

[32] Y. Jia and C. Huang. Beyond spatial pyramids: Receptive field learning for pooled image features. In *NIPS 2011 Workshop Deep Learning and Unsupervised Feature Learning*, 2012.

[33] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*, 2012.

[34] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[35] D. D. Souza, P. C. Polina, and R. V. Yampolskiy. Avatar captcha: Telling computers and humans apart via face classification. In *EIT*, pages 1–6, 2012.

[36] T. Yamasaki and T. Chen. Face recognition challenge: Object recognition approaches for human/avatar classification. In *ICMLA (in press)*, 2012.