IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 23, 2019, accepted January 12, 2020, date of publication January 27, 2020, date of current version February 5, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2969783*

# An Enhancement to the Spatial Pyramid Matching for Image Classification and Retrieval

**PRIYABRATA KARMAKAR, SHYH WEI TENG, GUOJUN LU, AND DENGSHENG ZHANG**
School of Science, Engineering and Information Technology, Federation University Australia, Gippsland Campus, Churchill, VIC 3842, Australia
Corresponding author: Priyabrata Karmakar (p.karmakar@federation.edu.au)

This work was supported in part by the Australian Research Council Discovery Projects Scheme under Grant DP130100024.

**ABSTRACT** Spatial pyramid matching (SPM) is one of the widely used methods to incorporate spatial information into the image representation. Despite its effectiveness, the traditional SPM is not rotation invariant. A rotation invariant SPM has been proposed in the literature but it has many limitations regarding the effectiveness. In this paper, we investigate how to make SPM robust to rotation by addressing those limitations. In an SPM framework, an image is divided into an increasing number of partitions at different pyramid levels. In this paper, our main focus is on how to partition images in such a way that the resulting structure can deal with image-level rotations. To do that, we investigate three concentric ring partitioning schemes. Apart from image partitioning, another important component of the SPM framework is a weight function. To apportion the contribution of each pyramid level to the final matching between two images, the weight function is needed. In this paper, we propose a new weight function which is suitable for the rotation-invariant SPM structure. Experiments based on image classification and retrieval are performed on five image databases. The detailed result analysis shows that we are successful in enhancing the effectiveness of SPM for image classification and retrieval.

**INDEX TERMS** Spatial pyramid matching, rotation invariance, image classification, image retrieval.

## I. INTRODUCTION

Over the last decade, bag of words (BOW) [1] has become one of the most successful image representations to be used in image classification and retrieval tasks. In BOW, local descriptors, like scale invariant feature transform (SIFT) [2], are extracted from all the images in a database, followed by clustering the local descriptors of training images to obtain a visual word dictionary. By encoding the local descriptor set of each image with the visual word dictionary, each image is represented by a histogram of visual words. Although BOW is a popular approach, it lacks spatial information. To overcome this issue, spatial pyramid matching (SPM) [3] was proposed. SPM divides an image into multiple partitions at different levels to form a pyramid of grid partitions, such that $n^{th}$ pyramid level has $2^{2n}$ number of partitions. Each grid partition is then represented by a histogram of visual words. Concatenated histograms obtained from all the grid partitions of a particular pyramid level is the image-level representation of that level. Level-wise similarity scores are obtained by applying the histogram intersection kernel between the corresponding pyramid levels of two images. Finally, similarity scores obtained from each pyramid level are aggregated using a weight function to get the final matching between two images.

SPM has established itself as a popular method in image processing applications for its simplicity and computational efficiency. Subsequently, many researchers have worked on SPM to enhance its performance. Although several images in a database could be similar except for the orientation of their visual contents, research on effectively making SPM robust to rotation is limited. Based on our knowledge, the only such work is proposed in [4]. However, in [4], the proposed method has limitations that will be addressed in this paper. Specifically, the main contribution of this paper is to make enhancements to [4] by addressing its limitations for more effective image classification and retrieval.

The preliminary results of our work have been presented in [5]. Since then, more work has been carried out. The main difference between this paper and [5] are (1) Two additional partitioning schemes are investigated in search of better rotation-invariant image representation. (2) A new weight function is proposed to accurately apportion the similarity scores obtained from each pyramid level to the final

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

similarity score. (3) To test the effectiveness of rotation-invariant SPM, along with image classification, image retrieval is also used as the measure of performance evaluation. (4) More extensive image databases are used for the experiments.

Recently, different deep learning architectures have shown higher effectiveness in various image processing applications compared to the traditional feature descriptors (e.g. SIFT [2], HOG [6] ). However, the traditional feature descriptors have their advantages. First of all, the deep learning architectures are often huge with many layers of neurons with millions of parameters and require large set of training images for the appropriate tuning and weight adjustments. In contrast, traditional feature descriptors can perform satisfactorily with a small number of training images and fewer parameters need to be tuned [7], [8]. In addition, deep learning is generally treated as a black-box and therefore, it is still not clear what visual information is being represented in the complex features derived from a deep learning method. In some application domains, e.g. in crime investigation and presentation of evidence in the court, it is essential to explain how a computer algorithm or method derives its results. Traditional feature descriptors are easier to explain and visualize [9]. Due to the aforementioned reasons, further research and development of traditional feature descriptors are still important and essential. Therefore, in this paper, our focus is to improve an existing feature descriptor framework.

The structure of the rest of the paper is as follows. Section II summarises the existing research works to improve the traditional SPM. The limitations of existing rotation invariant SPM [4] are discussed in Section III. Section IV presents the investigation carried out to find the appropriate partitioning scheme which is suitable to design a rotation invariant SPM structure. A new weight function is proposed in Section V. Section VI provides the details of experimental study. Finally, Section VII concludes the paper.

## II. RELATED WORK

In this section we discuss the concept of traditional SPM (TrSPM) followed by a review of relevant literature and TrSPM's limitation towards rotation.

### A. OVERVIEW OF TRADITIONAL SPM

A pictorial representation of TrSPM is given by Figure 1. The main stages of TrSPM are as follows.

1) **Image partitioning:** To represent images using SPM, each image is partitioned into increasing number of grids in the increasing order of grid levels. At Grid Level $n$, an image is partitioned into $2^{2n}$ numbers of grid partitions. In Figure 1, image partitions are shown for three grid levels (up to Grid Level 2) which is optimum as per [3]. In general, grid levels are called pyramid levels.
2) **Image-level descriptor extraction:** After partitioning images in different grid levels, each of the partitions is represented with a histogram of visual words using a fixed size dictionary. Representation of each partition is the same as representing each partition with BOW. After that, histograms belonging to a particular Grid Level are concatenated to form the corresponding image-level descriptors.
3) **Image matching:** After representing grid levels, the matching between two images at each grid level is performed using histogram intersection given by (1). Next, the match sores obtained by matching each grid levels are summed up using (2) to obtain the final match between two images.

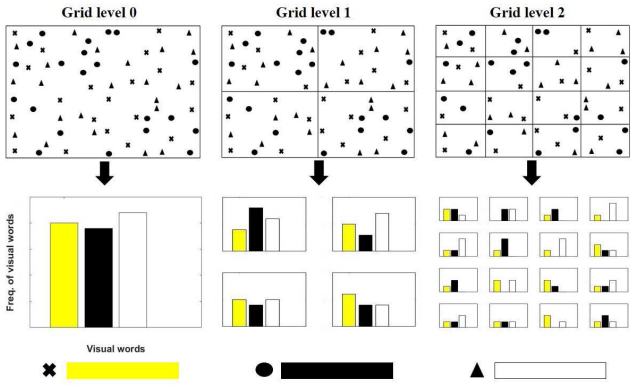



**FIGURE 1.** The concept of TrSPM.

$$M(A, B) = \sum_{j=1}^{r} min\ (A_j, B_j) \tag{1}$$

where $A$ and $B$ are histograms with $r$ bins, and $A_j$ (or $B_j$) denotes the count of the $j^{th}$ bin of $A$ or $B$.

$$K(Image1,\ Image2) = \sum_{i=0}^{L} (w_i N_i) \tag{2}$$

where $w_i$ represents the weight associated with Grid Level $i$ given by (3), $L$ represents the corresponding numerical value of the highest resolution level, $N_i = m_i - m_{i+1}$; $m_i$ and $m_{i+1}$ represent the matches found at Grid Levels $i$ and $i+1$ respectively. As the matches found at Grid Level $i$ also contains all the matches found at Grid Level $i+1$. Therefore, the new matches found at Grid Level $i$ is given by $N_i$.

$$w_i = 1/2^{L-i} \tag{3}$$

### B. RELEVANT LITERATURE

After [3], several works have been proposed to modify TrSPM. In this section, the main ones are discussed. In TrSPM, spatial information is incorporated using grid sampling (i.e. the way images are partitioned in different grid levels). Subsequently, various researchers have proposed other methods of sampling to achieve better representation. It has been shown that the performance of SPM increases as the number of grid levels is increased up to four [10].

A different pyramidal structure was proposed in [11] and it was later also adopted by [12] and [13], where the first two levels are like the TrSPM, but the final level consists of only three horizontal partitions. Grid sampling was extended beyond the fixed spatial pyramids in [14] where a comprehensive set of grids are densely sampled over location, size and aspect ratio. To improve the image representation, scene geometry [15] is used as an input parameter for generating the spatial pyramid definitions. Whereas, in [16] apart from the geometric information, photometric aspects of the images are also captured to distinguish different images more effectively. A fast-deformable spatial pyramid matching algorithm [17] was introduced for computing dense pixel correspondences to enforce both appearance agreements between the matched pixels as well as the geometric smoothness between the neighbouring pixels. To characterise the image layout by various patterns, randomized spatial partition is proposed in [18] to extract the most descriptive image layout pattern for each category and combine them thereafter by training a discriminative classifier.

To increase the effectiveness of SPM, sparse coding [19] was used in SPM framework. Based on SIFT and sparse coding, a hierarchical spatial pyramid max-pooling method was proposed in [20]. Another approach of sparse coding was proposed in [21] where different weights are assigned to the patches of different levels . Whereas in [22], the reconstruction error which is the result of sparse coding in SPM framework is eliminated. When the database is very large or the dictionary size is too high, the resulting image representation turns out to be very high dimensional. As high-dimensional descriptors are inefficient to process, many researchers have focused on dimension reduction of descriptors by sacrificing the overall performance to a least extent [13], [23]–[28].

### C. LIMITATION OF TRADITIONAL SPM

TrSPM shows good performance in image processing applications and there are many subsequent improvements proposed for it to achieve better performance in various aspects as discussed above. However, it is not robust to any kind of image or object rotation. This limitation is discussed in detail with the use of Figure 2.

Consider the two images in Figure 2, both images contain a common scene, but the second one is 180 degrees rotated from the first one. A star object exists in both images. In the first image, the star object exists near to the left top area but in the second image, the same object exists near the bottom right area. When matching these two images using the traditional SPM, there is no problem with the matching at Level 0. This is because all descriptors representing the objects are still within the corresponding partitions of the two images. However, when matching at Level 1, the object of interest is in the first partition of Image 1 but in the fourth partition of Image 2. Thus, the TrSPM may indicate that the two images are very different even though visually they are very similar. The same limitation arises at Level 2 as well.

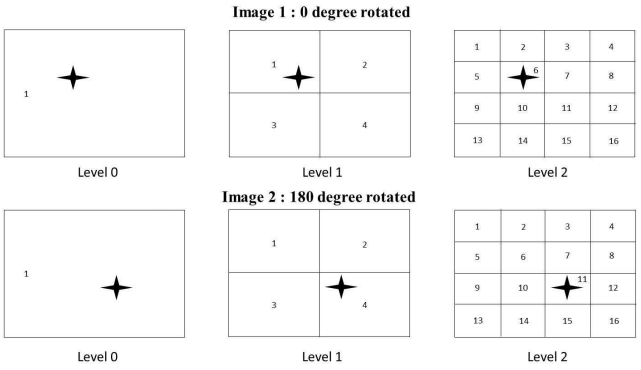



**FIGURE 2. Limitation of TrSPM; Image 1: 0 degree rotated; Image 2 : 180 degrees rotated.**

## III. LIMITATIONS OF THE EXISTING ROTATION INVARIANT SPM

To deal with the rotational issue of TrSPM, the main concern is to preserve the spatially close descriptors in the corresponding partitions of similar images with different rotations. To do this, in [4] authors have proposed spatial pyramid ring (SPR) approach which partitions images into circular concentric rings in different pyramid levels.

Although the authors [4] have claimed that SPR addressed the rotational issue, it still has limitations regarding the effectiveness. Moreover, the result analysis is also not elaborated. The limitations of SPR are given as follows.

1) In SPR, each of the concentric rings is captured over the entire image region and at each level, the number of rings is doubled to its previous level. Hence, at Level 4, there are 8 rings. The probability that the rings may too small and sensitive to any object translation and movement is very high. Therefore, we will investigate the use of a linear increase of rings in the successive pyramid levels.
2) In [4], there is no discussion of using a weight function in the image matching process. Authors have either used the same approach as in [3] or the match scores obtained from the individual pyramid levels are given equal importance to the final matching. In both cases, image matching is not as accurate as it should be with an appropriate weight function. Therefore, to apportion the contribution of descriptors resulting from individual pyramid levels to the final matching between two images, a suitable weight function is proposed in this paper.
3) In SPR, it was aimed to address the SPM's limitation caused by image-level rotation. We have explained this issue using Figure 2. The authors of SPR have used conventional key point-based SIFT which is rotation invariant. However, being a local descriptor, its rotation invariance property is confined in an image patch or a local image region only. Therefore, conventional SIFT has a limited contribution to the image-level rotation.

**TABLE 1. Details of proposed partitioning schemes.**

| RRP | CORP | CIRP |
|---|---|---|
| For a $N \times M$ image, at Level $i$ ( $i > 0$), a $N/(i+1) \times M/(i+1)$ block is first extracted from the centre of the image. This partition is referred as $r\_P_1^i$ which is the first partition at Level $i$. After that, a rectangular ring (if $i = 1$) or several rectangular rings (if $i > 1$) are extracted from the rest of image. The $j^{th}(j > 1)$ partition $r\_P_j^i$ at Level i of an image will be a rectangular ring between the $jN/(i+1) \times jM/(i+1)$ block and $(j-1)N/(i+1) \times (j-1)M/(i+1)$ block. Both blocks are centred at the image centre and $jN/(i+1) \times jM/(i+1)$ block is the entire image when $j = i+1$. RRP scheme of an image up to Level 2 is shown by Figure 3(a). | For a $N \times M$ image, at Level $i$ (for $i > 0$), a circular partition with radius $(\sqrt{N^2+M^2}/2(i+1))$ is extracted with referenced to centre of the image. This partition is referred as $co\_P_1^i$ which is the first partition at Level $i$. Next, a circular ring (if $i = 1$) or several circular rings (if $i > 1$) are extracted from rest of the image. The $j^{th}$ $(j > 1)$ partition $co\_P_j^i$ of the image at Level $i$ (for $i > 0$) will be the circular ring between the circular sections with radii $j\sqrt{N^2+M^2}/2(i+1)$ and $(j-1)\sqrt{N^2+M^2}/2(i+1)$ respectively from the centre. CORP scheme is illustrated using Figure 3(b) where partitions are shown up to Level 2. | For a $N \times M$ image ($N < M$), at Level $i$ (for $i > 0$), a circular partition with radius $N/2(i+1)$ is extracted with referenced to centre of the image. This partition is referred as $ci\_P_1^i$ which is the first partition at Level $i$. Next, a circular ring (if $i = 1$) or several circular rings (if $i > 1$) are extracted from rest of the image. The $j^{th}(j > 1)$ partition $ci\_P_j^i$ of an image will be a circular ring between the circular sections with radii $jN/2(i+1)$ and $(j-1)N/2(i+1)$ respectively from the centre. CIRP scheme up to Level 2 is demonstrated using Figure 3(c) (considering $N = M$). |

In contrast, dense SIFT descriptors which are extracted at every location in an image, are likely to carry more discriminative information [29] compared to the conventional SIFT which sparsely captures information. In addition, in the SPM framework, images are partitioned in different pyramid levels. If conventional SIFT is used, there is a high probability that some keypoints are detected in one partition but the neighbouring pixels required to describe the detected keypoints fall into different partitions. This will result in a less informative descriptor and therefore, the result provided in [4] is unsatisfactory.

In this paper, we have enhanced the effectiveness of SPM. This is done by investigating three concentric ring partitioning schemes which are used to build effective rotation-invariant SPM. In addition, a new weight function is proposed for the better contribution of pyramid levels to the final match score between two images. Also, dense SIFT is used as the local descriptor and finally, the enhanced effectiveness of SPM is evaluated for image classification as well as for image retrieval.

## IV. INVESTIGATION OF DIFFERENT PARTITIONING SCHEMES

In this section, we will carry out a thorough investigation to build a robust spatial pyramid structure which is effective to match images with rotation. To do this, three different concentric ring partitioning schemes which partition images with reference to the image centre are investigated here. The aim of these three partitioning schemes is to achieve a rotation-invariant image representation. Specifically, these partitioning schemes, unlike TrSPM, preserve spatially close descriptors in the corresponding partitions of two images at each level of the spatial pyramid. The three proposed partitioning schemes which are given in Figure 3 are rectangular ring partitioning (RRP), circular outer ring partitioning (CORP) and circular inner ring partitioning (CIRP). For each of these three partitioning schemes, Level 0 represents the entire image and each higher pyramid levels have one additional partition than the previous lower level. Details of the partitioning schemes are provided in Table 1.

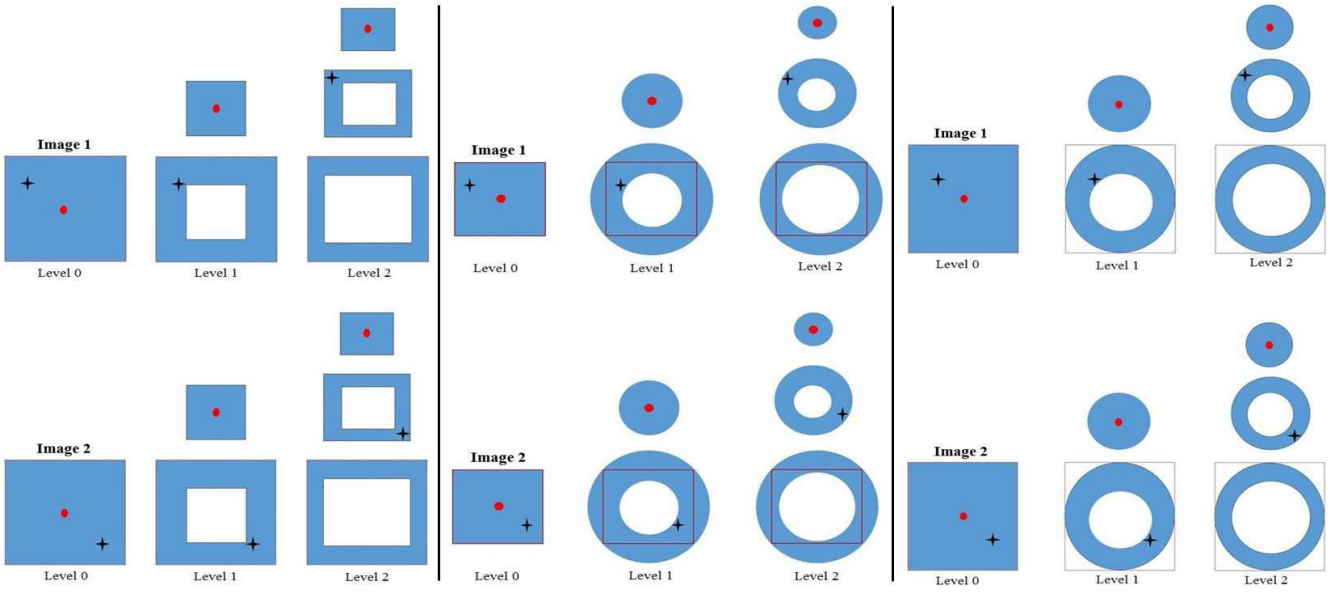
In Figure 3, it is visible that for RRP, CORP and CIRP schemes, object of interest (star) is preserved in the same corresponding partitions of original and rotated images. In addition, the red circular object which exists at the image centre, by using all the three rotation-invariant (RI) partitioning schemes, it is always preserved in a single partition irrespective of different pyramid levels. From here on, the SPM structures built with RI-partitioning schemes will be referred as RI-SPMs in the rest of this paper.

Among the proposed partitioning schemes, CORP and CIRP are based on concentric circular rings similar to the SPR. The authors of SPR [4] attempted to address the rotational issue of SPM. However, the experimental results are not satisfactory. Therefore, in this paper, we aim to investigate the various way of building the concentric ring partitions to make SPM rotation invariant and comprehensively compare the effectiveness of them.

## V. PROPOSAL OF A NEW WEIGHT FUNCTION

In TrSPM, image matching is performed using (2). The similarity scores obtained at the higher pyramid levels are given more importance than the lower levels as the higher-level descriptors contain more location-specific information. In practice, this is done by incorporating a weight function given by (3) which we refer here as the conventional weight function (CWF).

As per CWF, the weight associated with the similarity scores at each level is inversely proportional to the square root of the number of grids (or partitions) at that level. As the number of image partitions at each level of RI-SPMs is different compared to what it is in TrSPM, therefore, the way spatial information is incorporated into the descriptors at different levels of the RI-SPMs is very different from TrSPM. For this reason, CWF is not appropriate to provide weight assignments to the RI-SPM structures. Therefore, there is a need for the proposal of a new weight function and the motivation behind this is two-fold: (1) The proposed weight function should be suitable to the RI-SPM structures. (2) The proposed weight function should not violate CWF when it is applied in the TrSPM scenario. By satisfying these two conditions, we propose a new weight function which is




(a). RRP (b). CORP (c). CIRP

**FIGURE 3.** Pictorial representation of three partitioning schemes.

named as the generalized weight function (GWF) and it is given by (4).

$$w_i = 1/\sqrt{P_f/P_i} \tag{4}$$

where $P_f$ is the total number of image partitions at the highest pyramid level and $P_i$ is the total number of image partitions at the $i^{th}$ pyramid level. While proposing GWF, the focus is to determine how much location-specific information (in terms of the number of image partitions) individual level carries with respect to the highest level which carries the maximum location-specific information.

In the three-level TrSPM, if GWF is applied as per (4), then the weights assigned to Levels 2, 1 and 0 are 1, 0.5 and 0.25 respectively. This set of weights is the same if CWF is applied. Therefore, GWF satisfies CWF in TrSPM scenario. In contrast, from Figure 4, it can be seen that in the RI-SPM scenario, the weights assigned to the different levels using GWF are completely different from what it is for CWF.

Most of the SPM related literature where different partitioning schemes are proposed compared to the traditional SPM, still use the CWF as weight function. This may lead to an improper degree of contribution of each partition to the final matching. Therefore, GWF provides a pathway to calculate more accurate weights associated with the similarity scores obtained at each SPM level.

## VI. EXPERIMENTAL STUDY

This section presents the following studies: (1) The effectiveness of three proposed RI-SPMs (where the number of image partitions is linearly increased with the pyramid levels and dense SIFT is used as the local descriptor) in representing images for image classification and retrieval. (2) The effectiveness of GWF compared to CWF in the RI-SPM scenario.

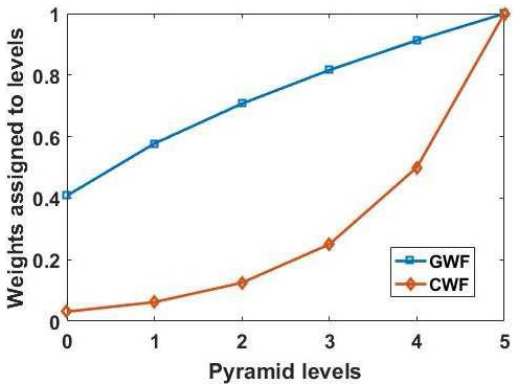



**FIGURE 4.** Comparison of weights to be assigned at different levels of RI-SPMs as per GWF and CWF.

### A. TEST DATABASES

To achieve the above-mentioned goals, the following databases are used.

#### 1) SCENE CATEGORIES DATABASE

Scene Categories database [3] (https://figshare.com/articles/15-Scene_Image_Dataset/7007177) contains 15 different classes of grayscale images. The number of images in the classes varies from 200 to 400 and in total, the database consists of 4485 images. The average size of the images in this database is 300 × 250 pixels. A set of sample images from this database is shown in Figure 5.

Along with the original, a rotated version of this database is used here. To form the rotated database, each of the images

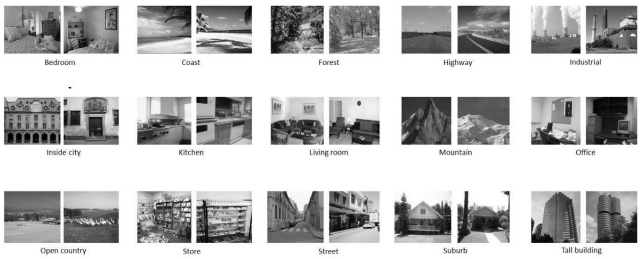



**FIGURE 5.** Sample images from scene categories database.

is manually rotated by 30, 45, 60, 135, 150, 210, 225, 255 and 315 degrees. These rotated images along with the original unrotated images are kept together in the rotated database. Thus the rotated database consists of a total of 44,850 images.

#### 2) 21 LAND USE DATABASE

This database (http://weegee.vision.ucmerced.edu/datasets/landuse.html) consists of 21 different classes of images [30]. Each class consists of 100 images and overall, the database contains 2100 images. Mostly, the images of this database are of 256 × 256 pixels. The main characteristic of this database is that the images belonging to different classes are captured with different camera angles. Therefore, images are naturally rotated. A set of sample images from this database is shown in Figure 6.


**FIGURE 6.** Sample images from 21 Land use database.

#### 3) CALTECH 101 DATABASE

Caltech 101 database [31] (http://www.vision.caltech.edu/Image_Datasets/Caltech101/) contains 101 object categories and 'Google background' category. The number of images in each category varies from 31 to 800 and in total, the database consists of 9146 images. Image resolutions in this database vary from very low to very high. However, most of the images are of 300 × 300 pixels on average. A set of sample images from this database is given in Figure 7.

#### 4) CALTECH 256 DATABASE

Caltech 256 database [32] (http://www.vision.caltech.edu/Image_Datasets/Caltech256/) contains 256 object categories. It contains 30,608 images in total and each class has at least 80 images. This database is superior than Caltech 101 which has issues like left-right alignment, rotation artefacts etc. A set of sample images from this database is given in Figure 8.

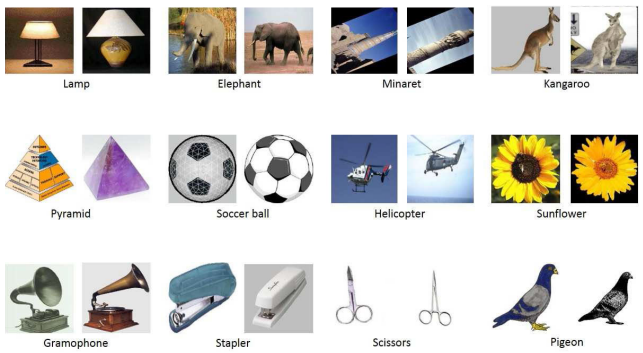



**FIGURE 7.** Sample images from Caltech 101 database.

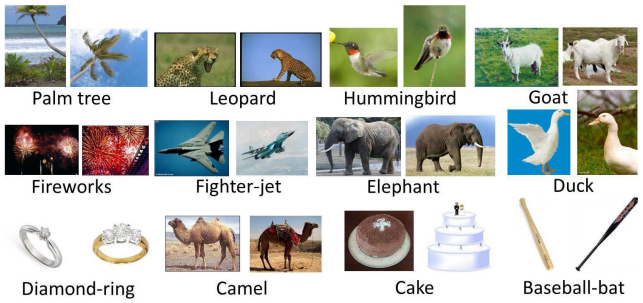



**FIGURE 8.** Sample images from Caltech 256 database.

### B. DESCRIPTOR EXTRACTION AND DICTIONARY CONSTRUCTION

SIFT descriptors are extracted from the training set images of each database over a dense regular grid of 16 × 16 pixel patches with 8 pixels of spacing. By applying k-means clustering on the training-image descriptor sets, separate dictionaries for each database are formed. To increase the discriminative power of image representation [33], the dictionary size of each database considered is 1000. Next, all the images of each database are partitioned by both RI and traditional schemes. SIFT descriptors are extracted from all the partitions at all levels and encoded by corresponding visual words after comparing them with their respective dictionaries. Now each partition is represented by a histogram of 1000 bins which is the image-level descriptor of the corresponding partition. Experiments were performed on a computer with an Intel Core i7 processor running at 3.2 GHz using 16 GB of RAM.

### C. IMAGE CLASSIFICATION

LIBSVM [34] is used for SVM based classification over Matlab platform. The kernel function which SVM classifiers use is histogram intersection. 10-fold cross-validation is performed on each database by randomly splitting individual databases to 10 training and test sets. To obtain a fair result, in each iteration, the training and test sets are completely different from each other. The final classification accuracy is the average accuracy over 10 iterations. In the results, legends 'RRP', 'CIRP' and 'CORP' represent the RI-SPM built with the corresponding RI-partitioning schemes. TrSPM

represents traditional SPM [3] and SPR represents the existing rotation invariant SPM [4].

#### 1) IMAGE CLASSIFICATION RESULTS TO COMPARE THE EFFECTIVENESS OF RI-SPMs WITH THE EXISTING ROTATION INVARIANT SPM (SPR)

In this section, the performances of RI-SPMs using the three partitioning schemes are compared with the performance of SPR [4]. For a fair comparison, the same settings of [4] are used to evaluate SPR. The performance of SPR is also compared with the performance of TrSPM. The comparison result is provided in Table 2. While evaluating three RI-SPMs, it is observed that the classification accuracies stop increasing at Level 6 and the highest classification accuracies are obtained at Level 5. For the three RI-SPMs, results are only provided with GWF. A detailed performance comparison between GWF and CWF is provided in the latter section. For the results in Table 2, 21 Land Use database is used for the testing as this database has images with in-built rotation.

**TABLE 2. Comparison of classification accuracies (%) between SPR, TrSPM and RI-SPMs on 21 Land use database.**

| SPR | TrSPM | RRP | CORP | CIRP |
|---|---|---|---|---|
| 72.14 | 75.64 | 86.82 | 86.57 | 85.74 |

From the results, it is clear that the three RI-SPMs under investigation perform better than SPR. Moreover, the classification accuracy of SPR is even worse than TrSPM. This is because, in SPR, conventional SIFT is used. The partitioning approach in SPR is coarse and causes descriptors to become over-discriminative. Moreover, there is no use of appropriate weight function. Due to these reasons, later in the result analysis, the performances of RI-SPMs are only compared with the TrSPM and not SPR.

#### 2) IMAGE CLASSIFICATION RESULTS TO COMPARE THE EFFECTIVENESS OF RI-SPMs WITH TrSPM AND TO VALIDATE THE EFFECTIVENESS OF GWF

In this section, image classification results are provided to compare how robust RI-SPMs are with respect to the TrSPM. Also, the effectiveness of GWF compared to CWF to the proposed structures of RI-SPM is investigated. Image classification is performed for both RI-SPMs and TrSPM using the databases considered. Experiments are conducted in two ways, i.e. 'Single-level' and 'Pyramid'. In the 'Single-level' experiment, descriptors of a level are tested separately and in the 'Pyramid' experiment, the descriptors up to a level are tested together. The experiments on the RI-SPMs are conducted up to Level 6 and for the TrSPM, the experiment is conducted only up to Level 2 (as according to [3], Level 2 is optimum for the TrSPM).

Performance comparison of all the three RI-SPMs along with TrSPM in terms of classification accuracies for all five databases considered are given in Table 3. The table contains only the highest classification accuracies of each SPM using both GWF and CWF. From the results, it is clear that the classification accuracies of all three RI-SPMs are higher than the TrSPM.

In addition, for each RI-SPM on each database, GWF performs better than CWF. This is because GWF provides more appropriate weights to the pyramid levels of RI-SPMs compared to the CWF. The performance progression of RI-SPMs with the increase of pyramid levels is shown in Figure 9 for the Scene Categories database. As it is already observed that GWF performs better compared to CWF in the RI-SPM scenario, therefore, to avoid redundancy, in Figure 9 pyramid performances are shown only with GWF. The performance progression of RI-SPMs on other databases follow a similar trend as in Figure 9.

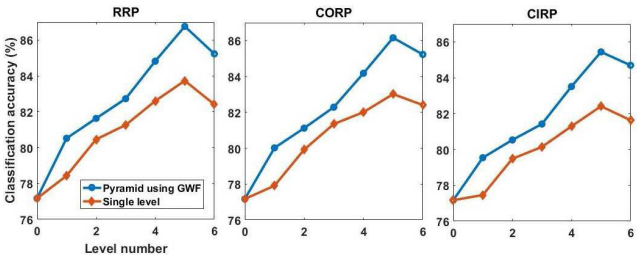



**FIGURE 9. Level wise classification accuracies on scene categories database.**

#### 3) COMPARISON OF CLASSIFICATION ACCURACIES WITH THE EXISTING POPULAR METHODS

To the best of our knowledge, the experiment settings we have used to test our proposed methods are the same as the experiment settings followed in the TrSPM. In the literature, some of the existing popular methods have been compared their classification accuracies with the TrSPM. So, it would be a fair comparison if we compare our proposed method (RRP with GWF) in terms of classification accuracy with the existing popular methods with reference to the TrSPM based on the databases we have considered in this paper. These comparisons are given in Tables 4, 5, 6, 7 for Scene Categories, 21 Land Use, Caltech 101 and Caltech 256 databases respectively and by observing these tables, we can conclude that proposed method outperforms all the other compared methods.

### D. IMAGE RETRIEVAL

In this section, image retrieval results are shown in terms of mean average precision (MAP) values and recall-precision curves. Retrieval performances are shown to compare the effectiveness of RI-SPMs with TrSPM. Individual images are used as a query to retrieve similar images from the corresponding database. Image-level descriptors (pyramid of histograms) of query image find the similarity with the image-level descriptors of the database images using (2). For the sake of simplicity, to apportion the contribution of similarity scores obtained from each level to the final similarity score, only GWF is used as it is already proven that GWF performs better compared to CWF. For the experiment, each image in the database is used as a query to

**TABLE 3.** Comparison of classification accuracies (%).

| Database | RRP | | CORP | | CIRP | | TrSPM |
|---|---|---|---|---|---|---|---|
| | GWF | CWF | GWF | CWF | GWF | CWF | GWF/CWF |
| Scene Categories | 86.76 | 85.20 | 86.16 | 84.51 | 85.44 | 83.79 | 84.10 |
| Rotated Scene Categories | 77.62 | 76.34 | 77.07 | 75.36 | 76.23 | 74.72 | 60.25 |
| 21 Land Use | 86.82 | 85.45 | 86.57 | 85.15 | 85.74 | 84.34 | 75.64 |
| Caltech 101 | 76.48 | 75.05 | 76.07 | 74.62 | 75.13 | 73.77 | 74.62 |
| Caltech 256 | 38.65 | 37.14 | 38.12 | 36.61 | 37.26 | 35.85 | 36.45 |

**TABLE 4.** Comparison of classification accuracies (%) on Scene categories database.

| Method | Accuracy (%) |
|---|---|
| RRP | 86.76 |
| OSRP [18] | 83.90 |
| SP-RSC [22] | 83.67 |
| HSPMP [20] | 81.46 |
| PlsSPR [27] | 81.40 |
| ScSPM [37] | 86.76 |
| SPM [3] | 84.10 |

**TABLE 5.** Comparison of classification accuracies (%) on 21 Land use database.

| Method | Accuracy (%) |
|---|---|
| RRP | 86.82 |
| SPCK [16] | 73.12 |
| OSRP [18] | 75.5 |
| SPCK++ [16] | 77.38 |
| Color-HLS [31] | 81.19 |
| SPM [3] | 76.54 |

**TABLE 6.** Comparison of classification accuracies (%) on Caltech 101 database.

| Method | Accuracy (%) |
|---|---|
| RRP | 76.48 |
| ScSPM [37] | 73.20 |
| PlsSPR [27] | 67.21 |
| ML+CORR [29] | 69.60 |
| SPM [3] | 74.62 |

**TABLE 7.** Comparison of classification accuracies (%) on Caltech 256 database.

| Method | Accuracy (%) |
|---|---|
| RRP | 38.65 |
| SP-RSC [22] | 36.86 |
| ScSPM [37] | 34.22 |
| SPM [3] | 36.45 |

retrieve the rest of the images from the individual databases. MAP and recall-precision curves are used here to evaluate the image retrieval performances. To obtain the MAP and recall-precision curves, for each query, the top $k$ images are considered.

Table 8 shows the MAP values for the top $k$ retrieved images. Also, Figure 10 shows the recall-precision (R-P) curve for Rotated Scene Categories database. To avoid redundancy, R-P curves are not shown for other databases. However, they follow the similar trend as Figure 10. From the analysis of MAP values and R-P curves, it is clear that irrespective of databases, RI-SPMs perform better than TrSPM. Furthermore, for the Rotated Scene Categories and 21 Land Use database, the performance gaps are broader between RI-SPMs and TrSPM compared to the other three

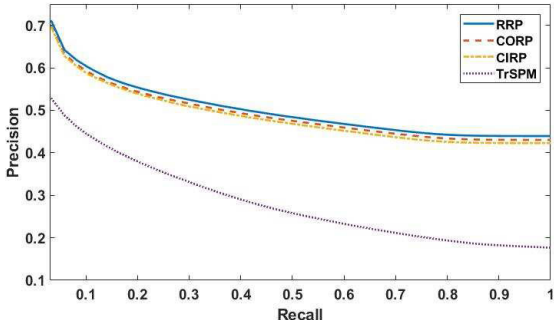



**FIGURE 10.** Recall-Precision curve on rotated scene categories database.

(a) Living room (b) Office

**FIGURE 11.** Intra-class images from scene categories database exhibit rotational behaviour.

databases. This is because the images of Rotated Scene Categories and 21 Land Use databases are affected by rotation and RI-SPMs effectively dealt with the rotational issue that TrSPM failed to do.

### E. QUALITATIVE ANALYSIS

It can be observed from the result analysis that the performance of three RI-SPMs is consistently higher than the TrSPM for all the databases considered here. RI-SPMs are expected to perform better for Rotated Scene Categories and 21 Land Use databases. In contrast, for the other two databases, RI-SPMs also perform better than TrSPM. This is because some classes from these databases contain a couple of images with similar objects but the objects exhibit rotational behaviour. For example, consider sample images in Figure 11 from the Scene Categories database where two images from (a) 'Living room' and (b) 'Office' classes are shown. The images from the 'Living room' class contain a 'couch', and with respect to the centre of the image 'couch' positions are almost 180 degrees apart from each other. The same thing happens to 'Office' class images where 'computer' is the object of concern. Therefore, RI-SPMs have always performed better than TrSPM.

Here, with the help of a retrieval example, the effectiveness of RI-SPMs are further analysed. Specifically, an image, which has a 'storage tank' as the distinct object, from 21 Land

**TABLE 8. Comparison of MAP (%) based on top k retrieved images.**

| Database | k | TrSPM | RRP | CORP | CIRP |
|---|---|---|---|---|---|
| Scene Categories | 100 | 55.66 | 57.63 | 57.45 | 56.55 |
| Rotataed Scene Categories | 100 | 54.58 | 65.52 | 65.33 | 64.19 |
| 21 Land Use | 99 | 57.59 | 65.12 | 64.97 | 63.72 |
| Caltech 101 | 30 | 53.96 | 56.39 | 55.60 | 54.81 |
| Caltech 256 | 80 | 45.28 | 47.86 | 46.99 | 46.18 |

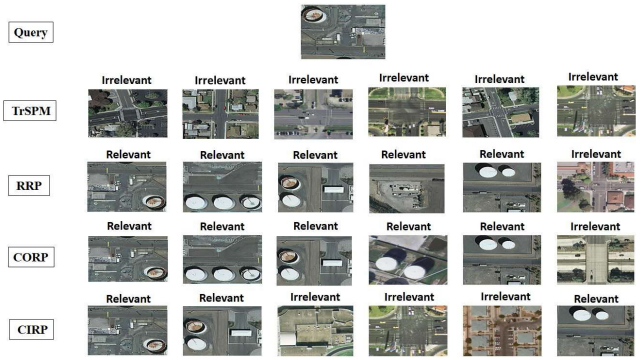



**FIGURE 12. Retrieved images based on a query.**

Use database is used as the query to retrieve images using three RI-SPMs as well as TrSPM. The retrieval results are provided in Figure 12 where the top 6 retrieved images are shown. Based on the relevance with the query, the retrieved images are labelled accordingly. The query image is rotated by 180 degrees and used as one of the database images which can be also retrieved by the query image. The motivation for rotating the query image and use it as a database image is to test the effectiveness of three RI-SPMs and TrSPM by investigating whether the rotated version of the query image is retrieved by any of the SPMs or not. As expected, Figure 12 clearly shows that, compared to the three RI-SPMs, TrSPM performs worst as no image with 'storage tank' is retrieved. Specifically, within the top 6, retrieved images using TrSPM, the rotated version of the query is not found. Whereas, the rotated version of the query is retrieved by all the three RI-SPMs as the top-ranked image.

The classification accuracies and the retrieval performances for each of the three RI-SPMs show a consistent trend for all the databases. RRP is based on concentric rectangular rings and CORP and CIRP are based on concentric circular rings. Although the three RI-SPMs perform better than the TrSPM, RI-SPM with CIRP performs worse. This is because, when an image is partitioned using CIRP, some parts (corner parts) of the image are not considered from Level 1 onwards and in some cases, the content not captured within the circular rings may play an important role to characterise that image. Therefore, the image representation using CIRP carries less information and results in lesser performance. The performances of RI-SPMs structured with RRP and CORP are comparable. The dense patches on which SIFT descriptors are extracted are of square-shaped. Therefore, theoretically, RRP which is based on rectangular concentric rings is more appropriate to build an RI-SPM. Thus, RRP is our recommended approach.

## VII. CONCLUSION

In this paper, the performance of spatial pyramid matching is enhanced. The three RI-SPMs investigated are robust to any kind of rotational changes that occurred in an image and perform better than the TrSPM. Among the three RI-SPMs, the one built with RRP is selected further to work with. Our proposed weight function, GWF assigns appropriate weights to the similarity scores obtained at each level of RI-SPMs. Our experiment results show that GWF apportions the similarity scores obtained from each SPM level, more accurately than that of CWF. Since SPM is widely used, our proposed improvement will have a significant positive impact on a wider number of applications.

## REFERENCES

[1] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2005, pp. 524–531.

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

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2006, pp. 2169–2178.

[4] X. Li, Y. Song, Y. Lu, and Q. Tian, "Spatial pooling for transformation invariant image representation," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, 2011, pp. 1509–1512.

[5] P. Karmakar, S. W. Teng, G. Lu, and D. Zhang, "Rotation invariant spatial pyramid matching for image classification," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2015, pp. 1–8.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2005, pp. 886–893.

[7] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4357–4366, Sep. 2018.

[8] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 4961–4970.

[9] P. Karmakar, S. W. Teng, G. Lu, and D. Zhang, "A kernel-based approach for content-based image retrieval," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2018, pp. 1–6.

[10] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao, "Group-sensitive multiple kernel learning for object categorization," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 436–443.

[11] M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[12] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Comput. Vis. (ECCV)*, 2010, pp. 143–156.

[13] J. Yang, K. Yu, and T. Huang, "Efficient highly over-complete sparse coding using a mixture model," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 113–126.

[14] S. Yan, X. Xu, D. Xu, S. Lin, and X. Li, "Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification," in *Proc. Comput. Vis. (ECCV)*, 2012, pp. 473–487.

[15] H. E. Tasli, R. Sicre, T. Gevers, and A. A. Alatan, "Geometry-constrained spatial pyramid adaptation for image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1051–1055.
[16] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1465–1472.
[17] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2307–2314.
[18] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2012, pp. 730–743.
[19] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3517–3524.
[20] H. Han, J. Gu, X. Li, and Q. Han, "Hierarchical spatial pyramid max pooling based on SIFT features and sparse coding for image classification," *IET Comput. Vis.*, vol. 7, no. 2, pp. 144–150, Apr. 2013.
[21] X. Wang, J. Ma, and M. Xu, "Image classification using sparse coding and spatial pyramid matching," in *Proc. Int. Conf. e-Educ., e-Bus. Inf. Manage.*, 2014, pp. 81–84.
[22] C. Zhang, S. Wang, Q. Huang, J. Liu, C. Liang, and Q. Tian, "Image classification using spatial pyramid robust sparse coding," *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1046–1052, Jul. 2013.
[23] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 1800–1807.
[24] N. M. Elfiky, J. Gonzàlez, and F. X. Roca, "Compact and adaptive spatial pyramids for scene recognition," *Image Vis. Comput.*, vol. 30, no. 8, pp. 492–500, Aug. 2012.
[25] N. M. Elfiky, F. S. Khan, J. Van De Weijer, and J. González, "Discriminative compact pyramids for object and scene recognition," *Pattern Recognit.*, vol. 45, no. 4, pp. 1627–1636, Apr. 2012.
[26] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. Comput. Vis. (ECCV)*, 2010, pp. 448–461.
[27] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi, "Discriminative spatial pyramid," in *Proc. CVPR*, Jun. 2011, pp. 1617–1624.
[28] P. Jain, B. Kulis, and K. Grauman, "Fast image search for learned metrics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[29] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[30] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (GIS)*, 2010, pp. 270–279.
[31] 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," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 59–70, Apr. 2007.
[32] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. 7694, 2007.
[33] P. Karmakar, S. W. Teng, D. Zhang, Y. Liu, and G. Lu, "Combining pyramid match kernel and spatial pyramid for image classification," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–8.
[34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.
[35] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.

**PRIYABRATA KARMAKAR** received the B.Tech. degree from the West Bengal University of Technology, in 2009, the M.Tech. degree from the Sikkim Manipal Institute of Technology, in 2012, and the D.Phil. degree from Federation University Australia, in 2018. He is currently associated with the Faculty of Science, Engineering and Information Technology, Federation University Australia. He frequently reviews journal articles and conference articles. His research interests include image processing, computer vision, and machine learning.

**SHYH WEI TENG** received the bachelor's degree (Hons.) in computing and the Ph.D. degree in information technology (IT) from the Faculty of IT, Monash University, in 1998 and 2004, respectively. He also successfully completed the Graduate Certificate in Higher Education, Higher Education Development Unit, Monash University, Australia, in 2005. He was a Senior Lecturer with the Gippsland School of IT, Monash University. He is currently an Associate Professor and Deputy Dean of IT with the School of Science, Engineering and Technology, Federation University Australia. His research interests include image processing, machine learning, and bioinformatics. He also received the Prestigious Australian Research Council (ARC) Discovery Project Grant, in 2013. He taught a number of subjects such as database, systems analysis and design, programming, and project management. He was also nominated for the 2007 Staff Excellence Award in Teaching and the 2008 Excellence Award in Gippsland Community Engagement Team. He regularly serves on the reviewer boards of a number of top international journals (e.g., *Pattern Recognition* and *Neurocomputing*) and as a Program Member for several international conferences (e.g., IEEE CEC and ICME).

**GUOJUN LU** received the B.Eng. degree from South East University, China, in 1984, and the Ph.D. degree from Loughborough University, in 1990. He has held positions at Loughborough University, the National University of Singapore, and Deakin University. He is currently a Professor and the Associate Dean by research with the School of Science, Engineering and Information Technology, Federation University Australia. He is also a Leading Researcher with the Centre for Multimedia Computing, Communications, and Artificial Intelligence Research (MCCAIR), where he is also working on an ARC DP Project and supervising an ARC DECRA Project. He has published more than 200 refereed journal and conference articles in these areas and wrote two books *Communication and Computing for Distributed Multimedia Systems* (Artech House 1996), and *Multimedia Database Management Systems* (Artech House 1999). His main research interests are in multimedia information processing, indexing, and retrieval.

**DENGSHENG ZHANG** received the B.Sc. and B.A. degrees in 1985 and 1987, respectively, and the Ph.D. degree in computing from Monash University, Australia, in 2002. He is currently a Senior Lecturer with the School of Science, Engineering and Information Technology, Federation University Australia. He has more than 20 years of research experience in the field of artificial intelligence and big data. He has published nearly 90 refereed articles on international journals and conference proceedings in his career. His main research interests include pattern recognition, machine learning, big multimedia data classification, and retrieval. He is on the reviewer board of several top international journals including IEEE TPAMI, TOM, TIP, CSVT; ACM TOMCCAP, TMIS, and *Pattern Recognition*. He has also served on the program committees and organization committees of 15 international conferences. He is a regular PM Member for IEEE mainstream conferences of ICME, ICASSP, and ICIP. He also currently serves as an Associate Editor of WSEAS (The World Scientific and Engineering Academy and Society) *Transactions on Signal Processing*.