




Article

# A Novel Discriminating and Relative Global Spatial Image Representation with Applications in CBIR

Bushra Zafar <sup>1,2\*</sup> , Rehan Ashraf <sup>1</sup>, Nouman Ali <sup>3,4</sup> , Muhammad Kashif Iqbal <sup>5</sup>,  
Muhammad Sajid <sup>6</sup>, Saadat Hanif Dar <sup>3</sup> and Naeem Iqbal Ratyal <sup>6</sup> 

<sup>1</sup> Department of Computer Science, National Textile University, Faisalabad 38000, Pakistan; rehan@ntu.edu.pk

<sup>2</sup> Department of Computer Science, Government College University, Faisalabad 38000, Pakistan;

<sup>3</sup> Department of Software Engineering, Mirpur University of Science & Technology, Mirpur AJK 10250, Pakistan; nali@caa.tuwien.ac.at (N.A.); saadat.dar@gmail.com (S.H.D.)

<sup>4</sup> Computer Aided Automation, Computer Vision Lab, Vienna University of Technology, A-1040 Vienna, Austria

<sup>5</sup> Department of Mathematics, Government College University, Faisalabad 38000, Pakistan; kashifiqbal@gcuf.edu.pk

<sup>6</sup> Department of Electrical Engineering, Mirpur University of Science & Technology, Mirpur AJK 10250, Pakistan; sajid.ee@must.edu.pk (M.S.); naeemratyal@hotmail.com (N.I.R.)

\* Correspondence: bkgcuf@gmail.com

Received: 18 October 2018; Accepted: 12 November 2018; Published: 14 November 2018



**Abstract:** The requirement for effective image search, which motivates the use of Content-Based Image Retrieval (CBIR) and the search of similar multimedia contents on the basis of user query, remains an open research problem for computer vision applications. The application domains for Bag of Visual Words (BoVW) based image representations are object recognition, image classification and content-based image analysis. Interest point detectors are quantized in the feature space and the final histogram or image signature do not retain any detail about co-occurrences of features in the 2D image space. This spatial information is crucial, as it adversely affects the performance of an image classification-based model. The most notable contribution in this context is Spatial Pyramid Matching (SPM), which captures the absolute spatial distribution of visual words. However, SPM is sensitive to image transformations such as rotation, flipping and translation. When images are not well-aligned, SPM may lose its discriminative power. This paper introduces a novel approach to encoding the relative spatial information for histogram-based representation of the BoVW model. This is established by computing the global geometric relationship between pairs of identical visual words with respect to the centroid of an image. The proposed research is evaluated by using five different datasets. Comprehensive experiments demonstrate the robustness of the proposed image representation as compared to the state-of-the-art methods in terms of precision and recall values.

**Keywords:** image analysis; image retrieval; spatial information; image classification; computer vision

## 1. Introduction

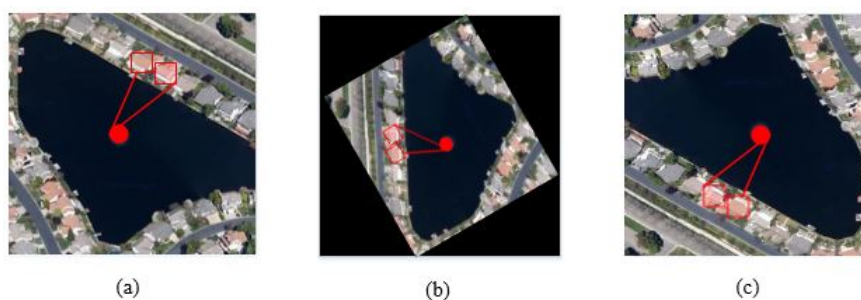
In recent years, with the rapid development of imaging technology, searching or retrieving a relevant image from an image archive has been considered an open research problem for computer vision based applications [1–4]. Higher retrieval accuracy, low memory usage and reduction of semantic gap are examples of common problems related to multimedia analysis and image retrieval [3,5]. The common applications of multimedia and image retrieval are found in the fields of video surveillance, remote sensing, art collection, crime detection, medical image processing and image retrieval in real-time applications [6]. Most of the retrieval systems, both for multimedia and images, rely on the matching

of textual data with the desired query [6]. Due to the existing semantic gaps, the performance of these systems suffers [7]. The appearance of a similar view in images belonging to different image categories, results in the closeness of the feature vector values, and degrades the performance of image retrieval [6]. The main focus of the research in Content-Based Image Retrieval (CBIR) is to retrieve images that are in a semantic relationship with a query image [8]. CBIR provides a framework that compares the visual feature vector of a query image to the images places in the dataset [9].

The Bag of Visual Words (BoVW), also known as Bag of Features (BoF) [10], is commonly used for video and image retrieval [11]. The local features or interest point detectors are extracted from a group of training images. To achieve a compact representation, the feature space is quantized to construct a code-book that is also known as visual vocabulary or visual dictionary. The final feature vector, which consists of histograms of visual words, is orderless with respect to the sequence of co-occurrences in the 2D image space. The performance of the BoVW model suffers as the extraction of spatial information is beneficial in image classification and retrieval-based problems [6,12].

Various approaches have been proposed to enhance the performance of image retrieval, such as soft assignments, computation of larger codebooks and visual word fusion [8]. All of these techniques do not contain any information about the visual word's locations in the final histogram-based representation [13]. There are two common techniques that can compute the spatial information from the image. These are based on (1) the construction of histograms from different sub-regions of image, and (2) visual word co-occurrence [13–15]. The first approach is to split the image into different cells for the histogram's computation; it is reported to be robust for content-based image matching applications [16]. Spatial Pyramid Matching (SPM) [16] is considered as a notable contribution for the computation of spatial information for BoVW-based image representation. In SPM, an image is divided into different sizes of rectangular regions for the creation of level-0, level-1 and level-2 histograms of visual words. However, SPM is sensitive to image transformations (i.e., rotation, flipping and translation) and loses its discriminative power, resulting in the misclassification of two similar scene images [17,18].

The second approach to the computation of spatial layout is based on relationships among visual words [19–21]. This paper proposes a novel approach to extracting the image spatial layout based on global relative spatial orientation of visual words. This is achieved by computing the angle between identical visual word pairs with respect to the centroid in the image. Figure 1 provides an illustration to better understand the proposed approach. The image in Figure 1 is rotated at varying angles. It can be seen that the same angle is computed between visual words irrespective of the image orientation.



**Figure 1.** Angle between identical visual word pairs with respect to the centroid. Here (a) represents the original image, (b) the image rotated by  $120^\circ$ , and (c) the image rotated by  $180^\circ$ .

The main contributions of this research are the following: (1) the addition of the discriminating relative global spatial information to the histogram of BoVW model and (2) reduction of the semantic gap. An efficient image retrieval system must be capable to retrieve images that meet user preferences and their specific requirements. The reduction of the semantic gap specifies that the related categories are given higher similarity scores than unrelated categories. The proposed representation is capable of handling geometric transformations, i.e., rotation, flipping and translation. Extensive experiments on

five standard benchmarks demonstrate the robustness of the proposed approach and a remarkable gain in the precision and recall values over the state-of-the-art methods.

The structure of the paper is as follows. Section 2 contains the literature review and related work; Section 3 is about the BoVW model and proposed research; and Section 4 deals with the experimental parameters and image benchmarks, while also presenting a comparison with the existing state-of-the-art techniques. Section 5 provides a discussion, while Section 6 concludes the proposed research with future directions.

## 2. Related Work

According to the literature [6], SIMPLicity, Blobworld and Query by Image Content (QBIC) are examples of computer vision applications that rely on the extraction of visual features such as color, texture and shape. Image Rover and WebSeek are examples of image search systems that rely on a query-based or keyword-based image search [6]. The main objective of any CBIR system is to search for relevant images that are similar to the query image [22]. Overlapping objects, differences in the spatial layout of the image, changes in illumination and semantic gaps make CBIR challenging for the research community [8]. Wang et al. [23] propose the Spatial Weighing BOF (SWBOF) model to extract the spatial information by using three approaches, i.e., local variance, local entropy and adjacent block distance. This model is based on the concept of the different parts of an image object contributing to image categorization in varying ways. The authors demonstrate significant improvement over the traditional methods. Ali et al. [9] extract the visual information by dividing an image into triangular regions to capture the compositional attributes of an image. The division of the image into triangular cells is reported as an efficient method for histogram-based representation. Zeng et al. [24] propose spatiogram-based image representation that consists of a color histogram that is quantized by using Gaussian Mixture Models (GMMs). The quantized values of GMMs are used as an input for the learning of the Expectation-Maximization (EM). The retrieval is performed on the basis of the closeness of the feature vector values of two spatiograms which are obtained by using the Jensen–Shannon Divergence (JSD) [24]. Yu et al. [25] investigate the impact of the integration of different mid-level features to enhance the performance of image retrieval. They investigate the impact of the integration of SIFT descriptors with LBP and HOG descriptors respectively, in order to address the problem of the semantic gap. Weighed  $k$ -means clustering is used for quantization, and best performance is reported with SIFT-LBP integration.

To reduce the semantic gap between the low-level features and the high-level image concepts, Ali et al. [8] propose image retrieval based on the visual words integration of Scale Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF). Their approach acquires the strength of both features, i.e., invariance to scale and rotation of SIFT and robustness to illumination of SURF. In another recent work, Ali et al. [26] propose a late fusion of binary and local descriptors i.e., FREAK and SIFT to enhance the performance of image retrieval. Filliat et al. [27] present an incremental and interactive localization and map-learning system based on BoW. Hu et al. [28] propose a real-time assistive localization approach that extracts compact and effective omnidirectional image features which are then used to search a remote image feature-based database of a scene, in order to help indoor navigation.

In another recent work, Li et al. [29] propose a hybrid framework of local (BoW) and global image features for efficient image retrieval. According to Li et al. [29], a multi-fusion based on two lines of image representation can enhance the performance of image retrieval. The authors [29] extract the texture information by using Intensity-Based Local Difference Patterns (ILDLP) and by selecting the HSV color space. This scheme is selected to capture the spatial relationship patterns that exist in the images. The global color information is extracted by using the H and S components. The final feature vector is constituted by combining the H, S feature space and ILDP histograms. The experimental result validates that the fusion of color and texture information enhances the performance of image retrieval [29]. According to Liu et al. [30], the ranking and incompatibility of the image feature

descriptor is not considered much in the domain of image retrieval. The authors address the problem of incompatibility by using gestalt psychology theory and manifold learning. A combination of gradient direction and color is used to imitate human visual uniformity. The selection of a proposed feature scheme [30] enhances the image retrieval performance. According to Wu et al. [31], ranking and feature representation are two important factors that can enhance the performance of image retrieval and they are considered separately in image retrieval models. The authors propose a texton uniform descriptor and apply an intrinsic manifold structure through visualizing the distribution of image representations on the two-dimensional manifold. This process provides a foundation for subsequent manifold-based ranking and preserves intrinsic neighborhood structure. The authors apply a Modified Manifold Ranking (MMR) to enhance and propagate adjacent similarity between the images [31]. According to Varish et al. [32], a hierarchical approach to CBIR based on a fusion of color and texture can enhance the performance of image retrieval. The color feature vectors are computed on the basis of quantized HSV color space, and texture values are computed to achieve rotation invariance on the basis of Value (V) component of HSV space. The sub-band of various Dual Tree Complex Wavelet Transform (DT-CWT) is applied to compute the principal texture direction.

Zou et al. [33] propose an effective feature selection approach based on Deep Belief Networks (DBN) to boost the performance of image retrieval. The approach works by selecting more reconstructible discriminative features using an iterative algorithm to obtain the optimized reconstruction weights. Xia et al. [34] perform a systematic investigation to evaluate factors that may affect the retrieval performance of the system. They focus the analysis on the visual feature aspect to create powerful deep feature representations. According to Wan et al. [7], a pre-trained deep convolution neural network outperforms the existing feature extraction techniques at the cost of high training computations for large-scale image retrieval. It is important to mention that the approaches based on deep networks may not be an optimal selection as they require large-scale training data with a lot of computations to train a classification-based model [21,35].

### 3. Proposed Methodology

The basic notations for the BoVW model are discussed in this section. This is then followed by a discussion of the proposed Relative Global Spatial Image Representation (RGSIR) and the details of its implementation.

#### 3.1. BoVW Model

The Bag-of-Words (BoW) methodology was first proposed in textual retrieval systems [11] and was further applied in the form of BoVW representation for image analysis. In BoVW, the final image representation is a histogram of visual words. It is termed a bag, as it counts how many times a word occurs in a document. A histogram does not have any order and does not retain any information regarding the location of visual words in the 2D image space [9,16]. The similarity of two images is determined by histogram intersection. In the case of dissimilar images, the result of the intersection is small.

As a first step in BoVW, the local features are extracted from the image  $Im$ , and the image is represented as a set of image descriptors, such as  $Im = \{d_1, d_2, d_3, \dots, d_I\}$ , where  $d_i$  denotes the local image features and  $I$  represents total image descriptors. The feature extraction can be done by applying some local descriptors such as SIFT descriptors [36]. The key points can be acquired automatically by using interest point detectors or by applying dense sampling [16].

Consequently, there are numerous local descriptors created for each image for a given dataset. The extracted descriptors are vector quantized by applying  $k$ -means [11] clustering technique to construct the visual vocabulary, as in

$$v = \{w_1, w_2, w_3, \dots, w_K\} \quad (1)$$

where  $K$  shows the specified number of clusters or visual words and  $v$  denotes the constructed visual vocabulary.

The assignment of each descriptor to the nearest visual word is done by computing the minimum distance as follows:

$$w(d_j) = \underset{w \in v}{\operatorname{argmin}} \operatorname{Dist}(w, d_j) \tag{2}$$

here,  $w(d_j)$  represents the visual word mapped to  $j$ th descriptor and  $\operatorname{Dist}(w, d_j)$  depicts the distance between the descriptor  $d_j$  and visual word  $w$ .

The histogram representation of an image is based on the visual vocabulary. The number of histogram bins equates the number of visual words in the code book or dictionary (i.e.,  $K$ ). Each histogram bin  $\operatorname{bin}_i$  represents a visual word  $w_i$  in  $v$  and signifies the number of descriptors mapped to a particular visual word as shown in (3)

$$\operatorname{bin}_i = \operatorname{card}(D_i) \text{ where } D_i = \{d_j, j \in 1, \dots, n \mid w(d_j) = w_i\} \tag{3}$$

$D_i$  is the set of descriptors mapped to a particular visual word  $w_i$  in an image, and the cardinality of this set is given by  $\operatorname{Card}(D_i)$ . The final histogram representation for the image is created by repeating the process for each word in the image. The histograms hence created do not retain the spatial context of the interest points.

### 3.2. The Proposed Relative Global Spatial Image Representation (RGSIR)

In the BoVW model the final image representation is created by mapping identical image patches to the same visual word. In [20], Khan et al. capture the spatial information by modeling the global relationship between identical visual word pairs (PIWs). Their approach exhibits invariance to translation and scaling but is sensitive to rotation [20,37], since the relative relationship between PIWs is computed with respect to the x-axis. Anwar et al. [37] propose an approach to acquire rotation invariance by computing angles between Triplets of Identical Visual Words (TIWs). Although the approach of [37] acquires rotation invariance, it significantly increases computation complexity due to the increase in the number of possible triplet combinations. For instance, if the number of identical visual words is 30, the number of distinct pair combinations is 435 and the number of possible distinct triplet combinations is 4060.

This paper proposes a novel approach to acquiring spatial information for transformation invariance by computing the global geometric relationship between pairs of identical visual words. This is accomplished by extracting the spatial distribution of these pairs with respect to a centroid in an image as shown in Figure 2.

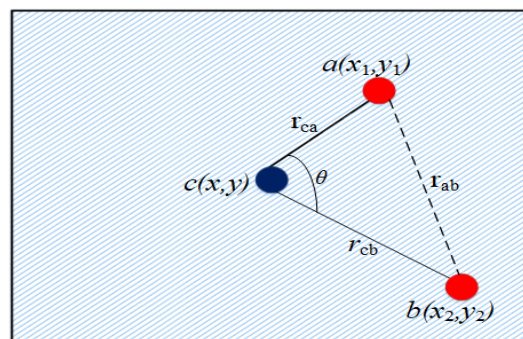


Figure 2. Angle between identical visual word pairs with respect to the centroid.

Hence we define the set of all pairs (PW) of identical visual words related to a visual word  $w_i$  as:

$$PW_i = \{(a, b) \mid (d_a, d_b) \in D_i^2, d_a \neq d_b\} \tag{4}$$

where  $a(x_1, y_1)$  and  $b(x_2, y_2)$  are the spatial locations of the descriptors  $d_a$  and  $d_b$ , respectively. Since the  $i$ th histogram bin signifies the descriptor  $d_i$ , its value determines the total occurrences of the word  $w_i$ . The cardinality of the set  $PW_i$  is  ${}^b C_2$ . The centroid  $c = (x, y)$  of an image  $Im$  of size  $R \times C$  is calculated as

$$x = \frac{1}{|Im|} \sum_{i=1}^{|Im|} x_i, \quad y = \frac{1}{|Im|} \sum_{i=1}^{|Im|} y_i \tag{5}$$

where  $Im = \{(x_i, y_i) \mid 1 \leq x_i \leq R, 1 \leq y_i \leq C\}$  and  $|Im|$  is the number of elements in  $Im$ . Let  $r_{ab}$  be the Euclidean distance between  $a$  and  $b$ , then

$$r_{ab} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{6}$$

Similarly, the Euclidean distances of  $a$  and  $b$  from  $c$  are calculated as

$$r_{ca} = \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$$

$$r_{cb} = \sqrt{(x_2 - x)^2 + (y_2 - y)^2}$$

Using the Law of cosines, we have

$$\theta = \arccos \left( \frac{(r_{ca})^2 + (r_{cb})^2 - (r_{ab})^2}{2(r_{ca})(r_{cb})} \right) \tag{7}$$

where  $\theta = \angle acb$ .

The  $\theta$  angles obtained are then concatenated to create the histogram representation with bins equally distributed between 0–180°. The optimal number of bins used for histogram representation is determined empirically. The  $RGSIR_i$  represents the spatial distribution for a particular visual word  $w_i$ . The  $RGSIR_i$  obtained from all the visual words in an image are concatenated to create the global image representation. A bin replacement technique is used to transform the BoVW representation to RGSIR. This is achieved by replacing each bin of the BoVW histogram with the associated  $RGSIR_i$  related to a particular  $w_i$ . To add the spatial information while keeping the frequency information intact, the sum of all bins of  $RGSIR_i$  is normalized to the size of the bin  $bin_i$  of the BoVW histogram that is being replaced. The image representation for RGSIR is hence formulated as:

$$RGSIR = (\alpha_1 RGSIR_1, \alpha_2 RGSIR_2, \dots, \alpha_K RGSIR_K) \tag{8}$$

where  $\alpha_i$ , the coefficient of normalization, is given by  $\alpha_i = \frac{bin_i}{\|RGSIR_i\|}$ . If the size of the visual vocabulary is  $K$  and the number of histogram bins is  $H$ , then the dimensions of RGSIR are  $K \times H$ .

### 3.3. Implementation Details

The histogram representations for all of the datasets are created by following the same sequence of steps as shown in Figure 3. As a preprocessing step, the images are converted to gray-scale mode by using the available standard resolution, and the dense SIFT features are extracted on six multi-scales, i.e., {2,4,6,8,10,12} for the computation of codebook [38]. The step size of 5 is applied to compute the Dense SIFT features [38]. Dense features are selected, as the dense regular grid has shown to possess better discriminative power [16]. To save computation time for clustering, 40% of the features (per image) are selected by applying a random selection on a training set to compute the codebook.

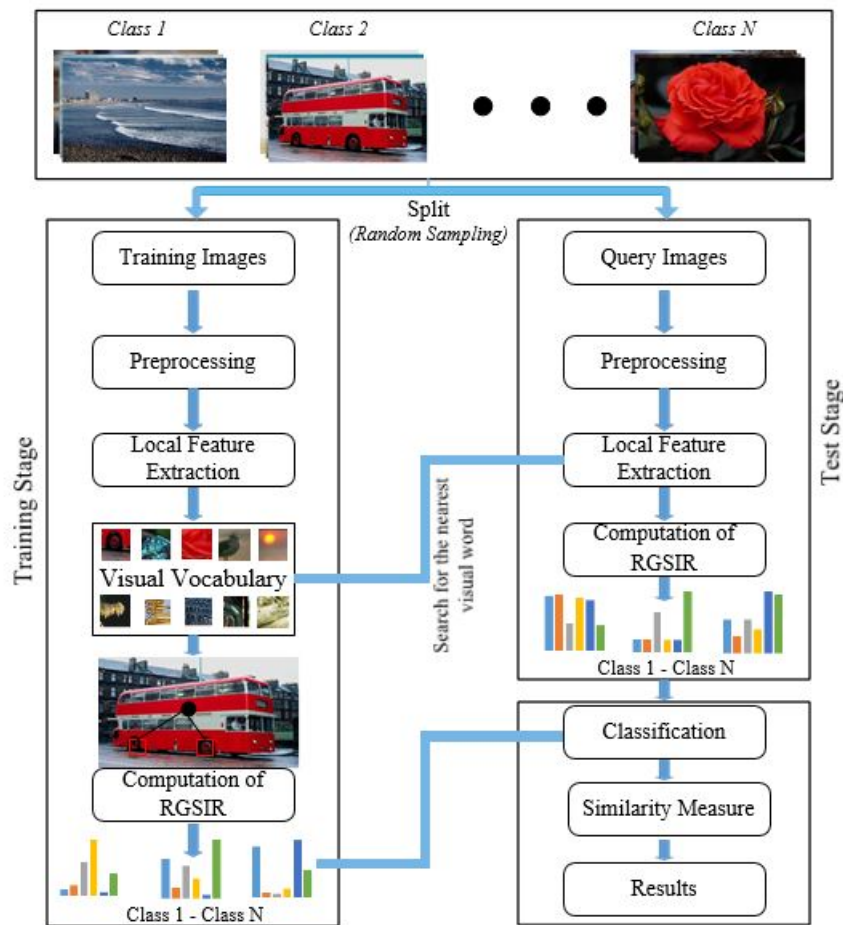
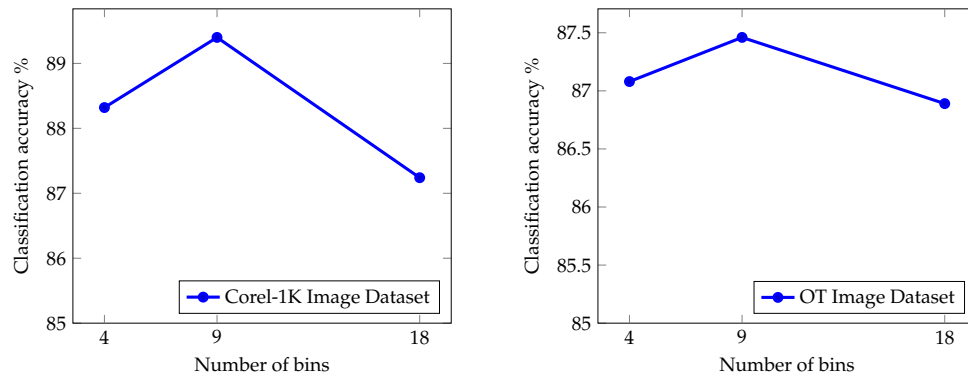


Figure 3. Block diagram of the proposed work.

To quantize the descriptors, *k-means* clustering is applied to generate visual vocabulary. Since the size of the codebook is one of the major factors that affects the performance of image retrieval, the proposed approach is evaluated by using different sizes of codebook to sort out the best retrieval performance. The visual vocabulary is constructed from the training set and the evaluation is done using the test set. The experiments are repeated in 10 trials to remove the ambiguity created by the random initialization of cluster centers by *k-means*. For each trial, the training and test images are stochastically selected and the average retrieval performance is reported in terms of precision and recall values, which are considered as standard image retrieval measures [8,39].

The calculation of RGSIR involves computing subsets of pairs from sets of identical visual words. To accelerate computation, a threshold value is set and a random selection is applied to limit the number of identical words used for creating the pair combinations. We use a nine-bin RGSIR representation for the results presented in Section 4. Figure 4 gives the empirical justification for the number of bins on two different image benchmarks used in our experiments. Support Vector Machine (SVM), a supervised learning technique, is used for classification. The SVM Hellinger Kernel is applied to the normalized RGSIR histograms. The optimal value for the regularization parameter is determined by applying 10-fold cross validation on the training dataset. As we have used a classification-based framework for image retrieval, the class of the image is predicted by using the classifier labels; similarity among the images of the same class is determined on the basis of distance in decision values [8]. The results obtained from the evaluation metrics are normalized and average values are reported in tables in graphs. MATLAB is used to simulate the research by using Corei7, a 7th generation processor with 16 GB RAM.



**Figure 4.** The influence of the number of bins on the performance of RGSIR.

#### 4. Datasets and Performance Evaluation

This section provides a description of the datasets, measures used for evaluation, and the details of the experiments conducted for the validation of the proposed research.

##### 4.1. Dataset Description

To assess the effectiveness of the proposed research for image retrieval, experiments are conducted on the benchmark datasets used extensively in the literature. The first dataset used in our experiments is the Corel-1K [40] image dataset. The Wang’s image dataset is comprised of a total of 1000 Corel images from diverse contents such as beach, flowers, horses, mountains, food, etc. The images are grouped into 10 categories with image sizes of  $256 \times 384$  or  $384 \times 256$  pixels. The second dataset is the Corel-1.5K image benchmark comprised of 15 classes with 100 images per category [40]. Figure 5 shows sample images from Corel-1K and Corel-1.5K, respectively.



**Figure 5.** Randomly selected images from each class of Corel-1K and Corel-1.5K image datasets [40].

The third dataset used to validate the efficacy of the proposed RGSIR is the Corel-2K image benchmark. Corel-2K is a subset of Corel image dataset and is comprised of 2000 images classified into 20 semantic categories. Example images from this dataset are shown in Figure 6.



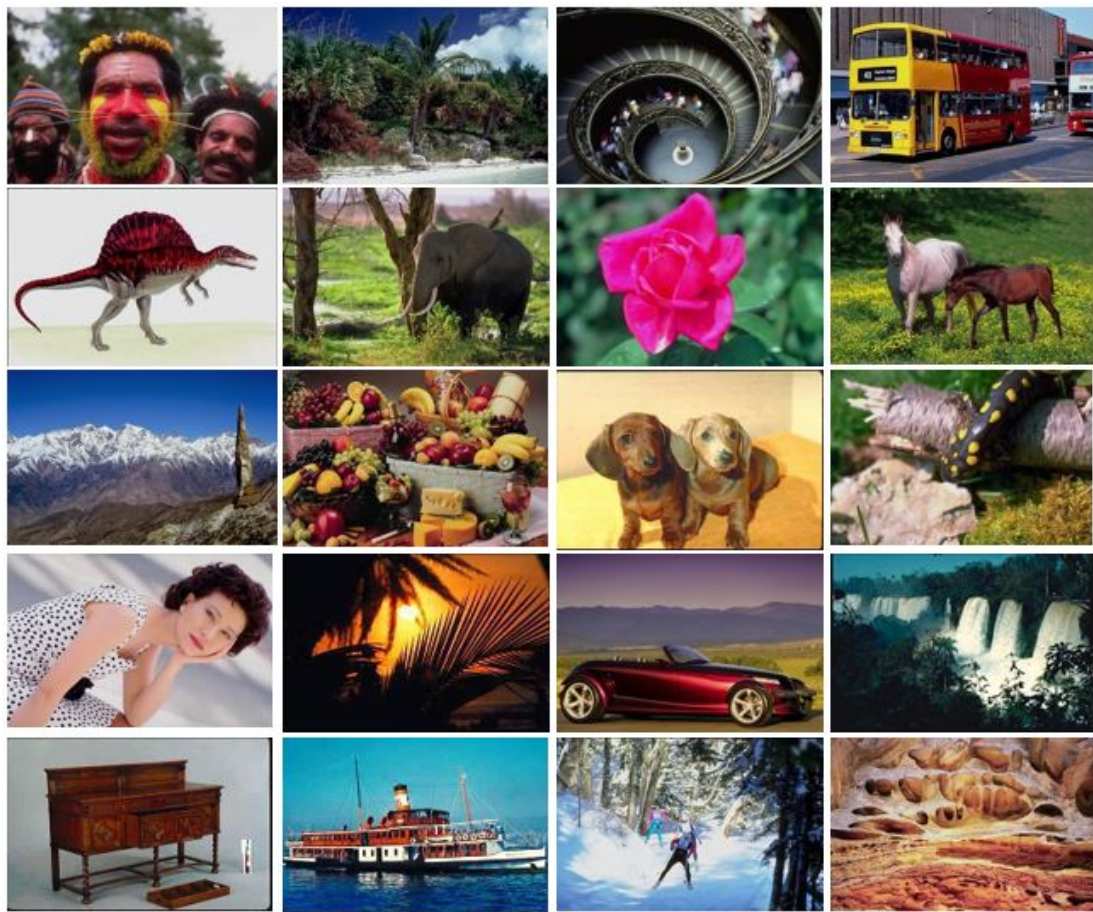


Figure 6. Class representatives from the Corel 2000 image dataset [40].

The fourth dataset is the Oliva and Torralba (OT) dataset [41], which includes 2688 images classified into 8 semantic categories. This dataset exhibits high inter and intra-class variability, as the river and forest scenes are all considered as forest. Moreover, there is no specific sky category, since all the images contain the sky object. The average image size is  $250 \times 250$  pixels and the images are collected from different sources (i.e., commercial databases, digital cameras, websites). This is a challenging dataset as the images are sampled from different perspectives, varying rotation angles, different spatial patterns and different seasons. Figure 7 shows the photo gallery of images for the OT image dataset.



Figure 7. Class representatives from the OT image dataset [41].

The last dataset used in our experiments is the RSSCN image dataset [33], released in 2015, comprised of images collected from Google Earth. It consists of 2800 images categorized into 7 typical scene categories. There are 400 images per class, and each image has a size of  $400 \times 400$  pixels. It is a

challenging dataset, as the images in each class are sampled at 4 different scales, with 100 images per scale under varied imaging angles. Consistent with related work [33], the dataset is stochastically split into two equal image subsets for training and testing, respectively. Example images from this dataset are shown in Figure 8.



**Figure 8.** Class representatives from the RSSCN image dataset [33].

#### 4.2. Evaluation Measures

Let the database  $I_1, \dots, I_n, \dots, I_N$  be a set of images represented by the spatial attributes. To retrieve an image identical to the query image  $Q$ , each image from the database  $I_n$  is compared with  $Q$ , using the appropriate distance function  $(Q, I_n)$ . The database images are then sorted based on the distances such that  $(d(Q, I_{n_i}) \leq (d(Q, I_{n_{i+1}}))$  holds for each pair images  $I_{n_i}$  and  $I_{n_{i+1}}$  of distances in the sequence  $I_{n_1}, \dots, I_{n_i}, \dots, I_{n_N}$ .

##### 4.2.1. Precision

The performance of the proposed method is measured in terms of precision  $P$  and recall  $R$ , which are the standard measures used to evaluate CBIR. Precision measures the specificity of the image retrieval, and it gives the number of relevant instances retrieved in response to a query image. The Precision ( $P$ ) is defined as

$$P = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \quad (9)$$

##### 4.2.2. Recall

The Recall is the fraction of the relevant instances retrieved to the total number of instances of that class in the dataset. It measures the sensitivity of the image and is given by

$$R = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images}} \quad (10)$$

##### 4.2.3. Mean Average Precision (MAP)

Based on  $P$  and  $R$  values, we also report results in terms of precision vs recall curve ( $P$ - $R$  curve) and the mean average precision (MAP). The  $P$ - $R$  curve represents the tradeoff between precision and recall for a given retrieval approach. It reflects more information about retrieval performance that is determined by the area under the curve. If the retrieval system has better performance, the curve is as far from the origin of coordinates as possible. The area between the curve and the  $X$ - $Y$  axes should be larger, which is usually measured and is approximate to MAP [42]. In other words, the most common

way to summarize the  $P$ - $R$  curve in one value is  $P$ - $R$ .  $P$ - $R$  is the mean of the average precision ( $AP$ ) scores of all queries and is computed as follows:

$$MAP = \frac{1}{|T|} \sum_{Q \in T} AP(Q) \tag{11}$$

where  $T$  is the set of test images or queries  $Q$ . An advantage of MAP is that it contains both precision and recall aspects and is sensitive to the entire ranking [43].

### 4.3. Performance on Corel-1K Image Dataset

The Corel-1K image benchmark is extensively used to evaluate CBIR research. To ensure fair comparison experiments, the dataset is stochastically partitioned into training and test subsets with a ratio of 0.5:0.5. The image retrieval performance of the proposed image representation is compared with the existing state-of-the-art CBIR approaches. In order to obtain a sustainable performance, the mean average precision of RGSIR is evaluated by using visual vocabulary of different sizes [50, 100, 200, 400, 600, 800]. The best image retrieval performance for Corel-1K is obtained for a vocabulary of size 600, as can be seen in Figure 9.

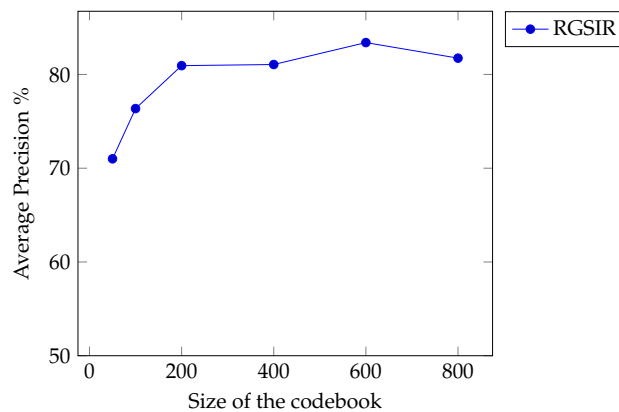


Figure 9. Average Precision as a function of vocabulary size.

The class-wise comparison obtained from the proposed research in terms of precision and recall is presented in Tables 1 and 2. It can be seen that the proposed approach outperforms the state-of-the-art image retrieval approaches. The proposed RGSIR provides 17.7% higher precision compared to Yu et al. [25]. Our proposed representation outperforms SWBOF [23] by {13.7%, 2.74%} in terms of average precision and recall values for the top 20 retrieval. RGSIR yields {8.23%, 1.65%} higher performance compared to [8] in terms of average retrieval precision and recall values.

Table 1. Comparison of precision when using Corel-1K image dataset.

Class Name/ Method	RGSIR	Li et al. [29]	Level-1 RBF-NN [9]	Visual Words Integration SIFT-SURF [8]	SWBOF [23]	SIFT-LBP [25]
African People	72.80	76.55	73.06	60.08	64.00	57.00
Beach	69.40	63.70	69.98	60.39	54.00	58.00
Building	66.20	69.05	76.76	69.66	53.00	43.00
Bus	97.16	87.70	92.24	93.65	94.00	93.00
Dinosaur	100.00	99.40	99.35	99.88	98.00	98.00
Elephant	80.80	91.05	81.38	70.76	78.00	58.00
Flower	94.60	91.70	83.40	88.37	71.00	83.00
Horse	90.80	95.40	82.81	82.77	93.00	68.00
Mountain	76.20	83.40	78.60	61.08	42.00	46.00
Food	86.00	65.80	82.71	65.09	50.00	53.00
<b>Mean</b>	<b>83.40</b>	<b>82.36</b>	<b>82.03</b>	<b>75.17</b>	<b>69.70</b>	<b>65.70</b>

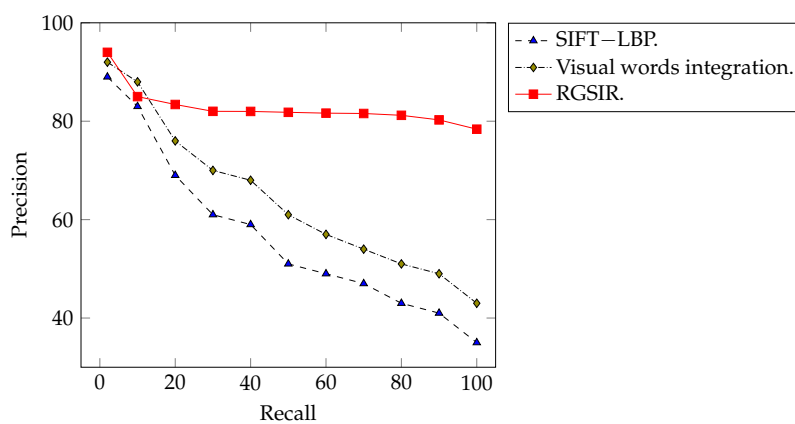
**Table 2.** Comparison of recall when using Corel-1K image dataset.

Class Name/ Method	RGSIR	Li et al. [29]	Level-1 RBF-NN [9]	Visual Words Integration SIFT-SURF [8]	SWBOF [23]	SIFT-LBP [25]
African People	14.56	15.31	14.61	12.02	12.80	11.4
Beach	13.88	12.74	14.00	12.08	10.80	11.6
Building	13.24	13.81	15.35	13.93	10.30	8.6
Bus	19.43	17.54	18.45	18.73	18.80	18.6
Dinosaur	20.00	19.88	19.87	19.98	19.60	19.6
Elephant	16.16	18.21	16.28	14.15	15.60	11.6
Flower	18.92	18.34	16.68	17.67	14.20	16.6
Horse	18.16	19.08	16.56	16.55	18.60	13.6
Mountain	15.24	16.68	15.72	12.22	8.40	9.2
Food	17.20	13.16	16.54	13.02	10.00	10.6
<b>Mean</b>	16.68	16.48	16.41	15.03	13.94	13.14

The proposed RGSIR results in {1.04%, 0.2%} higher precision and recall values compared to the work of Li et al. [29]. Experimental results validate the robustness of the proposed approach against the state-of-the-art retrieval methods.

The comparative analysis of the proposed research with the existing state-of-the-art verifies the effectiveness of RGSIR for image retrieval. The average precision depends on the total number of relevant images retrieved, and hence is directly proportional to the number of relevant images retrieved in response to a given query image. It is evident from the Figure that the proposed approach attains the highest number of relevant images against a given query image as compared to the state-of-the-art approaches. Similarly, the average recall is directly proportional to the number of relevant images retrieved to the total number of relevant images of that class present in the dataset. The proposed approach outperforms the state-of-the-art methods by attaining the highest precision and recall values.

The *P-R* curve obtained for the Corel-1K image benchmark is shown in Figure 10. The *P-R* curve demonstrates the ability of the retrieval system to retrieve relevant images from the image database in an appropriate similarity sequence. The area under the curve illustrates how effectively different methods perform in the same retrieval scenario. The results indicate that the proposed spatial features enhance the retrieval performance as compared to the state-of-the-art image retrieval approaches.



**Figure 10.** *P-R* curve obtained using Corel-1K image benchmark.

The image retrieval results for the semantic classes of Corel-1K image dataset are shown in Figures 11 and 12 (which reflects the reduction of the semantic gap). The image shown in the first row is the query image and the remaining 20 images are images retrieved by applying a similarity measure that is based on image classification score values. Here a classification label is used to determine the class of the image, while the similarity with-in the same class is calculated on the basis of similarity among classification scores of images of the same class from the test dataset.

Figure 11 shows that, for a given query image, all images of the related semantic category are retrieved. In Figure 12 it can be seen that in a search based on a flower query image, an image from a different semantic category containing flowers is also displayed in the 3rd row in addition to images from the flower image category. The experimental results demonstrate that the proposed approach achieves much higher performance compared to the state-of-the-art complementary approaches [9,23,25].

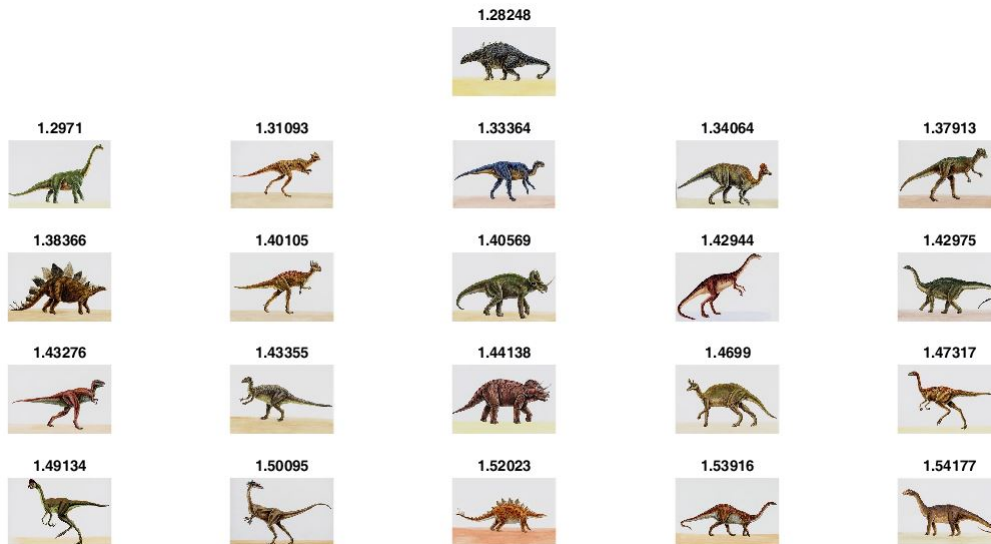


Figure 11. Result of image retrieval for the semantic class “Dinosaurs”.



Figure 12. Result of image retrieval for the semantic class “Flowers”.

#### 4.4. Performance on Corel-1.5K Image Dataset

To further assess the effectiveness of the proposed method, experiments are conducted on Corel-1.5 image benchmark. The image retrieval performance of Corel-1.5 dataset is analyzed using the visual vocabulary of different sizes. The optimal performance is obtained for a vocabulary size of 400. Table 3 provides a comparison of the mean average precision for the top 20 retrievals with the state-of-the-art image retrieval approaches [8,24,26].

It is evident from the table that the proposed RGSIR provides better retrieval performance compared to the state-of-the-art approaches with higher retrieval precision values than those of the existing research. Experimental results demonstrate that the proposed approach provides {18.9%,

3.78%) better performance compared to the method without soft assignment, i.e., SQ + Spatiogram [24] and {8.75%, 2.77%}, than the probabilistic GMM + mSpatigram [24] in terms of precision and recall, respectively.

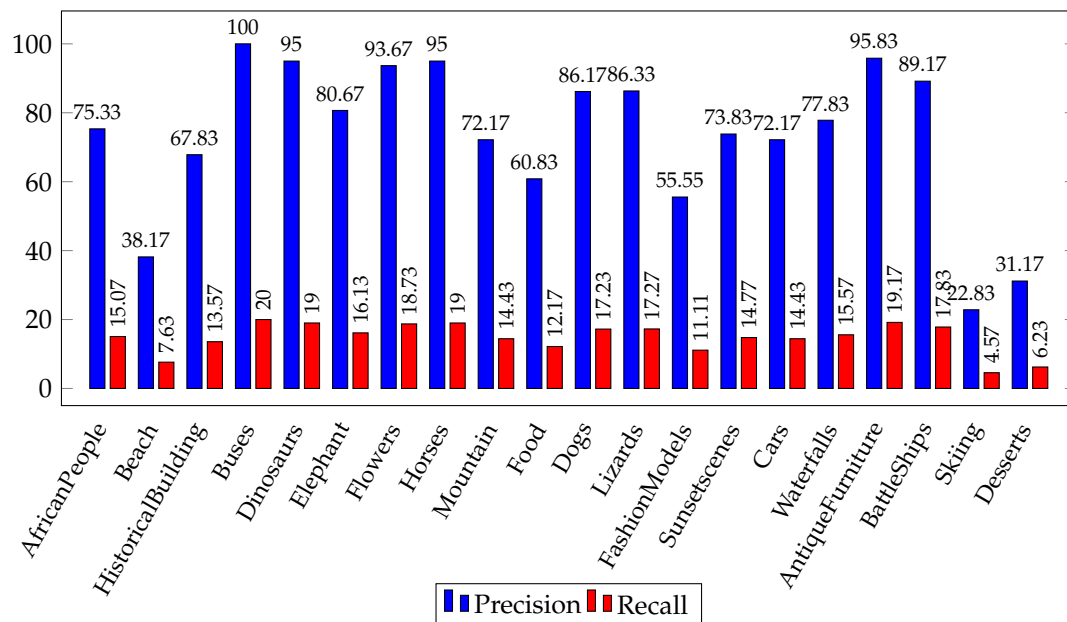
**Table 3.** Comparison of Average Retrieval Precision and Recall when using Corel-1.5K image dataset.

Performance and Name of Method	RGSIR	Ali et al. [26]	Visual words Integration SIFT-SURF [8]	GMM + mSpatigram [24]	SQ + Spatiogram [24]
Precision	82.85	72.60	74.95	74.10	63.95
Recall	16.57	14.52	14.99	13.80	12.79

The proposed approach based on relative spatial feature extraction achieves 7.9% higher retrieval precision compared to the image retrieval based on visual words integration of SIFT and SURF [8]. Our proposed approach provides {10.25%, 2.05%} better precision and recall results compared to the late fusion based approach [26]. The experimental results demonstrate that our proposed approach significantly improves the retrieval performance compared to the state-of-the-art image retrieval techniques.

#### 4.5. Performance on Corel-2K image Dataset

The optimal performance for the Corel-2K image dataset is obtained for a vocabulary size of 600. Table 4 provides a comparison of Corel-2K with the state-of-the-art image retrieval approaches. It is evident that the proposed approach yields the highest retrieval accuracy. The proposed approach provides 13.68% highest mean retrieval precision compared to the second best method. Figure 13 illustrates the average precision and recall values for the top 20 image retrievals. The experimental results validate the efficacy of the proposed approach for content-based image retrieval.



**Figure 13.** Average Precision and Recall of the proposed RGSIR for the top 20 retrievals using Corel-2K image benchmark.

**Table 4.** Comparison of the mean average precision using Corel-2K image benchmark.

Performance/Method	RGSIR	Visual Words Integration SIFT-SURF [8]	MissSVM [44]	MI-SVM [45]
MAP	79.09	65.41	65.20	54.60

The image retrieval results for the semantic classes of Corel-2K image dataset are shown in Figures 14 and 15 (which reflect the reduction of the semantic gap). The image displayed in the first row is the query image and the remaining images are the results of the top 20 retrievals selected on the basis of the image classification score displayed at the top of each image.



Figure 14. Result of image retrieval for the semantic class “Lizards”.

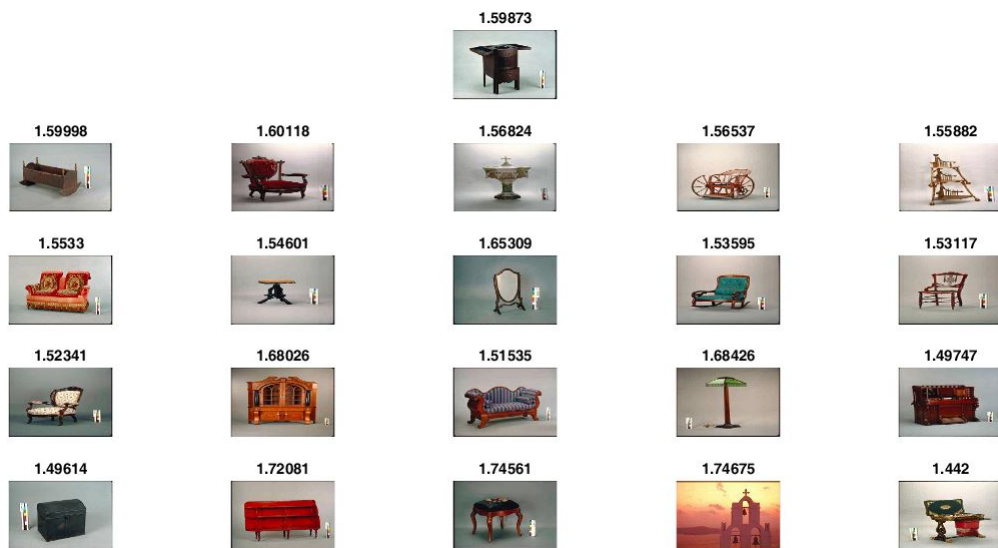
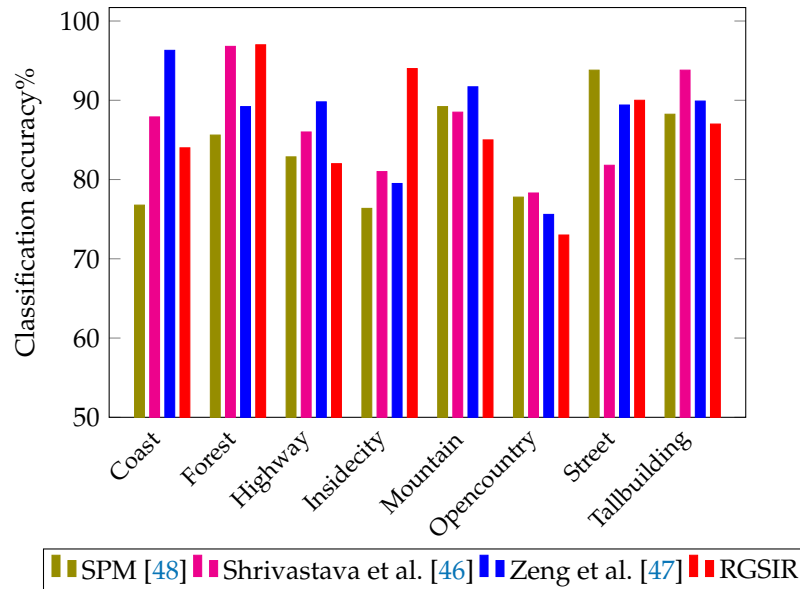


Figure 15. Result of image retrieval for the semantic class “Antique Furniture”.

#### 4.6. Image Retrieval Performance While Using Oliva and Torralba (OT-Scene) Dataset

To demonstrate the effectiveness of the proposed research, experiments are performed on the challenging OT image dataset. The best performance for the proposed research is obtained for a vocabulary size of 600. As the proposed approach has been designed on a classification-based framework, Figure 16 provides a class-wise comparison of the classification accuracy of the proposed approach with the recent state-of-the-art classification approaches [46,47]. Shrivastava et al. [46] propose a fusion of color, texture and edge descriptors to enhance the performance of image classification and report an accuracy of 86.4%. Our proposed approach outperforms SPM by 3.85% [48] and yields 1.06% higher accuracy compared to [46]. Zang et al. [47] use the Object Bank (OB) approach to construct powerful image descriptors and boost the performance of OB-based scene image classification. The best mean classification accuracy for the proposed RGSIR is 87.46%, while the accuracy reported by Zang et al. [47] is 86.5%. The proposed

approach provides 0.96% higher accuracy compared to their work. It is observed that the performance of the proposed approach is low for the natural coast and the open country category due to high variability in these classes. The proposed approach based on spatial features provides better performance compared to the state-of-the-art retrieval approaches.



**Figure 16.** Class-wise comparison between of the proposed research with the state-of-the-art methods for OT scene image dataset.

The comparison of the proposed research with existing research [8] in terms of precision is presented in Table 5. The proposed approach provides 13.17% higher accuracy compared to the second best method in comparison. The experimental results validate the efficacy of the proposed approach for content based image retrieval.

**Table 5.** Comparison of the mean average precision using OT-Scene image benchmark.

Performance/ Method	RGSIR	Visual Words Integration SIFT-SURF [8]	Log Gabor + OC-LBP Technique [49]	Late Fusion (SIFT + FREAK) [26]	Feature Extraction with Morphological Operators
MAP	82.92	69.75	63.74	63.14	60.70

#### 4.7. Performance on the RSSCN Image Dataset

To evaluate the effectiveness of proposed approach for scene classification, experiments are conducted on the challenging high resolution remote sensing scene image dataset. The training test ratio of 0.5:0.5 is used for the RSSCN image dataset as is followed in the literature [33]. The training set comprises 1400 stochastically selected images and the remaining images are used to assess the retrieval performance. The optimal retrieval performance is obtained for a visual vocabulary size of 200. As we have used a classification based framework for image retrieval, it is important to note here that the classification accuracy for the proposed RGSIR is 81.44% and the accuracy reported by the dataset creator is 77%. Our proposed representation provides 4.44% higher accuracy compared to the deep learning technique, i.e., the DBN adopted by the Zou et al. [33].

Table 6 provides a comparison of the retrieval performance of RSSCN with the state-of-the-art image retrieval approaches. We have computed MAP for the top 100 retrievals using the proposed RGSIR. Xia et al. [34] perform an extensive analysis to develop a powerful feature representation to enhance image retrieval. They consider different CNN representative models, i.e., CaffeNet [50],

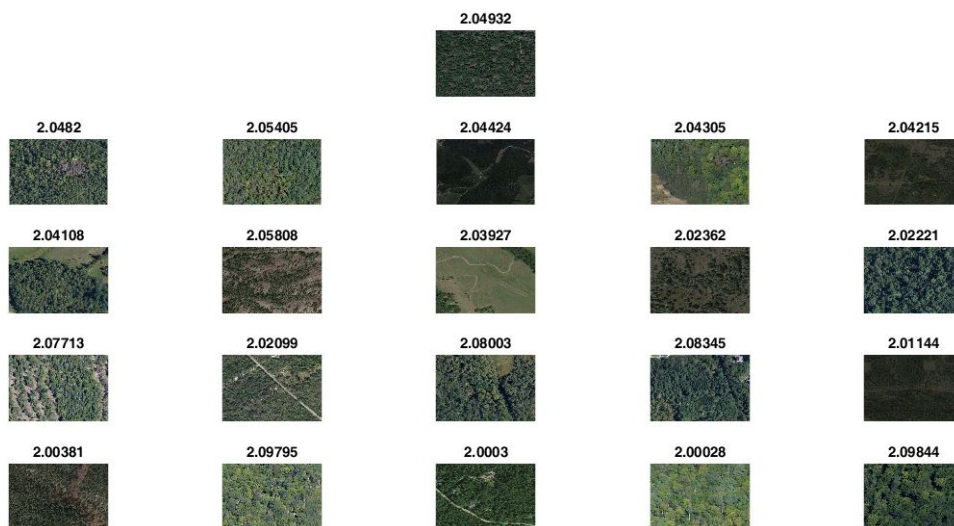


VGG-M [51], VGG-VD19 [52] and GoogLeNet [53], in combination with different feature extraction approaches. As our proposed approach is based on mid-level features, we have selected BoW based aggregation methods for comparison. Mid-level features are more resilient to various transformations such as rotation, scale and illumination [34]. The proposed approach provides 16.63% higher accuracy compared to VGG-M (IFK). The proposed RGSIR outperforms the GoogLeNet (BoW), VGG-VD19 (BoW) and CaffeNet (BoW) by 18.56%, 19.2% and 20.88 %, respectively.

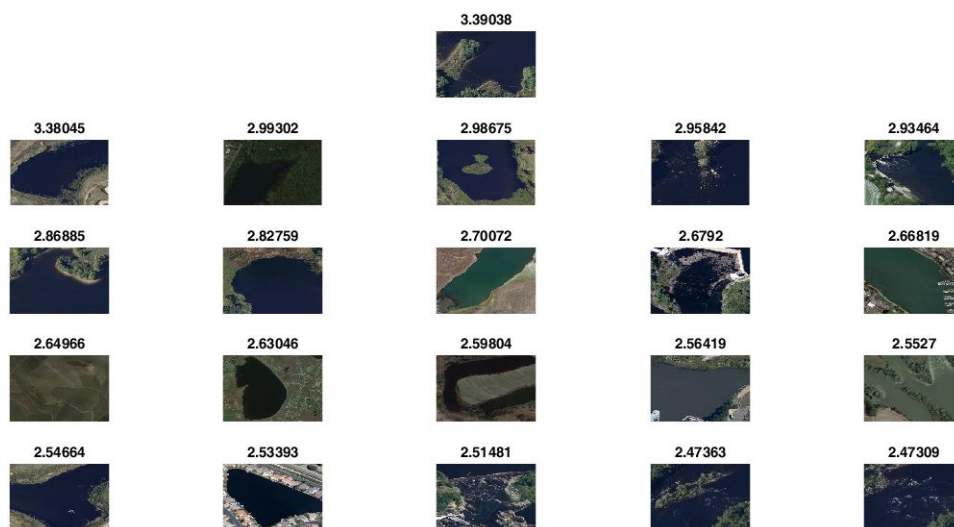
**Table 6.** Comparison of the mean average precision when using RSSCN image benchmark.

Performance/ Method	RGSIR	CaffeNet (BoW) [34]	VGG-VD19 (BoW) [34]	GoogLeNet (BoW) [34]	VGG-M (IFK) [34]
MAP	72.42	51.54	53.22	53.86	55.79

It is important to note here that we have selected the RSSCN image dataset as the images are captured at varying angles and exhibit significant rotation differences. Hence the robustness of the proposed approach to rotation in-variance is also illustrated to some extent. The top 20 retrieval results against the “Forest” and “River & Lake” semantic categories of the RSSCN image dataset are shown in Figures 17 and 18.



**Figure 17.** Results of image retrieval for the semantic class “Forest”.



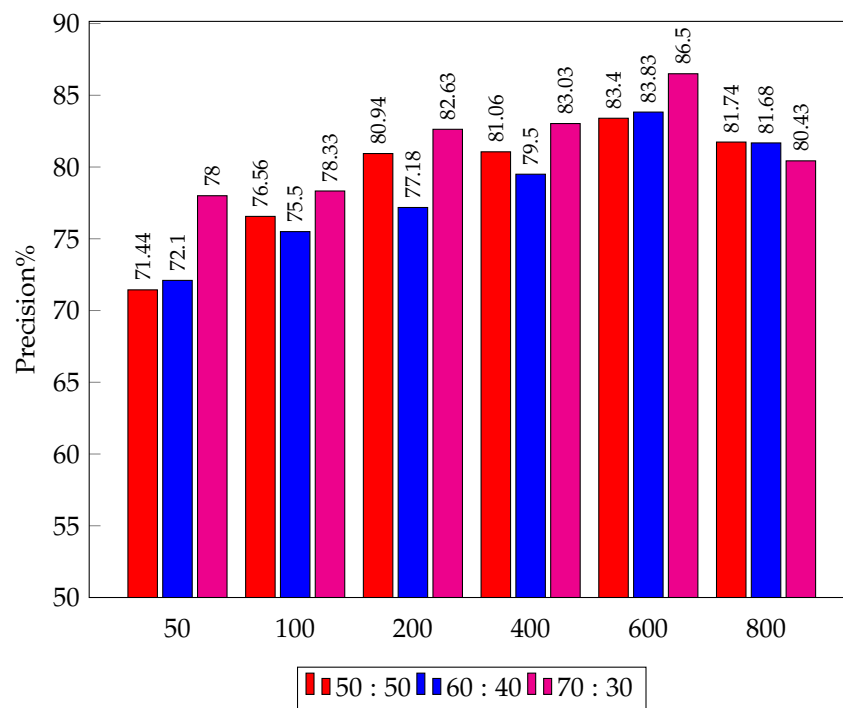
**Figure 18.** Results of image retrieval for the semantic class “River & Lake”.

## 5. Discussion

In this paper, we have proposed an image retrieval approach based on relative geometric spatial relationships between visual words. Extensive experiments on challenging image benchmarks demonstrate that the proposed approach outperforms the concurrent and the state-of-the-art image retrieval approaches based on feature fusion and spatial feature extraction techniques [8,23,24].

### 5.1. Factors Affecting the Performance of the System

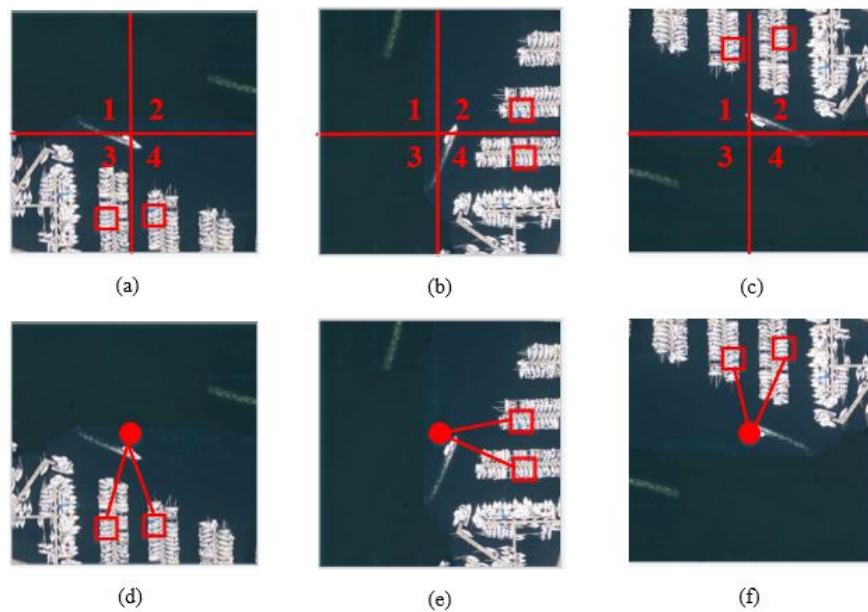
One of the factors affecting the retrieval performance is the size of the visual vocabulary. We have conducted experiments with visual vocabulary of different sizes to determine the optimal performance of the proposed representation as discussed in the preceding sections. Another factor affecting the performance of the system is the ratio of the training images used to train the classifier. Figure 19 provides a comparison of different training test ratios i.e., 70:30, 60:40, 50:50 for the Corel-1K image dataset. It can be seen that the performance of the system increases at higher training test ratios. However, to be consistent with related approaches [8], 50:50 is used to report the precision and recall retrieval results for the experimental comparisons presented in Section 4.



**Figure 19.** Average precision of the proposed RGSIR on visual vocabulary of different sizes using different training test ratios.

### 5.2. Invariance to Basic Transformations

Spatial Pyramid Matching (SPM) [16] is the most notable contribution to incorporate spatial context into the BoVW model. SPM captures the absolute spatial distribution of visual words. However, SPM is sensitive to image transformations such as rotation, flipping and translation. For images that are not well-aligned, SPM may lose its discriminative power. An object may rotate by any angle on the image plane (rotation), it may be flipped horizontally or vertically (flipping), or the object may appear anywhere in an image (translation). The proposed approach is capable of addressing various transformations, by encoding the global relative spatial orientation of visual words. This is achieved by computing the angle between identical visual word pairs with respect to the centroid in image. Figure 20 provides an illustration to better understand our approach.

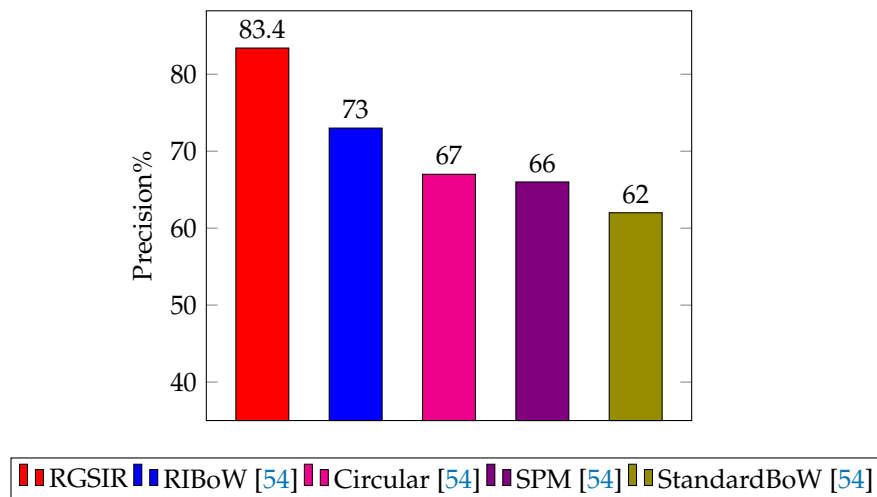


**Figure 20.** SPM (a,b,c) vs.the proposed approach (d,e,f). Here (a,d) represent the original images, (b,e) the images rotated by  $90^\circ$ , and (c,f) vertically flipped images.

The upper region of Figure 20a–c represents the idea of histograms constructed with SPM [16], while the lower region demonstrates the proposed approach Figure 20d–f. In the figures, we can see Figure 20a,d the original image, Figure 20b,e the image rotated by  $90^\circ$  and Figure 20c,f the vertically flipped image. The performance of SPM [16] degrades in this case, as the objects occupy different regions in the original and transformed images. In Figure 20a the identical visual words are located in the 3rd and 4th regions, in Figure 20b they are found in the 2nd and 4th regions, while in Figure 20c they are in the 1st and 2nd regions, respectively. Hence the three histogram representations will be different for the same image. In the case of the proposed RGSIR, the same histogram representation will be generated for the original and for the transformed images, as the angle between identical visual words with respect to the centroid remains the same.

Figure 21 presents a graphical comparison of the average precision for the top 20 retrievals with the concurrent state-of-the-art approaches. Chaturani et al. [54] propose a Rotation Invariant Bag of Visual Words (RIBoW) approach to encode the spatial information using circular image decomposition in combination with a simple shifting operation using global image descriptors. They report improved performance to existing BoVW approaches. Although SPM [16] encodes the spatial information, it is sensitive to rotation, translation and scale variance of an image. The circular decomposition approach [54] partitions the image into sub-images, and features which are then extracted from each sub-image are used for feature representation. The proposed RGSIR provides 10.4% higher retrieval precision compared to the second best method.

Experimental results demonstrate the superiority of the proposed approach to the concurrent state-of-the-art approaches. It is important to note here that some approaches incorporate the spatial context prior to the visual vocabulary construction step, while others do so after it [9]. The proposed approach adds this information after the visual vocabulary construction step. In future, we intend to enhance the discriminative power of the proposed approach by extracting rotation-invariant features at the feature extraction step, prior to the construction of the visual vocabulary.



**Figure 21.** Average precision comparison of the proposed RGSIR with the state-of-the-art approaches for the Corel-1K image benchmark.

## 6. Conclusions and Future Directions

The final feature vector for the BoVW model contains no information regarding the distribution of visual words in the 2D image space. Due to this reason, the performance of a computer vision application suffers, as spatial information of visual words in the histogram-based feature vector enhances the performance of image retrieval. This paper presents a novel approach to image representation to incorporate the spatial information to the inverted index of the BoVW model. The spatial information is added by calculating the global relative spatial orientation of visual words in a transformation-invariant manner. This is established by computing the geometric relationship between pairs of identical visual words with respect to the centroid of an image. The experimental results and quantitative comparisons demonstrate that our proposed representation significantly improves the retrieval performance in terms of precision and recall values. The proposed approach outperforms other concurrent methods and provides competitive performance as compared with the state-of-the-art approaches.

Furthermore, the proposed approach is not confined to the retrieval task but can be applied to other image analysis tasks, such as object detection. This is because we incorporate the invariant spatial layout information into the BoVW image representation, thereby ensuring seamless application of follow-up techniques.

In future, we would like to enhance the discriminative power of the proposed approach by extracting rotation invariant low-level features at descriptor level. We intend to create a unified representation, tolerant to all kinds of layout variances. As the proposed method has shown excellent results on five image benchmarks, in future we aim to apply a pre-trained deep convolution neural network for the computation of histogram of visual words for learning of classifier to a large scale image dataset. Combining our image representation with a complementary absolute feature extraction method and enriching it with other cues such as color and shape is another possible direction for future research.

**Author Contributions:** Conceptualization, B.Z., R.A. and N.A.; Data curation, B.Z., R.A. and N.A.; Formal analysis, B.Z., R.A. and N.A.; Investigation, B.Z., R.A., N.A. and S.H.D.; Methodology, B.Z. and N.A.; Project administration, N.A., M.S., S.H.D. and N.I.R.; Resources, M.K.I., M.S., S.H.D. and N.I.R.; Software, B.Z., N.A. and M.S.; Supervision, R.A. and N.A.; Validation, B.Z., N.A., M.K.I. and N.I.R.; Visualization, B.Z., M.K.I., S.H.D. and N.I.R.; Writing—original draft, B.Z., R.A., N.A., M.K.I., M.S., S.H.D. and N.I.R.; Writing—review and editing, R.A., N.A., M.K.I., M.S., S.H.D. and N.I.R.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Irtaza, A.; Adnan, S.M.; Ahmed, K.T.; Jaffar, A.; Khan, A.; Javed, A.; Mahmood, M.T. An Ensemble Based Evolutionary Approach to the Class Imbalance Problem with Applications in CBIR. *Appl. Sci.* **2018**, *8*, 495. [[CrossRef](#)]
2. Ye, J.; Kobayashi, T.; Toyama, N.; Tsuda, H.; Murakawa, M. Acoustic Scene Classification Using Efficient Summary Statistics and Multiple Spectro-Temporal Descriptor Fusion. *Appl. Sci.* **2018**, *8*, 1363. [[CrossRef](#)]
3. Piras, L.; Giacinto, G. Information fusion in content based image retrieval: A comprehensive overview. *Inf. Fusion* **2017**, *37*, 50–60. [[CrossRef](#)]
4. Nazir, A.; Ashraf, R.; Hamdani, T.; Ali, N. Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor. In Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 3–4 March 2018; pp. 1–6.
5. Zhu, L.; Shen, J.; Xie, L.; Cheng, Z. Unsupervised visual hashing with semantic assistant for content-based image retrieval. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 472–486. [[CrossRef](#)]
6. Alzu'bi, A.; Amira, A.; Ramzan, N. Semantic content-based image retrieval: A comprehensive study. *J. Vis. Commun. Image Represent.* **2015**, *32*, 20–54. [[CrossRef](#)]
7. Wan, J.; Wang, D.; Hoi, S.C.H.; Wu, P.; Zhu, J.; Zhang, Y.; Li, J. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In Proceedings of the ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 157–166. [[CrossRef](#)]
8. Ali, N.; Bajwa, K.B.; Sablatnig, R.; Chatzichristofis, S.A.; Iqbal, Z.; Rashid, M.; Habib, H.A. A novel image retrieval based on visual words integration of SIFT and SURF. *PLoS ONE* **2016**, *11*, e0157428. [[CrossRef](#)] [[PubMed](#)]
9. Ali, N.; Bajwa, K.B.; Sablatnig, R.; Mehmood, Z. Image retrieval by addition of spatial information based on histograms of triangular regions. *Comput. Electr. Eng.* **2016**, *54*, 539–550. [[CrossRef](#)]
10. O'Hara, S.; Draper, B.A. Introduction to the bag of features paradigm for image classification and retrieval. *arXiv* **2011**, arXiv:1101.3354.
11. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1470–1477.
12. Liu, P.; Miao, Z.; Guo, H.; Wang, Y.; Ai, N. Adding spatial distribution clue to aggregated vector in image retrieval. *EURASIP J. Image Video Process.* **2018**, *2018*, 9. [[CrossRef](#)]
13. Anwar, H.; Zambanini, S.; Kampel, M.; Vondrovec, K. Ancient Coin Classification Using Reverse Motif Recognition: Image-based classification of Roman Republican coins. *IEEE Signal Process. Mag.* **2015**, *32*, 64–74. [[CrossRef](#)]
14. Ali, N.; Zafar, B.; Riaz, F.; Dar, S.H.; Ratyal, N.I.; Bajwa, K.B.; Iqbal, M.K.; Sajid, M. A Hybrid Geometric Spatial Image Representation for scene classification. *PLoS ONE* **2018**, *13*, e0203339. [[CrossRef](#)] [[PubMed](#)]
15. Zafar, B.; Ashraf, R.; Ali, N.; Ahmed, M.; Jabbar, S.; Naseer, K.; Ahmad, A.; Jeon, G. Intelligent Image Classification-Based on Spatial Weighted Histograms of Concentric Circles. *Comput. Sci. Inf. Syst.* **2018**, *15*, 615–633. [[CrossRef](#)]
16. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2169–2178.
17. Li, X.; Song, Y.; Lu, Y.; Tian, Q. Spatial pooling for transformation invariant image representation. In Proceedings of the 19th ACM International Conference on Multimedia, Scottsdale, AZ, USA, 28 November–1 December 2011; pp. 1509–1512.
18. Karmakar, P.; Teng, S.W.; Lu, G.; Zhang, D. Rotation Invariant Spatial Pyramid Matching for Image Classification. In Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Adelaide, Australia, 23–25 November 2015; pp. 1–8.
19. Liu, D.; Hua, G.; Viola, P.; Chen, T. Integrated feature selection and higher-order spatial feature extraction for object categorization. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.

20. Khan, R.; Barat, C.; Muselet, D.; Ducottet, C. Spatial orientations of visual word pairs to improve bag-of-visual-words model. In Proceedings of the British Machine Vision Conference, Surrey, UK, 3–7 September 2012; pp. 89.1–89.11.
21. Zafar, B.; Ashraf, R.; Ali, N.; Ahmed, M.; Jabbar, S.; Chatzichristofis, S.A. Image classification by addition of spatial information based on histograms of orthogonal vectors. *PLoS ONE* **2018**, *13*, e0198175. [[CrossRef](#)] [[PubMed](#)]
22. Ahmed, K.T.; Irtaza, A.; Iqbal, M.A. Fusion of local and global features for effective image extraction. *Appl. Intell.* **2017**, *47*, 526–543. [[CrossRef](#)]
23. Wang, C.; Zhang, B.; Qin, Z.; Xiong, J. Spatial weighting for bag-of-features based image retrieval. In *Integrated Uncertainty in Knowledge Modelling and Decision Making*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 91–100.
24. Zeng, S.; Huang, R.; Wang, H.; Kang, Z. Image retrieval using spatiograms of colors quantized by gaussian mixture models. *Neurocomputing* **2016**, *171*, 673–684. [[CrossRef](#)]
25. Yu, J.; Qin, Z.; Wan, T.; Zhang, X. Feature integration analysis of bag-of-features model for image retrieval. *Neurocomputing* **2013**, *120*, 355–364. [[CrossRef](#)]
26. Ali, N.; Mazhar, D.A.; Iqbal, Z.; Ashraf, R.; Ahmed, J.; Khan, F.Z. Content-Based Image Retrieval Based on Late Fusion of Binary and Local Descriptors. *arXiv* **2017**, arXiv:1703.08492.
27. Filliat, D. A visual bag of words method for interactive qualitative localization and mapping. In Proceedings of the 2007 IEEE International Conference on Robotics and Automation, Roma, Italy, 10–14 April 2007; pp. 3921–3926.
28. Hu, F.; Zhu, Z.; Mejia, J.; Tang, H.; Zhang, J. Real-time indoor assistive localization with mobile omnidirectional vision and cloud GPU acceleration. *AIMS Electron. Electr. Eng.* **2017**, *1*, 74–99. [[CrossRef](#)]
29. Li, L.; Feng, L.; Wu, J.; Sun, M.X.; Liu, S.I. Exploiting global and local features for image retrieval. *J. Cent. South Univ.* **2018**, *25*, 259–276. [[CrossRef](#)]
30. Liu, S.; Wu, J.; Feng, L.; Qiao, H.; Liu, Y.; Luo, W.; Wang, W. Perceptual uniform descriptor and ranking on manifold for image retrieval. *Inf. Sci.* **2018**, *424*, 235–249. [[CrossRef](#)]
31. Wu, J.; Feng, L.; Liu, S.; Sun, M. Image retrieval framework based on texton uniform descriptor and modified manifold ranking. *J. Vis. Commun. Image Represent.* **2017**, *49*, 78–88. [[CrossRef](#)]
32. Varish, N.; Pradhan, J.; Pal, A.K. Image retrieval based on non-uniform bins of color histogram and dual tree complex wavelet transform. *Multimedia Tools Appl.* **2017**, *76*, 15885–15921. [[CrossRef](#)]
33. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [[CrossRef](#)]
34. Xia, G.S.; Tong, X.Y.; Hu, F.; Zhong, Y.; Datcu, M.; Zhang, L. Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation. *arXiv* **2017**, arXiv:1707.07321.
35. Vassou, S.A.; Anagnostopoulos, N.; Amanatiadis, A.; Christodoulou, K.; Chatzichristofis, S.A. Como: A compact composite moment-based descriptor for image retrieval. In Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, Florence, Italy, 19–21 June 2017; p. 30.
36. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
37. Anwar, H.; Zambanini, S.; Kampel, M. Encoding spatial arrangements of visual words for rotation-invariant image classification. In Proceedings of the 36th German Conference, GCPR 2014, Münster, Germany, 2–5 September 2014; pp. 443–452.
38. Tuytelaars, T. Dense interest points. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2281–2288.
39. Mehmood, Z.; Anwar, S.M.; Ali, N.; Habib, H.A.; Rashid, M. A novel image retrieval based on a combination of local and global histograms of visual words. *Math. Probl. Eng.* **2016**, *2016*, 8217250. [[CrossRef](#)]
40. Li, J.; Wang, J.Z. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 985–1002. [[PubMed](#)]
41. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [[CrossRef](#)]
42. Zhou, J.; Liu, X.; Liu, W.; Gan, J. Image retrieval based on effective feature extraction and diffusion process. *Multimedia Tools Appl.* **2018**, 1–28. [[CrossRef](#)]

43. Deselaers, T.; Keysers, D.; Ney, H. Features for image retrieval: An experimental comparison. *Inf. Retr.* **2008**, *11*, 77–107. [[CrossRef](#)]
44. Zhou, Z.H.; Xu, J.M. On the relation between multi-instance learning and semi-supervised learning. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 1167–1174.
45. Andrews, S.; Tsochantaridis, I.; Hofmann, T. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2002; pp. 577–584.
46. Shrivastava, P.; Bhoyar, K.; Zadgaonkar, A. Image Classification Using Fusion of Holistic Visual Descriptions. *Int. J. Image Graph. Signal Process.* **2016**, *8*, 47. [[CrossRef](#)]
47. Zang, M.; Wen, D.; Liu, T.; Zou, H.; Liu, C. A pooled Object Bank descriptor for image scene classification. *Expert Syst. Appl.* **2018**, *94*, 250–264. [[CrossRef](#)]
48. Yin, H. Scene Classification Using Spatial Pyramid Matching and Hierarchical Dirichlet Processes. MSc Thesis, Rochester Institute of Technology, Rochester, NY, USA, 2010.
49. Walia, E.; Verma, V. Boosting local texture descriptors with Log-Gabor filters response for improved image retrieval. *Int. J. Multimedia Inf. Retr.* **2016**, *5*, 173–184. [[CrossRef](#)]
50. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, OR, Florida, USA, 3–7 November 2014; pp. 675–678.
51. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
52. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
53. Mousavian, A.; Kosecka, J. Deep convolutional features for image based retrieval and scene categorization. *arXiv* **2015**, arXiv:1509.06033.
54. Chathurani, N.; Geva, S.; Chandran, V.; Cynthujah, V. Content-Based Image (Object) Retrieval with Rotational Invariant Bag-of-Visual Words Representation. In Proceedings of the 2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, 18–20 December 2015; pp. 152–157.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).