Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 8217250, 12 pages
http://dx.doi.org/10.1155/2016/8217250

Research Article

A Novel Image Retrieval Based on a Combination of Local and Global Histograms of Visual Words

Zahid Mehmood,[1,2] Syed Muhammad Anwar,[2] Nouman Ali,[1] Hafiz Adnan Habib,[3] and Muhammad Rashid[4]

[1]*Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan*
[2]*Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan*
[3]*Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan*
[4]*Department of Computer Engineering, Umm Al-Qura University, Makkah 21421, Saudi Arabia*

Correspondence should be addressed to Zahid Mehmood; zahid.mehmood@uettaxila.edu.pk

Received 19 April 2016; Revised 14 June 2016; Accepted 19 June 2016

Academic Editor: Jinyang Liang

Content-based image retrieval (CBIR) provides a sustainable solution to retrieve similar images from an image archive. In the last few years, the Bag-of-Visual-Words (BoVW) model gained attention and significantly improved the performance of image retrieval. In the standard BoVW model, an image is represented as an orderless global histogram of visual words by ignoring the spatial layout. The spatial layout of an image carries significant information that can enhance the performance of CBIR. In this paper, we are presenting a novel image representation that is based on a combination of local and global histograms of visual words. The global histogram of visual words is constructed over the whole image, while the local histogram of visual words is constructed over the local rectangular region of the image. The local histogram contains the spatial information about the salient objects. Extensive experiments and comparisons conducted on Corel-A, Caltech-256, and Ground Truth image datasets demonstrate that the proposed image representation increases the performance of image retrieval.

1. Introduction

CBIR is used to search the images from an image archive that are in a semantic relationship with the query image [1–3]. Occlusion, overlapping objects, spatial layout, image resolution, variations in illumination, semantic gap, and exponential growth in multimedia contents make CBIR a challenging problem [1–3]. In CBIR, the feature vector is used to represent the image in the form of low-level visual features [1, 2]. The feature vector of a query image is computed and compared with the feature vectors of the images placed in an image archive [4]. The closeness of the feature vector values determines the output. The appearance of a similar view in the images belonging to the different classes result in the closeness of the feature vector values and it decreases the performance of image retrieval [4]. The main focus of the research in CBIR is to retrieve the images that are in a semantic relationship with the query image [3, 5].

In the standard BoVW model [6], an image is represented as an orderless global histogram of visual words by ignoring the spatial layout of 2D image space. The spatial attributes of an image carry information that enhances the performance of image retrieval [7]. The approaches based on query expansion [8], large vocabulary size [7], and soft quantization [9] are applied to enhance the performance of CBIR. All of these approaches ignore the spatial layout that provides the discriminating details [7]. According to Anwar et al. [10], two approaches add the spatial information to the inverted index of the BoVW based image representation. The first approach deals with the visual words cooccurrence and requires computational cost with the larger size vocabulary [11]. The second approach divides an image into subregions and construct histograms from each of the subregions [12]. Various types of image representations are proposed by selecting different semantic regions from the images [12–16].

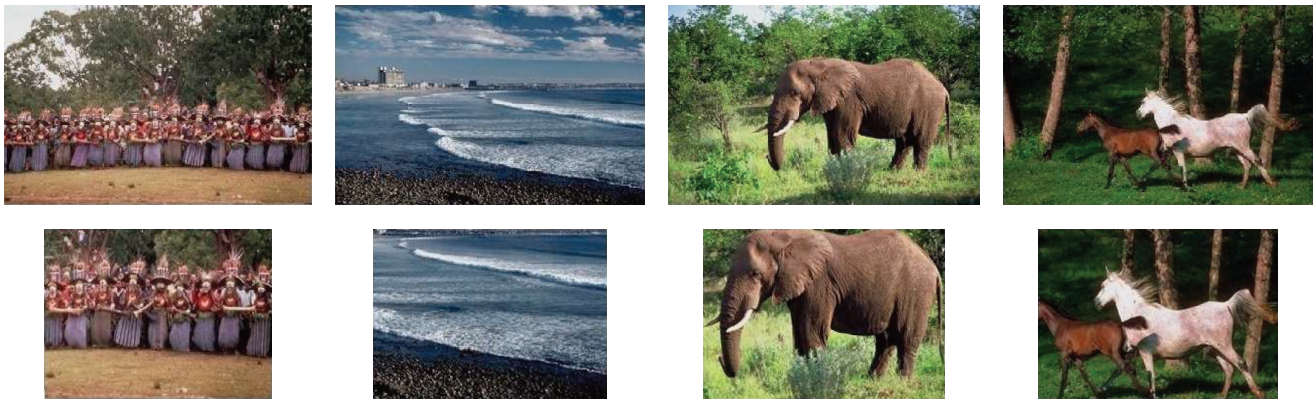


Figure 1: Images from four different classes of Corel-A image benchmark.

Keeping in view the robust performance of the second approach [12, 14], the proposed research is constructing two histograms from a single image. The global histogram of visual words is constructed over the whole image, while the local histogram of visual words is constructed over the local rectangular region of the image. The details about the selection of the local rectangular area for the construction of local histogram are mentioned in Section 3. Figure 1 is representing the images from four different classes (Africa, Beach, Elephants, and Horses) of Corel-A image benchmark. Grass, sky, and trees are in the global appearance of all these images, while the objects of interest or salient objects lie in the central region of the images (like people, beach, elephant, and horses). There is no semantic relationship between these images, as they belong to the different classes. The discriminating information is likely to be available in the central region of an image. The second row of Figure 1 is representing the extracted local rectangular regions of these images (area of the image that is selected for the construction of local histogram of visual words). The extracted local rectangular regions of all of these images are discriminating, as they contain the information about the salient objects of the image. Figure 2 is representing an image from the semantic class Elephant of the Corel-A image benchmark. The global appearance of the image contains sky, trees, and grass, while the main object of interest in the image is elephant. The construction of a local histogram from the image central area adds the spatial attributes of the salient object to the inverted index of the BoVW representation.

Keeping these facts in view, dense features are extracted from a set of training and test images; the feature space is quantized to construct a visual vocabulary. Images are represented as histograms of visual words (by using a combination of local and global histograms of visual words). The local histogram that is constructed over the local rectangular region of an image contains the spatial information about the salient objects. The global and local histograms of visual words are concatenated and this information is added to the inverted index of the BoVW representation. The main contributions of this paper are

(1) image representation as a combination of local and global histograms of visual words;

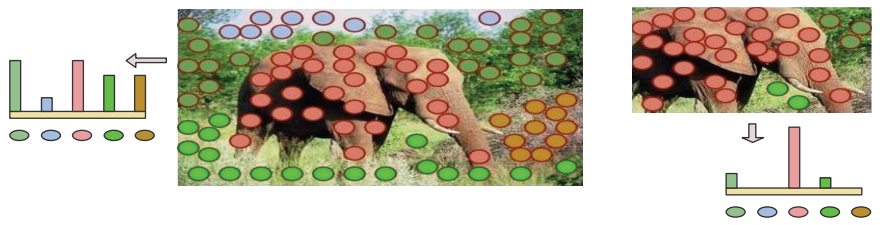
(2) the addition of spatial information from the central area of the image to the inverted index of BoVW representation;

(3) reduction of semantic gap between the low-level features of an image and high-level semantic concepts.

The rest of the paper is organized as follows. Section 2 is about the related research. Section 3 describes the proposed research methodology. Section 4 is about the experimental details and results conducted on three image benchmarks. Section 5 concludes our research work and points towards the future directions.

2. Related Work

Query by Image Content (QBIC) is the first system launched by IBM for image search [2]. After that, a variety of image retrieval techniques are proposed that are based on color, texture, and shape [1–3, 5]. Interest points detectors like Histogram of Oriented Gradients (HOG) [22], Scale-Invariant Feature Transform (SIFT) [23], Maximally Stable Extremal Regions (MSER) [24], Speeded-Up Robust Features (SURF) [25], and BRISK features [26] are used for robust content-based image matching [27]. Ashraf et al. [28] proposed an image retrieval by using a combination of color features and bandlet transformation. The bandlet transformation-based representation was selected to extract the salient objects from the images. Artificial Neural Networks (ANN) with an inverted index was selected for an efficient image retrieval. Zeng et al. [29] proposed an image representation based on spatiogram that was generalized histogram of quantized colors. In the first step, the color space was quantized by applying the Gaussian Mixture Models (GMM) and Expectation-Maximization (EM) algorithm. The number of quantized color bins was determined on the basis of Bayesian Information Criterion (BIC). A spatiogram was based on a histogram with a spatial distribution of color and each bin represents the weighted distribution of pixels. The similarity between the two images was determined on the basis of the similarity between the respective spatiograms. Walia and Pal [30] proposed a framework for color image retrieval by using a combination of low-level features. The Color Difference Histogram (CDH) was used to extract the color



(a) (b)

Figure 2: (a) is presenting the procedure for computation of global histogram, while (b) is presenting the procedure for the extraction of local histogram.

and texture. The shape features were extracted by applying an Angular Radial Transform (ART). The retrieval effectiveness of the proposed framework was enhanced by modifying the CDH algorithm. Irtaza and Jaffar [31] proposed a combination of Genetic Algorithm (GA) and Support Vector Machines (SVM) to reduce the semantic gap. Images were represented as a combination of curvelet, wavelet, and Gabor features. Relevance Feedback (RF) was applied to enhance the accuracy of the proposed work. Tian et al. [19] proposed a rotation-invariant and scale-invariant Edge Orientation Difference Histogram (EODH) descriptor. Steerable filter and vector sum were applied to obtain the main orientation of the pixels. The effectiveness of the feature space was improved by selecting a combination of Color SIFT and EODH. Codebook was constructed by applying the weighted average of Color SIFT and EODH. Yu et al. [32] proposed an image retrieval by using a combination of Local Binary Pattern (LBP) with HOG and SIFT. The midlevel features' combination was selected to achieve high performance for the complex background. Visual features were separately extracted from the images by using SIFT and LBP. Weighted average clustering was applied for the codebook construction to obtain the integration of two midlevel features. According to the experimental results [32], the best performance of image retrieval was obtained by applying the feature integration of SIFT and LBP. Wang et al. [15] proposed Spatial Weighting Bag-of-Features (SWBOF) framework and extracted spatial information from the subblocks of the images. Local entropy, local variance, and adjacent blocks distance were selected to calculate the spatial information. According to experimental results [15], SWBOF with spatial information performs better than the traditional BoF representation. Yildizer et al. [33] proposed CBIR by using multiple Support Vector Machines ensemble. SVM is used for regression and Support Vector Regression (SVR) ensemble is selected for classification. According to the experimental results, the proposed technique improves the accuracy of image retrieval. Cardoso et al. [20] proposed iterative technique for CBIR that is based on Multiple SVM Ensembles. Discrete Cosine Transform (DCT) is selected for feature extraction and multi-SVM is used for classification. Yildizer et al. [21] integrated wavelets for an effective content-based image retrieval. k-means with B+-tree data structure is used for clustering with Daubechies wavelet transform that has excellent spatial and spectral locality properties, which make it very useful for CBIR; it is applied to partitioning the images into different levels. Youssef [17] proposed a novel Integrating Curvelet Transform with Enhanced Dominant Colors extraction and Texture (ICTEDCT) analysis for efficient CBIR. Curvelet multiscale ridgelets were integrated with region-based vector codebook subband clustering to enhance dominant colors extraction and texture analysis.

3. Proposed Methodology

The lack of spatial information is the main problem in the standard BoVW representation [7, 15]. Visual words are represented in a histogram without considering their locations in the 2D image plane. The spatial information carries discriminating details that enhance the performance of CBIR [7, 15]. The block diagram of proposed research methodology is presented in Figure 3.

(1) In the BoVW representation, a raw image I is represented as

$$I = \left(a_{i,j}\right), \tag{1}$$

where $a_{i,j}$ is the pixel at the position (i, j).

(2) Dense SIFT features are extracted from the image and an image I is represented as

$$I = \{d_1, d_2, \ldots, d_m\}, \tag{2}$$

where d_1 to d_m are image descriptors.

(3) Quantization algorithms like k-means clustering is applied to construct a visual vocabulary (codebook) consisting of n visual words, represented as

$$\text{voc} = \{w_1, w_2, \ldots, w_n\}, \tag{3}$$

where w_1 to w_n are visual words.

(4) For the construction of global histogram of visual words, mapping of each visual word is done over the whole image. For the construction of local histogram of visual words, mapping of each visual word is done by extracting the image central area by using (5). The nearest visual words are assigned to the quantized descriptors by using the following equation:

$$w\left(d_k\right) = \underset{w \in \text{voc}}{\operatorname{argmin}} \operatorname{Dist}\left(w, d_k\right), \tag{4}$$

where $w(d_k)$ is representing the visual word assigned to the kth descriptor d_k, while $\text{Dist}(w, d_k)$ is the

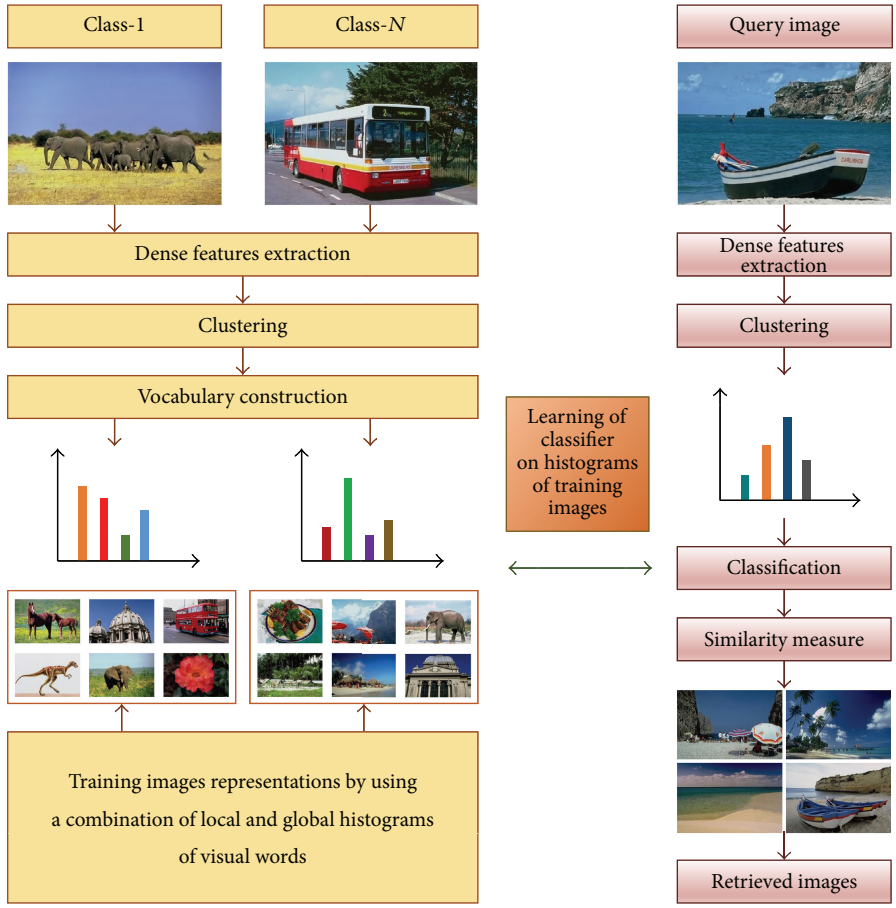


FIGURE 3: Block diagram of proposed research based on a combination of local and global histograms of visual words.

distance between the descriptor d_k and visual word w. Each image is represented as a collection of patches and each patch is represented by visual words.

(5) Two histograms of N visual words are extracted from a single image. The computed local and global histograms are concatenated and this information is added to the inverted index of BoVW representation. The visual words of the local histogram are mapped to a rectangular area that is defined by image width (W) and height (H). Four points are defined to extract the optimized local rectangular area as follows:

$$\begin{aligned} p_1 &= W \times x, \\ p_2 &= (W \times y) + p_1 = (W \times y) + (W \times x) \\ &= W \times (x + y), \\ p_3 &= H \times x, \\ p_4 &= (H \times y) + p_3 = (H \times y) + (H \times x) \\ &= H \times (x + y), \end{aligned} \tag{5}$$

where x is the normalized starting point and y is the ending point of the local rectangular area.

(6) Consider N as the number of visual words of the vocabulary. Let D_i be the set of the descriptors that are mapped to the visual word w_i; then the ith bin of the histogram of visual words b_i is the cardinality of the set D_i:

$$\begin{aligned} b_i &= \text{Card}(D_i), \\ D_i &= \{d_k,\ k \in (1, \ldots, N) \mid w(d_k) = w_i\}. \end{aligned} \tag{6}$$

3.1. Image Classification. SVM is a state-of-the-art supervised learning classification algorithm [5]. The linear SVM separates two classes by using a hyperplane. The dataset with two classes is represented as

$$\{(x_i, y_i)\}_{i=1}^{N}\ y_i = \{+1, -1\}, \tag{7}$$

where x_i and y_i are input datasets and +1 and −1 are the correspondence labels of the classes, respectively. The hyperplanes are generated by finding the values of the coefficients:

$$w^T \cdot x + b = 0, \tag{8}$$

where w is weight vector and b is bias. The maximum margin is determined by $2/\|w\|$ hyperplanes and the two classes

are separable from each other according to the following equations:

$$w^T \cdot x_i + b = 1,$$
$$w^T \cdot x_i + b = -1. \quad (9)$$

This can be expressed equivalently as

$$y_i \left(w^T \cdot x_i + b \right) \geq 1. \quad (10)$$

The kernel method [34] is used in SVM to compute the dot product in the high-dimensional feature space that provides the ability to generate nonlinear decision boundaries. The kernel function permits using the data with no obvious fixed dimensions. The histograms of visual words constructed over the local and global areas of the image are normalized and SVM Hellinger kernel [35] is applied on the normalized histograms by using following equation:

$$K\left(h, h'\right) = \sum_i \sqrt{h\left(i\right) h'\left(i\right)}, \quad (11)$$

where h and h' are the normalized histograms.

The SVM Hellinger kernel is selected because of its low computational cost and instead of computing the kernel values, it explicitly computes the feature map and the classifier remains linear. The one-versus-one rule is applied for k number of classes; $k \cdot (k-1)/2$ classifiers are constructed and each classifier trains the data by using two classes.

4. Experiments and Results

This section is about the details of experiments conducted for the evaluation of the proposed image representation based on a combination of local and global histograms of visual words. The proposed research is evaluated on Corel-A image benchmark, Caltech-256, and Ground Truth image datasets. The images are randomly divided into training and test datasets. The visual vocabulary (codebook) is constructed from the training images and retrieval precision is calculated by using the test dataset. Keeping in view the unsupervised nature of clustering by applying k-means, each experiment is repeated 10 times and average values of precision are reported. In order to evaluate the performance of proposed research, we determined the relevant images retrieved in response to a query image. A computer simulation is used to select images randomly from test dataset and use them as a query image. The response to the query image is evaluated on the basis of relevant images retrieved. Precision determines the number of relevant images retrieved in response to a query image and it shows the specificity of the image retrieval system:

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}}. \quad (12)$$

Recall measures the sensitivity of the image retrieval system. Recall is calculated by the ratio of correct images retrieved to the total number of images of that class in the image benchmark:

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images}}. \quad (13)$$

The details about experimental parameters are given below.

(1) Vocabulary Size. The size of visual vocabulary is a major parameter that affects the performance of content-based image matching [36, 37], increasing the size of visual vocabulary at certain level, and increases the performance and larger size visual vocabulary tends to overfit. Different sizes of visual vocabulary are constructed for the Corel-A (20, 50, 100, 200, and 400), Caltech-256 (20, 50, 100, 200, 400, and 600), and Ground Truth (10, 20, 30, 40, and 50) image benchmarks in order to evaluate the best performance of the proposed image representation.

(2) Dense Pixel Stride. The dense SIFT features are extracted from the training and test images. For a precise content-based image matching, we extracted dense features using three different scales. The dense pixel stride or the step size is used to control the spatial resolution of the dense grid. A smaller pixel stride results in a finer grid, while a larger pixel stride makes the grid coarser. With a pixel stride of 5, 15, and 25, we considered every 5th, 15th, and 25th pixel, respectively, as a features to calculate the SIFT descriptor from local and global regions of each image.

(3) Feature Percentage for the Vocabulary Construction. The feature percentage to construct a visual vocabulary from a training dataset is one of the parameters that affect the performance of image retrieval. According to [36], increasing the percentage of features for visual vocabulary construction increases the performance of content-based image matching and vice versa. In experiments, different percentages (10%, 25%, 50%, 75%, and 100%) of dense features per image are used to construct visual vocabulary.

(4) Value of x and y. x and y are used for the extraction of local rectangular area. After a number of experiments, the optimized normalized starting value of $x = 0.22$ and ending value of $y = 0.60$ are selected for training and test images.

4.1. Retrieval Performance on Corel-A Image Benchmark. The Corel-A image benchmark (http://wang.ist.psu.edu/docs/related/) is selected for the evaluation of the proposed image representation and results are compared with the standard BoVW representation [6] and the state-of-the-art CBIR research [4, 15, 17–19]. The Corel-A image benchmark contains 1000 images that are divided into ten semantic categories, namely, Africa, Buildings, Beach, Dinosaurs, Buses, Elephants, Horses, Flowers, Mountains, and Food. Each semantic category consists of 100 images with a resolution of 256×384 pixels or 384×256 pixels. Figure 4 is representing the images from all the semantic classes of Corel-A image benchmark.

Figure 4: Samples of images from each category of Corel-A image benchmark.

Table 1: MAP of the proposed image representation on a vocabulary size of 200 visual words and pixel stride of 5.

Vocabulary size & features% used	20	50	100	200	400
10%	77.40	80.27	81.98	84.01	82.57
25%	77.74	80.92	82.02	84.12	83.16
50%	77.85	81.06	82.53	84.25	83.01
75%	78.08	81.16	83.07	84.33	83.24
100%	78.36	81.32	83.27	84.38	83.29
MAP	77.88	80.95	82.57	84.21	83.05
Standard deviation	0.3609	0.4050	0.5899	**0.1522**	0.2905
Confidence interval	77.43–78.33	80.44–81.44	81.84–83.30	**84.02–84.40**	82.69–83.41
Standard error	0.1614	0.1811	0.2638	**0.0680**	0.1299

In order to maintain a balance, 50% images are used for the training and the remaining 50% images are used for the testing. Different sizes of visual vocabulary (20, 50, 100, 200, and 400) are constructed from a set of training images. Mean Average Precision (MAP) is calculated by a random selection of 500 images from the test dataset. MAP for top-20 image retrievals as a function of vocabulary size and percentage of dense features per image used in vocabulary construction are presented in Table 1.

The additional statistical investigation in Table 1 reinforces our described results. We have calculated standard error as description and estimated 95% confidence interval as inferential results. From these results, we can easily conclude that both the lower and the upper bounds at 5% level of significance exceed 84% of MAP and the smaller value of standard error further enhances the fact that the visual vocabulary of size 200 visual words shows precise and consistent result as compared to other sizes of visual vocabulary on different features percentages (10%, 25%, 50%, 75%, and 100%).

The MAP obtained by using proposed image representation based on a combination of local and global histograms of visual words (by using a vocabulary size of 200 visual words) on the pixel strides of 5, 15, and 25 is 84.21%, 82.12%, and 78.29%, respectively. The MAP of the proposed image representation is compared with the standard BoVW representation (without considering the spatial information of local histogram). The MAP comparison of the proposed research and standard BoVW as a function of vocabulary size on the pixel strides of 5, 15, and 25 is graphically represented in Figure 5. The proposed image representation outperforms the standard BoVW representation on all vocabulary sizes (pixel strides of 5, 15, and 25). According to the experimental results, the MAP for all vocabulary sizes on a pixel stride of 5 is greater than pixel strides of 15 and 25.

Table 2: Comparison of MAP for top-20 image retrievals on Corel-A image benchmark.

Class & method	Proposed research	Youssef [17]	Irtaza et al. [4]	Poursistani et al. [18]	Tian et al. [19]	Wang et al. [15]
Africa	**73.03**	63.5	65	70.24	**74.6**	64
Beach	**74.58**	**64.2**	60	44.44	37.8	54
Buildings	**80.24**	69.8	62	**70.8**	53.9	53
Buses	**95.84**	91.5	85	76.3	**96.7**	94
Dinosaurs	97.95	**99.2**	93	**100**	99	98
Elephants	**87.64**	78.1	65	63.8	66	78
Flowers	85.13	**94.8**	**94**	92.4	92	71
Horses	86.29	**95.2**	77	**94.7**	87	93
Mountains	**82.43**	73.8	73	56.2	58.5	42
Food	78.96	**80.6**	**81**	74.5	62.2	50

Table 3: Comparison of recall for top-20 image retrievals on Corel-A image benchmark.

Class & method	Proposed research	Youssef [17]	Irtaza et al. [4]	Poursistani et al. [18]	Tian et al. [19]	Wang et al. [15]
Africa	**14.61**	12.70	13.00	**14.05**	14.92	12.80
Beach	**14.92**	**12.84**	12.00	8.89	7.56	10.80
Buildings	**16.05**	13.96	12.40	**14.16**	10.78	10.60
Buses	**19.17**	18.30	17.00	15.26	**19.34**	18.80
Dinosaurs	19.59	**19.84**	18.60	**20.00**	19.80	19.60
Elephants	**17.53**	**15.62**	13.00	12.76	13.20	15.60
Flowers	17.03	**18.96**	**18.80**	18.48	18.40	14.20
Horses	17.26	**19.04**	15.40	**18.94**	17.40	18.60
Mountains	**16.49**	**14.76**	14.60	11.24	11.70	8.40
Food	15.79	**16.12**	**16.20**	14.90	12.44	10.00
Mean	**16.84**	**16.21**	15.10	14.57	14.55	13.94

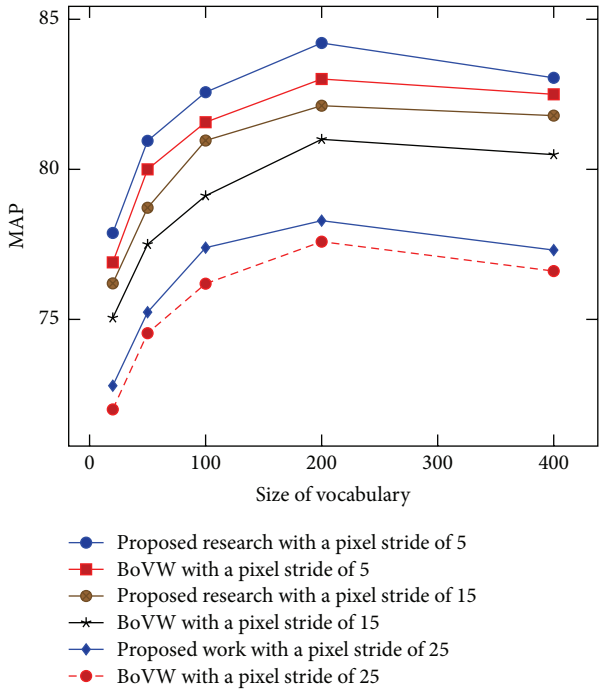


Figure 5: MAP as a function of vocabulary size.

In order to present a sustainable performance of the proposed research, the image retrieval precision and recall for the top-20 image retrievals are compared with the state-of-the-art research of CBIR [4, 15, 17–19]. Tables 2 and 3 are presenting the classwise comparison of average precision and recall of the proposed research (with a vocabulary size of 200 words on a pixel stride of 5 and by using 100% features per image) with the existing state-of-the-art techniques of CBIR. The MAP comparison is shown in Figure 6, while precision-recall curve is shown in Figure 7.

The experimental results conducted on Corel-A image benchmark prove the robustness of the proposed image representation. The MAP of the proposed research is higher than the existing state-of-the-art research [4, 15, 17–19]. The MAP obtained from the proposed image representation is 84.21% (with a vocabulary size of 200 visual words on a pixel stride of 5 and by using 100% features per image).

The image retrieval results obtained by using proposed image representation for the semantic class "Buses" and "Beach" are shown in Figures 8 and 9, respectively, in response to the query images that shows reduction of semantic gap in terms of classifier decision value (score). The classifier decision label determines the class of the image, while classifier decision value (score) is used to find the similar

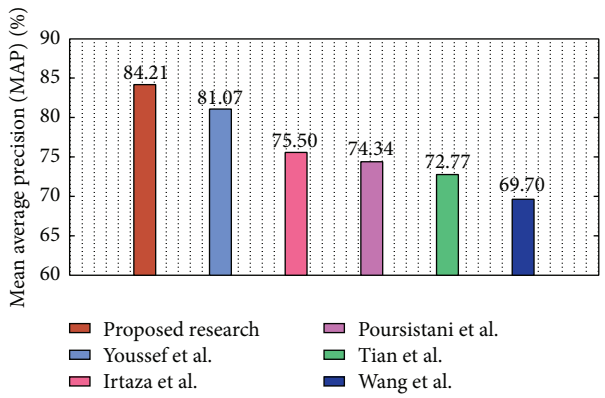


Figure 6: Comparison of MAP with the state-of-the-art methods on Corel-A image benchmark.

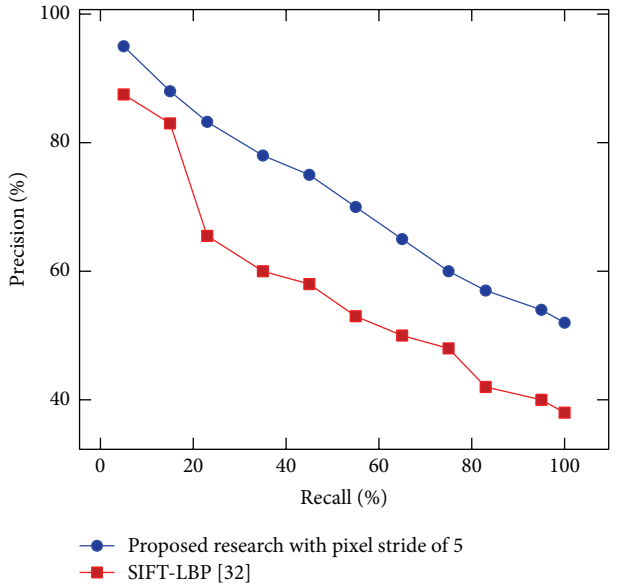


Figure 7: Precision-recall curve for Corel-A image benchmark.

images. The real values shown at the top of each image are the classifier decision value (score) of the respective image, computed by applying the Euclidean distance between score of the query image and scores of the retrieved images. Top-20 retrieved images, whose score is close to the score of the query image are more similar to the query image and vice versa (due to limited space, the retrieval results in response to the query images are shown for two semantic classes only).

4.2. Performance on Caltech-256 Image Benchmark. Caltech-256 image benchmark (http://www.vision.caltech.edu/) consists of 256 semantic classes and each semantic class contains 80 to 827 images. For the simplicity, we randomly selected 10 classes from the Caltech-256 image benchmark. The selected classes are Motorbike, Faceeasy, Fireworks, Bonsai, Butterfly, Leopard, Airplanes, Ketch, and Hibiscus as shown in Figure 10.

Different sizes of visual vocabulary (20, 50, 100, 200, 400, and 600) are constructed from a set of training images. 50%

Figure 8: Image retrieval result shows reduction of semantic gap for the semantic class "Buses."

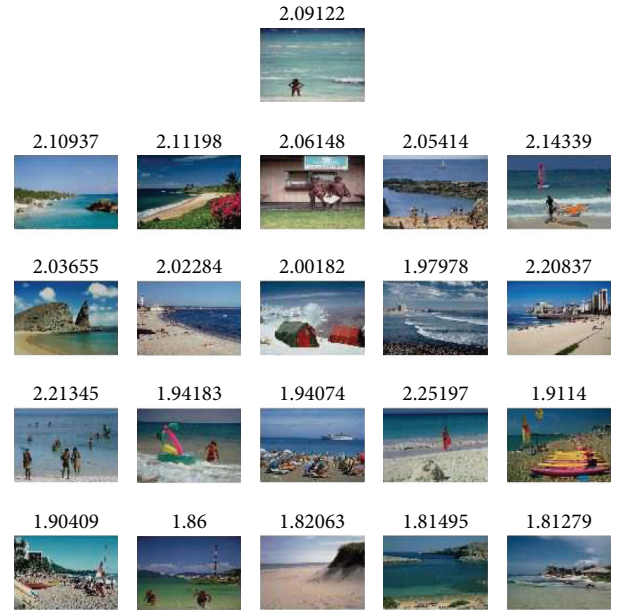


Figure 9: Image retrieval result shows reduction of semantic gap for the semantic class "Beach."

images are used for the training and the remaining 50% images are used for the testing. The MAP of the proposed research is compared with standard BoVW representation (without considering the spatial information) on different pixel strides (5, 15, and 25), vocabulary sizes (20, 50, 100, 200, 400, and 600), and feature percentages (10%, 25%, 50%, 75%, and 100%) for vocabulary construction. The MAP obtained from the proposed image representation and the standard BoVW representation by using different parameters is graphically presented in Figure 11.

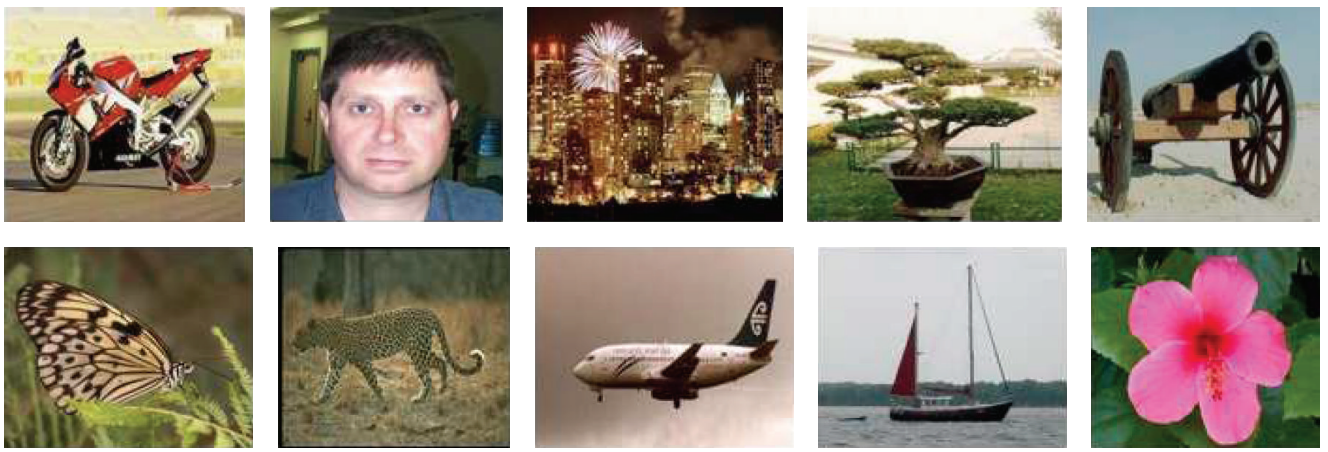


Figure 10: Samples of images from 10 semantic classes of Caltech-256 image benchmark.

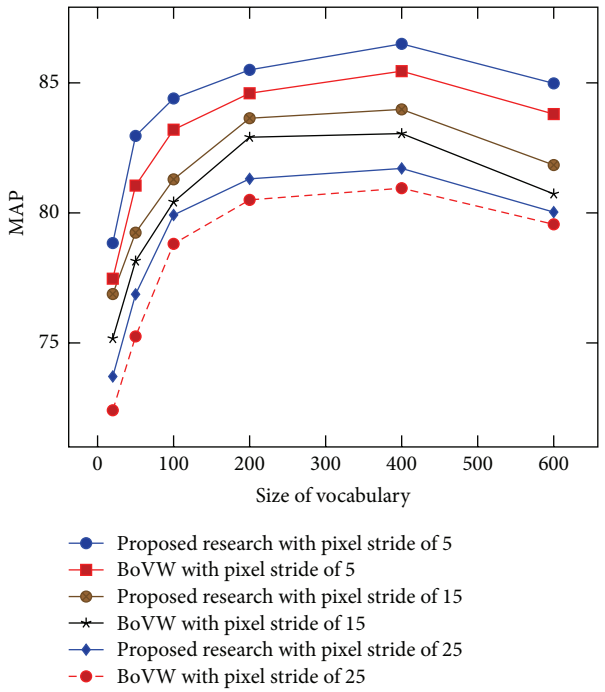


Figure 11: MAP as a function of vocabulary size for the Caltech-256 image benchmark.

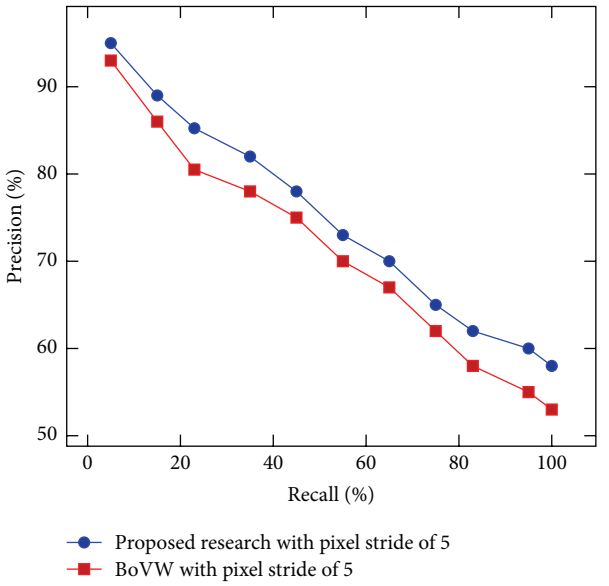


Figure 12: Precision-recall curve for Caltech-256 image benchmark.

The experimental results conducted on 10 semantic classes of Caltech-256 image benchmark prove the robustness of the proposed image representation. The MAP obtained by using the proposed image representation (with a vocabulary size of 400 visual words and by using 100% features per image) on the pixel strides of 5, 15, and 25 is 86.50%, 83.98%, and 81.71%, respectively, while MAP obtained by using the same experimental parameters for standard BoVW representation is 85.45%, 83.05%, and 80.95%, respectively. According to the experimental results, the MAP for all vocabulary sizes decreases by increasing the pixel stride as shown in Figure 11, while precision-recall curve is shown in Figure 12.

4.3. Performance on Ground Truth Image Benchmark. Ground Truth image benchmark (http://imagedatabase.cs.washington.edu/groundtruth/) is a publicly available image benchmark and is used for the evaluation of CBIR research [20, 21, 33]. There are a total of 1109 images that are divided into 22 semantic classes. We manually selected all the images from 5 different semantic classes of Ground Truth image benchmark. The selected classes for the evaluation of the proposed image representation are Abro green, Cherries, Football, Green Lake, and Swiss Mountains as shown in Figure 13. The 5 classes are selected in order to compare the performance of the proposed image representation with existing state-of-the-art research [20, 21, 33] because these researchers also reported their results on the same number of classes of Ground Truth image benchmark.

50% images are used for the training and the remaining 50% images are used for the testing. Different sizes (10, 20,

Figure 13: Samples of images from 5 semantic classes of Ground Truth image benchmark.

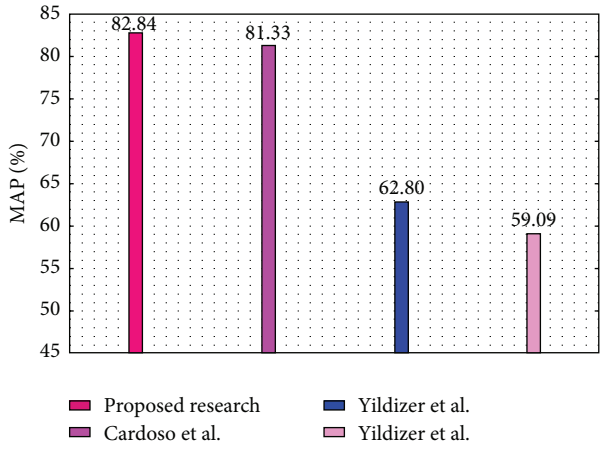


Figure 14: Comparison of MAP with the state-of-the-art methods on Ground Truth image benchmark.

Table 4: Classwise comparison of average precision.

Class & method	Proposed method	[20]	[21]
Abro green	75.25	**80**	66.67
Cherries	71.94	**80**	50
Football	90.18	**100**	75
Green Lake	**84.23**	80	50
Swiss Mountains	**92.63**	60.67	50

30, 40, and 50) of visual vocabulary are constructed and MAP on the vocabulary of these different sizes is calculated. According to the experimental results, the MAP of 82.84% is obtained by using the proposed image representation (with a vocabulary size of 40 visual words on pixel stride of 5 and by using 100% features per image). The classwise average precision obtained from the proposed image representation is presented in Table 4, while MAP is graphically presented in Figure 14.

Table 5: Computational cost (in seconds) of the proposed algorithm.

Number of images retrieved	Proposed research
Top-5	0.2872
Top-10	0.3972
Top-15	0.6470
Top-20	0.7837
Top-25	0.9184
Top-30	1.1103

Experimental results and comparisons conducted on the Ground Truth image benchmark prove the robustness of the proposed research based on a combination of local and global histograms of visual words. The MAP obtained from the proposed image representation is higher than the existing state-of-the-art research [20, 21, 33].

4.4. Complexity Performance. The computational cost of the proposed algorithm is calculated using Intel(R) Core i3 (fourth generation) 1.7 Ghz CPU with 4 GB RAM (DDR3), 3 MB L3 cache, and Windows 7 operating system. The proposed algorithm is implemented in MATLAB and visual vocabulary is constructed offline using a training dataset and tested at run time using a test dataset. The average CPU time (in seconds) required from features extraction to image retrieval is presented in Table 5.

5. Conclusion and Future Directions

In this paper, we proposed a novel image representation based on a combination of local and global histograms of visual words. The local histogram of visual words adds the spatial information of the central area to the inverted index of BoVW representation. The combination of local and global histograms of visual words is a possible solution to capture the semantic information from the image. The performance

of the proposed image representation is evaluated on three challenging image datasets. The proposed image representation outperforms the state-of-the-art research including the standard BoVW representation. For the future work, we plan to replace the BoVW model with either Fisher kernel framework or the Vector of Locally Aggregated Descriptors (VLAD) model to evaluate a large-scale image retrieval.

Competing Interests

The authors declare that they have no competing interests.

References

[1] A.-M. Tousch, S. Herbin, and J.-Y. Audibert, "Semantic hierarchies for image annotation: a survey," *Pattern Recognition*, vol. 45, no. 1, pp. 333–345, 2012.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, article 5, 2008.

[3] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[4] A. Irtaza, M. A. Jaffar, E. Aleisa, and T.-S. Choi, "Embedding neural networks for semantic association in content based image retrieval," *Multimedia Tools and Applications*, vol. 72, no. 2, pp. 1911–1931, 2014.

[5] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, no. 1, pp. 346–362, 2012.

[6] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, pp. 1470–1477, IEEE, 2003.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, IEEE, Minneapolis, Minn, USA, June 2007.

[8] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: automatic query expansion with a generative feature model for object retrieval," in *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, IEEE, Rio de Janeiro, Brazil, October 2007.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: improving particular object retrieval in large scale image databases," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, IEEE, Anchorage, Alaska, USA, June 2008.

[10] H. Anwar, S. Zambanini, M. Kampel, and K. Vondrovec, "Ancient coin classification using reverse motif recognition: image-based classification of roman republican coins," *IEEE Signal Processing Magazine*, vol. 32, no. 4, pp. 64–74, 2015.

[11] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao, "Generating descriptive visual words and visual phrases for large-scale image applications," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2664–2677, 2011.

[12] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 59–77, 2012.

[13] Z. Mehmood, S. M. Anwar, M. Altaf, and N. Ali, "A novel image retrieval based on rectangular spatial histograms of visual words," *Kuwait Journal of Science*, In press.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 2169–2178, New York, NY, USA, June 2006.

[15] C. Wang, B. Zhang, Z. Qin, and J. Xiong, "Spatial weighting for bag-of-features based image retrieval," in *Integrated Uncertainty in Knowledge Modelling and Decision Making*, Z. Qin and V.-N. Huynh, Eds., vol. 8032 of *Lecture Notes in Computer Science*, pp. 91–100, Springer, Berlin, Germany, 2013.

[16] N. Ali, K. B. Bajwa, R. Sablatnig, and Z. Mehmood, "Image retrieval by addition of spatial information based on histograms of triangular regions," *Computers & Electrical Engineering*, 2016.

[17] S. M. Youssef, "ICTEDCT-CBIR: integrating curvelet transform with enhanced dominant colors extraction and texture analysis for efficient content-based image retrieval," *Computers and Electrical Engineering*, vol. 38, no. 5, pp. 1358–1376, 2012.

[18] P. Poursistani, H. Nezamabadi-pour, R. Askari Moghadam, and M. Saeed, "Image indexing and retrieval in JPEG compressed domain based on vector quantization," *Mathematical and Computer Modelling*, vol. 57, no. 5-6, pp. 1005–1017, 2013.

[19] X. Tian, L. Jiao, X. Liu, and X. Zhang, "Feature integration of EODH and Color-SIFT: application to image retrieval based on codebook," *Signal Processing: Image Communication*, vol. 29, no. 4, pp. 530–545, 2014.

[20] D. N. M. Cardoso, J. Muller, F. Alexandre, L. A. P. Neves, P. M. G. Trevisani, and G. A. Giraldi, "Iterative technique for content-based image retrieval using multiple SVM ensembles," in *A Treatise on Electricity and Magnetism*, J. Clerk Maxwell, Ed., vol. 2, pp. 68–73, Cambridge University Press, 1873.

[21] E. Yildizer, A. M. Balci, M. Hassan, and R. Alhajj, "Efficient content-based image retrieval using multiple support vector machines ensemble," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2385–2396, 2012.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 886–893, IEEE, San Diego, Calif, USA, June 2005.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[24] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: speeded up robust features," in *Computer Vision—ECCV 2006*, pp. 404–417, Springer, 2006.

[26] D. J. Mankowitz and S. Ramamoorthy, "BRISK-based visual feature extraction for resource constrained robots," in *RoboCup 2013: Robot World Cup XVII*, S. Behnke, M. Veloso, A. Visser, and R. Xiong, Eds., vol. 8371 of *Lecture Notes in Computer Science*, pp. 195–206, Springer, Berlin, Germany, 2014.

[27] S. Krig, "Interest point detector and feature descriptor survey," in *Computer Vision Metrics*, pp. 217–282, Springer, Berlin, Germany, 2014.

[28] R. Ashraf, K. Bashir, A. Irtaza, and M. T. Mahmood, "Content based image retrieval using embedded neural networks with bandletized regions," *Entropy*, vol. 17, no. 6, pp. 3552–3580, 2015.

[29] S. Zeng, R. Huang, H. Wang, and Z. Kang, "Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models," *Neurocomputing*, vol. 171, pp. 673–684, 2016.

[30] E. Walia and A. Pal, "Fusion framework for effective color image retrieval," *Journal of Visual Communication and Image Representation*, vol. 25, no. 6, pp. 1335–1348, 2014.

[31] A. Irtaza and M. A. Jaffar, "Categorical image retrieval through genetically optimized support vector machines (GOSVM) and hybrid texture features," *Signal, Image and Video Processing*, vol. 9, no. 7, pp. 1503–1519, 2014.

[32] J. Yu, Z. Qin, T. Wan, and X. Zhang, "Feature integration analysis of bag-of-features model for image retrieval," *Neurocomputing*, vol. 120, pp. 355–364, 2013.

[33] E. Yildizer, A. M. Balci, T. N. Jarada, and R. Alhajj, "Integrating wavelets with clustering and indexing for effective content-based image retrieval," *Knowledge-Based Systems*, vol. 31, pp. 55–66, 2012.

[34] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, NY, USA, 2004.

[35] A. Vedaldi and A. Zisserman, "Sparse kernel approximations for efficient classification and detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2320–2327, June 2012.

[36] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classifiation," in *Computer Vision—ECCV 2006*, pp. 490–503, Springer, 2006.

[37] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of the Workshop on Statistical Learning in Computer Vision (ECCV '04)*, vol. 1, pp. 1–2, Prague, Czech Republic, 2004.