

RESEARCH

Open Access



Scene search based on the adapted triangular regions and soft clustering to improve the effectiveness of the visual-bag-of-words model

Zahid Mehmood¹, Naila Gul², Muhammad Altaf², Toqeer Mahmood³, Tanzila Saba⁴, Amjad Rehman⁵ and Muhammad Tariq Mahmood^{6*}

Abstract

The storage size of the image and video repositories are growing day by day due to the extensive use of digital image acquisition devices. The position of an object within an image is obtained by analyzing the content-based properties like shape, texture, and color, while compositional properties present the image layout and include the photographic rule of composition. The high-quality images are captured on the basis of the rule of thirds that divide each image into nine square areas. According to this rule, salient objects of an image are placed on the intersection points or along the imagery lines of the grid to capture the position of the salient objects. To improve image retrieval performance, visual-bag-of-words (VBoW) framework-based image representation is widely used nowadays. According to this framework, the spatial relationship between salient objects of an image is lost due to the formation of a global histogram of the image. This article presents a novel adapted triangular area-based technique, which computes local intensity order pattern (LIOP) features, weighted soft codebooks, and triangular histograms from the four triangular areas of each image. The proposed technique adds the spatial contents from four adapted triangular areas of each image to the inverted index of the VBoW framework, solve overfitting problem of the larger sizes of the codebook, and overwhelmed the problem of the semantic gap. The experimental results and statistical analysis performed on five image collections show an encouraging robustness of the proposed technique that is compared with the recent CBIR techniques.

Keywords: Content-based image retrieval, Support vector machine, Deep learning, Adapted triangular features

1 Introduction

The extensive use of the social networking sites and technological improvements of the image acquisition devices, as well as the size of image repositories, are increased exponentially. It has gained the researchers' interest to find a better approach to search images from huge image collections using an effective and efficient mechanism [1–3]. The commonly used image annotation methods are based on the mapping of image descriptors with few keywords. These methods cannot describe the diversity of contents within images due to

the lack of discriminative capability. Content-based image retrieval (CBIR) gained significance in research over the years [4, 5]. The focus of any CBIR technique is to compute low-level visual features from the images and extend the association between them in terms of co-occurrence of similar visual contents. The visual contents of the image are represented in terms of the low-level visual features such as texture and shape [6]. Texture-based features are capable to find the spatial variations between intensity values and surface attributes of an object within an image. However, segmentation of texture is a challenging task in order to fulfill human perception [7]. Color-based features are invariant to scale and rotation with high computational cost. Shape-based features are not able to provide a

* Correspondence: tariq@koreatech.ac.kr

⁶School of Computer Science and Engineering, Korea University of Technology and Education, Cheonan 330-708, Republic of Korea

Full list of author information is available at the end of the article

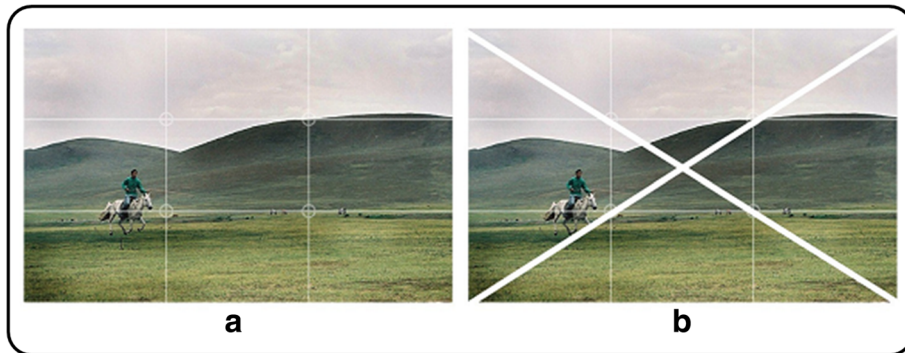


Fig. 1 Image **a** <http://mikemellizaphotography.blogspot.com/> is showing the photographic rule of thirds, while image **b** is showing its division into four adapted triangular areas

mathematical foundation for image deformation and generally consistent with the intuitive feeling [4]. Therefore, image representation by using only low-level features cannot describe the semantic relationship between images efficiently. The resemblance in the pictorial look of the images belonging to various categories consequence in the closeness of low-level visual features and it decreases the performance of CBIR [8, 9].

In VBoW framework [10], features are computed from each image, after that clustering is applied to formulate codebook that consists of visual words. These visual words are used to build a global histogram for each image, which result in loss of the spatial information and after that classification is performed, and the similarity is calculated between inquiry image and archive images in order to retrieve images. The robustness of the VBoW-based image illustration suffers due to ignorance of spatial context among local features [11–14]. Different techniques such as geometric coding [15] and co-occurrence of visual words [16] are introduced to add the spatial context of visual words to the VBoW framework. These techniques required high computational complexity on larger sizes of the codebook or dictionary [11]. The spatial information is available in the sub-areas of an image. The technique of Lazebnik et al.

[17] formulates spatial histogram from each square area of the grid by splitting an image into different square areas. Keeping in view the effective performance of [17] to incorporate spatial information, the proposed technique of this article is to split an image into four adapted triangular regions instead of square regions [17] and extracting adapted local features, formulating weighted soft codebooks, and computing histograms over the each triangular area of an image.

The content-based properties of an image describe the presence of salient objects in the image, while the compositional properties describe the image layout and it includes the photographic rule of composition [18]. The principle of the rule of thirds is to divide each image into nine square areas and place the salient objects at the intersection points of the grid to acquire content-based and compositional properties which improve the CBIR performance [18]. Figure 1a represents an image that is captured according to the photographic rule of thirds, and salient objects are placed at the intersection points of the square areas. Figure 1b also represents an image that is separated into four triangular areas to extract features over four dense scales (from four distributed triangular areas) for the addition of spatial information to the VBoW



Fig. 2 a Image with very close visual appearance but with a different semantic meaning. **b** The image on the left is the inquiry image, while the image of building and lady are the retrieved images in reply to the inquiry image due to close visual appearance

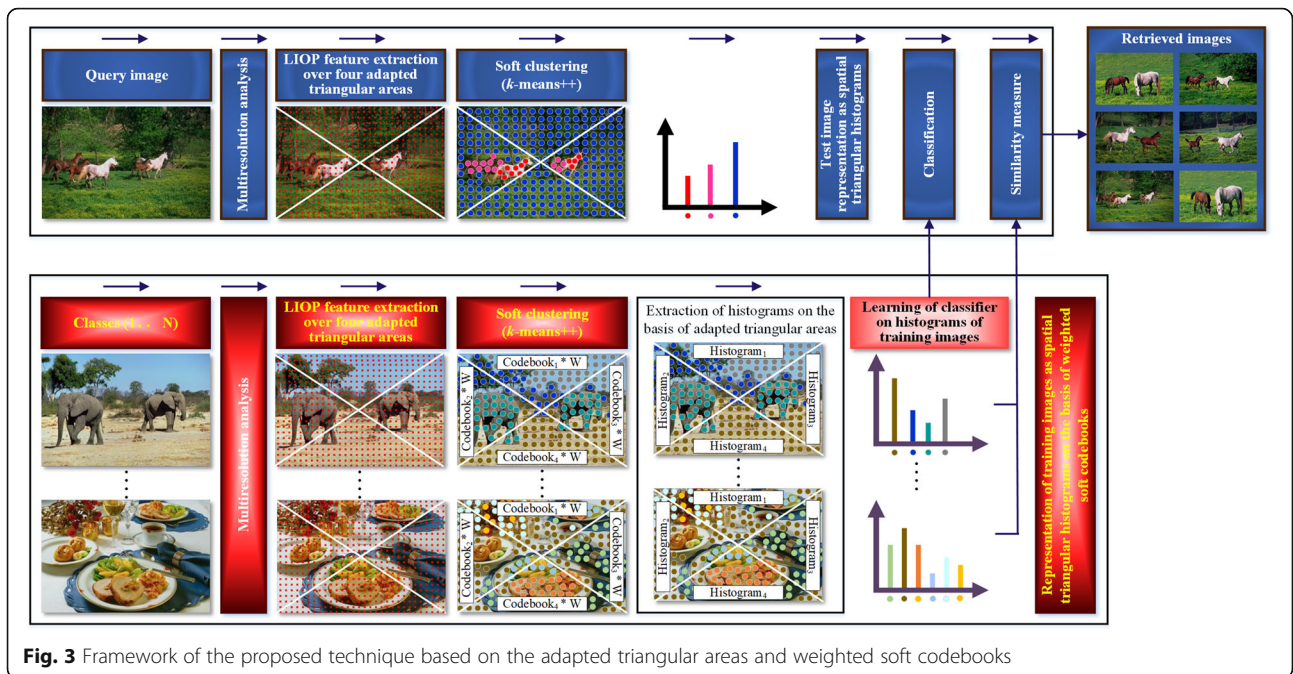


Fig. 3 Framework of the proposed technique based on the adapted triangular areas and weighted soft codebooks

framework and to reduce the semantic gap. In an image of a scene, there are different regions or objects that are located in different sub-regions. The water or grass is likely to be located at the bottom and clouds or sky is positioned at the top, while salient objects are positioned at left or right side. The detachment of an image into four adapted triangular areas represents this triangular association.

Figure 2a presents two images with a very close visual similarity but different semantic meanings. The feature map of the dog and the lady are visually analogous, but their semantic implications are entirely diverse. In Fig. 2b, the image on the left is the inquiry image, the

output of image retrieval can give emphasis on the contemporary building so the preminent contest could be a modern building like the one in the middle image, while in some cases the emphasis may be on the prettiness, so the rightmost image is the preminent contest. The best CBIR system is the one which retrieves accurate images according to the user requirements. Different regions like the ground, water, sky, clouds, and grass are located within these different triangular areas, while the salient objects like peoples and horses are positioned at left or right side. Keeping these facts into view, we computed dense LIOP features from four adapted triangular areas; feature space is quantized to formulate weighted soft

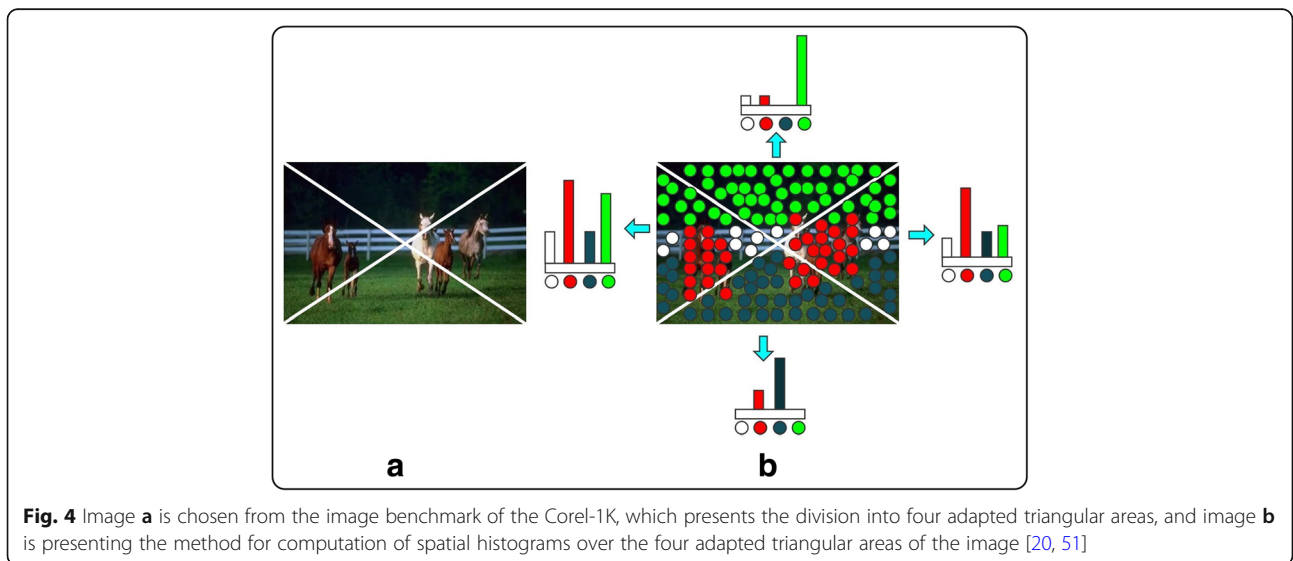


Fig. 4 Image a is chosen from the image benchmark of the Corel-1K, which presents the division into four adapted triangular areas, and image b is presenting the method for computation of spatial histograms over the four adapted triangular areas of the image [20, 51]

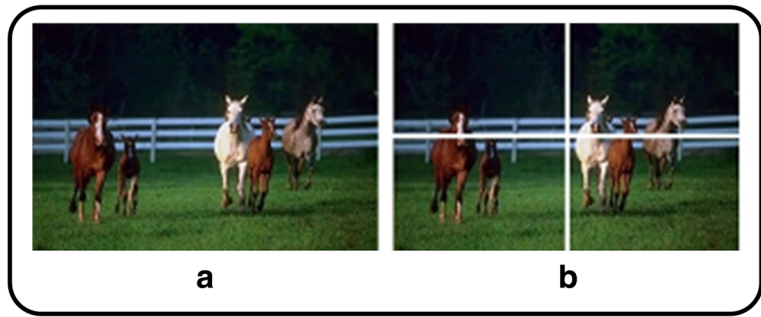


Fig. 5 Image **a** is chosen from the semantic class “Horses” of the Corel-1K image benchmark, and image **b** is presenting the division of the image into four rectangular areas for computation of histogram from each rectangular area [40]

codebook from each triangular area and computed spatial histogram from each adapted triangular area of the image. The proposed technique provides an option to extract the spatial properties of an image from each of the triangular areas and solve the problem of overfitting on larger sizes of the codebook by formulating weighted soft codebooks. Following are the key contributions of this article:

1. The accumulation of the spatial information to the VBoW framework.
2. An adapted triangular regions based technique for the feature extraction over four adapted triangular regions of an image, resolve the problem of overfitting on larger sizes of the codebook by formulating weighted soft codebook over each triangular region.
3. Lessening of semantic gap between high-level semantic perceptions and low-level features of the image.

2 Related work

The query by image content (QBIC) system [19] is the first CBIR system introduced by IBM. Many systems after QBIC are developed by IBM. Common domain for all of these was to enhance the image searching techniques and similarity matching in order to increase the performance of CBIR. In existing literature [20, 21], several methods have been implemented to overcome the

limitations in CBIR, i.e., to reduce the semantic gap between low-level features and high-level semantic concepts. To address these issues, focus of research is local features such as color, boundary contour, texture, and spatial layout, and different discriminative feature extraction techniques were introduced to enhance the performance of CBIR.

An optimized technique for image retrieval is introduced by Zhong and Defée [22], which relies on the quantized histograms of discrete cosine transform (DCT) blocks and uses different global parameters such as scalar quantization, histograms size, difference vectors, and integrated AC-pattern and DC-DirecVec histograms. Histograms are optimized through the factor of quantization together with the count of DCT blocks that are normalized under luminance in order to improve CBIR performance. Yuan et al. [23] propose a local descriptor that integrates SIFT and LBP features to obtain a high-dimensional feature vector for each image. Two fusion models, i.e., patch-level and image-level, are employed for feature fusion. For compact representation of high-dimensional feature vector, clustering technique based on *k*-means is used to formulate a codebook. According to the semantic category of the query image, images are retrieved and ranked based on the similarity measure. Yu et al. [24] present feature fusion technique which uses a histogram of oriented gradients (HOG), SIFT, and local binary pattern (LBP) features in order to achieve effective results for CBIR. A high-dimensional



Fig. 6 Sample images associated with image benchmark of the Corel-1K [40]

Table 1 MAP performance on the image benchmark of the Corel-1K with a step size = 10

Features % per image for dictionary formation	MAP performance analysis on different dictionary sizes				
	20	60	100	200	400
10%	80.98	83.42	85.14	87.03	86.81
25%	81.08	83.58	85.21	87.14	86.68
50%	81.41	83.34	85.33	87.19	86.54
75%	81.78	83.78	85.22	87.22	86.64
100%	81.92	83.86	85.46	87.21	86.51
MAP	81.43	83.59	85.27	87.15	86.63
Std. dev.	0.41	0.22	0.12	0.16	0.11
Conf. interval	80.91–81.94	83.31–83.87	85.11–85.42	87.01–87.42	86.48–86.78
Std. error	0.18	0.10	0.05	0.07	0.05

feature descriptor is formed by the fusion of separately computed visual features using SIFT, LBP, and HOG features. After that, these fused features are encoded into visual words by applying a k-means clustering technique to form a dictionary and each image is characterized as a distribution of these visual words. Raja and Bhanu [25] propose an improve CBIR technique on the basis of image local features like color, texture, shape, wavelet-based histogram, and incorporate relevance feedback to achieve better accuracy for CBIR. A visual similarity matching technique known as adaptive region matching (ARM) is proposed by Yang and Cai [26] which uses region-based image retrieval (RBIR). A semantic meaningful region (SMR) and region important index (RII) are built to reduce the adverse consequence of interference regions and loss of spatial information. Images are compared conforming to whether the given image has an SMR and it performs SMR-to-image matching to improve the performance of the CBIR.

Wang et al. [27] propose a spatial weighting bag-of-features (SWBoF) model of visual words by applying texture measure. The spatial information is extracted from diverse areas of the image. The dissimilarity between groups of pixels is selected to compute the useful information. The spatial information is computed by applying local entropy, adjacent blocks distance and local variance. According to the experimental results of [27], SWBOF model performs better than traditional BoF approach. According to Liu et al. [11], the spatial information among local features carries significant information for content verification. A rotation and scale-invariant edge orientation difference histogram (EODH) descriptor are proposed by Tian et al. [28]. The steerable filter and vector sum are applied to obtain the main orientation of pixels. The color-SIFT and EODH descriptors are integrated to improve the effectiveness of the feature space and to reduce the semantic gap. The dictionary is constructed by applying a weighted average of

Color-SIFT and EODH. According to the experimental results [28], weighted average distribution enhances the performance of the image retrieval.

Rashno et al. [29] propose an effective technique for CBIR which relies on the discrete wavelet transform (DWT) and color features. In this technique, visual contents of each image are represented through a feature vector which comprises of texture feature by applying wavelet transform and color feature obtained by converting each image from RGB and HSV space. In wavelet transform, each image is decomposed into four

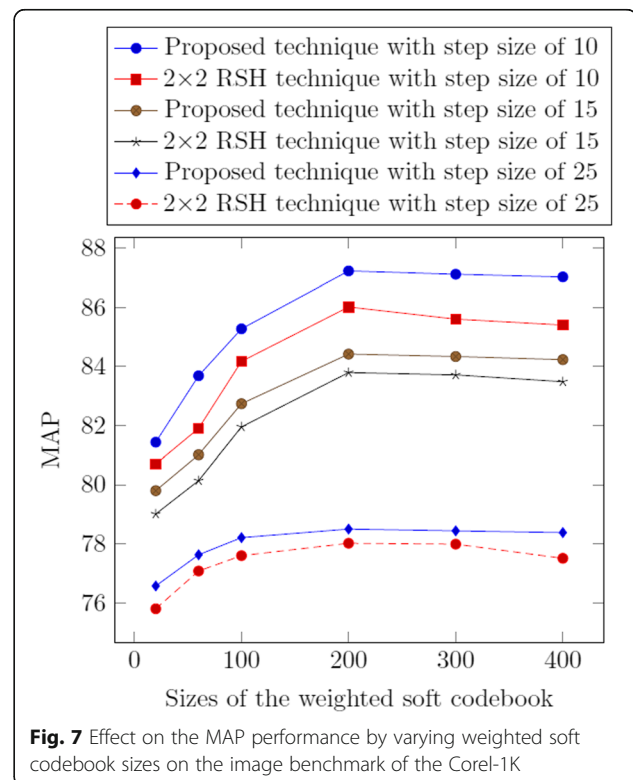
**Fig. 7** Effect on the MAP performance by varying weighted soft codebook sizes on the image benchmark of the Corel-1K

Table 2 MAP performance comparison of the proposed technique with recent CBIR techniques on the image benchmark of the Corel-1 K

Class name	Proposed technique with step size = 10	[43]	[42]	[41]	[28]	[44]
Africa	70.22	73	65	63.50	74.60	58.73
Beach	78.95	72	70	64.20	37.80	48.94
Buildings	80.89	79	75	69.80	53.90	53.74
Busses	97.87	100	95	91.50	96.70	95.81
Dinosaurs	99.29	97	100	99.20	99	98.36
Elephants	94.65	75	80	78.10	65.90	64.17
Flowers	89.11	86	95	94.80	91.20	85.64
Horses	97.04	82	90	95.20	86.90	80.31
Mountains	81.84	69	75	73.80	58.50	54.27
Food	82.25	90	75	80.60	62.20	63.14
MAP	87.22	82.30	82.00	81.07	72.67	70.31

sub-bands and then low-frequency sub-band is used as a texture-feature. For color-features, dominant color descriptor (DCD) is used for quantization of the image to achieve color statistics and histogram features. Ant colony optimization technique is used for selecting relevant and unique features from the entire feature set consisting of both color and texture features. The images are retrieved by applying Euclidean distance to find a resemblance between inquiry image and database images. Rahimi and Moghaddam [30] introduce a CBIR technique which uses intraclass and interclass features to improve the performance of the CBIR. The distribution of the color tone is used as an intraclass feature, whereas singular value decomposition (SVD) and complex wavelet transform are used as inter-class features. Self-organizing map (SOM) is produced using these features by applying artificial neural network (ANN) to increase the proficiency of the CBIR. Yan et al. [31] propose a novel technique using the deep convolutional neural network to analyze image contents to learn high-quality binary codes, known as a one-stage supervised deep hashing framework (SDHP). The proposed technique assign similar binary codes to the similar images and vice versa. The learned codes in this technique are evenly distributed, and during the conversion process of the Euclidean space to Hamming space, quantization loss is reduced. The discriminative power of the learned binary codes is further improved by extending SDHP to SDHP+, which significantly improve the search accuracy as compared with state-of-the-art hashing algorithms. Yan et al. [32] present another novel framework for recognition of Uyghur language text in case of intricate background images. For detecting text regions, the maximally stable extremal regions (MSERs) technique is introduced but one of its shortcomings is that in case of blur and low contrast images, it does not perform well. Due to this reason, another technique is

introduced, known as channel-enhanced MSERs. This technique outperforms the traditional MESRs technique, but one limitation occurs in this case that is noise and overlapping regions. The HOG and SVM are employed for extracting non-text overlapping regions and noise. One of the most important outcomes of this technique is the usefulness for detecting Uyghur language as well as other languages can be identified by changing some empirical rules and parameters. Different efficient techniques are introduced to analyze the image and video contents in a variety of applications [33, 34].

3 Proposed methodology

The framework of the proposed technique is presented in Fig. 3. We obtained the spatial information by separating an image into four adapted triangular areas. This enables to extract the visual features based on LIOP

Table 3 Average-recall performance comparison of the proposed technique with recent CBIR techniques on the image benchmark of the Corel-1K

Class name	Proposed technique with step size = 10	[43]	[42]	[41]	[28]	[44]
Africa	14.04	14.6	13	12.7	14.92	11.75
Beach	15.79	14.4	14	12.84	7.56	9.79
Buildings	16.18	15.8	15	13.96	10.78	10.75
Busses	19.57	20	19	18.3	19.34	19.16
Dinosaurs	19.86	19.4	20	19.84	19.8	19.67
Elephants	18.93	15	16	15.62	13.18	12.83
Flowers	17.82	17.2	19	18.96	18.24	17.13
Horses	19.41	16.4	18	19.04	17.38	16.06
Mountains	16.37	13.8	15	14.76	11.7	10.85
Food	16.45	18	15	16.12	12.44	12.63
Average	17.44	16.46	16.40	16.21	14.53	14.06

features, weighted soft codebooks, and spatial histograms from the top, down, left, and right areas of the image. Figure 3 is presenting the procedure for computation of dense LIOP features, weighted soft codebooks, and spatial histograms over the four triangular areas of the image. The description of each step of the proposed technique is as follows:

1. The approximation coefficient of each image (represented by IMG) of the training and test sets after applying level-2 decomposition using discrete wavelet transform (DWT) is divided into the four adapted triangular areas, which are extracted by applying following mathematical equations:

$$R_{ttp1} = IMG(1, 1), R_{ttp2} = IMG(1, w), R_{ttp3} = IMG\left(\frac{h}{2}, \frac{w}{2}\right) \tag{1}$$

$$R_{btp1} = IMG(h, 1), R_{btp2} = IMG(h, w), R_{btp3} = IMG\left(\frac{h}{2}, \frac{w}{2}\right) \tag{2}$$

$$R_{ltp1} = IMG(1, 1), R_{ltp2} = IMG(h, 1), R_{ltp3} = IMG\left(\frac{h}{2}, \frac{w}{2}\right) \tag{3}$$

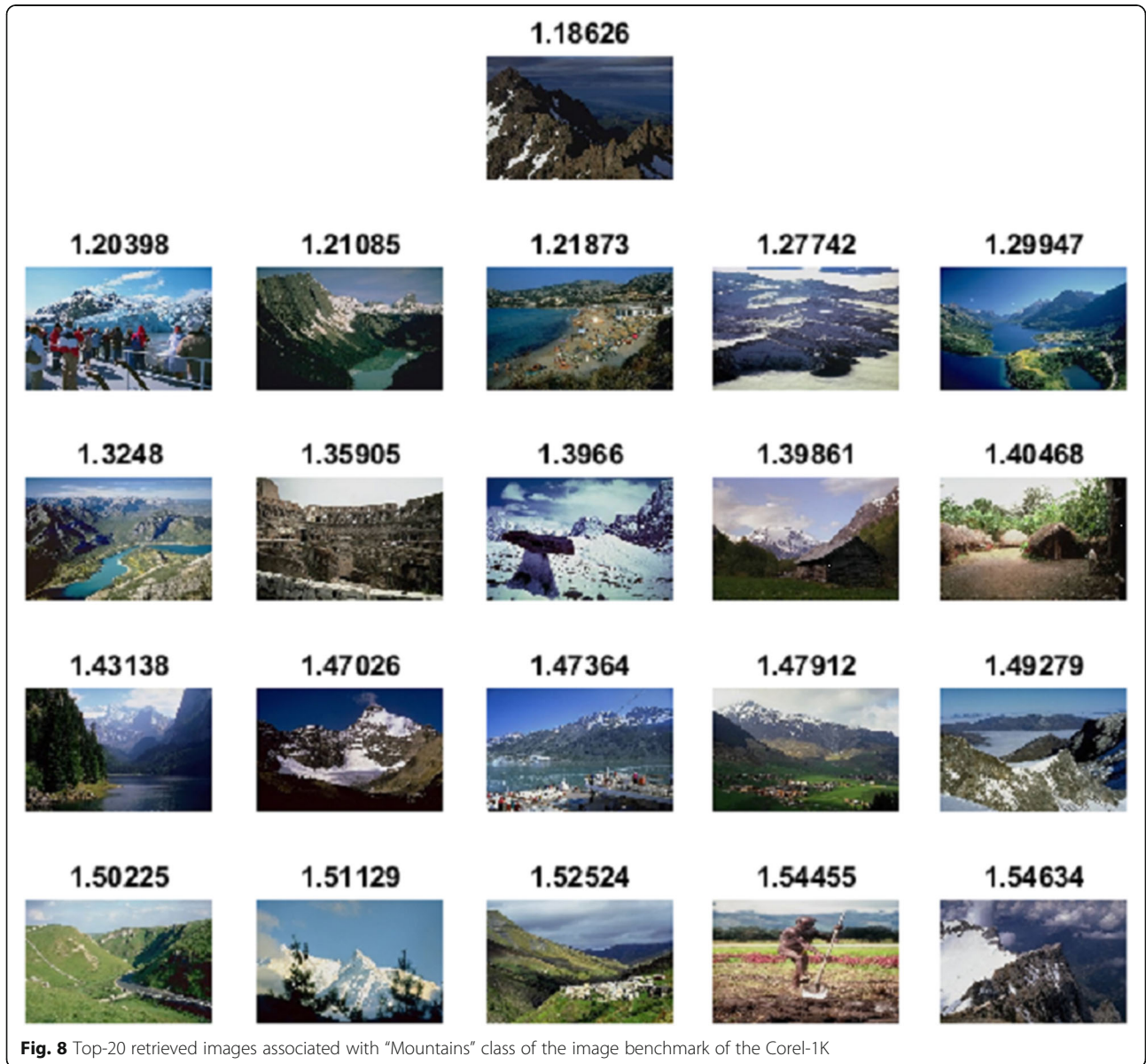


Fig. 8 Top-20 retrieved images associated with “Mountains” class of the image benchmark of the Corel-1K

$$\begin{aligned}
 R_{rtp1} &= \text{IMG}(1, w), \quad R_{rtp2} = \text{IMG}(h, w), \quad R_{rtp3} \\
 &= \text{IMG}\left(\frac{h}{2}, \frac{w}{2}\right)
 \end{aligned}
 \tag{4}$$

$$\text{des}_l = \sum_{x \in \text{bin}_l} w(x) \text{LIOP}(x)
 \tag{6}$$

where

$$\text{LIOP}(x) = \phi(\gamma(P(x)))
 \tag{7}$$

where

$$P(x) = (I(x_1), I(x_2), \dots, I(x_n)) \in P^N
 \tag{8}$$

2. The LIOP features [35] are computed from each adapted triangular area of the image by applying following mathematical equations:

$$\text{LIOP descriptor} = (\text{des}_1, \text{des}_2, \dots, \text{des}_i)
 \tag{5}$$

and ϕ is a feature mapping function that map the permutation π to an $N!$ -dimensional feature vector $V_{N!}^i$ whose all the elements are 0 except for the i^{th} element which is 1. The feature mapping function ϕ is defined by the following mathematically equation:

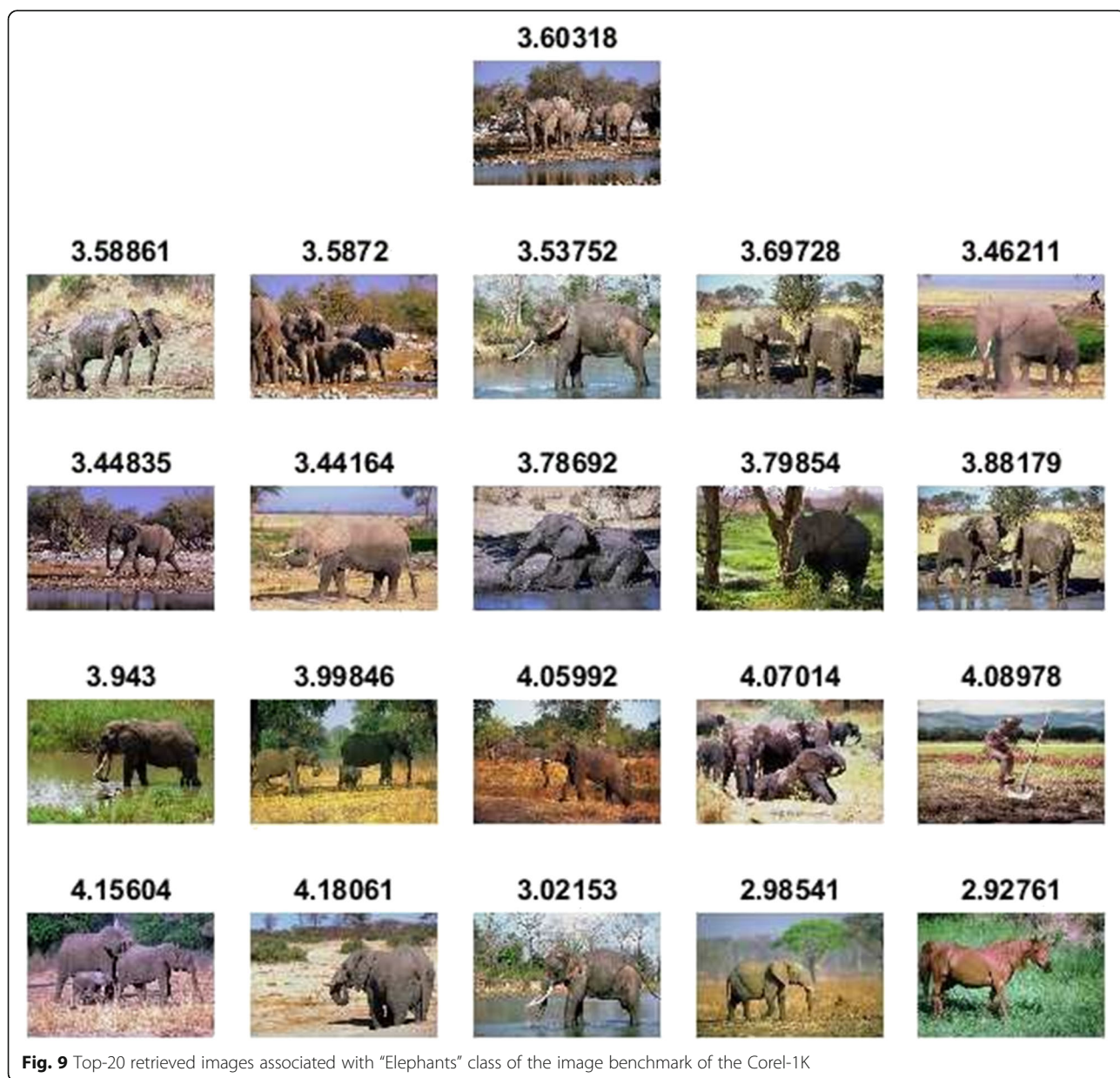


Fig. 9 Top-20 retrieved images associated with “Elephants” class of the image benchmark of the Corel-1K



Fig. 10 Sample images associated with image benchmark of the Corel-1.5K [40]

$$\phi(\pi) = V_{N!}^{\text{Ind}(\pi)}, \pi \in \prod^N \quad (9)$$

where $V_{N!}^{\text{Ind}(\pi)} = (0, \dots, 0, 1_{\text{Ind}(\pi)}, 0, \dots, 0)$ and $\text{Ind}(\pi)$ represents the index of π in the index table.

$$\text{LIOP}(x) = V_{N!}^{\text{Ind}(\gamma(P(x)))}$$

$$\text{LIOP}(x) = (0, \dots, 0, 1_{\text{Ind}(\gamma(P(x)))}, 0, \dots, 0) \quad (10)$$

$$\text{and } w(x) = \sum_{i,j} \text{sgn} (|I(x_i) - I(x_j)| - T_{lp}) + 1 \quad (11)$$

In the above equations, for a sample point x_n , $I(x_n)$ represents the intensity of the n^{th} neighboring sample, preset threshold is represented by T_{lp} , sign function is represented by sgn , $w(x)$ represents the weighted function of the LIOP descriptor, the feature mapping function is represented by ϕ , and i, j represent the coordinate position of the sample point x_n .

3. The combination of visual words known as a codebook. In order to formulate four weighted soft codebooks, the clustering technique based on the k -means++ [36] is applied on the extracted features from each adapted triangular area, which produces four soft codebooks. In order to resolve the problem of overfitting on a codebook of larger sizes, each soft codebook is multiplied by the weight of 0.25 (as four weighted soft codebooks are formulated, so 1/4 result in a weight (w) of 0.25), which produces four weighted soft codebooks. The clustering technique based on the k -means++ is chosen for auto selection of initial seed for clustering to improve the clustering results. The four weighted soft codebooks are represented by the following mathematical equations:

$$C_{wst} = 0.25 \times \{v_{t1}, v_{t2}, v_{t3}, v_{t4}, v_{t5}, \dots, v_{tx}\} \quad (12)$$

$$C_{wsb} = 0.25 \times \{v_{b1}, v_{b2}, v_{b3}, v_{b4}, v_{b5}, \dots, v_{bx}\} \quad (13)$$

$$C_{wsl} = 0.25 \times \{v_{l1}, v_{l2}, v_{l3}, v_{l4}, v_{l5}, \dots, v_{lx}\} \quad (14)$$

$$C_{wsr} = 0.25 \times \{v_{r1}, v_{r2}, v_{r3}, v_{r4}, v_{r5}, \dots, v_{rx}\} \quad (15)$$

where C_{wst} , C_{wsb} , C_{wsl} and C_{wsr} represent the weighted soft codebooks formulated from top, bottom, left, and right adapted triangular areas of the image, respectively. The v_{t1}

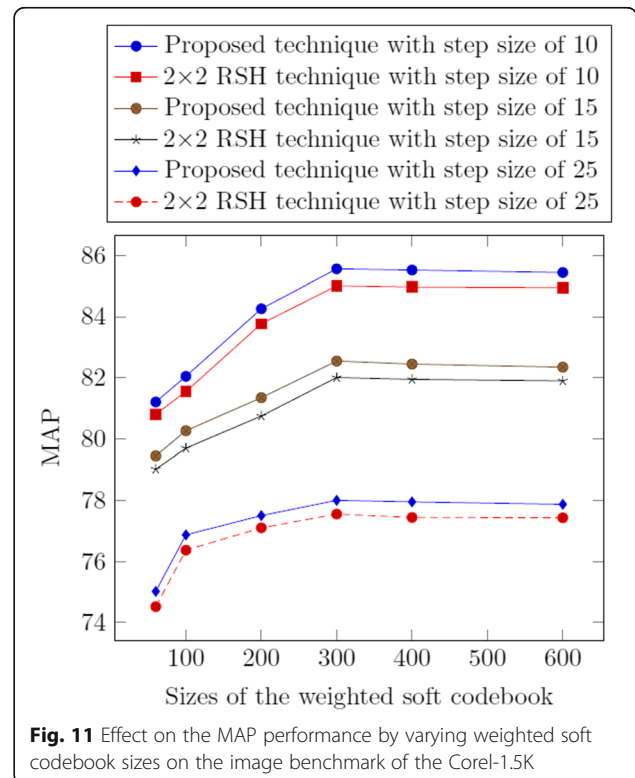


Fig. 11 Effect on the MAP performance by varying weighted soft codebook sizes on the image benchmark of the Corel-1.5K

Table 4 MAP performance and recall comparison of the proposed technique with recent CBIR techniques on the image benchmark of the Corel-1.5K

Performance evaluation parameters	Proposed technique with step size = 10	SQ + Spatiogram [45]	GMM + mSpatigram [45]
MAP	85.56	63.95	74.10
Recall	17.11	12.79	13.80

to v_{tx} , v_{b1} to v_{bx} , v_{l1} to v_{lx} and v_{r1} to v_{rx} represent the visual words of the top, bottom, left, and right adapted triangular areas of the weighted soft codebooks, respectively.

- The visual words of each adapted triangular area of the image are mapped to the associated quantized feature descriptor by applying following mathematical equation:

$$v(a_n) = \operatorname{argmin}_{v \in c_{atr}} \operatorname{Dist}(v, a_n) \tag{16}$$

where $v(a_n)$ represents the associated visual word of the n^{th} feature descriptor a_n . The distance between visual word v and feature descriptor a_n is represented by $\operatorname{Dist}(v, a_n)$. The c_{atr} represents the weighted soft codebook of the associated triangular area of the image.

- The spatial histogram is formulated using an x number of visual words of each adapted triangular area of the weighted soft codebook as shown in Fig. 4. The four adapted triangular areas of the image produce four resultant spatial histograms.
- The spatial histogram of each adapted triangular area of the image is concatenated together, and spatial information is added to the VBoW framework. Let total number of visual words of each weighted soft

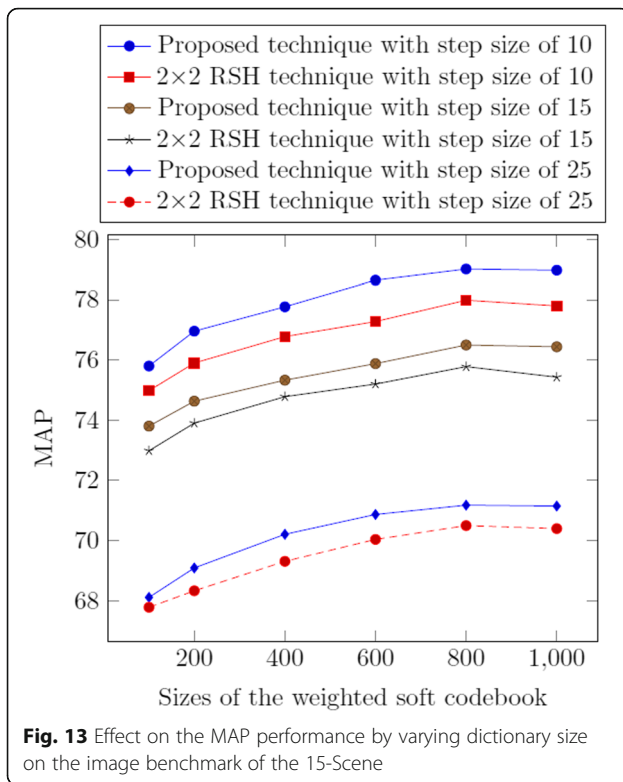
codebook (C_{ws}) are represented by x . If visual word v_i is mapped to the d_{si} descriptor, then i^{th} bin of each spatial histogram h_i is the cardinality of the set d_s is represented as follows:

$$h_i = \operatorname{Card}(d_{si}) \quad \text{and} \quad d_{si} = \{v_{dsi}, ds_i \in (1, \dots, x) | C_{ws}(v_{dsi}) = v_i\} \tag{17}$$

- The resultant spatial histogram of each image is normalized by applying the Hellinger kernel function [37] of the support vector machine (SVM). The total images of each reported image benchmark are categorized into training and test sets. The SVM classifier is trained using normalized histograms of the training images. The best values of the regularization parameters (C and Γ) are determined by applying 10-fold cross validation using the images of the training set for each reported image benchmark.
- The Euclidean distance [38] is used as a similarity measure technique. The images are retrieved by measuring the similarity between a score of the inquiry image and scores of the archive images for each reported image benchmark.



Fig. 12 Sample images associated with image benchmark of the 15-Scene



4 Experimental results and discussions

To evaluate the performance of proposed methodology, we have selected five challenging image benchmarks and results are compared with recent CBIR techniques. The images are randomly divided into training and test sets. Visual features are extracted by applying dense-LIOP feature descriptor, and all the processing is performed on the grayscale images.

Training images are used to formulate the weighted soft codebooks, and test images are used to compute the retrieval precision. Due to the unsupervised nature of clustering using *k*-means++, each experiment is repeated 10 times and the average values are reported and for every iteration, images are randomly divided into training and test sets. Lazebnik et al. [17] proposed a CBIR technique based on the spatial pyramid matching, which divides an image into several rectangular grids and formulate histograms from each region of the grid. The proposed technique of image representation based on adapted triangular areas is compared with 2 × 2 RSH technique of CBIR, which divides an image into 2 × 2 rectangular grids. Figure 5 is

presenting the division of an image into four rectangular areas. The histograms are computed from each rectangular region to perform a comparison between the spatial rectangular and adapted triangular histograms of visual words.

4.1 Experimental parameters and performance evaluation metrics

The details about the parameters used for the experimental research are given below:

- a) *Codebook size*: The total number of images of each reported image collection is divided into two sets known as test and training sets. The images of the training set are used to formulate the dictionary for each image benchmark. The performance of the CBIR techniques based on the VBoW model is affected by varying sizes of the codebook.
- b) *Step size*: Dense-LIOP is used for features extraction. For a precise content-based image matching, we extracted dense features from four divided triangular regions of each image (at four different scales). Step size is used to control the sampling density, which is the vertical and horizontal displacement of each feature center to the next. The proposed technique is evaluated using pixel step sizes of 10, 15, and 25. For a step size of 10, every 10th pixel is selected to compute the LIOP descriptor.
- c) *Features % per image for dictionary learning*: According to [39], the number of features percentage per image for codebook or dictionary learning from the training set is an important parameter that affects the performance of CBIR. We formulated dictionary using different feature percentages [10, 25, 50, 75, and 100%] per image of the training set. Precision, recall, average precision (AP), and mean average precision (MAP) are the standard performance evaluation metrics to evaluate the performance of the CBIR system. The performance of the proposed technique is also evaluated using these metrics.
- d) *Precision*: The specificity of the image retrieval model is evaluated by the precision \bar{P} , which is mathematically defined as follows:

$$P = \frac{C_r}{R_t} \tag{18}$$

Table 5 MAP performance and recall comparison of the proposed technique with recent CBIR techniques on the image benchmark of the 15-Scene

Performance measures	Proposed technique	RSH technique [28]	HOG 2 × 2 technique [52]	Spatial level-2 technique (DBN) [51]
Precision	79.02	81.10	81.00	79.70
Recall	15.80	16.22	16.20	15.94



Fig. 14 Sample images associated with image benchmark of the Ground-truth

- e) *Recall*: The recall ‘ R ’ evaluate the sensitivity of the image retrieval model, which is mathematically defined as follows:

$$R = \frac{C_r}{T_c} \tag{19}$$

where R_r , C_r , and T_c represent the total retrieved, correctly retrieved, and total per class images, respectively.

- f) *Average precision (AP)*: The AP for a set of image queries is the average of the precision of particular class of the image benchmark, which is mathematically defined as follows:

$$AP = \frac{\sum_{j=1}^M P(j)}{M} \tag{20}$$

- g) *Mean average precision (MAP)*: For a set of image queries, the MAP is the mean of the average precision values for each image query, which is mathematically defined as follows:

$$MAP = \frac{\sum_{j=1}^M AP(j)}{M} \tag{21}$$

where M is the total number of the image queries.

4.2 Analysis of the evaluation metrics on the image benchmark of the Corel-1K

The image benchmark of the Corel-1K is a subset of the WANG image benchmark [40]. The proposed technique is evaluated using image benchmark of the Corel-1K, and recent CBIR techniques [28, 41–44] are used for the performance comparison of the proposed technique. The image benchmark of the Corel-1K comprises of 1000 images, which are organized into 10 semantic classes. The sample image associated with each class of the Corel-1K is presented in Fig. 6. The test set and a training set of the Corel-1K are split into 300 and 700 images, respectively. The MAP performance for top 20 image retrievals for a step size of 10 using different sizes of the weighted soft codebooks and feature percentage for weighted soft codebook learning are shown in Table 1.

The MAP is calculated by taking the mean of the column-wise values of Table 1. The comparison of the MAP performance using proposed technique and 2×2 RSH technique for the dense pixel strides of 10, 15, and 25 is presented in Fig. 7. According to the experimental results, the MAP obtained by using the proposed research with a pixel stride of 10 is 87.22%, while the MAP obtained from the pixel strides of 15 and 25 is 84.27 and 78.13%, respectively (with a weighted soft codebook size of 200 visual words). This shows that increasing the pixel stride decreases the MAP performance and vice versa. In order to present a sustainable performance of the proposed research, the MAP for top-20

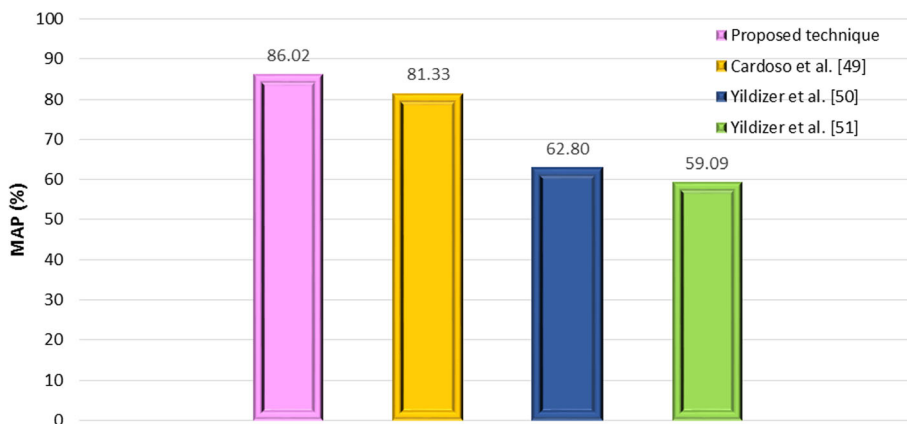


Fig. 15 Analysis of MAP performance comparison of the proposed technique with recent CBIR techniques [47–49] on the image benchmark of the Ground-truth

Table 6 MAP performance and recall comparison of the proposed technique with recent CBIR techniques on the image benchmark of the Caltech-256

Performance evaluation parameters	Proposed technique	MN-ARM [53]	MN-MIN [54]	DCT [22]
MAP	31.19	28.21	26.89	23.91
Recall	06.24	05.64	5.37	04.78

retrievals is calculated and compared with recent CBIR techniques [28, 41–44]. Table 2 and Table 3 are presenting the class-wise comparisons of average precision and recall on the image benchmark of the Corel-1K.

Experimental results and the comparisons are conducted on the image benchmark of the Corel-1K prove the robustness of the proposed technique. The mean precision and recall values obtained using proposed technique are higher than the recent CBIR techniques [28, 41–44]. The image retrieval results for the semantic classes of “Mountains” and “Elephants” are presented in Figs. 8 and 9, respectively in response to the query images that show a reduction of semantic gap in terms of classifier decision value (score). 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.

4.3 Analysis of the evaluation metrics on the image benchmark of the Corel-1.5K

There are 15 semantic classes in image benchmark of the Corel-1.5K, and each semantic class contains 100 images. The image benchmark of the Corel-1.5K is also

a subset of the WANG image benchmark [40] and used for performance comparison of the proposed technique with [45]. Figure 10 is presenting the sample of images from each semantic class of the image benchmark of the Corel-1.5K. The MAP performance as a function of the weighted soft codebook size is graphically presented in Fig. 11. The precision and recall values obtained using the proposed technique are compared with recent CBIR technique of [45] and are presented in Table 4.

According to the experimental results, the MAP performance of 85.56% is obtained using the proposed technique on a weighted soft codebook size of 400 visual words (with pixel step size of 10), while the MAP performance obtained using 2 × 2 RSH technique is 84.97%. The proposed image representation outperforms 2 × 2 RSH-based CBIR technique as well as recent CBIR technique of [45].

4.4 Analysis of the evaluation metrics on the image benchmark of the 15-Scene

The 15-Scene image benchmark [46] comprises of 4485 images that are organized into 15 categories. Each category comprises of 200 to 400 images containing outdoor and indoor scenes as shown in Fig. 12. The resolution of each image in this image collection is 250 × 300 pixels. The MAP performance of the proposed technique on different sizes of the dictionary is shown in Fig. 13, which is compared with 2 × 2 RSH-based CBIR technique.

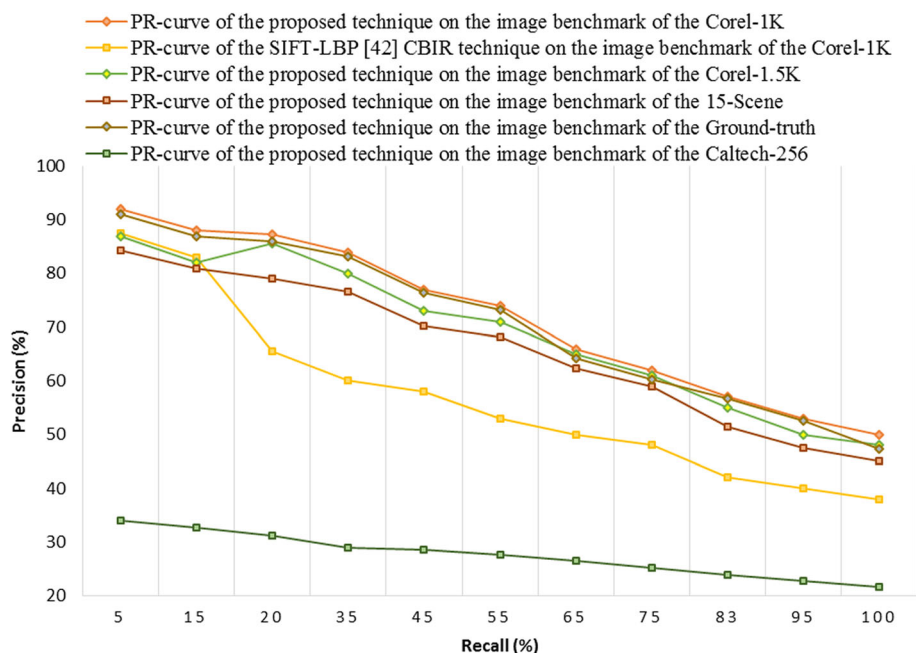


Fig. 16 PR-curve of the proposed technique on the image benchmarks of the Corel-1K, Corel-1.5K, 15-Scene, Ground-truth, and Caltech-256

The experimental details are shown in Fig. 13 and Table 5 that indicates the efficiency of the proposed technique on the reported weighted soft codebook sizes as compared with 2×2 RSH-based CBIR technique and recent CBIR techniques [26, 40, 41]. The proposed technique and 2×2 RSH technique gives best MAP performance of 79.02 and 77.99% by formulating the weighted soft codebook size of 800 visual words using a step size of 10, respectively.

4.5 Analysis of the evaluation metrics on the image benchmark of the Ground-truth

In Ground-truth image benchmark, a total number of images is 1109 which are organized into 22 semantic classes. The Ground-truth image benchmark is commonly used for performance evaluation of the recent CBIR techniques [47–49]. For a clear comparison, 5 semantic classes which comprise a total number of 228 images are chosen as the performance of the recent CBIR techniques [47–49] are also evaluated for the same classes. The sample images of the chosen categories of the Ground-truth image benchmark are shown in Fig. 14, while MAP performance comparison of the proposed technique with recent CBIR techniques [47–49] is presented in Fig. 15, which proves the robustness of the proposed technique by formulating a weighted soft codebook size of 60 words using a step size of 10.

4.6 Analysis of the evaluation metrics on the image benchmark of the Caltech-256

The image benchmark of the Caltech-256 is deposited on 2007 and a successor of the image benchmark of the Caltech-101. The total number of images in the Caltech-256 are 30,607 which are categorized into 256 semantic classes. Every semantic class contains a minimum of 80 images with no artifacts, performance is halved, shows diversity in image representation [50]. The best MAP performance of the proposed technique is 31.19%, which is achieved on a weighted soft codebook size of 1400 visual words. The performance of the proposed technique is compared with recent CBIR techniques [20, 44, 45], which also proved its robustness as presented in Table 6. The performance analysis in terms of the precision-recall (PR)-curve of the proposed technique is presented in Fig. 16 for the image benchmarks of the Corel-1K, Corel-1.5K, 15-Scene, Ground-truth, and Caltech-256.

4.7 Requirement of the computational resources

The performance of the proposed technique of this article is measured on a computer, whose hardware specifications are as follows: RAM with 8 GB storage capacity, GPU with 2 GB storage capacity, and Intel Pentium (R) Core i7 microprocessor with 2.4 GHz clock frequency.

Table 7 Computational complexity (time in seconds) of the proposed technique required for feature extraction only

Proposed technique	RSHD technique [55]			EODH technique [28]	Spatial level-2 technique [51]	Standard VBoW technique
	RSHD	CDH	SHE			
0.0788	0.3750	1.7090	0.1860	5.6	0.0821	0.0641

The required software resources for the implementation of the proposed technique are Microsoft Windows 7 64-bit operating system and MATLAB 2015a. The computational complexity required for feature extraction of the proposed technique is presented in Table 7, while computational complexity of the proposed image retrieval framework (i.e., complete framework) is presented in Table 8. The computational complexity is reported by selecting the image benchmark of the Corel-1 K, which comprises of each image resolution of 384×256 or 256×384 .

5 Conclusions

In this article, we have proposed a novel image representation based on the adapted triangular areas and weighted soft codebooks. The dense LIOP features, weighted soft codebooks, and spatial histograms are extracted over the four triangular areas of the image. The proposed technique adds the spatial context of information to the inverted index of VBoW model. The collection of dense LIOP features and spatial histograms over the four adapted triangular areas of an image is a possible solution to add the spatial information to the VBoW model and reduction of the semantic gap between low-level features of the image and high-level semantic concepts. The problem of overfitting on the codebook of larger sizes is reduced by the weighted soft codebooks to further improve CBIR performance. The Hellinger kernel of the SVM is selected for image classification. The proposed technique is evaluated on five challenging image benchmarks and results are compared with recent CBIR techniques and 2×2 RSH technique of the CBIR. The proposed image representation outperforms in terms of the performance comparison with recent CBIR techniques as well as RSH technique of the

Table 8 Computational complexity (time in seconds) of the proposed image retrieval framework and its comparisons with recent CBIR techniques

Total number of retrieved images	Proposed technique	Spatial level-2 technique [51]	Standard VBoW technique
Top-05	0.3601	0.3726	0.3415
Top-10	0.4829	0.5178	0.4674
Top-15	0.6901	0.7050	0.6589
Top-20	0.8513	0.8882	0.8213
Top-25	1.0243	1.0599	0.9913

CBIR. In future, we plan to replace VBoW model with a vector of locally aggregated descriptors (VLAD) or Fisher kernel framework to evaluate the proposed technique for a large-scale image retrieval.

Abbreviations

ANN: Artificial neural network; ARM: Adaptive region matching; CBIR: Content-based image retrieval; DCD: Dominant color descriptor; DCT: Discrete cosine transform; DWT: Discrete wavelet transform; EODH: Edge orientation difference histogram; HOG: Histogram of oriented gradients; LBP: Local binary pattern; LIOP: Local intensity order pattern; MAP: Mean average precision; MSERs: Maximally stable extremal regions; QBIC: Query-by-image-content; RBIR: Region-based image retrieval; RII: Region important index; RSH: Rectangular spatial histogram; SDHP: Supervised deep hashing framework; SMR: Semantic meaningful region; SOM: Self-organizing map; SVD: Singular value decomposition; SVM: Support vector machine; SWBoF: Spatial weighting bag-of-features; VBoW: Visual-bag-of-words; VLAD: Vector of locally aggregated descriptors

Funding

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B03933860).

Availability of data and materials

Data sharing is not applicable to this article as authors have used the publically available datasets, whose details are included in Section 4 of this article. Please contact authors for further requests.

Authors' contributions

All the authors contributed equally. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan. ²Department of Basic Sciences, University of Engineering and Technology, Taxila 47050, Pakistan. ³Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan. ⁴College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia. ⁵College of Computer and Information Systems, Al-Yamamah University, Riyadh 11512, Saudi Arabia. ⁶School of Computer Science and Engineering, Korea University of Technology and Education, Cheonan 330-708, Republic of Korea.

Received: 20 July 2017 Accepted: 25 May 2018

Published online: 13 June 2018

References

1. Alzu'bi, A, Amira, N, Ramzan, Semantic content-based image retrieval: a comprehensive study. *J. Vis. Commun. Image Represent.* **32**, 20–54 (2015).
2. A-M Tusch, S Herbin, J-Y Audibert, Semantic hierarchies for image annotation: a survey. *Pattern Recogn.* **45**(1), 333–345 (2012).
3. D Zhang, MM Islam, G Lu, A review on automatic image annotation techniques. *Pattern Recogn.* **45**(1), 346–362 (2012).
4. Hiremath, P, and J. Pujari. Content based image retrieval using color, texture and shape features. in *Advanced Computing and Communications, 2007. ADCOM 2007. International Conference on.* 2007. IEEE.
5. T Wan, Z Qin, An application of compressive sensing for image fusion. *Int. J. Comput. Math.* **88**(18), 3915–3930 (2011).
6. T Ojala, M Pietikainen, T Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002).
7. Takala, V., T. Ahonen, and M. Pietikäinen, Block-based methods for image retrieval using local binary patterns. *Image Analysis, 2005:* p. 13–181.
8. Z Mehmood, SM Anwar, N Ali, HA Habib, M Rashid, A novel image retrieval based on a combination of local and global histograms of visual words. *Math. Probl. Eng.* **2016**(2016), 1–12 (2016).
9. Z Mehmood, S Anwar, M Altaf, A novel image retrieval based on rectangular spatial histograms of visual words. *Kuwait J. Sci.* **45**(1), 54–69 (2018).
10. Sivic, J. and A. Zisserman. *Video Google: a text retrieval approach to object matching in videos.* in *null.* 2003. IEEE.
11. Z Liu, H Li, W Zhou, R Zhao, Q Tian, Contextual hashing for large-scale image search. *IEEE Trans. Image Process.* **23**(4), 1606–1614 (2014).
12. Philbin, J., O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. in *Computer Vision and Pattern Recognition, 2007. CVPR07. IEEE Conference on.* 2007. IEEE.
13. M Yousuf, Z Mehmood, HA Habib, T Mahmood, T Saba, A Rehman, M Rashid, A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval. *Math. Probl. Eng.* **2018**, 1–13 (2018).
14. S Jabeen, Z Mehmood, T Mahmood, T Saba, A Rehman, MT Mahmood, An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model. *PLoS One* **13**(4), e0194526 (2018).
15. Zhou, W., H. Li, Y. Lu, and Q. Tian, SIFT match verification by geometric coding for large-scale partial-duplicate web image search. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2013. 9(1): p. 4.
16. Khan, R., C. Barat, D. Muselet, and C. Ducottet. *Spatial orientations of visual word pairs to improve bag-of-visual-words model.* in *Proceedings of the British Machine Vision Conference.* 2012. BMVA Press.
17. Lazebnik, S., C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. in *Computer vision and pattern recognition, 2006 IEEE computer society conference on.* 2006. IEEE.
18. H Zhang, M Gönen, Z Yang, E Oja, Understanding emotional impact of images using Bayesian multiple kernel learning. *Neurocomputing* **165**, 3–13 (2015).
19. M Flickner, H Sawhney, W Niblack, J Ashley, Q Huang, B Dom, M Gorkani, J Hafner, D Lee, D Petkovic, Query by image and video content: the QBIC system. *Computer* **28**(9), 23–32 (1995).
20. Z Mehmood, T Mahmood, MA Javid, Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine. *Appl. Intell.* **48**(1), 166–181 (2018).
21. Mehmood, Z., F. Abbas, T. Mahmood, M.A. Javid, A. Rehman, and T. Nawaz, Content-based image retrieval based on visual words fusion versus features fusion of local and global features. *Arab. J. Sci. Eng.*, 2018: p. 1–20.
22. D Zhong, I Defée, DCT histogram optimization for image database retrieval. *Pattern Recogn. Lett.* **26**(14), 2272–2281 (2005).
23. X Yuan, J Yu, Z Qin, T Wan, in *IEEE International Conference on Image Processing.* A SIFT-LBP image retrieval model based on bag of features (2011).
24. J Yu, Z Qin, T Wan, X Zhang, Feature integration analysis of bag-of-features model for image retrieval. *Neurocomputing* **120**, 355–364 (2013).
25. NMK Raja, KS Bhanu, Content bases image search and retrieval using indexing by KMeans clustering technique. *Int. J. Adv. Res. Comp. Comm. Eng.* **2**(5), 2181–2189 (2013).
26. X Yang, L Cai, Adaptive region matching for region-based image retrieval by constructing region importance index. *IET Comput. Vis.* **8**(2), 141–151 (2013).
27. Wang, C., B. Zhang, Z. Qin, and J. Xiong. *Spatial weighting for bag-of-features based image retrieval.* In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making.* 2013. Springer..
28. X Tian, L Jiao, X Liu, X Zhang, Feature integration of EODH and Color-SIFT: application to image retrieval based on codebook. *Signal Process. Image Commun.* **29**(4), 530–545 (2014).
29. Rashno, A., S. Sadri, and H. SadeghianNejad. An efficient content-based image retrieval with ant colony optimization feature selection schema based on wavelet and color features. in *Artificial Intelligence and Signal Processing (AISP), 2015 International Symposium on.* 2015. IEEE.
30. M Rahimi, ME Moghaddam, A content-based image retrieval system based on Color Ton Distribution descriptors. *SIVIP* **9**(3), 691–704 (2015).
31. C Yan, H Xie, D Yang, J Yin, Y Zhang, Q Dai, Supervised hash coding with deep neural network for environment perception of intelligent vehicles. *IEEE Trans. Intell. Transp. Syst.* **19**(1), 284–295 (2018).

32. C Yan, H Xie, S Liu, J Yin, Y Zhang, Q Dai, Effective Uyghur language text detection in complex background images for traffic prompt identification. *IEEE Trans. Intell. Transp. Syst.* **19**(1), 220–229 (2018).
33. C Yan, Y Zhang, J Xu, F Dai, J Zhang, Q Dai, F Wu, Efficient parallel framework for HEVC motion estimation on many-core processors. *IEEE Trans. Circuits Syst. Video Technol.* **24**(12), 2077–2089 (2014).
34. C Yan, Y Zhang, J Xu, F Dai, L Li, Q Dai, F Wu, A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. *IEEE Signal Processing Lett.* **21**(5), 573–576 (2014).
35. Wang, Z, B. Fan, and F. Wu. Local intensity order pattern for feature description. in *Computer Vision (ICCV), 2011 IEEE International Conference on*. 2011. IEEE.
36. Arthur, D. and S. Vassilvitskii. *k-means++: The advantages of careful seeding*. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007. Society for Industrial and Applied Mathematics.
37. J Shawe-Taylor, N Cristianini, *Kernel Methods for Pattern Analysis* (Cambridge University Press, 2004).
38. R Short, K Fukunaga, The optimal distance measure for nearest neighbor classification. *IEEE Trans. Inf. Theory* **27**(5), 622–627 (1981).
39. Csurka, G., C. Dance, L. Fan, J. Willamowski, and C. Bray. *Visual categorization with bags of keypoints*. In *Workshop on statistical learning in computer vision, ECCV. 2004*. Prague.
40. J Li, JZ Wang, Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(6), 985–1002 (2008).
41. SM Youssef, ICTEDCT-CBIR: Integrating curvelet transform with enhanced dominant colors extraction and texture analysis for efficient content-based image retrieval. *Comput. Electr. Eng.* **38**(5), 1358–1376 (2012).
42. R Ashraf, K Bashir, A Irtaza, MT Mahmood, Content based image retrieval using embedded neural networks with bandletized regions. *Entropy* **17**(6), 3552–3580 (2015).
43. A Irtaza, MA Jaffar, Categorical image retrieval through genetically optimized support vector machines (GOSVM) and hybrid texture features. *SIVIP* **9**(7), 1503–1519 (2015).
44. GA Montazer, D Giveki, An improved radial basis function neural network for object image retrieval. *Neurocomputing* **168**, 221–233 (2015).
45. S Zeng, R Huang, H Wang, Z Kang, Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models. *Neurocomputing* **171**, 673–684 (2016).
46. Fei-Fei, L. and P. Perona. A bayesian hierarchical model for learning natural scene categories. in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. 2005. IEEE.
47. E Yildizer, AM Balci, TN Jarada, R Alhaji, Integrating wavelets with clustering and indexing for effective content-based image retrieval. *Knowl.-Based Syst.* **31**, 55–66 (2012).
48. E Yildizer, AM Balci, M Hassan, R Alhaji, Efficient content-based image retrieval using multiple support vector machines ensemble. *Expert Syst. Appl.* **39**(3), 2385–2396 (2012a).
49. DNM Cardoso, DJ Muller, F Alexandre, LAP Neves, PMG Trevisani, GA Giraldi, Iterative technique for content-based image retrieval using multiple SVM ensembles. *J Clerk Maxwell, A Treatise Electricity Magnetism* **2**, 68–73 (2013).
50. Griffin, G., A. Holub, and P. Perona, Caltech-256 Object Category Dataset. 2007.
51. N Ali, KB Bajwa, R Sablatnig, Z Mehmood, Image retrieval by addition of spatial information based on histograms of triangular regions. *Comput. Electr. Eng.* **54**, 539–550 (2016).
52. Xiao, J., J. Hays, KA. Ehinger, A. Oliva, and A. Torralba. Sun database: large-scale scene recognition from abbey to zoo. in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. 2010. IEEE.
53. G-H Liu, Z-Y Li, L Zhang, Y Xu, Image retrieval based on micro-structure descriptor. *Pattern Recogn.* **44**(9), 2123–2133 (2011).
54. C Carson, S Belongie, H Greenspan, J Malik, Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(8), 1026–1038 (2002).
55. SR Dubey, SK Singh, RK Singh, Rotation and scale invariant hybrid image descriptor and retrieval. *Comput. Electr. Eng.* **46**, 288–302 (2015).