Bibi et al. EURASIP Journal on Image and Video Processing (2020) 2020:27
https://doi.org/10.1186/s13640-020-00516-4

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

BoVW model based on adaptive local and global visual words modeling and log-based relevance feedback for semantic retrieval of the images

Ruqia Bibi[1], Zahid Mehmood[2*], Rehan Mehmood Yousaf[1], Muhammad Tahir[3], Amjad Rehman[4], Muhammad Sardaraz[3] and Muhammad Rashid[5]

* Correspondence: zahid.mehmood@uettaxila.edu.pk
[2]Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan
Full list of author information is available at the end of the article

Abstract

The core of a content-based image retrieval (CBIR) system is based on an effective understanding of the visual contents of images due to which a CBIR system can be termed as accurate. One of the most prominent issues which affect the performance of a CBIR system is the semantic gap. It is a variance that exists between low-level patterns of an image and high-level abstractions as perceived by humans. A robust image visual representation and relevance feedback (RF) can bridge this gap by extracting distinctive local and global features from the image and by incorporating valuable information stored as feedback. To handle this issue, this article presents a novel adaptive complementary visual word integration method for a robust representation of the salient objects of the image using local and global features based on the bag-of-visual-words (BoVW) model. To analyze the performance of the proposed method, three integration methods based on the BoVW model are proposed in this article: (a) integration of complementary features before clustering (called as non-adaptive complementary feature integration), (b) integration of non-adaptive complementary features after clustering (called as a non-adaptive complementary visual words integration), and (c) integration of adaptive complementary feature weighting after clustering based on self-paced learning (called as a proposed method based on adaptive complementary visual words integration). The performance of the proposed method is further enhanced by incorporating a log-based RF (LRF) method in the proposed model. The qualitative and quantitative analysis of the proposed method is carried on four image datasets, which show that the proposed adaptive complementary visual words integration method outperforms as compared with the non-adaptive complementary feature integration, non-adaptive complementary visual words integration, and state-of-the-art CBIR methods in terms of performance evaluation metrics.

Keywords: Query-by-image, Visual feature integration, Adaptive weighting features, Robust learning, Relevance feedback

Springer Open

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

1 Introduction

Due to a staggering increase in globalization, communication, and advancement in technology, the world has become a global village in its true sense. Digital image libraries are exponentially expanding because of the proliferation of social media and other information-sharing mediums. To extract meaningful information from such a huge repository requires certain techniques that can perform retrieval effectively and within minimum computational cost. Traditional text-based approaches retrieve images based on information that is annotated manually, which now has become impractical for such huge image repositories [1]. Another reason for opting for content-based image retrieval (CBIR) is a language dependency of textual annotations. CBIR has been a rapidly progressing area since 1990, and it retrieves images having similar contents/features, i.e., colors, shapes, and textures. It is categorized into two stages: (1) feature extraction and (2) feature matching. The purpose of the first stage is to get a feature vector that can effectively represent the visual contents of images. Features are categorized as global or local features. Global features encapsulate characteristics of an entire image as a single vector. Even though they are robust and computationally efficient, they may overlook the pixel's spatial relationship and local details [2]. On the contrary, local features preserve local characteristics of an image as they are extracted from patches of an image and are consider scale and rotation-invariant. Research has been done in the recent past to explore CBIR [3–7] and its applicability in different fields such as artificial intelligence (AI), human–computer interaction (HCI), and medical imaging. With the advent of deep learning approaches, research concern has now shifted towards deep features that can be learned by algorithms on their own. The ability of artificial neural networks to classify images either through supervised or unsupervised learning explored by Krizhevsky et al. [8] has taken the research inclination to a new dimension with their breakthrough results. Several feature descriptors are being developed for CBIR, but a selection of appropriate image representation is still challenging due to different issues such as illumination changes, viewing angles, and variation in image scale. As shown in Fig. 1, visual similarity between semantically different objects is also an intriguing issue that results in misclassification of an object, which affects the overall performance of the CBIR system. Another barrier for the retrieval system is an accurate feature matching. Most CBIR systems use similarity measures, whose performance highly depends on the selected feature descriptor and distance measure being used [10–12]. The research concern of today is to lessen the semantic gap concerning the

Fig. 1 An image having similar visual contents [9]

images' low-level visual features and user high-level semantics to improve the accuracy of image retrieval systems.

This study presents an innovative method for CBIR to advance the performance of image retrieval. The proposed methodology is categorized into two sections, namely training and testing sections. In the training section, complementary features are extracted using BFGF-HOG and GSURF feature descriptors from the images of the training group. To get optimized feature vectors, latent semantic analysis (LSA) followed by an adaptive feature weighting (AFW) method based on self-paced learning (SPL) is applied to each feature vector. Afterward, a visual vocabulary is constructed by applying adaptive fuzzy *k*-means (AFKM) clustering on each optimized feature vector, which represents the contents of the images in a more compact form. These two visual vocabularies are concatenated to get a resultant visual vocabulary that contains complementary features of both descriptors, which is termed as an adaptive complementary visual word integration in the proposed method of CBIR. In the next step, a histogram is formed using visual words of each image from the resultant complementary visual vocabulary. These histograms along with training labels are used as an input to quadratic kernel-based support vector machine (QSVM) for classification. In the testing section, the aforementioned steps are carried out on a query image taken from the testing group of images, which outputs a histogram-based visual representation of a query image. Afterward, a relevance score is computed between images residing in datasets and query image by applying Euclidean distance. For further improving the performance of image retrieval, the proposed method also uses a log-based relevance feedback mechanism.

The major contributions of this study are as follows:

a. An innovative image representation method by integrating adaptive local and global visual words along with log-based relevance feedback based on the BoVW model.
b. Non-adaptive complementary visual words integration for the principal objects of the images based on the BoVW model.
c. Non-adaptive complementary feature integration for the principal objects of the image based on the BoVW model.

The remaining sections of this paper are structured as follows: Section 2 provides a detailed review of the relevant CBIR methods. Section 3 presents a detailed methodology of the proposed method. Section 4 provides detail of the experimental parameters for performance evaluations of the proposed method along with experimental results and discussion. Section 5 presents the conclusion and future directions of the research work.

2 Literature review

Numerous techniques have been developed to efficiently and effectively retrieve images from repositories having an immense and diverse collection of images from users around the globe, hence uncovering a field that makes computers understand or learn, enabling them to compete with the human brain, in short working towards AI and going deep down to imitate the working of neurons. CBIR has gained immense

recognition in the recent past and motivates researchers to innovate new techniques to recognize objects or areas under consideration with the highest possible accuracy.

Singh et al. [2] presented a novel low-dimensional color texture descriptor named as a local binary pattern for color images (LBPC). A plane is used for a 3-dimensional RGB color space. LBP of color pixels is selected across a circularly symmetric neighbor lying within radius 2. Pixels having values above the plane are termed as 1 and below the plane as 0. A combination of hue component of HIS color space with LBP, i.e., LBPH and its fusion with color histogram (CH), is also analyzed to improve the discerning capability of the descriptor. To further reduce the dimension of the proposed descriptor, a uniform pattern with 59 bins is also calculated. In terms of performance, a fusion of LBPC, LBPH, and CH achieves better retrieval accuracy when an intra-class variation is highest. Meanwhile, uniform patterns of the proposed descriptor have achieved somewhat similar retrieval accuracy with a lower computational cost. LBP's for multi-channel color images are mostly calculated individually for each channel, thus results in loss of cross-channel information and higher computational cost. Misale et al. [10] presented an efficient CBIR system based on local tetra pattern (LTrP) features and bag-of-words (BoW) model. Initially, interest points are detected through SURF, and features are extracted locally through LTrP. Dataset images are classified in a 33: 33:34 ratio for training, validation, and testing, respectively. In the testing phase, a trained neural network is employed to classify images according to semantic categories. The performance of the proposed approach highlights better retrieval accuracy and reduced computational expense. A novel feature descriptor called as multi-trend binary code descriptor (MTBCD) is proposed by Yu et al. [13], which addresses some of the common issues faced by local feature descriptors in CBIR such as the change in pixel patterns, semantic gap, and lack of spatial information. The MTBCD descriptor works on the intensity component of the HSV model and identifies a change in trend among pixels along with four symmetrical directions (0°, 45°, 90°, 135°). The change in trend is classified as parallel, if the values of pixels within an assigned radius are in increasing or decreasing order, and as non-parallel, if values are equal or greater/smaller than the center pixels. To preserve the spatial relation among pixels, a co-occurrence matrix is also constructed. Experimental analysis depicts robustness of this framework against competitive methods.

Mistry et al. [14] designed and developed a robust CBIR system by integrating various spatial and frequency-based features. This method uses color moments, auto-correlogram and HSV histogram as spatial features and stationery, and Gabor wavelet transforms as frequency domain features. Apart from these, the approach also combines features extracted through color and edge directivity descriptor (CEDD) and binarized statistical image features (BSIF) descriptor. The feature vectors of 6-D and 64-D are extracted in case of color moments and color auto-correlogram, respectively. For the CEDD-BSIF feature set, 144-D CEDD and 256-D BSIF feature vectors are generated. Frequency domain features lead to better accuracy than spatial domain features when city block and Euclidean distance are utilized for measuring similarity while CEDD and BSIF features achieve the highest precision among all. However, this method is computationally expensive because of the high-dimensional feature vector. An innovative technique based on spatial histograms (spatiograms) is presented by Zeng et al. [15] to address issues faced by generalized histograms in CBIR, i.e., loss of

spatial information, high dimensionality, and semantic gap. It quantizes the color space by using the Gaussian mixture model (GMM) learned through the expectation maximization-Bayesian information criterion (EM-BIC) algorithm, which automatically identifies the number of Gaussians (color bins) and associate pixels to multiple bins based on probability. Spatiograms are computed and incorporated with GMM. For determining a distance between spatiograms, a new measure based on Jensen–Shannon (JS) divergence is also proposed in this method. The experimental analysis highlights the robustness of the method for image retrieval. Roy et al. [16] presented a novel and highly discriminative rotation invariant texture descriptor named as a local directional zigzag pattern (LDZP). The proposed framework first reduces the noise of textured images by generating a local directional edge map (LDEM) through Kirsch compass mask along 6 directions from 0 to 150° with a 30° interval. Zigzag patterns and corresponding uniform histograms are extracted from each LDEM and concatenated to obtain rotation invariance. In terms of performance, LDZP efficiently encodes recurrent changes in local texture patterns and has better texture classification accuracy because of its zigzag sampling structure as compared to LBP which suffers from unreliable texture information because of its circular sampling structure. Amato et al. [17] investigated the application of aggregation methods to binary local features and presented a CBIR method based on Fisher kernels, Bernoulli mixture models, and CNN. The method is two times faster in extracting binary features as compared to the traditional SIFT method and can be used as an alternative to direct matching in CBIR. The information that we get from images may be insufficient to build a feature vector so Li et al. [18] suggested a re-ranking mechanism called discriminative multi-view interactive image re-ranking (DMINTIR) that integrates relevance feedback with complementary features. The feature set is encoded by utilizing neural code, VLAD+, and triangulation embedding. The proposed mechanism shuffles the images based on updated scores obtained through learned weight vector. To maximize precision, a new similarity learning method named maximum top precision similarity (MTPS) for the CBIR system is proposed [19]. The precision achieved after initial retrievals can be maximized by tuning parameters of similarity function. For that, similarity function is exhibited by hinge loss and designed as a linear function; squared Frobenius norm for each query is minimized to prevent overfitting problems. The experimental evaluation highlighted a shorter running time. Similarity measures have been evaluated in detail in [20]. The study concluded by suggesting a new matching measure by integrating relevance feedback and sequential forward selector.

Retrieving images based on regions usually results in the repetitive matching of similar regions and loss of spatial information. To overcome this issue, Meng et al. [21] presented a novel method for extracting and matching regions. Firstly, segments are identified and merged using statistical region merging and affinity propagation (SRM-AP). Instead of incorporating local descriptors, the method utilizes a CNN-based feature extraction method named as regional convolution mapping feature (RCMF) to preserve the spatial layout of the key objects of the image. Layer 5 of VGGNet19 is used as a feature layer, which outputs a 256-dimensional feature vector. For effective image representation, a number of regions and their locations are also incorporated with the RCMF method. Images are matched based on integrated category matching (ICM), which utilizes centroids rather than area or center-based methods. The method exhibits

superior performance against benchmark methods but suffers from higher dimensionality of the feature vector. Another retrieval method based on the region is presented by Song et al. [22]. In this method of CBIR, the foreground and background parts of the HSV color space image are segmented by applying the Otsu algorithm. For extracting color, the hue component is quantized into 3 bins and the saturation component is quantized into 2 bins. The intensity component (V) of HSV space is utilized to generate diagonal texture structure descriptor (DTSD), which efficiently describes the edges and preserves spatial resolution and finer details of an image. The DTSD treats an image as a 4×4 grid and computes the difference between the center and neighboring pixels. Afterward, diagonal pixels are multiplied and evaluated based on a threshold. The resultant matrix is weighted, and values are accumulated to represent diagonal texture structure. The histograms of three components of both regions combinedly form a feature vector. In terms of performance, this method surpassed many competitive methods. A hybrid method for region-based image retrieval is presented by Ahmed et al. [23], which integrates local and global features for effective image representation. In this method, interest points of the image are assembled using connected stable regions method and described using the histogram of oriented gradients. For extracting texture, uniform local binary patterns are used. The resultant higher-dimensional features are transformed into compact vectors by applying the principal component analysis (PCA) method. Experimental analysis shows improved accuracy as compared with competitive CBIR methods. Other than the semantic gap, one of the major setbacks for CBIR is edge-based object identification, which only uses edges to differentiate objects having visually similar content and spatial invariance problem, which arises because of the varied spatial position of objects within images. Pradhan et al. [24] addressed these problems by incorporating a color edge map for extracting color and shape features simultaneously and a novel image block re-ordering method based on texture direction. Initially, foreground and background regions are extracted through saliency maps. Edges from the foreground part are first extracted through a combined edge map (canny edge, fuzzy edge) and later through color edge map by accumulating the pixels into 9 groups based on orientations. For texture, the Y component of the YC_bC_r color space is divided into 24 non-overlapping blocks and rearranged using principal texture direction, which is based on the largest eigenvalue of the intensity covariance matrix. In terms of performance, this rearrangement scheme resulted in better retrieval accuracy because the objects within images became more comparable to each other irrespective of their position. The compact detail of the competitive methods of CBIR is presented in Table 1.

3 Methodology

In this section, the methodology of the proposed method is presented in detail. The proposed method adopted the BoVW model that has been one of the most dominant and frequently used methods for classifying and retrieve images. The BoVW model (as shown in Fig. 2) inherited its basic concept from the bag-of-features (BoF) model, which is particularly developed for retrieving similar documents. To get representations of visual contents of the images based on the BoVW model, images undergo following transformations: (1) firstly, local features are extracted by detecting keypoints and their corresponding descriptors are computed; (2) the extracted features are then organized

Table 1 Compact detail of competitive methods of CBIR

Method	Problem addressed	Extracted features	Dimension reduction	Clustering	Classification	Similarity measure	Shortcomings
SIFT-FREAK [25]	Semantic gap	SIFT-FREAK/color and texture features	Not applied	*k*-means	SVM	*L2*-norm	Hand-crafted features, computationally expensive
Optimized TPTSSR [26]	The computational expense of sparse classifiers, semantic gap	TPTSSR features	Not applied	Not applied	TPTSSR/NN classifier	*L2*-norm	Unification of parameters obtained through proposed strategies is unfeasible
MO-BoF [27]	Optimizing the BoF model by exploiting the manifold structure of the histogram spaces, semantic gap	HoG features	PCA	Spectral clustering	Not applied	X^2 distance	Hand-crafted features, computationally expensive
EODH-color SIFT [28]	Semantic gap	EODH features and color-SIFT descriptor	Not applied	*k*-means/weighted distribution/ unweighted distribution	Not applied	*L2*-norm	Computationally expensive
Modified VLAD [29]	Unequal contribution of residual vectors/ insufficient burst patterns because of power-law normalization, semantic gap	VLAD/local coordinate system/ residual normalization	PCA/product quantization	k-means	Not applied	*L2*-norm	Computationally expensive
Attribute features+ Fisher vectors [30]	Enhancing image retrieval by incorporating attribute features along with FV, semantic gap	Fisher vectors/ attributes/text	Random selection of attributes/ selection with cross-validation/ PCA/product quantization	Nearest neighbor search	SVM	*L2*-norm	Limited impact of semantic attributes on image retrieval
Fisher kernel-GMM [31]	The high computational expense of Fisher vectors, semantic gap	SIFT/Fisher vectors	PCA/simple binarization, local sensitivity hashing/spectral hashing	GMM	Maximum likelihood estimation	Cosine similarity	Exhaustive database search
Spatial L2 method [32]	Lack of spatial information in the BoF model, semantic gap	Dense SIFT	Not applied	*k*-means	RBF-NN/ SVM/DBN	*L2*-norm	Hand-crafted equal weights for all triangular histograms
RSHD method [33]	Semantic gap	Quantized RGB/ local neighboring structure pattern	Not applied	Not applied	Not applied	Canberra distance	Hand-crafted feature area, non-robust performance
WATH method [34]	Lack of spatial information in BoVW model, semantic gap	Dense SIFT, weighted triangular	Not applied	Hard clustering based on *k*-	SVM	*L2*-norm	Overfitting problem, computationally expensive

Table 1 Compact detail of competitive methods of CBIR *(Continued)*

Method	Problem addressed	Extracted features	Dimension reduction	Clustering	Classification	Similarity measure	Shortcomings
		histograms		means++			
Hybrid [35]	Scene categorization, semantic gap	GIST/HOG/LBP/ dense SIFT	Not applied	*k*-means	SVM	X^2 distance	A smaller overlap threshold results in overlapped annotations for objects within images.

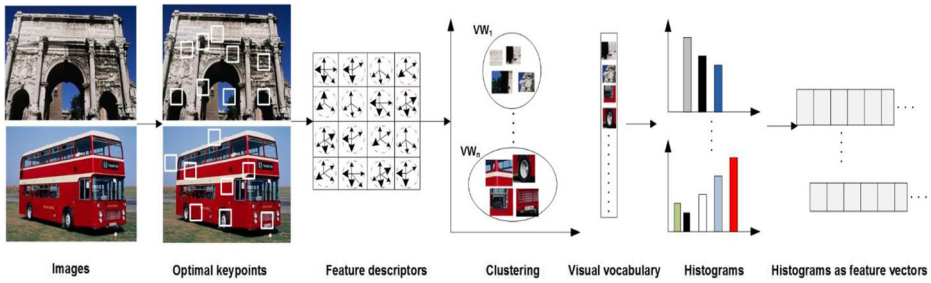

Fig. 2 The BoVW model-based visual representation of the images

into clusters by applying clustering algorithms, each cluster head then termed as a visual word which accumulates into visual vocabulary or codebook; (3) for each image, a signature is formed by representing visual words in terms of a histogram, (4) histograms are normalized to retain fine details, and (5) these signatures are then fed into the classifier for training purposes. Apart from exhibiting remarkable performance in several image retrieval applications [36–38], the BoVW model still has certain limitations that need to be addressed, i.e., lack of spatial information, extraction of redundant, and insignificant features (background regions), and most importantly, it lacks from effective, efficient feature representation and feature weighting method as some features are of greater importance than others. The proposed method of image retrieval addresses the aforementioned issues of the BoVW model to improve the performance of image retrieval.

The detail of each module of the training and testing sections of the proposed method is discussed in the following subsequent sections and its complete framework is shown in Fig. 3.

3.1 The training section of the methodology

This section presents the detail of the different modules of the proposed method, which are complementary feature extraction, adaptive feature weighting, clustering, histogram formation, and image classification. The detail of these modules is presented in the following subsequent sections.

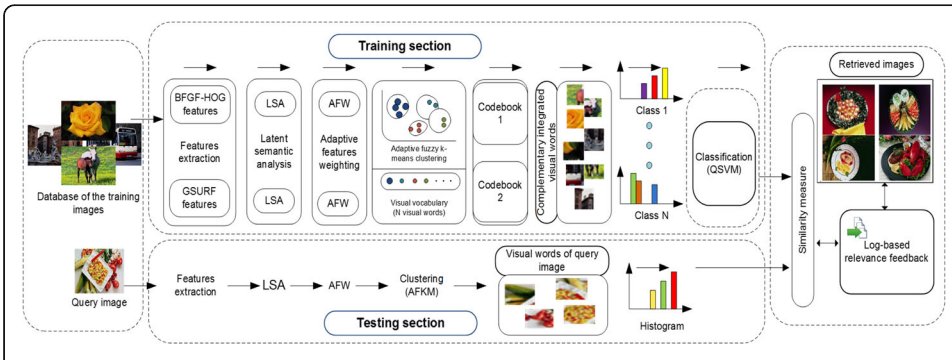

Fig. 3 Framework of the proposed adaptive complementary visual words integration based on the BoVW methodology

3.2 Feature extraction using BFGF-HOG descriptor

This step comprises extracting features from each image by using the BFGF-HOG descriptor, which is a variant of the HOG descriptor. The HOG descriptor [39] has been used widely in machine vision tasks for detecting objects within images, humans, etc. It is a window-based descriptor and works by capturing the edge directions or local intensity gradients. A window is focused on interest points and partitioned into $n \times n$ cells. For each pixel in a cell, gradient direction $\theta(x, y)$ and magnitude $M(x, y)$ are mathematically calculated as follows:

$$M(x,y) = \sqrt{\left(\partial I \Big/ \partial x\right)^2 + \left(\partial I \Big/ \partial y\right)^2} \tag{1}$$

$$\theta(x,y) = \tan^{-1} \frac{\partial I \Big/ \partial y}{\partial I \Big/ \partial x} \tag{2}$$

The computed gradient directions for each pixel are then quantized into 9 bin histogram of 45°, and the corresponding magnitudes are accumulated. The contrast of the resultant histogram is normalized to achieve illumination invariance.

Given an image $I(x, y)$, a non-iterative bilateral field (BF), which efficiently preserves edges, is applied. The bilateral filter is an alternative to low-pass filters, which reduces noise but fade edges too. To overcome this, BF computes weighted averages like low-pass filters but utilizes geometric closeness (spatial) as well as photometric information/similarity between a center pixel c and its neighboring pixels $(k - c)$ to calculate weights. Mathematically, it is expressed as follows:

$$h(c) = N^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(k) g(k,c)(I(k) - I(c))dk \tag{3}$$

$$N = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(k,c)(I(k) - I(c))dk \tag{4}$$

where N is a normalization constant, $g(k, c) = k - c$ represents geometric closeness, and $(I(k) - I(c))$ measures the similarity between the center pixel and its neighbors.

After that, feature vector of the BF-based GF-HOG feature descriptor is computed, which represents image structure as dense gradient field (GF), interpolated by neighboring sparse edge pixels. Begin with binary canny edge map I_e, edge orientations and magnitudes are calculated. Pixels having smaller magnitudes are discarded to obtain a set of sparse orientation edge pixels $S = \{\theta(x, y)_{M > t}\}$ against a certain threshold t. The gradient field G_{R^2} is dense orientation field interpolated from sparse set S. Issue of smoothness of dense gradient field is solved by the Poisson equation with Dirichlet boundary conditions. The Poisson approximates $\Delta G = 0$ by using a 3×3 Laplacian window, which results in a linear equation (Eq. (5)) with Dirichlet boundary conditions (Eq. 6).

$$G(x,y) = G(x-1,y) + G(x+1,y) + G(x,y-1) + G(x,y+1) \tag{5}$$

$$G(x,y) = \begin{cases} \theta(x,y) & f(x,y) \in S \\ 0 & f(x,y) \text{ is located on image boundaries} \end{cases} \tag{6}$$

After detecting keypoints by applying a Hessian detector on each image, a histogram of gradients (detail mentioned earlier) is then calculated over the density gradient field G and the range of orientations is quantized into m bins. The resultant vector is mn^2-dimensional vector for the entire window. A resultant feature vector of the BFGF-HOG descriptor is $64 \times J$ dimensional, where J represents a number of interest points of the features, which are automatically selected by the descriptor depending upon the contents of the image, and it is mathematically expressed as follows:

$$F_a = (a_{1d}, a_{2d}, a_{3d}, ..., a_{nd}) \tag{7}$$

where a_{1d} to a_{nd} are image descriptors of the BFGF-HOG feature vector.

3.3 Feature extraction using Gauge SURF descriptor

This step comprises extracting features by applying the Gauge SURF (GSURF) descriptor to each image. To locally adapt the blur within a region and to retain fine details or edges, GSURF [40] feature descriptor utilizes gauge coordinates. Instead of using first-order derivatives, GSURF detects keypoints from multiscale images using the determinant of the Hessian matrix. Hessian matrix is a result of convolving an integral image with second-order partial derivative Gaussian to obtain a maximum gradient. Give an image $I(x, y)$, Hessian matrix $H(z, \sigma)$ at point $z(x, y)$ and scale parameter σ are mathematically defined as follows:

$$H(z,\sigma) = \begin{bmatrix} L_{xx}(z,\sigma) & L_{xy}(z,\sigma) \\ L_{xy}(z,\sigma) & L_{yy}(z,\sigma) \end{bmatrix} \tag{8}$$

where L_{xx} is a convolution of second-order gauge derivative with image I at point z and is calculated as follows:

$$L_{xx}(z,\sigma) = I(z) * \frac{\partial^2 g(\sigma)}{\partial x^2} \tag{9}$$

and similarly $L_{yy}(z,\sigma) = I(z) * \frac{\partial^2 g(\sigma)}{\partial y^2}$ and $L_{xy}(z,\sigma) = I(z) * \frac{\partial^2 g(\sigma)}{\partial x\, \partial y}$.

The motivation behind using gauge coordinates is their ability to describe each pixel in an image by its 2D local structure. Even if an image is rotated, the structure will remain the same. Gauge coordinates comprise of a gradient vector $\overrightarrow{w}$ and its perpendicular vector $\overrightarrow{v}$, which are mathematically defined as follows:

$$\overrightarrow{w} = \left(\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}\right) = \frac{1}{\sqrt{L_x^2 + L_y^2}} \cdot (L_x, L_y) \tag{10}$$

$$\overrightarrow{v} = \left(\frac{\partial L}{\partial y}, -\frac{\partial L}{\partial x}\right) = \frac{1}{\sqrt{L_x^2 + L_y^2}} \cdot (L_y, -L_x)$$

where L denotes convolution of image I with Gaussian kernel having σ as scale parameter, i.e., $L(x, y, \sigma) = I(x, y) * g(x, y, \sigma)$.

Derivatives of any scale and order can be obtained using these coordinates. Second-order derivatives of these coordinates are of special interest and can be calculated by taking a product of 2×2 Hessian matrix with gradients in $\overrightarrow{w}$ and $\overrightarrow{v}$ directions. For building a descriptor of $64 \times J$ dimensions, first- and second-order Haar wavelet

responses in a horizontal and vertical direction are calculated over a 20×20 region, i.e., L_x, L_y, L_{xx}, L_{yy}, L_{xy}. The 20×20 window is further subdivided into 4×4 sub-blocks without any overlap and Haar wavelet of size 2σ is calculated. After fixing the gauge coordinates for each of these pixels, gauge invariants $|L_{ww}|$, $|L_{vv}|$ are computed. The parameters of the GSURF descriptor are mathematically defined as follows:

$$L_{ww} = \frac{1}{L_x^2 + L_y^2} \begin{pmatrix} L_x L_y \end{pmatrix} \begin{pmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{pmatrix} \begin{pmatrix} L_x \\ L_y \end{pmatrix} \tag{11}$$

$$L_{vv} = \frac{1}{L_x^2 + L_y^2} \begin{pmatrix} L_y - L_x \end{pmatrix} \begin{pmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{pmatrix} \begin{pmatrix} L_y \\ -L_x \end{pmatrix} \tag{12}$$

A resultant feature descriptor for each sub-region will be four-dimensional vector $V_d = (\Sigma L_{ww}, \Sigma L_{vv}, \Sigma |L_{ww}|, \Sigma |L_{vv}|)$. Resultant feature vector will be $64 \times J$ dimensional, where J represents a number of the interest points of the features that are chosen automatically by the descriptor depending upon the contents of the image, mathematically, it can be expressed as follows:

$$F_b = (b_{1d}, b_{2d}, b_{3d}, \ldots, b_{nd}) \tag{13}$$

where b_{1d} to b_{nd} are feature descriptors of the GSURF descriptor.

To detect objects within images, their location and spatial orientation of edges are of high significance. Using the HOG descriptor to extract such information results in poor performance because of difficulty in the selection of appropriate window size, as the window captures either too much or too less of local edge structure. Similarly, the standard SURF descriptor utilizes the Gaussian scale space, which incorporates blurring as a pre-processing step to remove noise. However, this step resulted in the removal of structure details such as edges. Therefore, a fusion of adaptive complementary visual words obtained through a bilateral filter (BF)-based gradient field HOG [25] and gauge SURF descriptors is proposed in this article to overcome said issues. In the next two steps, features from both the descriptors are weighted for optimal feature selection, which can reduce training time (computational cost) and improve the performance of the proposed method.

3.4 Latent semantic analysis as a dimension reduction mechanism

The feature vectors extracted in the previous steps exhibit high dimensionality, which generates issues in constructing compact feature interpretation of the image as there exist redundancy and multiple correlations among certain feature points. To get robust and discriminative features, a latent semantic analysis (LSA) method is applied to each feature vector to easily perceive and preserve data, while reducing storage and computational cost. Deerwester et al. [41] applied this method for document retrieval systems, which is based on a singular value decomposition (SVD) mechanism. The proposed method uses LSA to construct a term-context matrix A of dimension $r \times q$ for each extracted feature vector, which highlights the hidden relationship among semantically similar images. In the case of the proposed method of CBIR, each column A represents a resultant feature vector (i.e., refers to F_a (defined in Eq. (7)) in case of BFGF-HOG resultant feature vector, while it refers to F_b (defined in Eq. (13)) in case of GSURF resultant feature vector), while rows are distinct features. $A_{r \times q}$ indicates the association

between r^{th} term and q^{th} context. The key step of LSA is SVD, which decomposes the high-dimensional term-context matrix A into three matrices U, Z,and V of smaller dimensions d, represented mathematically as follows:

$$\mathrm{A} => UZV^T \tag{14}$$

where U, V are orthogonal matrices and Z is the diagonal matrix. The columns of U and V contain orthonormal eigenvectors of AA^T and A^TA,respectively, while the diagonal matrix contains singular values, which are square roots of eigenvalues from U or V. The values of the diagonal matrix Z are sorted in descending order, so the significant information can be retained by considering higher values while eliminating the lower values/noise. For dimension d, the reduced matrix can then be represented as follows:

$$A_d = U_d Z_d V_d^T \tag{15}$$

In the next step, reduced features from both descriptors are weighted for optimal feature selection.

3.5 Adaptive feature weighting based on self-paced learning

In computer vision-based applications, some features of the image are more significant than the others. The proposed method applies an adaptive feature weighting method to each reduced size feature vector to classify features as significant or insignificant based on the self-paced learning (SPL) method [42]. The SPL dynamically pick features and learn in an easy to hard learning fashion. Given a matrix of extracted LSA features $X = [X_1, X_2, X_3, \ldots, X_n]$ (where $X = > A_d$) and y as the corresponding class label, the objective function of SPL can be defined mathematically as follows:

$$\min_{\alpha} p(t) = \|y - Xt\| + \|\lambda(t)\| \tag{16}$$

where t and $\lambda(t)$ denote the representation coefficient and regularization parameter, respectively. A weight variable w is added in Eq. (16) to assign a higher or lower value of weights to each feature categorize as easy or hard. Equation (16) can then be mathematically transformed as follows:

$$\min_{w} p(t, w) = \sum_{i=1}^{n} \left((w^i)^{\frac{1}{2}} (y^i - X^i t) \right)^2 - \frac{1}{\gamma} \sum_{i=1}^{n} w^i + \lambda(t) \tag{17}$$

where γ, X^i, y^i are the learning parameters, which controls the selection of learning sample, vector of the i^{th} training feature, and i^{th} feature of a test sample, respectively. The value of l is higher for the initial learning sample, which yields smaller losses and decreases gradually when hard samples are selected. The process continues until all the samples are selected. The features are selected by setting a threshold which is mathematically described as:

$$w^i(f^i, l) = \begin{cases} 1, if\ f^i \le \dfrac{1}{l} \\ 0, if\ f^i > \dfrac{1}{l} \end{cases} \tag{18}$$

where $f^i = (y^i - X^i\alpha)^2$. In the next step, feature vectors of the adaptive feature weighting are clustered separately using an adaptive fuzzy k-means clustering algorithm, whose details are provided in the following section.

The framework of the first competitive method of non-adaptive complementary features integration method is shown in Fig. 4. While in the case of the non-adaptive complementary visual words integration method (second competitive method), all the framework is the same as shown in Fig. 3, except that it does not use the adaptive feature weighting (AFW) to analyze its image retrieval performance.

3.6 Adaptive fuzzy *k*-means clustering for complementary visual vocabulary formation

In this step, the visual vocabulary is built by applying adaptive fuzzy *k*-means (AFKM) clustering on the optimized adaptive features of BFGF-HOG and GSURF descriptors of the whole data of the training images. The AFKM clustering is an improved version of the *k*-means clustering algorithm. It is one of the frequently used unsupervised, non-deterministic, and iterative clustering algorithms. However, initialization of the cluster center, the number of clusters, sensitivity to noise, and outliers are some of the shortcomings of the standard *k*-means algorithm. To overcome these issues, the proposed method of CBIR uses the AFKM clustering algorithm [43]. It is a combination of moving *k*-means (MKM) [44] and fuzzy c-means (FCM) [45] clustering algorithms. The MKM clustering contributes to an assignment of data to its closest center and FCM allows data to belong to two or more clusters. For a point x and cluster center c, the objective function of AFKM clustering is calculated as follows:

$$F = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(E_{ij}^{m}\right)\left(x_j - c_i\right)^2 \tag{19}$$

where E_{ij}^{m} represent a fuzzy membership function and m represent a fuzziness exponent. The level of being in a specific group is inverse of the distance to clusters. The new position for each centroid is calculated as follows:

$$C_{new} = \frac{\sum_{j=1}^{n}\left(E_{ij}^{m}\right)x_j}{\sum_{j=1}^{n}\left(E_{ij}^{m}\right)} \tag{20}$$

In AFKM clustering, the concept of belongingness is introduced to improve clustering. The belongingness estimates the relationship between the cluster center and its members. The degree of belonging is calculated using the following mathematical equation:

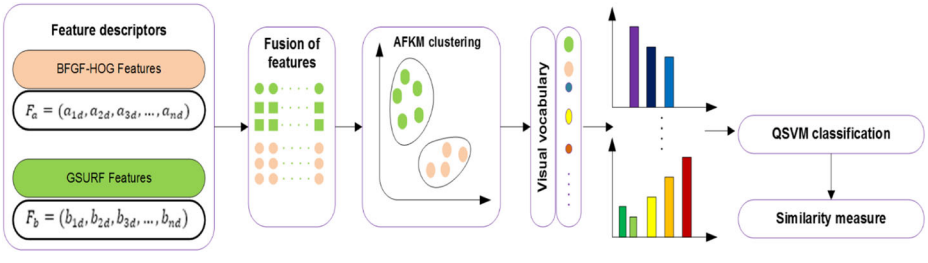

Fig. 4 Framework of the competitor method of non-adaptive complementary feature integration based on the BoVW methodology

$$B_i = \frac{C_i}{E_{ij}^m} \tag{21}$$

The proposed method of CBIR minimizes the AFKM's objective function, defined in Eq. (19). In AFKM, the clustering is iteratively performed until the center is converged and all data can be considered. In the AFKM clustering, cluster heads of the formed clusters are then termed as visual words, which are grouped to form a visual vocabulary. The proposed method of image retrieval formulates two visual vocabularies, which are represented by $W_A = \{a_1, a_2, a_3, \cdots, a_i\}$, where a_1 to a_i represent the visual words of BFGF-HOG feature vector and $W_B = \{b_1, b_2, b_3, \cdots, b_j\}$, where b_1 to b_j represent the visual words of the GSURF feature vector. After that, both visual vocabularies are concatenated vertically to form combined visual vocabulary denoted by $W_{AB} = W_A + W_B = \{W_A; W_B\}$ of size $i + j$ visual words to achieve complementary features by integrating visual words of both descriptors in the proposed method.

3.6.1 Image representation as a histogram

In this phase, salient objects of an image are transformed into a histogram, which is formed using fused visual words from the complementary visual vocabulary. Assume that the total no. of visual words in the complementary visual vocabulary (termed as W_{AB} in the previous step) are denoted by T. Consider D_j denote the number of descriptors, which are mapped to the j_{th} visual word ab_j, then the cardinality of D_j is the j_{th} bin of the histogram of visual word ab_j, which is mathematically denoted as follows:

$$ab_j = \text{Cardinality}\left(D_j\right), D_j = \left\{D_j, j\epsilon(1, \ldots, T) \;\middle|\; ab\left(D_j\right) = ab_j\right\} \tag{22}$$

The obtained histograms are then forward to a classifier for learning a model that can classify images semantically.

3.7 Image classification

In this step, the proposed method uses quadratic kernel-based SVM (QSVM) to perform image classification. The histograms of the training images along with labels of each class act as inputs to the QSVM for image classification in the proposed method. To improve retrieval efficiency and accuracy of any CBIR system, image classification is regarded as one of the vital steps. The SVM [46] is one of the frequently used classifiers and has been applied in various computer vision-based applications because of its outstanding generalization ability. Given a linear training set $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots. (x_n, y_n)\}$, where $y_{1,\ 2,\ \ldots,n} = \{1, -1\}$ are the corresponding class labels, SVM classifies linear data as defined in Eq. (23). It defines decision boundaries known as hyperplanes by focusing on data points that lie at the edges of classified class distributions, which are also known as support vectors. Mathematically, it is defined as:

$$f(x) = w^T x_i + b \tag{23}$$

where w, x, and b represent weight vector, sample point of the training set, and bias, respectively. For hyperplane to be optimal, SVM tries to (i) maximize the margin between support vectors and (ii) reduces misclassification by introducing slack variable ξas defined in Eq. (24):

$$\text{Minimize } \frac{1}{2}\|w\|^2 + R\sum_{i=1}^{n}\xi_i \tag{24}$$

$$\text{Subject to } y\left(w^T x_i + b\right) \geq 1 - \xi_i,$$

$$\xi_i \geq 0 \qquad i = 1, \cdots, n,$$

where ξ represents a misclassified sample of corresponding hyperplane and R represents a tradeoff between margin maximization and misclassification error. The higher the value of R, error reduction will be predominant, and for lower values of R, margin maximization will be emphasized. For non-linear data points, the traditional SVM algorithm fails to converge hence consumes more processing time and it also affects image retrieval accuracy. As a solution, SVM utilizes kernel functions k to map data points into new feature space also known as kernel space. The transformed equation for hyperplane is then represented as:

$$f(\mathrm{x}_i) = \sum_{m=1}^{n} y_i \alpha_i \mathrm{k}(x_m, \mathrm{x}_i) + \mathrm{b} \tag{25}$$

where $k(x_m, x_i) = \phi(x_m)$. $\phi(x_i)$ is a kernel function that uses a non-linear mapping ϕ, which maps the data points to kernel space and α_i is the Lagrange multiplier.

Among several available kernels, the proposed method uses the polynomial kernel of degree 2, also known as the quadratic kernel. It has low running costs as compared to RBF, sigmoid, and other higher-order kernels, and it also produces a robust performance of the image classification. It is mathematically represented as follows:

$$k(x_m, x_i) = (x_m . x_i + 1)^2 \tag{26}$$

3.8 Performance testing section of the methodology

As mentioned earlier, a query image from the test group of the images is selected that undergoes all the steps mentioned in the training section. The similarity between a query image and dataset images is computed using Euclidean distance. The retrieval accuracy of the proposed method is further improved by incorporating log-based RF. The details about these two modules are presented in the following subsequent sections.

3.8.1 Retrieval of the images based on the similarity measure

Given a query image q, a set of similar images are retrieved by computing relevance score between query image and images in the datasets denoted as I_{DB}. For this purpose, Euclidean distance is utilized as a measure of relevance score. Mathematically, it is defined as follows:

$$\mathrm{D}(q, \mathrm{I}_{DB}) = \sqrt{\left(\sum_{k=1}^{n} \left(I_{DB_k} - q_k\right)^2\right)^2} \tag{27}$$

3.8.2 Log-based relevance feedback

To improve the performance of image retrieval, the proposed method uses log-based relevance feedback (LRF) method for CBIR. It integrates user feedback along with low-

level features to further improve the learning process of a CBIR system. Traditional relevance feedback (RF) methods for CBIR require several iterations to return satisfactory results, which are considered time-consuming and tedious from a user perspective. In [47], an active learning approach is proposed that requires a user to label extra images retrieved by the system as most informative.

The CBIR methods based on RF have been studied immensely and the one based on RF logs is presented in [48]. The proposed method uses the LRF method, which starts with a query image (represented by q) and its corresponding retrieved images (represented by N), which are marked by a user as relevant or irrelevant. The user judgment is then saved in a history log, and a relevance matrix R is created from all log sessions. In the case of relevant, irrelevant, and non-judged images in log sessions, a cell in R is marked as +1, − 1, and 0, respectively. The LRF method aims to look for a function f_q that can map images to a relevance degree between 0 and 1.

$$f_q : I_{DB} \rightarrow [0, 1] \tag{28}$$

As LRF method utilizes low-level features (i.e., BFGF-HOG and GSURF features) and log sessions, so the overall function f_q can be defined mathematically as follows:

$$f_q(I_{DB}) = \frac{1}{2}\left(f_R(I_{DB}) + f_x(I_{DB})\right) \tag{29}$$

where f_R and f_x are relevance functions based on log-based data and low-level features of images, respectively. To find relevance between two images I_i and I_j, the correlation between log data l_i and l_j of these images is calculated, which is mathematically defined as:

$$corF_{i,j} = \sum\nolimits_k v_{k,i,j} \cdot l_{k,i} \cdot l_{k,j} \tag{30}$$

where

$$v_{k,i,j} = \begin{cases} 1, if\ l_{k,i} + l_{k,j} \geq 0 \\ 0, if\ l_{k,i} + l_{k,j} < 0 \end{cases}$$

The f_R for a log session k can be calculated using the following mathematical equation:

$$f_R(I_i) = \max\nolimits_{k \in \mathscr{L}^+} \left\{ \frac{corF_{k,i}}{\max corF_{k,j}} \right\} - \max\nolimits_{k \in \mathscr{L}^-} \left\{ \frac{corF_{k,i}}{\max corF_{k,j}} \right\} \tag{31}$$

where $corF_{k,\ i}$ is a correlation function, $\mathcal{L}^+$and $\mathcal{L}^-$denotes a set of relevant and irrelevant images, respectively.

4 Evaluation metrics, results of the experiments, and discussions

This section describes the chosen datasets, evaluation metrics, experimental results, and discussions of the proposed method. The experimental results of the proposed method are reported by performing each experiment 5 times for consistent performance. The comprehensive details of these metrics are presented in the following subsequent sections.

4.1 Evaluation metrics

To assess the performance of our proposed method, the evaluation metrics that we have used are described in detail in the subsequent sections.

4.2 Precision

The accuracy of a CBIR system in retrieving relevant images (images that belong to the same semantic class of the dataset) according to the visual contents of a query image is evaluated by precision (P), which is a ratio of images retrieved as relevant over total retrieved images. Mathematically, it is defined as follows:

$$\text{Precision} = P = \frac{\text{No.of retrieved relevant images}}{\text{No.of retrieved images}} \tag{32}$$

4.3 Average precision

Average precision (P_{avg}) computes an average of precision scores (P) of all relevant retrieved images. Mathematically, it is described as:

$$P_{avg} = \frac{1}{n}\sum\nolimits_{j=1}^{n} P(j) \tag{33}$$

where $P(j)$ represents the precision value of j^{th} iteration.

4.4 Mean average precision

The mean average precision (mAP) computes the average of P_{avg} values. Mathematically, it is expressed as follows:

$$\text{mAP} = \frac{1}{k}\sum_{i=1}^{k} P_{\text{avg}}(i) \tag{34}$$

where k represents a number of queries of the image.

4.5 Recall

The ratio of images retrieved as relevant over the number of relevant images available in the dataset is known as recall. It is defined as follows:

$$\text{Recall} = R = \frac{\text{Number of retrieved relevant images}}{\text{Number of relevant images in the dataset}} \tag{35}$$

4.6 F-measure

The overall success of an image retrieval system and its efficiency can also be assessed by utilizing F-measure, which is formalized by combining precision and recall as mentioned in the equation below:

$$\text{F–measure} = F = 2 \times \frac{(P \times R)}{(P + R)} \tag{36}$$

4.7 Datasets, experimental parameters, results, and discussions

The performance assessment of the proposed method and its competitor methods is accomplished on four standard image datasets of CBIR, which are Corel 1000, Corel

1500, Scene 15, and Holidays. The detail of these image datasets along with experimental results is presented in Section 4.2.1 to Section 4.2.4. Table 2 presents the detail of different experimental parameters, which are used to analyze the performance of the proposed method.

4.7.1 Comparative performance analysis on the Corel 1000 image dataset

The Corel-1000 [49] image dataset comprises a total of 1000 images, which are divided among 10 semantic categories, each having 100 images of resolution of 256 × 384 pixels or 384 × 256 pixels. The categories of the images included in this image dataset are buses, flowers, buildings, mountains, dinosaurs, human beings, food, landscape, elephants, and horses. Figure 5 presents the sample images, which are taken from each semantic category of the Corel 1000 image dataset.

The experimental results of the non-adaptive complementary feature integration (first competitor method), non-adaptive complementary visual words integration (second competitor method), and the proposed adaptive complementary visual words integration methods using different sizes of the visual vocabulary are presented in Figs. 6, 10, 14, and 16. After the analysis of experimental facts presented in these figures, it can be deduced that the proposed system that is based upon adaptive complementary visual word integration produces robust performance in contrast to its competitor methods of CBIR for all the specified datasets. The size of the visual vocabulary, which produces the best performance of the proposed method, is 600 visual words and achieved mAP performance on this visual vocabulary size is 89.91% for the Corel 1000 dataset. Tables 3, 4, 5, and 6 present the performance comparison of the proposed method with its state-of-the-art image retrieval methods. It can be concluded from experimental results that the proposed method gives promising results as compared to its competitor CBIR methods due to the following reasons: (a) firstly, it uses complementary visual feature representation for salient contents of the images; (b) it uses adaptive feature weighting method based on self-paced learning to select optimized features for each image; (c) it uses twice size complementary visual words to represent salient contents of each image; (d) it uses quadratic kernel-based SVM (QSVM) to achieve robust image classification results, which ultimately improve the similarity measure process in the proposed method of CBIR; and (e) lastly, the proposed method uses log-based relevance feedback

Table 2 Detail of experimental parameters of the proposed method

Datasets/visual vocabularies	Percentage of the images selected for training	Percentage of the images selected for testing
Corel 1000	70%	30%
Corel 1500	50%	50%
Scene 15	60%	40%
Holidays	60%	40%
Size of the BFGF-HOG visual vocabularies	10, 25, 50, 100, 200, 300, 400, 500, 600 (unit is visual word)	
Size of the GSURF visual vocabularies	10, 25, 50, 100, 200, 300, 400, 500, 600 (unit is visual word)	
Size of complementary visual vocabulary	20, 50, 100, 200, 400, 600, 800, 1000, 1200 (unit is visual word)	

Fig. 5 Sample images (one per category) of the Corel 1000 image dataset

(LRF) mechanism for CBIR, which integrates user feedback along with low-level complementary features to further improve the learning process of a CBIR system.

Figures 7 and 8 show the results of the image retrieval according to the salient objects of the query images. The query image (first row) of Figs. 7 and 8 are taken from the "Dinosaurs" and "Horses" categories of the Corel 1000 dataset, respectively. Furthermore, Fig. 7 shows the result of LRF-0 image retrieval. The integer value with LRF shows the iteration of the feedback. The images shown in Fig. 8 are the result of the image retrieval after applying LRF-1, which are semantically more relevant to the query image as compared to the LRF-0 retrieval result of the query image.

4.8 Comparative performance analysis on the Corel 1500 image dataset

The Corel 1500 [49] image dataset is a subset of the WANG image dataset. The images in the Corel 1500 image dataset are ordered into 15 semantic categories, and each

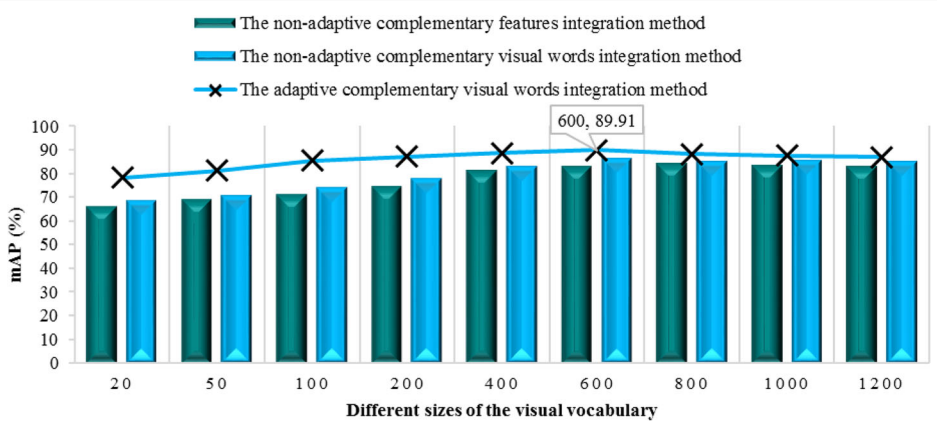

Fig. 6 Comparison of mAP of the proposed method with competitive CBIR methods on the Corel 1000 image dataset

Table 3 Comparative analysis of competitive methods with a proposed method on the Corel 1000 dataset

Semantic category		Color-shape-texture [24]	CNN-RCMF-ICM [21]	DTSD-MR [22]	Region-texture [23]	LBPC-LBPH-CH [2]	MTBCD [13]	Proposed method
African tribes	*P*	85.00	66.30	64.33	81.00	61.52	66.00	81.36
	R	17.00	13.26	06.43	16.00	12.30	13.20	16.27
	F	28.33	22.10	11.69	26.72	20.50	22.00	27.12
Beaches	*P*	75.00	52.70	28.82	75.00	43.36	64.00	78.74
	R	15.00	10.54	02.88	15.00	08.67	12.80	15.74
	F	25.00	17.57	05.24	25.00	14.45	21.33	26.24
Building	*P*	60.00	70.60	54.23	67.00	54.57	82.00	82.93
	R	12.00	14.12	05.42	13.00	10.91	16.40	16.58
	F	20.00	23.53	09.86	21.78	18.18	27.33	27.64
Buses	*P*	100.0	99.10	79.43	93.00	87.90	100.0	89.56
	R	20.00	19.82	07.94	19.00	17.58	20.00	17.91
	F	33.33	33.03	14.44	31.55	29.30	33.33	29.85
Dinosaurs	*P*	100.0	100.0	99.59	97.00	98.88	100.0	100.0
	R	20.00	20.00	09.96	19.00	19.78	20.00	20.00
	F	33.33	33.33	18.11	31.78	32.97	33.33	33.33
Elephants	*P*	80.00	90.80	40.09	91.00	44.54	83.00	93.09
	R	16.00	18.16	04.01	18.00	08.91	16.60	18.61
	F	26.67	30.27	07.29	30.06	14.85	27.67	31.03
Flowers	*P*	100.0	99.20	78.25	90.00	83.08	94.00	97.11
	R	20.00	19.84	07.83	18.00	16.62	18.80	19.42
	F	33.33	33.07	14.24	30.00	27.70	31.33	32.37
Horses	*P*	90.00	95.10	71.17	84.00	81.59	99.00	96.35
	R	18.00	19.02	07.12	17.00	16.32	19.80	19.27
	F	30.00	31.70	12.94	28.28	27.20	33.00	32.11
Mountain	*P*	85.00	71.50	37.55	59.00	39.89	67.00	86.73
	R	17.00	14.30	03.76	12.00	07.98	13.40	17.34
	F	28.33	23.83	06.84	19.94	13.30	22.33	28.91
Foods	*P*	85.00	85.30	58.40	79.00	56.31	95.00	93.24
	R	17.00	17.06	05.84	16.00	11.26	19.00	18.64
	F	28.33	28.43	10.62	26.61	18.77	31.67	31.08
mAP (%)	*P*	86.00	83.06	61.19	81.60	65.16	85.00	89.91
Avg. *R*	*R*	17.20	16.61	06.12	16.30	13.03	17.00	17.98
Avg. *F*	*F*	28.66	27.68	11.12	27.17	21.72	28.33	29.97

Table 4 Comparative analysis of competitive methods with a proposed method on the Corel 1500 dataset

Performance parameters	Proposed method	MTBCD [13]	SQ+Spatiogram [15]	GMM-mSpatiogram [15]	SIFT-FREAK [50]
mAP (%)	83.99	82.16	63.95	74.10	72.60
Avg. recall (*R*)	16.79	16.43	12.79	13.80	14.52
Avg. F-measure (*F*)	27.98	27.38	21.32	23.26	24.20

Table 5 Comparative analysis of competitive methods with the proposed method on the Scene 15 dataset

Performance parameters	Proposed method	MTBCD [13]	Optimized TPTSSR [26]	MO-BoF [27]	Hybrid [35]	EODH-color SIFT [28]
mAP (%)	83.11	80.88	60.20	36.57	81.00	81.10
Avg. recall (*R*)	16.62	16.17	12.04	07.314	16.20	16.22
Avg. F-measure (*F*)	27.70	26.95	20.06	12.19	27	27.03

contains a total of 100 images. The image resolution in this image dataset is either 256×384 pixels or 384×256 pixels. The sample images from each semantic category of the Corel 1500 image dataset are shown in Fig. 9.

By varying different sizes of the visual vocabulary, the mAP performance of the proposed method, and its comparison with competitor methods, is presented in Fig. 10. After analyzing experimental details, it can be deduced that the proposed method outperforms as compared to its competitor methods on the Corel 1500 image dataset. The best mAP performance of the proposed method is obtained on a visual vocabulary of size 1000 visual words, which is 83.99%. Table 4 presents the performance comparison of the proposed method against competitive methods in terms of performance evaluation metrics of the CBIR. Based on the experimental details shown in Table 4, it can also be concluded that the proposed method also outperforms its comparative methods due to the factors mentioned in Section 4.2.1. The results of the image retrieval using the proposed method according to the salient objects of the query images of the Corel 1500 image dataset are shown in Figs. 11 and 12 for the semantic categories "Sunset" and "Postcard," respectively.

4.9 Comparative performance analysis on the Scene 15 image dataset

The Scene 15 dataset [51] comprises of 4485 gray-scale images, divided into 15 scene categories. This dataset contains images of indoor as well as outdoor scenes. There are 200 to 400 images in each semantic class of this dataset, and the resolution of each image is 300×250 pixels. Figure 13 shows different sample images from each semantic class of the Scene 15 image dataset.

Figure 14 shows the performance comparison of the proposed method with its competitor methods in terms of the mAP performance on different sizes of the visual vocabulary. On the Scene 15 image dataset, the best mAP performance of the proposed method against its competitor CBIR methods is attained on a visual vocabulary of size 1000 visual words, which is 83.11%. To further analyze the robustness of the proposed method, its performance comparison is performed with state-of-the-art CBIR methods

Table 6 Comparative analysis of competitive methods with a proposed method on the Holidays dataset

Performance parameters	Proposed method	MTBCD [13]	BMM-FV CNN [17]	Modified VLAD [29]	Att. features+Fisher vectors [30]	Fisher kernel-GMM [31]
mAP (%)	72.85	67.09	54.70	65.80	69.90	70.50
Avg. recall (*R*)	14.57	13.41	10.94	13.16	13.98	14.10
Avg. F-measure (*F*)	24.28	22.35	18.23	21.93	23.30	23.50

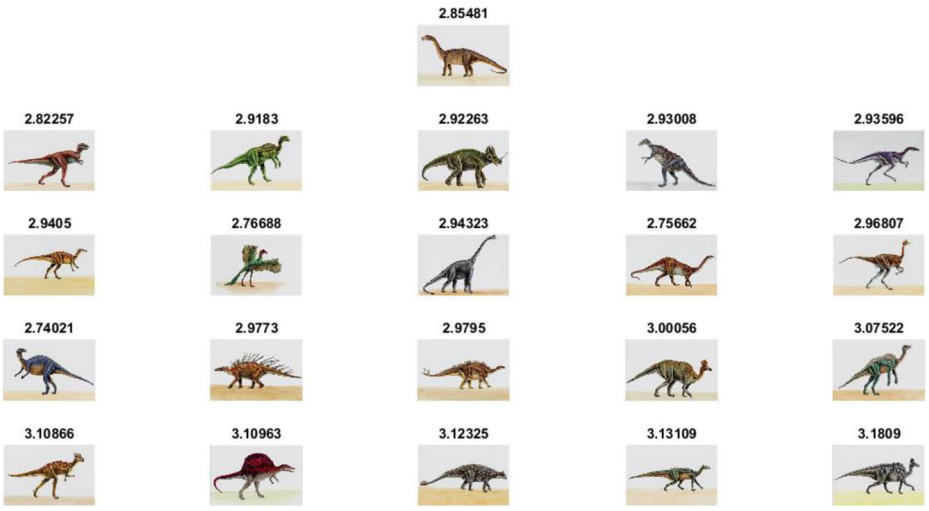

Fig. 7 Semantic category "Dinosaurs" of the Corel 1000 dataset shows retrieved images as a response of the query image (first row) (LRF-0)

in terms of standard performance evaluation metrics, whose details are presented in Table 5 for the Scene 15 image dataset. Different factors of the proposed method such as robust complementary image representation, efficient and effective adaptive feature weighting of visual words, twice size visual words for key objects of the image result in the robust performance of the proposed method as compared to its competitor CBIR methods.

4.10 Comparative performance analysis on the Holidays image dataset

The Holidays image dataset [52] contains 1491 images, out of which, 500 images are the query images and the remaining 991 are corresponding relevant images that are

Fig. 8 Semantic category "Horses" of the Corel 1000 dataset shows retrieved images as a response of the query image (first row) (LRF-1)

Fig. 9 Sample images (one per category) of the Corel 1500 image dataset [49]

classified into distinct semantic groups. Each semantic group of the images represents a distinct scene having various transformations on the images such as rotation, blurring, and viewpoint. The resolution of the image in this dataset is 2448 × 3204 pixels. The different sample images of the Holidays image dataset are shown in Fig. 15.

The experimental details and comparative analysis of the effect of varying different sizes of visual vocabulary on mAP performance of the proposed method with its competitor methods are presented in Fig. 16 for the Holidays image dataset. The proposed method produces the best mAP performance of 72.85% on the visual vocabulary of size 800 visual words against its competitive methods of CBIR. The second competitor method of non-adaptive complementary visual words integration of CBIR produces best mAP performance of 62.53% on a visual vocabulary of size 600 visual words as compared to its other reported sizes of the visual vocabulary. Similarly, the best mAP performance produces by the first competitor method of non-adaptive complementary features integration method is 57.14%, which is attained on a visual vocabulary size of 600 visual words as compared to its other reported sizes of the visual vocabulary on the Holidays image dataset. The performance comparison of the proposed method with state-of-the-art CBIR methods is presented in Table 6, which concludes that the

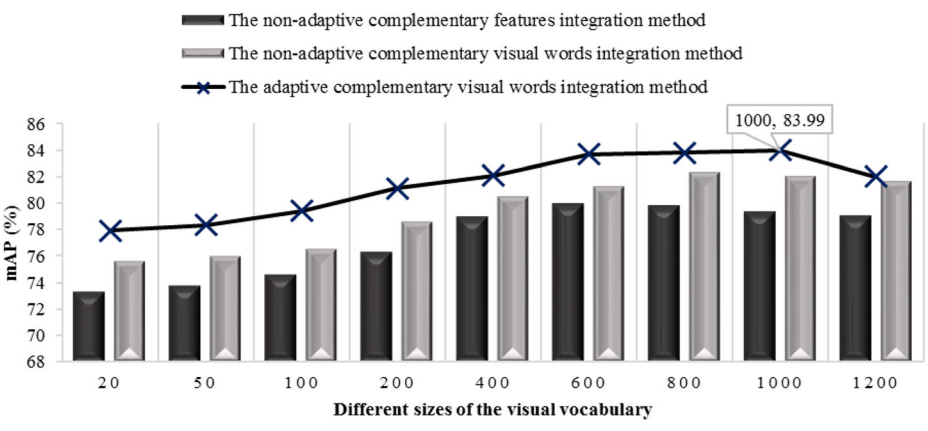

Fig. 10 Comparison of mAP of the proposed method with competitive CBIR methods on the Corel 1500 image dataset

Fig. 11 Semantic category "Sunset" of the Corel 1500 dataset shows retrieved images as a response of the query image (first row) (LRF-0)

proposed method produces robust performance as compared to recent CBIR methods in terms of performance evaluation metrics.

4.11 Required hardware/software resources and computational cost

The performance of the proposed method in terms of computational cost is measured using a desktop PC having following hardware and software requirements: Intel(R) Core(TM)-i3 CPU (frequency 2.1 GHz-series 2310 M), 8 GB of RAM, 120 GB SSD, Windows 7 Professional (64-bit), and MATLAB (2015b-x64 bit). The computational cost of the proposed method based on adaptive complementary visual words integration and its comparison with other competitive CBIR methods are presented in Table 7 for the Corel 1000 image dataset.

Fig. 12 Semantic category "Postcard" of the Corel 1500 dataset shows retrieved images as a response of the query image (first row) (LRF-0)

Fig. 13 Indoor and outdoor scenes of the sample images taken from the Scene 15 image dataset

5 Conclusion and future work

In this article, we explored the effect of adaptive feature weighting and adaptive fuzzy *k*-means clustering on the robust representation of the principal objects of the images by integrating complementary visual words of the local and global features based on the BoVW methodology. The latent semantic analysis is applied to the adaptive feature weighting to reduce the computational complexity of the proposed method, which is slightly increased due to the integration of the complementary visual words. The classification accuracy of the proposed method is improved using quadratic kernel-based SVM, which ultimately improved the similarity measure process of the CBIR. The log-based relevance feedback mechanism is also introduced in the proposed method to further improve the performance of the CBIR. The performance comparison of the proposed adaptive complementary visual words integration method is carried with a non-adaptive complementary feature integration method and non-adaptive complementary visual words integration method using the same local and global features as well as with state-of-the-art CBIR methods. It can be concluded that the integration of adaptive complementary visual words significantly improved the performance of the CBIR

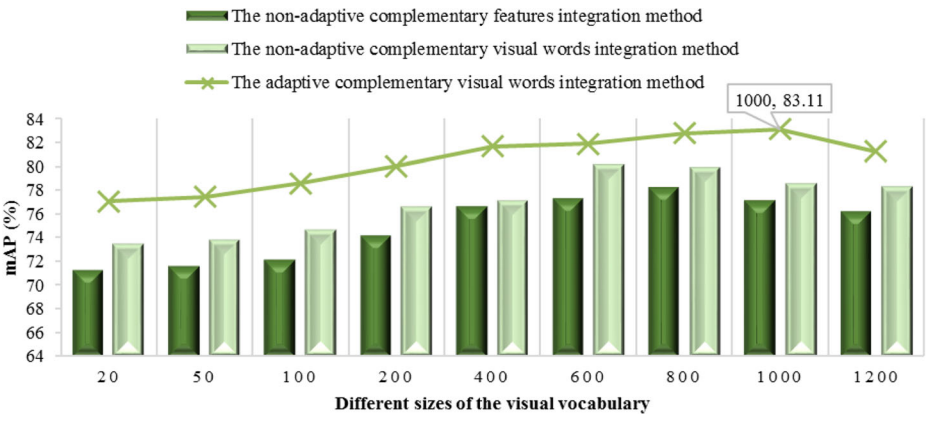

Fig. 14 Comparison of mAP of the proposed method with competitive CBIR methods on the Scene 15 image dataset

Fig. 15 Eight sample images (one per category) of the Holidays image dataset

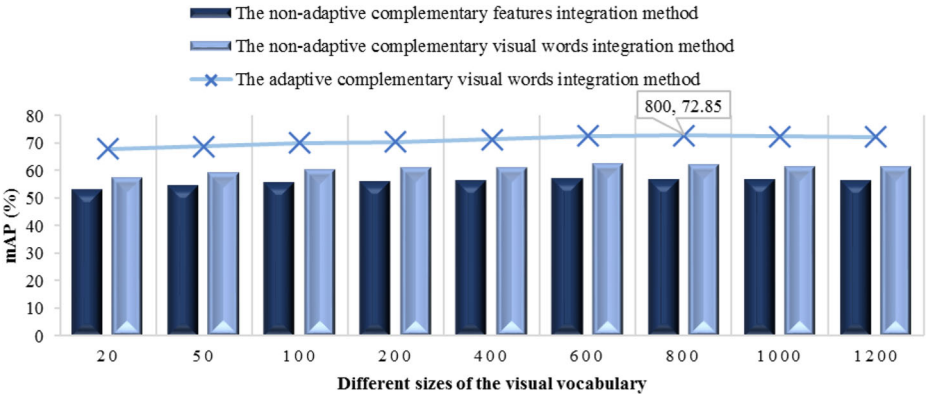

Fig. 16 Comparison of mAP of the proposed method with competitive CBIR methods on the Holidays image dataset

Table 7 Computational time (in seconds) of the proposed method as compared to competitive CBIR methods

Proposed method	ATR+SOFT method [9]	EODH method [28]	Spatial L2 method [32]	RSHD method [33]	WATH method [34]
0.06874	0.0788	5.6000	0.0821	0.3750	0.0745

as compared with the integration of non-adaptive complementary features and non-adaptive complementary visual words integration methods due to the assignment of twice size visual words to the salient objects of the images. In the future work, due to the radically increasing volume of the image and video databases, the performance of the proposed method can be analyzed using normalized discriminative deep learning-based compressed domain methods like JPEG-2000 and HEVC to improve the accuracy and efficiency of content-based video and image retrieval systems.

Abbreviations

CBIR: Content-based image retrieval; BoVW: Bag-of-visual-words; LRF: Log-based relevance feedback; QSVM: Quadratic kernel-based support vector machine; AFKM: Adaptive fuzzy *k*-means; SPL: Self-paced learning; BoF: Bag-of-features; LSA: Latent semantic analysis; HoG: Histogram of oriented gradient; LBPC: Local binary pattern for color images; CH: Color histogram; BF: Bilateral field; GF: Gradient field; LTrP: Local tetra pattern; GSURF: Gauge speeded-up robust features; CEDD: Color and edge directivity descriptor; BSIF: Binarized statistical image features; LDEM: Local directional edge map; DBSCAN: Density-based spatial clustering of applications with noise; DCD: Dominant color descriptor; PCA: Principal component analysis; FCM: Fuzzy c-means; VLAD: Vector of locally aggregated descriptors; SVD: Singular value decomposition; mAP: Mean average precision; AI: Artificial intelligence; HCI: Human–computer interaction

Acknowledgements

Not applicable.

Authors' contributions

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

Funding

Not applicable.

Availability of data and materials

Data sharing is not applicable to this article as the authors have used publicly available datasets, whose details are included in the "experimental results and discussions" section of this article. Please contact the authors for further requests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan. [2]Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan. [3]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan. [4]AIDA Lab, CCIS, Prince Sultan University, Riyadh 11586, Saudi Arabia. [5]Department of Computer Engineering, Umm Al-Qura University, Makkah 21421, Saudi Arabia.

Received: 25 November 2018 Accepted: 21 June 2020
Published online: 06 July 2020

References

1. M. Alkhawlani, M. Elmogy, H. El Bakry, Text-based, content-based, and semantic-based image retrievals: a survey. Int. J. Comput. Inf. Technol **4**(01), 58–66 (2015)
2. C. Singh, E. Walia, K.P. Kaur, Color texture description with novel local binary patterns for effective image retrieval. Pattern Recogn. **76**, 50–68 (2018)
3. A. Talib, M. Mahmuddin, H. Husni, L.E. George, A weighted dominant color descriptor for content-based image retrieval. J. Vis. Commun. Image Represent. **24**(3), 345–360 (2013)
4. A.T. Da Silva, A.X. Falcão, L.P. Magalhães, Active learning paradigms for CBIR systems based on optimum-path forest classification. Pattern Recogn. **44**(12), 2971–2978 (2011)
5. S. Murala, Q.J. Wu, Expert content-based image retrieval system using robust local patterns. J. Vis. Commun. Image Represent. **25**(6), 1324–1334 (2014)
6. M. Subrahmanyam, R. Maheshwari, R. Balasubramanian, Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking. Signal Process. **92**(6), 1467–1479 (2012)
7. R.S. Dubey, R. Choubey, J. Bhattacharjee, Multi feature content based image retrieval. Int. J. Comput. Sci. Eng. **2**(6), 2145–2149 (2010)
8. Krizhevsky, A., I. Sutskever, and G.E. Hinton, Imagenet classification with deep convolutional neural networks. ed. Advances in neural information processing systems, 2012, p. 1097-1105.

9. Z. Mehmood, N. Gul, M. Altaf, T. Mahmood, T. Saba, A. Rehman, M.T. Mahmood, Scene search based on the adapted triangular regions and soft clustering to improve the effectiveness of the visual-bag-of-words model. EURASIP Journal on Image and Video Processing **2018**(1), 48 (2018)
10. Misale, S. and A. Mulla, Learning visual words for content based image retrieval. ed. 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018, p. 580-585.
11. Z. Mehmood, S. Anwar, M. Altaf, N. Ali, A novel image retrieval based on rectangular spatial histograms of visual words. Kuwait Journal of Science **45**(1), 54–69 (2018)
12. Z. Mehmood, S.M. Anwar, N. Ali, H.A. Habib, M. Rashid, A novel image retrieval based on a combination of local and global histograms of visual words. Math. Probl. Eng. **2016**, 1–12 (2016)
13. L. Yu, L. Feng, H. Wang, L. Li, Y. Liu, S. Liu, Multi-trend binary code descriptor: a novel local texture feature descriptor for image retrieval. SIViP **12**(2), 247–254 (2018)
14. Mistry, Y., D. Ingole, and M. Ingole, Content based image retrieval using hybrid features and various distance metric. Journal of Electrical Systems and Information Technology, 2017.
15. S. Zeng, R. Huang, H. Wang, Z. Kang, Image retrieval using spatiograms of colors quantized by Gaussian Mixture Models. Neurocomputing **171**, 673–684 (2016)
16. S.K. Roy, B. Chanda, B.B. Chaudhuri, S. Banerjee, D.K. Ghosh, S.R. Dubey, Local directional ZigZag pattern: a rotation invariant descriptor for texture classification. Pattern Recogn. Lett. **108**, 23–30 (2018)
17. G. Amato, F. Falchi, L. Vadicamo, Aggregating binary local descriptors for image retrieval. Multimed. Tools Appl. **77**(5), 5385–5415 (2018)
18. J. Li, C. Xu, W. Yang, C. Sun, D. Tao, Discriminative multi-view interactive image re-ranking. IEEE Trans. Image Process. **26**(7), 3113–3127 (2017)
19. Liang, R.-Z., L. Shi, H. Wang, J. Meng, J.J.-Y. Wang, Q. Sun, and Y. Gu, Optimizing top precision performance measure of content-based image retrieval by learning similarity function. ed. Pattern Recognition (ICPR), 2016 23rd International Conference on, 2016, p. 2954-2958.
20. M. Mosbah, B. Boucheham, Distance selection based on relevance feedback in the context of CBIR using the SFS meta-heuristic with one round. Egyptian Informatics Journal **18**(1), 1–9 (2017)
21. F. Meng, D. Shan, R. Shi, Y. Song, B. Guo, W. Cai, Merged region based image retrieval. J. Vis. Commun. Image Represent. **55**, 572–585 (2018)
22. W. Song, Y. Zhang, F. Liu, Z. Chai, F. Ding, X. Qian, S.C. Park, Taking advantage of multi-regions-based diagonal texture structure descriptor for image retrieval. Expert Syst. Appl. **96**, 347–357 (2018)
23. K.T. Ahmed, M.A. Iqbal, Region and texture based effective image extraction. Clust. Comput. **21**(1), 493–502 (2018)
24. J. Pradhan, A.K. Pal, H. Banka, Principal texture direction based block level image reordering and use of color edge features for application of object based image retrieval. Multimed. Tools Appl. **78**(2), 1685–1717 (2019)
25. Hu, R., M. Barnard, and J. Collomosse, Gradient field descriptor for sketch based retrieval and localization. ed. 2010 IEEE International Conference on Image Processing, 2010, p. 1025-1028.
26. F. Dornaika, Y. El Traboulsi, Proposals for local basis selection for the sparse representation-based classifier. SIViP **12**(8), 1595–1601 (2018)
27. N. Passalis, A. Tefas, Information clustering using manifold-based optimization of the bag-of-features representation. IEEE transactions on cybernetics **48**(1), 52–63 (2016)
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. Delhumeau, J., P.-H. Gosselin, H. Jégou, and P. Pérez, Revisiting the VLAD image representation. ed. Proceedings of the 21st ACM international conference on Multimedia, 2013, p. 653-656.
30. Douze, M., A. Ramisa, and C. Schmid, Combining attributes and fisher vectors for efficient image retrieval. ed. CVPR 2011, 2011, p. 745-752.
31. Perronnin, F., Y. Liu, J. Sánchez, and H. Poirier, Large-scale image retrieval with compressed fisher vectors. ed. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, p. 3384-3391.
32. N. Ali, K.B. 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)
33. S.R. Dubey, S.K. Singh, R.K. Singh, Rotation and scale invariant hybrid image descriptor and retrieval. Comput. Electr. Eng. **46**, 288–302 (2015)
34. Z. Mehmood, T. Mahmood, M.A. 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)
35. Xiao, J., J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo. ed. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, p. 3485-3492.
36. Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, L. Zhang, Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geosci. Remote Sens. Lett. **13**(6), 747–751 (2016)
37. S. Zhang, Q. Tian, G. Hua, Q. Huang, W. Gao, Generating descriptive visual words and visual phrases for large-scale image applications. IEEE Trans. Image Process. **20**(9), 2664–2677 (2011)
38. S. Xu, T. Fang, D. Li, S. Wang, Object classification of aerial images with bag-of-visual words. IEEE Geosci. Remote Sens. Lett. **7**(2), 366–370 (2009)
39. Dalal, N. and B. Triggs, Histograms of oriented gradients for human detection. ed., 2005, p.
40. P.F. Alcantarilla, L.M. Bergasa, A.J. Davison, Gauge-SURF descriptors. Image Vis. Comput. **31**(1), 103–116 (2013)
41. S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
42. Kumar, M.P., B. Packer, and D. Koller, Self-paced learning for latent variable models. ed. Advances in Neural Information Processing Systems, 2010, p. 1189-1197.
43. S.N. Sulaiman, N.A.M. Isa, Adaptive fuzzy-K-means clustering algorithm for image segmentation. IEEE Trans. Consum. Electron. **56**(4), 2661–2668 (2010)
44. M.Y. Mashor, Hybrid training algorithm for RBF network. International Journal of the computer, the Internet and Management **8**(2), 50–65 (2000)

45. Dunn, J.C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. 1973.
46. Boser, B.E., I.M. Guyon, and V.N. Vapnik, A training algorithm for optimal margin classifiers. ed. Proceedings of the fifth annual workshop on Computational learning theory, 1992, p. 144-152.
47. M. Wang, X.-S. Hua, Active learning in multimedia annotation and retrieval: A survey. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(2), 10 (2011)
48. S.C. Hoi, M.R. Lyu, R. Jin, A unified log-based relevance feedback scheme for image retrieval. IEEE Trans. Knowl. Data Eng. **18**(4), 509–524 (2006)
49. J. Li, J.Z. Wang, Real-time computerized annotation of pictures. IEEE Trans. Pattern Anal. Mach. Intell. **30**(6), 985–1002 (2008)
50. Afifi, A.J. and W.M. Ashour, Content-based image retrieval using invariant color and texture features. ed. 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), 2012, p. 1-6.
51. 15 Scene. 2019 [cited 2019 19 June]; Available from: https://figshare.com/articles/15-Scene_Image_Dataset/7007177.
52. Holidays. 2019 [cited 2019 19 June]; Available from: http://lear.inrialpes.fr/people/jegou/data.php#holidays.

Publisher's Note

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

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com