Int J Multimed Info Retr (2016) 5:237–253
DOI 10.1007/s13735-016-0109-4

REGULAR PAPER

Improving content-based image retrieval with compact global and local multi-features

Ahmad Alzu'bi[1] · **Abbes Amira**[1] · **Naeem Ramzan**[1] · **Tareq Jaber**[2]

Received: 21 June 2016 / Revised: 18 August 2016 / Accepted: 7 September 2016 / Published online: 27 September 2016

Abstract The accuracy of content-based image retrieval (CBIR) systems is significantly affected by the discriminatory power of image features and distance measures. This paper performs an investigation towards finding the best local and global features and distance measures for content-based image retrieval. It provides insights into the trade-offs regarding computational costs, memory utilization and accuracy on several standard datasets which include MIRFLICKR, Corel, Holidays and ZuBuD. First, low-dimensional global and local features are extracted individually to generate a bank of small image features. Second, multilevel descriptor forms are utilized to produce highly discriminative image representations based on multi-features aggregation scheme. The relationship is highlighted between features (local and global) and other retrieval factors such as quantization approaches, visual codebooks, distance measures, vectorization methods, memory and retrieval speed. The resulting composite image representations are compact, i.e., only 32–64 vector dimension and 32–128 codebook size, and preserve high discriminative levels which further boost the retrieval accuracy and performance. The experimental results show that the presented multi-features image representations are efficient and outperform many competitive methods of the state-of-the-art.

Keywords Content-based image retrieval · Multi-feature analysis · Local detectors · Compact quantization · Similarity matching · Principle component analysis

✉ Ahmad Alzu'bi
ahmad.alzubi@uws.ac.uk

Abbes Amira
abbes.amira@uws.ac.uk

Naeem Ramzan
naeem.ramzan@uws.ac.uk

Tareq Jaber
tjaber1@uj.edu.sa

[1] School of Engineering and Computing, University of the West of Scotland, Paisley PA1 2BE, UK

[2] Faculty of Computing and Information Technology, University of Jeddah, Jeddah, Saudi Arabia

1 Introduction

Content-based image retrieval (CBIR) is one of the fast-advancing research areas in computer vision [1–4]. Feature extraction is a crucial part in the retrieval process and computer vision applications. The aim of feature extraction is to extract and formulate a meaningful and discriminative image representation to return the most similar and relevant images to query images. The majority of the works presented in typical CBIR systems/approaches are based on low-level image features [5–17], including color, texture, shape, and spatial information. Global features usually describe the whole visual contents of the image instead of only considering certain points of interests.

On the other hand, local image descriptors, e.g., scale-invariant feature transform (SIFT) [18], speeded-up robust features (SURF) [19], and histograms of oriented gradients (HOG) [20], have recently gained a great attention from the research community in computer vision. Local image descriptors generally describe local information using key points of some image parts such as region, object of interest, edges, or corners. Recently, local descriptors have shown their superiority in a range of computer vision applications, e.g., scene categorization, panoramic stitching, visual object classification, object tracking, and image retrieval. Local descriptors have many advantages over traditional global features due to their invariance to image scale and rotation, and

they provide a robust matching across a wide range of different situations [18].

As an integral part of this work, exploiting the benefits of both global and local image features, to strengthen the robustness and discrimination of image representation, is an interesting challenge. On one hand, local descriptors improve the system robustness against scale variance and many image deformations, e.g., noise, rotation, and viewpoint changes. On the other hand, global features consider the whole image structure, which is close to the human vision characteristics, including objects and spatial relations. Thus, it is an indispensable demand to extract proper features to gain high retrieval accuracy. Furthermore, the dimension of feature vector should be carefully considered as it substantially affects the CBIR system's performance in terms of memory and computation cost. Since this work targets generating very low-dimensional features and quantization, high-dimensionality reduction scheme is applied based on the principal component analysis (PCA)/whitening approach.

Basically, increasing the dimension of feature vectors means extracting more information from images. But, does that provide more distinctiveness on image representation? Will that improve the retrieval accuracy of CBIR system? And, importantly, how will different image features perform and relate with other retrieval aspects, e.g., vector size, quantization, and similarity measures/metrics? Exploring the relationship between these factors is one of the core parts of this paper. Distance measures, used in finding the similarity/dissimilarity between query images and all dataset images, affect the retrieval accuracy and image ranking. However, the majority of typical CBIR methods in the literature utilize one distance measure in a particular context. Accordingly, there is no sufficient evaluation on how different distance measures act under different retrieval settings, and its impact on the retrieval accuracy and system performance. This work employs many commonly used distance measures in the CBIR baseline system to enable a strong base for model optimization.

In this paper, a CBIR scheme is developed based on a multi-feature analysis to generate compact image signatures. The baseline CBIR evaluation is adopted using a multilevel retrieval scheme where each level is fed by the result obtained from the previous ones. The extracted features are formed into small-size image vectors including local descriptors. The system performance is also evaluated at all stages of development and optimization, which is observed under various POI detectors, similarity distances, vector dimensions, and quantization approaches. Our contribution presented in this paper is threefold:

- Employing a multilevel evaluation scheme based on image multi-feature analysis towards finding the best compact and discriminative local and global image representations for the CBIR task. First, we extract low-dimensional global features (color, texture, multiresolution, and local patterns) and local descriptors using different points of interest detectors, e.g., features from accelerated segment test (FAST) [21] and Harris. Second, best discriminative features are identified and passed to the next level of the CBIR system. Finally, the resulting image signatures are formed based on a robust combination between the extracted features to generate further discriminative image representations;
- investigating the relationship between image representation and other retrieval factors with trade-off analysis, including quantization approaches, various visual codebooks, similarity measures, retrieval speed, and memory usage. We provide an insight on when composite features can be expected to work efficiently, and how the benefit of using different vectorization setups can be exploited; and
- adapting the CBIR model to handle different queries based on a proper weighting of both image representations and distance measures. New insights are also provided into using particular distance measures such as Spearman for CBIR with thorough comparisons made between distance measures. The resulting CBIR model is evaluated on several standard image datasets.

The remaining part of this paper is organized as follows: Sect. 2 introduces the related works to this paper along with an overview on global and local image features; Sect. 3 illustrates the baseline framework of CBIR model and the methodology adopted for image retrieval, including feature extraction, image dataset, and evaluation protocol; Sect. 4 presents thorough experiments and results obtained by the baseline CBIR system; Sect. 5 demonstrates the procedure and results of compact quantization based on various retrieval factors utilized; Sect. 6 presents comparisons with the literature; and Sect. 7 concludes this paper.

2 Related work

This section introduces the most common global and local features used in the literature, main quantization approaches, and the most related works to this paper with certain limitations to be addressed.

Color feature is one of the most extensive vision characteristics due to its close relation with image objects, foregrounds, and backgrounds. Moreover, it does not depend on the state of image contents, e.g., direction, size and angle. The popular color representations are color histogram (e.g., linear color spaces such as RGB, XYZ, CMY, YIQ, YUV, and non-linear color spaces such as L/a/b, HSV, Nrgb, Nxyz,

L/u/v), color moments [5], color co-occurrence matrix [6], and dominant color descriptor (DCD) [7], which is one of the MPEG-7 color descriptors. Image global texture is another widely used feature for describing innate surface properties of a particular object and its relationship with the surrounding regions [8]. Some features commonly utilized as global texture descriptors are Gabor filters, Wavelet transforms, probabilistic texture retrieval [9], stochastic multivariate modeling [10], gray-level co-occurrence matrix (GLCM) [11], and Tamura features [12]. Image shape is also used as a global feature that basically carries semantic information, and it is broadly based on image boundary or regions. Some of the common shape descriptors include Fourier descriptors, deformable templates, invariant moments, B splines, aspect ratio, curvature scale space (CSS), circularity, and consecutive boundary segments [4,13].

However, most of the low-level features lack spatial information in the extracted representation, e.g., histograms and shape points. Consequently, using an abstract representation alone is not sufficient to represent the pictorial semantic content of images. Many spatial-based features have recently gained more attention, including regions of interest (ROIs) [14], graph/tree-based representations [15], strings-based [16], and matrices-based [17]. Moreover, none of a single global feature has a discriminative power under different image contents and deformations, e.g., lightening, noise, and rotation. Even though some research works [13,22–24] have been presented as a combination scheme between global features to increase the retrieval accuracy, there is no sufficient evaluation on the robustness and distinctiveness of those features. Therefore, handcrafted low-level visual features, e.g., SIFT and HOG, are extensively used to capture the local characteristics of image objects and to preserve some local patterns of image contents.

Among the most popular local point descriptors are SIFT, SURF and HOG. The discrimination level of local descriptors is mainly influenced by the detector type of point of interests (POI). The most commonly used detectors are blob-based detectors (e.g., SIFT), SURF detector, corner detectors (e.g., Harris and Hessian) [25], FAST [21], and maximally stable extremal regions (MSER) [26]. Investigating the impact of using different detectors to formulate different image descriptors is an interesting aspect in the domain of CBIR. Typically, a large number of local descriptors are extracted from each image and then used for direct matching between similar batches. This is not feasible in the CBIR context due to the high-speed equipment and memory storage needed. The PCA is well suited for representing key point patches, and it provides more compact image descriptor than the standard representations [27]. Therefore, local image descriptors are usually quantized using aggregation-based approaches.

Bag of words (BOW) [28] is one of the widely applied state-of-the-art approaches for image quantization. BOW assigns the extracted local descriptors from images to the closest visual words from a visual vocabulary, i.e., 'codebook'. The codebook's k centers are usually computed and learned by k-means clustering approach. This high-dimensional sparse vector represents the image which is then weighted using term frequency inverse document frequency (TF-IDF). Fisher vector (FV) [29] is another approach that provides a compact and dense representation which is probably more adequate for retrieval applications. Fisher kernel is a probabilistic model that identifies the similarity between objects using a set of measurements for each object with a higher order of statistics than BOW. Recently, a simplified non-probabilistic version of Fishers kernels has been introduced, referred as vector of locally aggregated descriptors VLAD.

VLAD [30] is also trained using k-means to accumulate the local descriptors, which is followed by L2 normalization. VLAD addresses the efficiency and memory constraints by aggregating local descriptors into a moderate fixed-size vector representation. Some recent research efforts have introduced different schemes on features combination and their impact on the retrieval accuracy [31–35]. Lakovidou et al. [31] evaluate a set of MPEG-7 and MPEG-7-like global descriptors, e.g., scalable color descriptor (SDC) and color edge directivity descriptor (CEDD), for CBIR tasks in conjunction with SIFT and SURF local descriptors. Their proposed combination scheme is tested using BOW quantization where the system performance is only evaluated in terms of retrieval accuracy. Elalami et al. [32] integrate the color co-occurrence matrix (CCM) and the difference between pixels of scan patterns (DBPSP) as texture representation, and the artificial neural network (ANN) is used as a classifier. Zhang et al. [33] combine SIFT with color histograms, moments, coherence, and autocorrelogram. Deselaers et al. [34] combine many features for image retrieval, including SIFT, color histogram, Tamura, MPEG7, and others. Walia and Verma [35] investigate several local texture descriptors (e.g., LBP) on Log-Gabor filters response for CBIR evaluation. Bosch et al. [36] use pyramid histograms of visual words (PHOW) descriptor that extracts dense SIFT descriptors at multiple scales on three HSV image channels and stacked them up. The work introduced by Alzu'bi et al. [37] also combines a single SIFT local feature with single color feature as image representation.

However, the aforementioned works only handle certain retrieval aspects to evaluate some extracted features in terms of accuracy. Limited types of feature representations are introduced with no sufficient involvement of both global and local features in the CBIR scheme. In addition, many important factors are not included or analyzed, e.g., quantization and similarity/dissimilarity analysis, which largely affects the retrieval performance, especially the retrieval speed and memory usage. Accordingly, our work is distinguished from

the other related works by addressing the aforementioned limitations along with obtaining low-dimensional but highly discriminative image representations. The presented CBIR model employs a multilevel retrieval structure to extract compact single and aggregated image features and highlights the relationship between all of the utilized factors in the retrieval model. System performance is evaluated and reported at all stages of model development and optimization, and it is observed under various POI detectors, similarity distances, vector dimensions, and quantization approaches.

3 CBIR framework

This section presents the baseline CBIR framework, the extraction scheme of global and local features, distance measures, benchmarking image dataset, and performance measures (i.e., accuracy, retrieval time, and memory usage) used throughout this work.

3.1 The general retrieval framework

Figure 1 demonstrates the adopted methodology in building the baseline CBIR system into a set of correlated blocks and steps. The process is summarized as follows: first, a query image is submitted to the system, i.e., query-by-example. Second, a set of individual global and local features is extracted from both query image and all of dataset images. Third, each feature type is separately evaluated at each stage of the multilevel procedure under certain setups. Specifically, the retrieval performance is tested on individual and combined global features; then, the best global feature obtained is used in the next level. Subsequently, the best local image descriptors, extracted by various POI detectors, are aggregated with the best performing global features obtained in the previous stage. Finally, a robust image signature is acquired and used as a base in the later optimization stages.

At every level of feature extraction and aggregation, different distance measures, e.g., Minkowski, Cityblock, correlation, and Spearman, are used and assessed against each feature or descriptor to identify its impact on system performance, accuracy, and results ranking. For the current stage, VLAD is used to quantize local descriptors into image vectors; but FV and BOW quantization approaches will be involved in the model optimization level. The returned images are sorted and ranked as a list of images from the most to the least similar images. The performance measures, i.e., accuracy, speed, and memory, are also reported at every stage of the CBIR system.

3.2 Feature extraction

In this section, all of the global and local image features are described, including features/descriptors extraction, local points detectors, and quantization methodology.

3.2.1 Global features

The proposed CBIR model utilizes ten global features that represent image color, texture, shape, and spatial features. The diversity of selected features provides more information on image representation. Particularly, the following global features are extracted from each image:

1. *Color* The RGB and HSV histograms are used to represent the color channels of image, and the color distribution of each channel is represented using the first two color moments [5].
2. *Texture* Extensive texture representations are extracted with a variety of global features; including, Gabor wavelets [38], wavelet moments [39], GLCM [11], SFTA [40], LBP [41,42], and GIST [43].

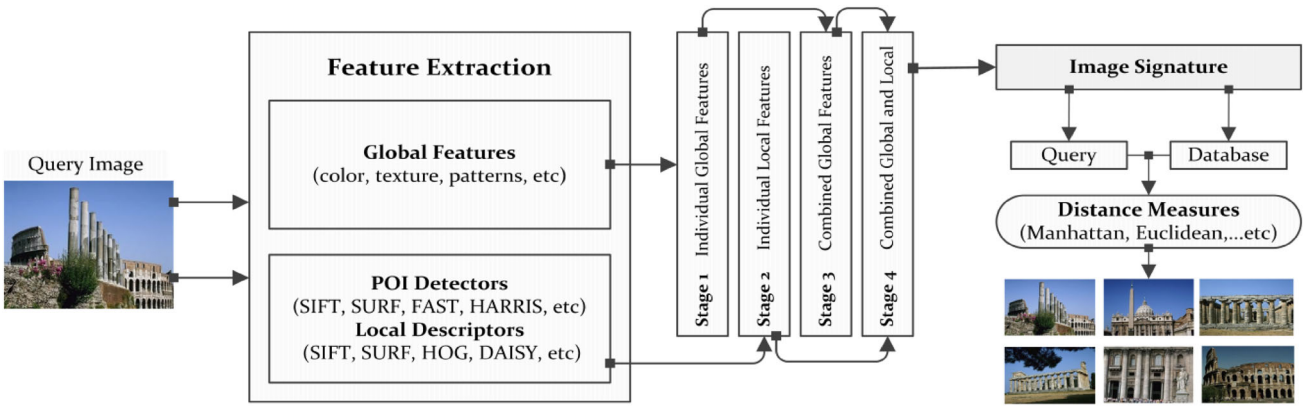


Fig. 1 The baseline CBIR framework

Table 1 Compact global features used in the baseline CBIR model

Feature	Category	Extraction procedure in this paper	Vector size
RGB histogram	Color	3 color channels are quantized into 3 histogram bins then normalized by L2 norm. The color value is: $C = 3R + 3G + 3B$	24
HSV histogram	Color	H component is quantized into 8 ranks non-uniformly and S and V components are quantized into 2 ranks uniformly. The color value is: $C = 8H + 2S + 2V$	32
Color moments	Color	Image distribution of color is represented as vector of first 2 color moments: *mean* and *standard deviation* for each color channel	40
Chromaticity moments	Shape	Statistics extracted from a bidimensional color distribution as implemented in [44]	10
Segmentation-based fractal texture analysis (SFTA)	Texture	A set of 4 thresholds is computed from the input image to get the final SFTA feature vector normalized by summation	24
Gabor wavelets	Texture	4 scales and 6 orientations are used. For each scale and orientation, mean and squared-mean are computed and concatenated into a feature vector	48
Wavelet moments	Texture	Two-dimensional discrete wavelet transform (DWT) is used to obtain the first two moments of normalized wavelet coefficients, i.e. standard deviation and mean	40
Gray-level co-occurrence matrix (GLCM)	Texture/spatial	4 statistical properties are extracted: energy, contrast, correlation, and homogeneity. Each property part is of size 20 that are concatenated together into a single GLCM feature vector	80
LBP	Texture	A normalized histogram of LBP is computed using 8 sampling points defined around the origin coordinates (0, 0) of a circle of radius 1	10
GIST	Texture	Standard GIST is applied using 4 scales, 8 orientations per scale, and 4 blocks	512

3. *Shape* The chromaticity moments feature [44] is used to extract image edges, which is based on the bidimensional distribution of image colors.

In addition, some spatial information is used from the statistical properties extracted in the GLCM algorithm. Table 1 summarizes the global features used in the baseline CBIR model along with the conducted extraction procedure for each type. It is clear that the size of the extracted image vectors is very small, which is an advantage for more efficiency and less memory usage.

3.2.2 Local points detectors

Since different points of interest (POI) or corner/region detectors may provide different levels of discrimination for local image descriptors, the proposed CBIR system is capable of handling a set of POI detectors to explore their effectiveness under different setups. The impact of detector–descriptor pairs on the final image signature is also evaluated. The following are the local detectors involved in our experiments:

1. *Blob-based detectors* This method aims at detecting regions of interest that are distinguished in properties and almost constant. One of the most common methods adopted to detect image blobs is based on scale-space interest point detectors such as Laplacian of the Gaussian (LoG) and difference of Gaussian (DoG), which is used in SIFT [18] image descriptors. The scale-normalized determinant of the Hessian calculated from Haar wavelet is also another successful method used as interest point detector in SURF [19] image descriptors.
2. *Corner-based and minimum eigenvalue detectors* These methods also detect local interest points for which there are two dominant and different edge directions, i.e., the intersection of two edges. Among the most commonly used approaches are: Harris–Stephens [25], FAST [21], and minimum eigenvalue [45]. These algorithms return the detected corners object that contains information about feature points detected in the 2-D grayscale input image.
3. *Region/intensity-based detectors* These types detect stable regions based on the intensity range of input images. The MSER [26] algorithm incrementally tests the variation of region area size between different intensity thresholds until a stable region is detected. On the other hand, LIOP [46] algorithm divides local image patches into subregions using affine covariant region detectors, which is based on the intensity order.

Table 2 Detectors of POI/regions/corners used in the baseline CBIR model

Detector	Parameters (I = input image)	Type
Gaussian blobs	Scales = 7, scale factor $\sqrt{2}$, bin size = 8 pixels, step = 4 pixels, $\sigma = 0$, contrast-threshold = 0	Blob-based
SURF	Threshold = 700, octaves = 3, scale levels = 4, ROI = [1 1 size (I, 1)]	Blob-based
FAST	Threshold = 0.10, contrast = 0.2, ROI = [1 1 size (I, 2) size (I, 1)]	Corner-based
Harris	Threshold = 0.01, Gaussian filter size = 5, ROI = [1 1 size (I, 2) size (I, 1)]	Corner-based
Minimum eigenvalue	Threshold = 0.01, Gaussian filter size = 5, ROI = [1 1 size (I, 2) size (I, 1)]	Corner-based
MSER	Threshold = 2, region size = [30 14000], variation=0.25, ROI = [1 1 size (I, 2) size (I, 1)]	Region/intensity-based
LIOP	Threshold = −0.02, neighbors sample = 4, sampling radius = 5, spatial bins = 6	Region/intensity-based

Table 2 briefly describes all detectors of local points/regions used to construct the local descriptors.

3.2.3 Local features

This section presents the local image descriptors used in the proposed CBIR system and the quantization methods (i.e., FV, VLAD, and BOW) used to encode them.

1. *SIFT* [18,47] It performs a reliable matching between different views of objects. To find the position and scale, SIFT detects the locations of key points in the scale space of image using scale space extrema in the difference-of-Gaussian (DoG) function included in the image I, $D(x, y, \sigma)$, which is calculated from the difference of two nearby scales (x, y) separated by a constant multiplicative factor k as follows:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \times I(x, y) \tag{1}$$

The standard SIFT uses a 4×4 array of histograms with 8 orientation bins in each, which yields a $(4 \times 4) \times 8 = 128$-dimensional descriptor for each key point. This vector is then normalized to unit L2 norm to reduce the effects of illumination change. In our proposed CBIR framework, the standard SIFT is densely extracted and computed at 7 scales by a factor $\sqrt{2}$ between successive scales, bin size of 8 pixels wide and a step of 4 pixels. Finally, the rootSIFT [47] is obtained that utilizes the square root Hellinger kernel instead of the standard Euclidean distance to measure the similarity between SIFT descriptors, which is proved to achieve a superior performance in most cases without increased processing or storage requirements.

The Hellinger kernel for two L1-normalized n-length histograms, x and y, is defined as:

$$H(x, y) = \sum_{i-1}^{n} \sqrt{x_i y_i} \tag{2}$$

Comparing rootSIFT descriptors using Euclidean distance (d_{E}) is equivalent to using the Hellinger kernel to compare the original SIFT vectors as follows:

$$d_{\mathrm{E}}\left(\sqrt{x}, \sqrt{y}\right)^2 = 2 - 2H(x, y) \tag{3}$$

2. *SURF* [19] It is mainly based on the sums of 2D Haar wavelet responses and makes an efficient use of integral images. The standard detector of interest points in the SURF is based on the Hessian matrix which detects blob-like structures at the locations where the determinant is the maximum. Unlike the Hessian–Laplace detector, this method relies on the determinant of the Hessian for the scale selection.
In addition, SURF descriptor identifies the distribution of intensity content within the interest point neighborhood, which is similar to the gradient information extracted by SIFT. The SURF also includes a new indexing step based on the Laplacian sign which reduces the time needed for feature computation and matching, and it increases the robustness simultaneously.
3. *HOG* [20] It characterizes the local object appearance and shape by edge directions or the distribution of local intensity gradients, even without an accurate knowledge of edge positions or the corresponding gradient. The HOG divides the image window into small spatial cells as regions, and accumulates edge orientations over the pixels of each cell or a local l-D histogram of gradient directions. As a result, the image representation is formed by the combined histogram entries. Furthermore, it accu-

mulates 'energy' as a measure of local histogram over 'blocks'.

4. *DAISY* [48] It is inspired by SIFT and GLOH and retains their robustness to reduce the computational requirements and convolutions using Gaussian filters. DAISY feature is densely extracted over all pixels. It computes the orientation maps of different sizes at low cost as convolutions with a large Gaussian kernel can be obtained from several consecutive convolutions with smaller kernels.
5. *Local intensity order pattern (LIOP)* [46] It encodes the local ordinal information which is used to divide the local patch into sub-regions. Then, a LIOP of each point is defined based on the relationships between the intensities of its neighboring sample points. Next, all LIOPs of points in each ordinal bin are accumulated and then concatenated together to construct the LIOP descriptor. LIOP descriptor is robust and invariant to image rotation, monotonic intensity changes, and geometric transformations, e.g., viewpoint change and image blur.

Table 3 summarizes all of local descriptors used in our CBIR system.

In this work, three quantization approaches are used to quantize local descriptors extracted from any input image, i.e., VLAD, FV, and BOW. The VLAD is used as a baseline method to evaluate the baseline CBIR system introduced in Sect. 4. In Sect. 5, FV and BOW are utilized to investigate their impact on the performance of the proposed image representations. As aforementioned, VLAD is formulated based on regions/corners/POIs that are extracted from images. Each descriptor is assigned to the closest cluster (i.e., image class) of a vocabulary of size k. For each cluster, differences between descriptors and cluster centers (i.e., residuals) are formed into a vector, and $k\times$ size(descriptor) sums of residuals are concatenated into a single vector. In the recent standard scheme, VLAD vectors are normalized by a signed square rooting (i.e., an element x_i is transformed into a $\text{sign}(x_i)\text{sqrt}(|x_i|)$, and the transformed vector is $L2$-normalized. In this work, the standard VLAD settings are adopted, and a visual vocabulary of 256 clusters built by k-means approach is used in all experiments.

Table 3 Local descriptors used in the baseline CBIR model

Descriptor	Vector size	Detectors
rootSIFT	128	Gaussian blobs
SURF	128	SURF/Harris/FAST/MSER/MinEign
HOG	144	SURF/Harris/FAST/MSER/MinEign
DAISY	200	SURF/Harris/FAST/MSER/MinEign
LIOP	144	LIOP batching

3.3 Distance measures

To achieve more retrieval accuracy and better performance, the evaluation scheme of CBIR system employs effective similarity matching measures to characterize and quantify the perceptual similarities. Distance measures are an integral part of system evaluation. The key advantage of using different distance measures is to find out adequate and robust measures under various retrieval setups. Table 4 describes all distance measures $d(X, Y)$ used in the experiments for any two images, X and Y, represented in a data space by two n-dimensional vectors $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$, respectively.

3.4 Performance measures

For each query initiated in the retrieval system, the top 20 images are retrieved and ranked according to the similarity

Table 4 The distance measures used in the baseline CBIR model

Measure	Equation	
Relative Manhattan	$d(X,Y)=\sum_{i=1}^{n}\frac{\lvert x_i-y_i\rvert}{1+x_i+y_i}$	(6)
Euclidean (L_2)	$d(X,Y)=\left(\sum_{i=1}^{n}\lvert x_i-y_i\rvert^2\right)^{1/2}$	(7)
Standard L_2	$d(X,Y)=\sum_{i=1}^{n}\text{weight}(x_i)\times(x_i-y_i)^2$	(8)
Chebyshev (L_∞)	Equation (6), where $r=\infty$	
Cosine	$d(X,Y)=1-\cos\theta=1-\frac{X\times Y}{\lVert X\rVert\times\lVert Y\rVert}$	(9)
Correlation	$d(X,Y)=1-\frac{d^2(X,Y)}{2n}$, where $d(X,Y)=\text{L}_2$ distance	(10)
Cityblock (L_1)	$d(X,Y)=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$	(11)
Spearman	$d(X,Y)=1-\frac{6\times\sum_{i=1}^{n}(\text{rank}(x_i)-\text{rank}(y_i))^2}{n(n^2-1)}$	(12)

Fig. 2 Image samples selected from each category

scores computed, and then, the precision of ranked images ($P(R_k)$) is computed to measure the retrieval accuracy for any query as follows:

$$P(R_k) = \frac{\#\,(\text{relevant images} \cap \text{retrieved images})}{\#\,(\text{retrieved images})}, \tag{4}$$

where the retrieved images are the top images retrieved (R_k) and the relevant images are only the relevant images to the query image, i.e., belongs to the same image category. A set of representative images is taken as queries from each image category. The average precision (AP) is then computed for each category, and the mean average precision (mAP) is reported for the whole dataset, i.e., the average of the AP of each image category. For a single query image, AP is the average of the precision value obtained for the set of top k images existing after each relevant image is retrieved, and this value is then averaged over all queries in the image category. Therefore, if the set of relevant images for a query $q_j \in Q$ is $\{I_1, \ldots, I_m\}$ where Q is the set of all queries, then mAP is defined as:

$$\text{mAP}(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m}\sum_{k=1}^{m}P(R_k) \tag{5}$$

Other performance measures are evaluated in terms of retrieval speed (in ms/s) and memory usage (in B/kB).

3.5 Image dataset

A subset of Corel image dataset[1] is used in the baseline CBIR system, which is widely utilized in computer vision applications. It consists of 1000 colorful images with 10 different image categories, i.e., 100 images per category. This semantic categorization of image dataset reflects the human perception of image similarity. Figure 2 shows a representative image for each image category.

[1] http://wang.ist.psu.edu/docs/related/.

4 Experiments and discussion on the baseline CBIR model

In this section, we present and discuss the system performance using three different image representations, i.e., individual global features, combined global features, and combined global and local features. Each representation is evaluated in terms of retrieval accuracy mAP, speed, vectorization time, and memory usage.

4.1 Individual global features

Figure 3 shows the retrieval accuracy (mAP) obtained by every single global feature at top 20 ranked images. It is evident that the color histograms achieve the highest retrieval accuracy among all global features under the same retrieval setup. Specifically, the HSV histogram outperforms all global features over all of distance measure, where the best results for HSV are 76 and 75 % reported using relative Manhattan and Cityblock similarity distances, respectively. Table 5 presents other performance measures achieved by all global features in terms of average vectorization time, speed, and memory usage.

The retrieval speed is the average time in seconds elapsed from the time of query submission until ranking and showing the top 20 images, which includes the average time of image vectorization. The memory size is the average actual memory required to store the image vector of the extracted global feature. As shown in Table 5, HSV feature still records a superiority according to the time elapsed for vectorization and search images with only 197 B of required memory. However, other features such as RGB, color moments, GIST, and LBP have comparable accuracy results (see Fig. 3), but they suffer from either the long search time or the large memory size. As a result, the HSV color feature is selected to be the basis of the combination between global features, i.e., HSV is combined with every individual global feature to assess its

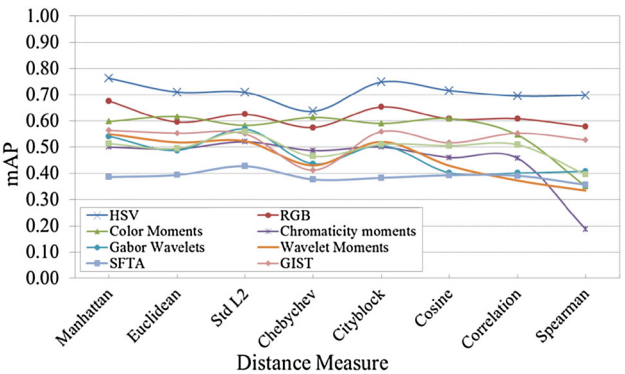


Fig. 3 Retrieval accuracy of global features

Table 5 System performance using individual global features

Feature	Vectorization time (s)	Retrieval speed (s)	Memory size (B)
HSV	0.093	0.273	197
RGB	0.115	0.577	126
Color moments	0.009	0.569	307
Chromaticity	0.144	0.494	73
Gabor wavelets	0.305	0.575	364
Wavelet moments	0.150	0.570	307
SFTA	0.585	0.643	184
GIST	0.125	0.623	1833
LBP	0.165	0.601	75

effectiveness in increasing the discrimination level of image signature as well as the retrieval performance.

4.2 Composite global features

In this section, the impact of adding more information to the extracted individual HSV features is examined. Figure 4 shows the mAP results obtained by conducting extensive experiments on combining HSV with every single global feature. As shown, a slight increment is gained by this combination; especially, the features plotted by solid lines achieve the best accuracy when combined with HSV. Other features do not positively influence the retrieval accuracy, shown in Fig. 4 as markers without lines.

The best combinations recorded using two global features are as follows: (HSV + RGB, 78 %, relative Manhattan vector_size = 59), (HSV + Colormoments, 78 %, relative Manhattan, vector_size = 72), (HSV + SFTA, 78 %, relative Manhattan,vector_size = 56), (HSV + GIST, 78 %, Euclidean, vector_size = 544), and (HSV + LBP, 78 %, relative Manhattan, vector_size = 42). However, these combinations necessarily require more time for vectorization and search as well as more memory storage. Furthermore, the amount of accuracy improved by these representations over a single HSV feature is negligible. As a result, only the best single global features (i.e., HSV, RGB, color moments, GIST, and LBP) are passed to the next level of combination (i.e., with local descriptors) in the proposed CBIR scheme, as illustrated in the next section.

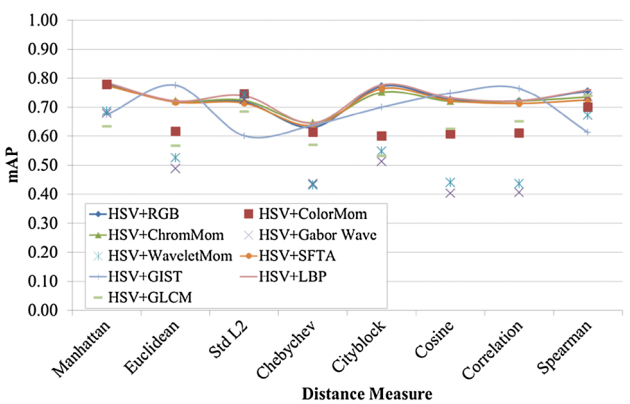


Fig. 4 Retrieval accuracy of composite global features with HSV

4.3 Composite global and local features

Based on the results reported on the system performance using single and combined global features, more experiments are conducted on using single local descriptors and combined global–local descriptor. The aim is to monitor how the CBIR system acts using different image signatures, and to figure out the impact of adding more data to the image vector. The main conclusion acquired in Sect. 3.2 is that adding many global features together does not necessarily yield an improvement in accuracy or performance. Accordingly, investigating different image signatures constituted of global and local descriptors is a worthy and an interesting test. First, the performance of single local descriptors is presented. Then, the performance of feature combinations between the best performing global and local descriptors is evaluated.

It is worth mentioning that SURF, HOG, and DAISY descriptors are formulated using five different detectors so that only the best result achieved among these detectors is used in the next development phases of the CBIR system. Figure 5 depicts the retrieval results using different detectors.

It is clear that all of the three descriptors (i.e., DAISY, HOG, and SURF) achieve the best accuracy using the minimum eigenvalue detector over almost all of distance measures. Therefore, these descriptors will be compared with SIFT and LIOP descriptors extracted by the minimum eigenvalue detector. Figure 6 shows the retrieval mAP for all local descriptors obtained by the baseline CBIR system. As shown, SIFT (rootSIFT and SIFT are used interchangeably) largely outperforms the other local descriptors under the same evaluation setups. Moreover, Table 6 compares between local descriptors in terms of time and storage. Here, we weigh between the accuracy shown in Fig. 6 and the performance metrics to pass only the best performing image local feature to the next phase, i.e., SIFT. However, SURF, HOG and LIOP provide better retrieval speed and lower memory consumption; while at the same time, they gain lower mAP accuracy compared to SIFT. Accordingly, SIFT will be used in the next combination level, i.e., local and global features. The vectorization and search speed of SIFT vector as well as memory usage will be considered later in the optimization phase.

For now, SIFT is selected from the best local descriptors, while HSV, RGB, color moments, GIST, and LBP are used as global features to be aggregated with SIFT. The CBIR system

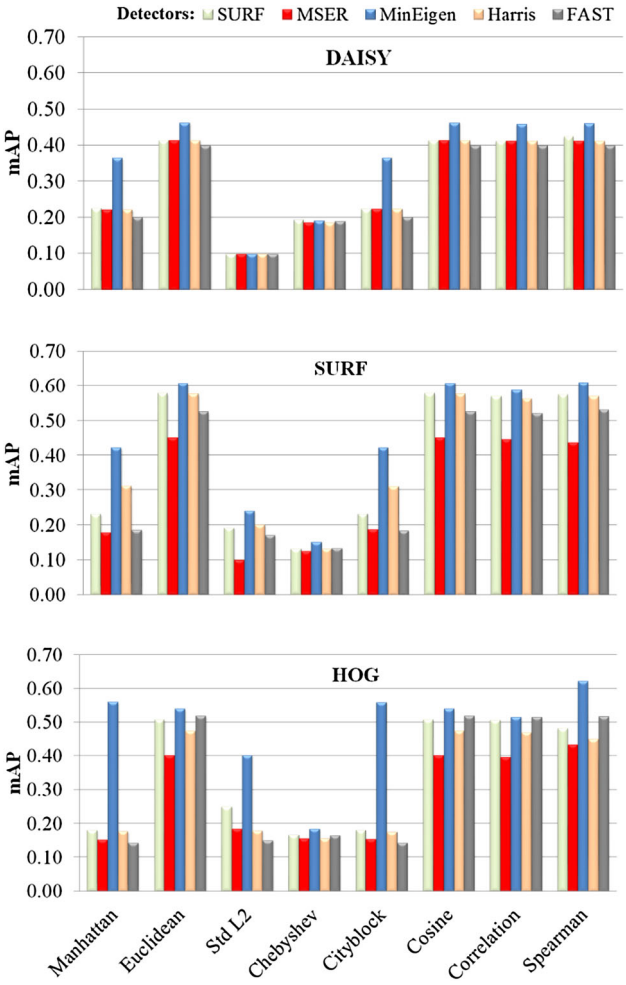


Fig. 5 The mAP of DAISY, SURF, and HOG descriptors using various detectors

performs two types of global–local combinations: one global feature is combined with SIFT, and two global features with SIFT.

(1) *Single global feature with rootSIFT* Since humans are usually interested in certain parts of images (e.g., objects), object-based retrieval that benefits from local descriptors tend to be more effective in satisfying the human needs. On the other side, global features provide a meaningful representation of the whole image which is very close to the human vision system. Therefore, one of the integral parts of this work is to construct a composite global and local feature signature to provide more representative descriptions for visual image contents. However, system efficiency and memory usage of this aggregation procedure will be closely handled.

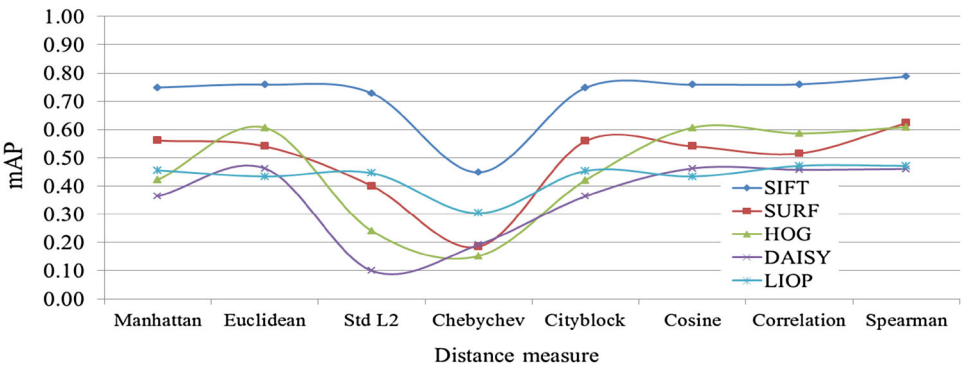
Table 6 System performance using individual local descriptors

Feature	Vectorization time (s)	Retrieval speed (s)	Memory size (kB)
SIFT	0.429	1.113	110.26
SURF	0.151	0.971	104.75
DAISY	0.889	1.566	102.18
HOG	0.376	0.704	83.16
LIOP	0.042	0.596	0.512

Figure 7 shows various combinations between the best global features (i.e., HSV, RGB, color moments, GIST, and LBP) with SIFT. Many conclusions can be drawn from these results. First, HSV feature has the superiority over other global features when it is combined with SIFT. Second, all of the combined representations improve the discrimination level of image signature compared to the single global or local features. Specifically, the retrieval accuracy (mAP) obviously is increased for all of image signatures. Third, some of global features outperform others if combined with SIFT; for example, GIST performs better than LBP as a single global feature (see Fig. 3), while LBP outperforms GIST when both combined with SIFT as shown in Fig. 7. In addition, image signatures of the combined features have different accuracy and efficiency levels using different distance measures, e.g., it varies with Euclidean but almost similar with standard Euclidean. Finally, the retrieval time and memory required for each combined features vary, which will be further analyzed in the optimization phase of the proposed CBIR system.

As a result, HSV combined with SIFT is the best image signature constructed in our CBIR system. The mAP rises

Fig. 6 The retrieval accuracy of single local descriptors



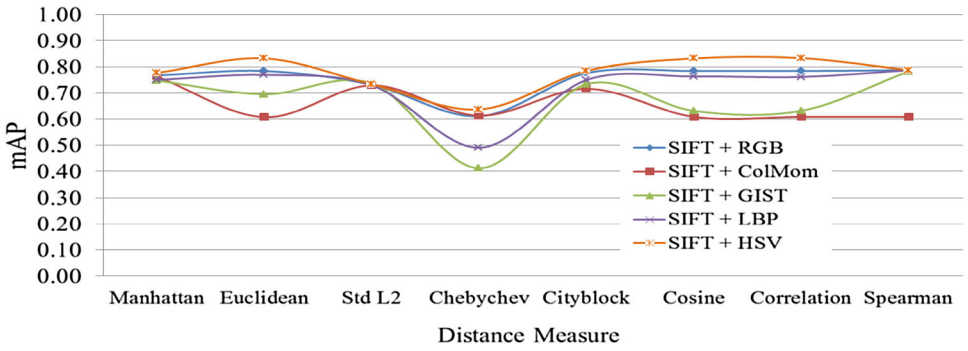


Fig. 7 The retrieval accuracy (mAP) of combined single global feature with SIFT

Table 7 System performance using two global features with rootSIFT

Image signature	Vector size[a]	Best mAP (%)	Distance measure	Vectorization time (s)	Retrieval speed (s)	Memory size (kB)
SIFT + HSV	$256 \times 128 + 32$	83	L2/cosine/correlation	0.517	0.868	110.35
SIFT + HSV + colMom	$256 \times 128 + 32 + 40$	79	Manhattan/Spearman	0.522	0.996	110.38
SIFT + HSV + GIST	$256 \times 128 + 32 + 512$	83	L2	0.709	0.895	112.14
SIFT + HSV + LBP	$256 \times 128 + 32 + 10$	83	L2/cosine/correlation	0.679	0.830	110.39

[a] Codebook size used in VLAD is 256

from 76 % for HSV and 79 % for SIFT to be 83 % for the combined HSV and SIFT. However, SIFT achieves 79 % of accuracy with higher vectorization time (∼3 s) using the Spearman distance measure, but it achieves 76 % of accuracy with a lower time (∼0.429 s). Consequently, the new combined feature improves the accuracy over the single HSV and SIFT by 7 %. However, the proposed CBIR system considers the best two signatures from different image representation, which means both SIFT + HSV and SIFT + LBP will be used later for system optimization in Sect. 5. The RGB color histogram is not involved since it is outperformed by the HSV color histogram so that the LBP is considered to involve different characteristic (i.e., texture). Next, another combination level will be examined in the system by adding the remaining best global features (i.e., color moments, GIST, and LBP) to the image vector SIFT + HSV to evaluate the retrieval performance from different aspects.

(2) *Multiple global features with rootSIFT* Table 7 summarizes the retrieval performance of integrating two global features with SIFT. It is very clear that only adding HSV features with SIFT still outperforms other features constituted of three features. Furthermore, it achieves 83 % of accuracy with lower search time and memory size. Generally, the experimental results on this level of combination shows that in most cases, the retrieval accuracy is decreased by adding more features to SIFT + HSV, e.g., color moments and GIST are lowering the accuracy with a noticeable performance degradation. These results confirm that adding more features to the image representation does not always guarantee improving the retrieval accuracy and performance of the CBIR system. As a result, only a single global feature is combined with SIFT and involved in the next stage of CBIR system development, i.e., SIFT + HSV and SIFT + LBP.

5 Compact quantization

In this section, a further analysis is established with regard to the efficiency of the CBIR system with minimum possible search time and memory storage. It is important to weigh between these constraints and the retrieval accuracy. Therefore, we consider two important factors that affect the retrieval performance: dimensionality reduction and quantization approaches. Here, only the best obtained signatures from the baseline system, i.e., SIFT + HSV and SIFT + LBP, are further evaluated under different optimization setups.

One of the conclusions reported in the previous sections is that different distance measures provide different accuracy results (mAP) and require different vectorization time, search time, and memory storage. In consequence, the CBIR system only involves the leading distance measures that provide the best performance through all previous extensive experiments on average. Specifically, relative Manhattan, Euclidean, Cityblock, and cosine distances are used for similarity matching in the current stage of multi-feature analysis. One of the main reasons to exclude some of distance measures such as Spearman, that achieved a comparable accuracy with other distances, is the long time required for image vectorization (∼3.0 s), while it is (∼0.5 s) in average using others. In the remaining part of this section, a thorough analysis is presented on the impact of reducing the dimension of image

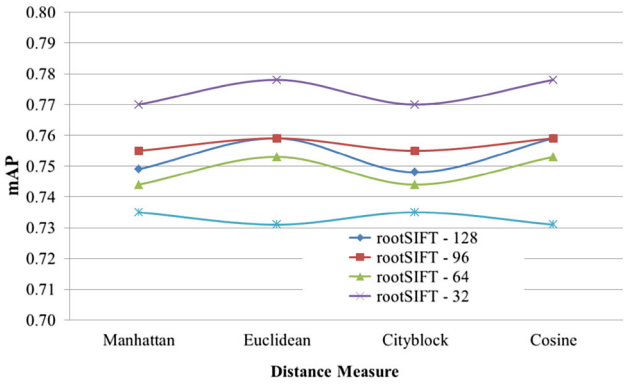


Fig. 8 The retrieval accuracy (mAP) of rootSIFT at different PCA-based reductions

vectors using different quantization methods (i.e., VLAD, FV, and BOW) on the system performance in terms of accuracy and efficiency.

The baseline CBIR system uses 128-D SIFT descriptor aggregated with 32-D HSV and 10-D LBP separately. Here, the PCA/whitening is utilized to gradually reduce the size of rootSIFT descriptor into four dimensions, i.e., 96, 64, 32, and 16. The retrieval accuracy may decrease while reducing the descriptor size down; thus, it is important to monitor both the accuracy and performance. Therefore, we attempt to exploit the correlation between some of image dimensions using the PCA/whitening approach. More precisely, the covariance matrix of image patches and its eigenvectors are computed. Then, the size of descriptor vectors is reduced by projecting their rotated version into a low-dimensional space based on the computed eigenvectors (components). Finally, the resulting reduced image descriptors are whitened by dividing each matrix component by the square root of its eigenvalue. This gives all of the image descriptors the same variance. Figure 8 shows the retrieval accuracy (mAP) of all rootSIFT image vectors at different dimensions. As shown, the retrieval accuracy directly proportionally decreases as the dimension size of rootSIFT is reduced except at size 32, which is unforeseen since the image vector is expected to lose some distinctive features. On average, the PCA-rootSIFT at 32-D outperforms other vectors by 2 %.

This is an interesting start for more optimization in the image representation. Distinctly, this reduction in vector size will substantially reduce the computation cost and memory requirements. Now, it is important to examine the impact of using the new rootSIFT dimension with HSV and LBP, i.e., rootSIFT(32) + HSV(32) and rootSIFT(32) + LBP(10).

As aforementioned, the new image representation with its compact size will be assessed using three quantization approaches: VLAD, FV, and BOW. All experiments are carried out using five different visual codebooks provided by k-means clustering: 512, 256, 128, 64, and 32. Figure 9 shows the retrieval accuracy achieved using each quantization approach at different codebook sizes and different distance measures.

Coming to the BOW, as shown in Fig. 9e, f, the mAP largely fluctuates using different codebook sizes. Unlike VLAD and FV, BOW has a noticeable drop in accuracy (∼10 %) of rootSIFT + HSV feature using Euclidean and cosine. However, BOW is still comparable with VLAD and FV using the rootSIFT + LBP feature, which shows more stability in the retrieval accuracy. The overall evaluation of image representations, based on the accuracy results obtained from the CBIR system, outweighs VLAD over FV and BOW. It shows higher accuracy at smaller sizes of visual codebook, which is a crucial factor that affects other performance criteria (i.e., retrieval time and memory usage). Therefore, rootSIFT + HSV and rootSIFT + LBP will be quantized using small VLADs in the final CBIR system.

Based on the conducted extensive multi-feature analysis, the CBIR system will be further optimized by: (1) exploiting the most beneficial conclusions reported; (2) involving rootSIFT + HSV and rootSIFT + LBP in the final CBIR system; and (3) weighting the similarity measures according to the relationship between image vector, image representation, query image, and distance measures.

The optimized CBIR system is capable of handling different types of query images with different sizes and structures. The selection and weighting of the new constructed image signatures (i.e., rootSIFT + HSV and rootSIFT + LBP) are processed automatically, which is based on the strength of the extracted local and global descriptors. The impact of using such small sizes of image features on the system performance is shown in Table 8. It is clear that the resulting image representation provides high and comparable retrieval accuracy (mAP) with the baseline representations, even a higher accuracy (86 %) achieved using rootSIFT(32)–HSV(32). This confirms that the resulting image representation in our CBIR system is not largely affected by reducing the dimension of rootSIFT as well as VLAD's codebook size. Moreover, it benefits from the substantial reduction reported in the vectorization time, retrieval time, and memory usage by 22, 35, and 96 %, respectively.

For example, Fig. 10 shows the top 20 relevant images to the elephant query image. Other image categories, e.g., Buses and Dinosaurs, achieve 100 % of retrieval accuracy (mAP) at top 20 ranked images. Furthermore, the top returned images are very similar in colors, direction, and contents (e.g., objects and background), which indicates a high quality of ranking due to the discriminating characteristic of image representation.

To evaluate the resulting model on a larger image dataset, we use MIRFlickr [49] dataset with Corel dataset. The MIRFlickr consists of 25,000 images and acts here as distractor images for the standard image categories used in the Corel,

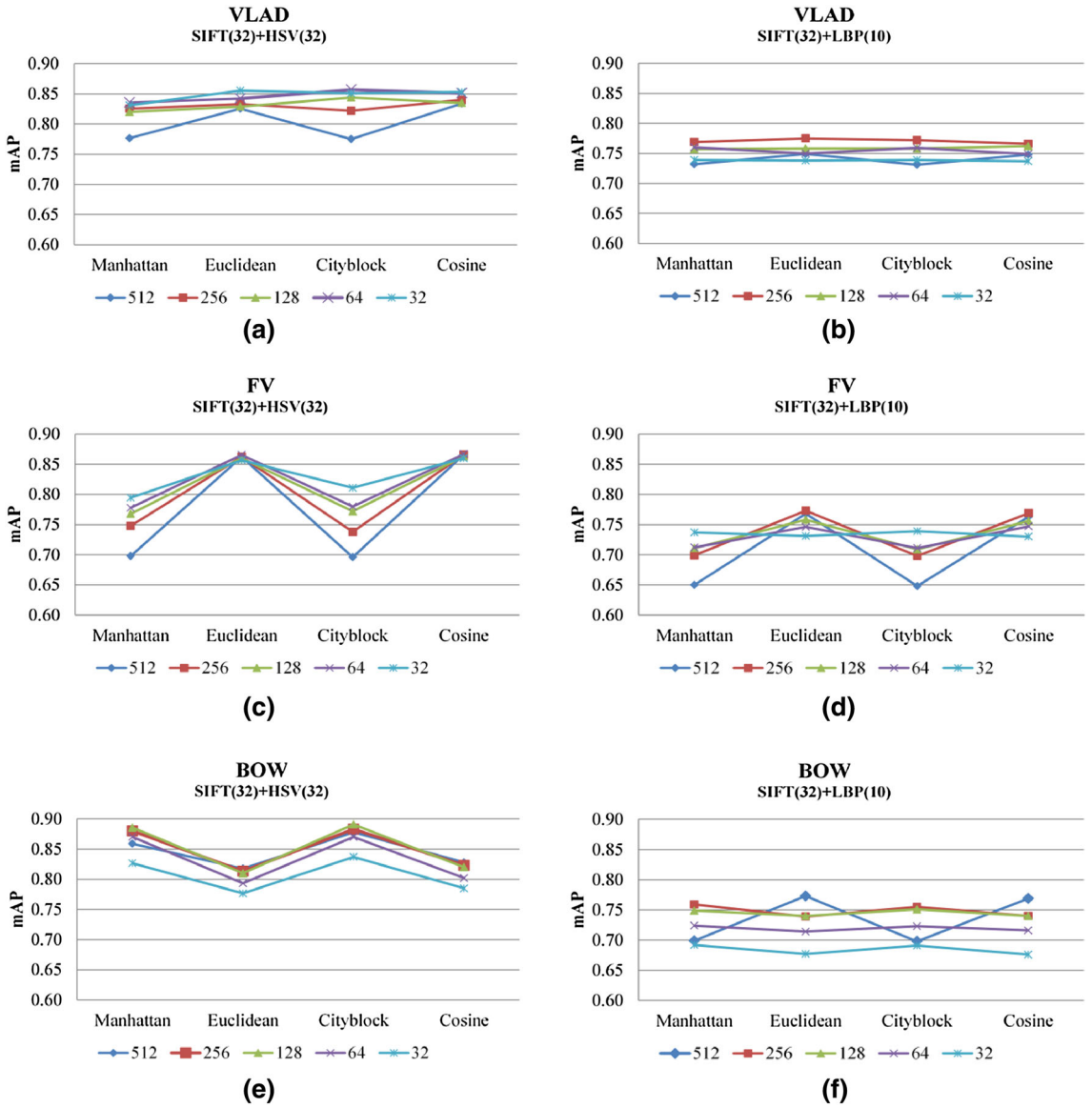


Fig. 9 The retrieval accuracy (mAP) of combined features using VLAD (**a**, **b**), FV (**c**, **d**), and BOW (**e**, **f**) at different codebooks

Table 8 System performance of optimized and baseline CBIR systems

Image signature	Vector size	mAP at		Vectorization time (s)	Retrieval speed (s)	Memory size (kB)
		Top 10 (%)	Top 20 (%)			
VLAD(256): SIFT(128) + HSV(32)	32800	89.6	83	0.517	0.868	110.35
VLAD(32): SIFT(32) + HSV(32)	1056	91.4	86	0.406	0.566	3.67
VLAD(256): SIFT(128) + LBP(10)	32778	80.4	77	0.564	0.882	110.30
VLAD(32): SIFT(32) + LBP(10)	1034	79.8	74	0.513	0.590	3.60

which makes the querying process more challenging. The MIRFlickr images represent a wide range of common daily life events, natural scenes, human, animals, and general objects. In addition, thousands of images are very similar to the Corel images and intersect with all of image categories, which makes the retrieval procedure of the standard queries more complicated. Figure 11 shows some sample images from the MIRFlickr image dataset.

The same experiment setup is applied on this dataset using the smallest vector dimension of image representation, i.e.,

Fig. 10 Sample results of our CBIR model on Corel dataset

Fig. 11 Sample images of MIRFlickr dataset

Table 9 The accuracy results on Corel with MIRFlicker dataset

rootSIFT(32) + HSV(32)	mAP (%)		
Codebook size	Euclidean	Manhattan	CityBlock
VLAD-512	51.7	60.8	62.2
VLAD-256	53.4	65.2	66.3
VLAD-128	55.8	61.3	62.9
VLAD-64	55.1	65.3	67.6
VLAD-32	57.3	63.2	65.1

rootSIFT(32) and HSV(32). The results are also reported on different sizes of visual codebooks ranging from 32 to 512. All of the experimental results are listed in Table 9.

The results reported for the MIRFlickr + Corel show a noticeable variance in performance using the three distance measures. The retrieval model is best performing with City-Block measure (67.6 % using codebook size 64) followed by relative Manhattan, and then, it has shown a degradation of 10 % in average using Euclidean measure. In general, the compact representation used for this composite dataset has shown high discrimination level with good performance even with high similarity overlapping between images. In addition, the accuracy is increased when the codebook size is decreased, emphasizing the high capability of the image signature in preserving the most representative descriptors of images.

As aforementioned, we only use the same queries of the standard Corel categories; therefore, the similar images in each of ten categories are only considered as relevant images for accuracy computations. Numerous images in the MIR-Flickr are very similar to the initiated queries but have not been included in their similar categories. This confirms the effectiveness of our image presentations on retrieving images from a large diverse dataset.

6 Comparisons

The retrieval model is compared with some related state-of-the-art methods on three standard dataset in image retrieval: (1) Wang's Corel, (2) Holidays, and (3) ZuBuD.

6.1 Corel dataset

Figure 12 shows that our proposed image representations outperform the other methods at top 20 ranked images. It is important to mention that the best retrieval accuracy achieved by Zhang et al. [33] is taken from a figure at top 20 with

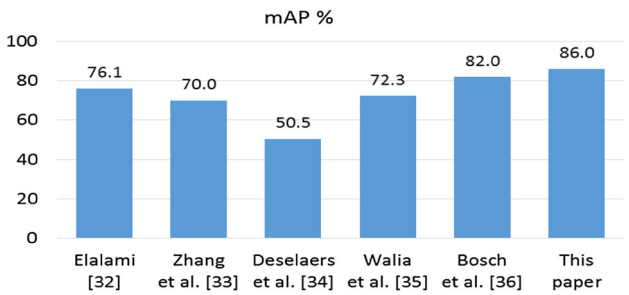


Fig. 12 A comparison of mAP obtained at top 20 images using Corel dataset

approximate error ±1. In addition, the implementation of PHOW in [50] is used for comparisons.

6.2 Holidays dataset

The Holidays dataset [51] is a set of images that contains some of Holidays photos. It is one of standard benchmarking dataset to measure the robustness against image rotations, viewpoint and illumination changes, blurring, etc. The dataset includes 1491 high-resolution images with a large variety of scene types (e.g., natural, man-made, water and fire effects, etc.), as shown in Fig. 13. The dataset contains 500 image groups, each of which represents a distinct scene or object. The first image of each image group is the query image and the correct relevant images are the other images of the group.

Table 10 compares the performance of our image representation with the related state-of-the-art methods on Holidays dataset. The proposed image representation outperforms the other related methods, even that we only use 32-D of root-SIFT and small codebook sizes. The mAP of 72.47 % improved the accuracy by ∼7 % over the best method.

6.3 ZuBuD dataset

The ZuBuD [55] consists of 1005 images of 201 different buildings (5 images of each building). For testing, 115 different and disjoint images from the database are used as queries. The ZuBuD is a challenging dataset, because the building images are taken by two cameras under different angles, illuminations, and weather conditions, as shown in Fig. 14. The mAP is computed at the top five images to evaluate the retrieval accuracy for all of 115 queries. Table 10 compares the performance of our CBIR system using both SIFT-HSV and SIFT-LBP with a related work achieving the best results on the ZuBuD image dataset. The best results reported in [31] are taken using different global or local features.

As shown, the main difference between the proposed system and their work is the size of both features and codebook. Using only 32-D of SIFT and 32-HSV, the retrieval accuracy mAP outperforms almost all other image representations. We only report the obtained results in [31] using a codebook of 512 centers, because they conducted their experiments on a codebook of size 512 and 2048. However, we take into account other performance factors, e.g., vectorization time and memory usage. Table 11 shows the results obtained by our proposed image representation over a range of visual codebooks, i.e., 32–512.

It is clear that the best accuracy achieved is 80 % using VLAD codebook of size 128-D which outperforms all of the results reported in [31] even that they use larger SIFT and BOW codebooks. As aforementioned, the best achieved result of mAP in our representation is 80 % using 32-D SIFT, 32-D HSV, and 128-D VLAD codebook. These accuracy results are reported based on the minimum similarity distance obtained using Euclidean, relative Manhattan, and CityBlock distance measures. In addition, the SIFT–HSV performs better than SIFT-LBP on ZuBuD even that both are increasing

Fig. 13 Sample images of Holidays dataset

Fig. 14 Sample images of ZuBuD dataset

Table 10 Comparisons of mAP obtained on Holidays dataset

Features	mAP (%)
BOW-200K [30]	54.00
FV [30]	59.50
Improved Fisher [52]	62.60
SIFT + VLADintra + innorm [53]	65.30
LCS + RN [54]	65.70
This paper (based on manhattan L_1 *similarity distance)*	
rootSIFT(32-D) + HSV/VLAD (64)	70.63
rootSIFT(32-D) + HSV/VLAD (128)	71.17
rootSIFT(32-D) + HSV/VLAD (256)	72.47
rootSIFT(32-D) + HSV/VLAD (512)	72.21

Table 11 Comparison of mAP obtained using ZuBuD dataset

Features (+size)	Codebook size	mAP (%)
SIFT(128) + CEDD [31]	512	67.26
SURF(128) + SLD [31]	512	79.01
Rnd(600) + CEDD [31]	512	76.75
GaussRnd(600) + CEDD [31]	512	77.29
This paper		
rootSIFT(32) + HSV	32	69.74
rootSIFT(32) + HSV	64	70.96
rootSIFT(32) + HSV	128	80.00
rootSIFT(32) + HSV	256	74.96
rootSIFT(32) + HSV	512	76.87

directly proportional to the size of VLAD codebook. Consequently, the proposed image representation still achieves high and comparable accuracy on ZuBuD.

7 Conclusions

This paper presents a thorough investigation towards finding compact and discriminative image representations using global and local multi-feature scheme. The conducted experiments provide insights into the relationship between image features and other retrieval factors, including distance measures, quantization and visual codebooks, retrieval speed, and memory requirements. A bank of image features is extracted and then formulated into compact image representations. All of the extracted features are evaluated against eight different distance measures for similarity matching. The experimental results show that different image features and combinations provide different performance. At the last evaluation phase, Euclidean, cosine, and correlation measures show almost the same impact on both retrieval accuracy and efficiency. The Spearman distance measure has shown the highest retrieval accuracy for single local descriptors compared to the combined global or local ones. However, it takes more matching time than other distance measures.

Also, the experimental results confirm that adding more features to the image representation does not guarantee gaining more distinctiveness or improving the system performance. The reported retrieval results show that aggregating three features together degrades the system performance by 3–20 % using certain distances, e.g., Euclidean and Chebyshev. Therefore, our CBIR model exploits the strength of image features extracted globally and locally using very low dimensionality. Moreover, it shows a high robustness against two important factors of image retrieval, dimensionality reduction and codebook size. First, the image signature is not largely affected by reducing the dimension of rootSIFT from 128 to only 32 using PCA/whitening. This reduces the vectorization time, retrieval time, and memory size of image signature by 22, 35, and 96 %, respectively.

The model performance is also evaluated using three different quantization methods for local descriptors, i.e., VLAD, FV, and BOW. The system quantized the final image signature (rootSIFT + HSV and rootSIFT + LBP) using only 32-D visual codebook of VLAD, which is another intrinsic utility for more improved performance in terms of retrieval speed and efficiency requirements. In addition, the minimum eigenvalue detector has shown the highest retrieval accuracy for local descriptors over all distance measures. Finally, our work presented in this paper proves that a proper selection and extraction of global and local features with a suitable aggregation scheme can compensate any performance degradation that may occur using different similarity measures, dimensionality reduction, and quantization.

References

1. Alzu'bi A, Amira A, Ramzan N (2015) Semantic content-based image retrieval: a comprehensive study. J Vis Commun Image Represent 32:20–54
2. Li J, Allinson NM (2013) Relevance feedback in content-based image retrieval: a survey. In: Handbook on neural information processing. Springer Berlin, Heidelberg, pp 433–469
3. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv (CSUR) 40(2):5
4. Liu Y, Zhang D, Lu G, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recognit 40(1):262–282
5. Duanmu X (2010) Image retrieval using color moment invariant. In: The seventh international conference on information technology: new generations (ITNG), 12–14, pp 200–203
6. Qiu G (2003) Color image indexing using BTC. IEEE Trans Image Process 12(1):93–101
7. Talib A, Mahmuddin M, Husni H, George LE (2013) A weighted dominant color descriptor for content-based image retrieval. J Vis Commun Image Represent 24(3):345–360
8. Shrivastava N, Tyagi V (2014) Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching. Inf Sci 259:212–224
9. Kwitt R, Uhl A (2010) Lightweight probabilistic texture retrieval. IEEE Trans Image Process 19(1):241–253
10. Lasmar NE, Berthoumieu Y (2014) Gaussian copula multivariate modeling for texture image retrieval using wavelet transforms. IEEE Trans Image Process 23(5):2246–2261
11. Haralick RM (1979) Statistical and structural approaches to texture. Proc IEEE 67(5):786–804
12. Tamura H, Mori S, Yamawaki T (1978) Textural features corresponding to visual perception. IEEE Trans Syst Man Cybern 8(6):460–473
13. Vogel J, Schiele B (2006) Performance evaluation and optimization for content-based image retrieval. Pattern Recognit 39(5):897–909
14. Lee J, Nang J (2011) Content-based image retrieval method using the relative location of multiple ROIs. Adv Electr Comput Eng 11(3):85–90
15. Hoàng N, Gouet-Brunet V, Rukoz M, Manouvrier M (2010) Embedding spatial information into image content description for scene retrieval. Pattern Recognit 43(9):3013–3024
16. Wang S, Liu D, Gu F, Feng Yang HL (2012) Similar matching for images with complex spatial relations. J Comput Inf Syst 8:8727–8734
17. Jaworska T, Kacprzyk J, Marín N, Zadrożny S (2010) On dealing with imprecise information in a content based image retrieval system. In: Computational intelligence for knowledge-based systems design. Springer, Berlin, Heidelberg, pp 149–158
18. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
19. Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359
20. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Computer Society conference on computer vision and pattern recognition, CVPR, vol 1, pp 886–893
21. Rosten E, Drummond T (2006) Machine learning for high-speed corner detection. In: Computer vision-ECCV, pp 430–443

22. Wang XY, Zhang BB, Yang HY (2014) Content-based image retrieval by integrating color and texture features. Multimed Tools Appl 68(3):545–569
23. Liu GH, Zhang L, Hou YK, Li ZY, Yang JY (2010) Image retrieval based on multi-texton histogram. Pattern Recognit 43(7):2380–2389
24. Huang ZC, Chan P, Ng W, Yeung DS (2010) Content-based image retrieval using color moment and Gabor texture feature. In: IEEE international conference on machine learning and cybernetics (ICMLC), vol 2, pp 719–724
25. Harris C, Stephens M (1988) A combined corner and edge detector. In: Alvey vision conference, vol 15, p 50
26. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of British machine vision conference, pp 384–393
27. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: Ninth IEEE ICCV, pp 1470–1477
28. Ke Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of IEEE CVPR, vol 2, p II-506
29. Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: IEEE CVPR'07, pp 1–8
30. Jégou H, Perronnin F, Douze M, Sanchez J, Perez P, Schmid C (2012) Aggregating local image descriptors into compact codes. IEEE Trans Pattern Anal Mach Intell 34(9):1704–1716
31. Iakovidou C, Anagnostopoulos N, Kapoutsis A, Boutalis Y, Lux M, Chatzichristofis SA (2015) Localizing global descriptors for content-based image retrieval. EURASIP J Adv Signal Process 1:1–20
32. ElAlami M (2014) A new matching strategy for content based image retrieval system. Appl Soft Comput 14:407–418
33. Zhang Y, Zhaoxing Z, Han X (2009) Category specific SIFT descriptor and its combination with color information for content-based image retrieval. In: Proceedings of the 2nd ACM international conference on interaction sciences: information technology, culture and human, pp 685–690
34. Deselaers T, Keysers D, Ney H (2008) Features for image retrieval: an experimental comparison. Inf Retr 11(2):77–107
35. Walia E, Verma V (2016) Boosting local texture descriptors with Log-Gabor filters response for improved image retrieval. Int J Multimed Inf Retr 5(4):173–184
36. Bosch A, Zisserman A, Munoz X (2007) Image classification using random forests and ferns. In: IEEE 11th ICCV, pp 1–8
37. Alzu'bi A, Amira A, Ramzan N, Jaber T (2015) Robust fusion of color and local descriptors for image retrieval and classification. In: IEEE international conference on systems, signals and image processing (IWSSIP), pp 253–256
38. Lee TS (1996) Image representation using 2D Gabor wavelets. IEEE Trans Pattern Anal Mach Intell 18(10):959–971
39. Mallat S (1998) A wavelet tour of signal processing. Academic Press, San Diego
40. Costa A F, Humpire-Mamani G,Traina A J (2012) An efficient algorithm for fractal analysis of textures. In: 25th IEEE SIBGRAPI, pp 39–46
41. Ojala T, Pietikäinen M, Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
42. Brahnam S, Jain LC, Nanni L, Lumini A (2014) Local binary patterns: new variants and applications. Springer, Berlin, Heidelberg
43. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
44. Bianconi F, Harvey R, Southam P, Fernández A (2011) Theoretical and experimental comparison of different approaches for color texture classification. J Electron Imaging 20(4):043006–043006
45. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comput Vis 60(1):63–86
46. Wang Z, Fan B, Wu F (2011) Local intensity order pattern for feature description. In: ICCV, pp 603–610
47. Arandjelović R, Zisserman A (2012) Three things everyone should know to improve object retrieval. In: IEEE CVPR, pp 2911–2918
48. Tola E, Lepetit V, Fua P (2010) An efficient dense descriptor applied to wide-baseline stereo. IEEE Trans Pattern Anal Mach Intell 32(5):815–830
49. Huiskes MJ, Thomee B, Lew MS (2010) New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. I:n Multimedia information retrieval, pp 527– 536
50. Vedaldi A, Fulkerson B (2010) VLFeat: an open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM conference on multimedia, pp 1469–1472
51. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search. In: ECCV, pp 304–317
52. Perronnin F, Liu Y, Sánchez J, Poirier H (2010) Large-scale image retrieval with compressed fisher vectors. In: Proceedings of CVPR, pp 3384–3391
53. Arandjelovic R, Zisserman A (2013) All about VLAD. In: CVPR, pp 1578–1585
54. Delhumeau J, Gosselin P H, Jégou H, Pérez P (2013) Revisiting the VLAD image representation. In: ACM multimedia, pp 653–656
55. Shao H, Svoboda T, Van Gool L (2003) Zubud-zurich buildings database for image based recognition. Technical report 260, Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland