Content based image retrieval using fusion of multilevel bag of visual words

Akbar Moghimian¹ · Muharram Mansoorizadeh¹ · MirHossein Dezfoulan¹

Received: 13 September 2019 / Accepted: 26 November 2019 / Published online: 30 November 2019
© Springer Nature Switzerland AG 2019

Abstract

Content based image retrieval (CBIR) is the art of finding visually and conceptually similar pictures to the given query picture. Usually, there is a semantic gap between low-level image features and high-level concepts perceived by viewers. Although features such as intensity and color enforce a good distinction between the images in terms of greater detail, they convey little semantic information. Therefore, employing higher-level features such as properties of regions and objects within the image could improve the retrieval performance. In this study, features are extracted at the pixel, region, object, and concept levels. The fusion step concatenates the four feature vectors and maps it to a lower-dimensional space using auto-encoders. The experiments confirm the efficiency of the proposed method over the individual feature groups and also the state of the art methods.

Keywords CBIR · Feature fusion · Bag of visual words

1 Introduction

Due to the growing interest in digital images and their application, the retrieval and search of images in both areas of research and commercial domains have been increasingly gaining attention. Without using image retrieval systems, it is very hard to search in the large and ever-growing image databases on the Internet. The basic methods of image retrieval are based on the labels and tags that are assigned to images by human experts. These approaches are costly and suffer from inconsistencies of the tags and labels composed by different experts. In addition, some images cannot be easily described by a few sets of keywords. Content-based image retrieval (CBIR) methods try to overcome these problems by learning models that map visual image features to well-defined objects and concepts [1].

Generally, CBIR systems are composed of indexing and searching steps. In the indexing step, feature vectors

representing the visual features of images are extracted and stored in a database. The search step processes a given query image by first extracting the same features as in the first step and looking it up in the database. The search procedure employs various similarity measures to rank and retrieve the most similar images.

The most important challenge in CBIR is the semantic gap which is the difference between the features extracted from an image and its latent high-level concepts. A high-level concept is an understanding that a human has from an image. In fact, two images may be very similar in terms of color and texture, but human perception of them may considerably be different. The semantic gap causes retrieval system to fetch images which are similar in terms of visual features but communicate quite different concepts.

Image feature extraction is done in two global and local methods. The global methods, extract visual features from the whole image without considering the similarity and

✉ Muharram Mansoorizadeh, mansoorm@basu.ac.ir; Akbar Moghimian, akbar.moghimian@gmail.com; MirHossein Dezfoulan, dezfoulan@basu.ac.ir | ¹Bu-Ali Sina University, Hamadan, Iran.

spatial relations of the pixels. Hence, the image is represented by a single feature vector that can be readily used in comparison and ranking procedures. Despite simplicity and robustness of these methods, their performance degrades with changes of light, scale, and viewpoint. Local methods extract features from several regions of the image. These regions can be obtained by a simple partitioning of the image, a color segmentation algorithm, or neighbourhoods around key points. Finally, the image is represented by a set of feature vectors corresponding to the regions. Due to the different number of regions in images, the set of local feature vectors must be further processed to induce a single descriptor for comparison purposes [2].

With the large and diverse number of available feature types, some works try to extract and use multiple groups of features and combine their results at feature or decision levels. The decision fusion strategy independently trains basic learners for each group. Results of the basic learners are then combined using functions such as majority voting [3]. The feature fusion, on the other hand, combines the extracted feature vectors to build a new feature vector. This strategy takes into account the interdependence and complementary role of the different feature groups. The fusion function ranges from a simple concatenation of the vectors to the more advanced belief networks [4].

According to a recent survey by Piras and Giacinto [5], most of the recent works that combine multiple sources of information are mainly based on decision level fusion. The adopted sources are visual image features, accompanied textual information, and relevance feedbacks. The major challenges of feature fusion are the intrinsic differences between the feature groups. Availability, robustness, scale, and dynamics of the pixel-based features are quite different from region-based features. For instance, a feature that is based on pixels (e.g. co-occurrence of pixel colors) is extracted from thousands of pixels which can be safely and robustly modeled by parametric probability distributions. However, a feature that is extracted from multiple regions (e.g. mean area of coherent color segments) is hardly based on more than 10 regions which limits the application of modeling and inferential methods. Furthermore, the different scale and dynamics of the feature groups require complex space transformations [6].

In this study, feature extraction is performed at four levels, namely pixel, region, object, and concept. A feature fusion step concatenates the four feature vectors and maps it to a lower dimensional space using an auto-encoder network. This final feature vector is the basis for similarity estimation and retrieval. As a decision level approach, in our earlier work [7], four basic retrieval systems are built over the four feature groups. Each system reports a ranked list of related images. The lists are then

combined and a final ranking is generated. Besides its simplicity of setup and extension, it ignores feature dependencies and overlaps. In this research, we extend our work in a new dimension that implicitly captures inter-feature information to build a more accurate and robust system.

The present work systematically differs from *deep image learning* and *image hashing* as two currently active topics in the area of image retrieval. At first, we emphasize that while the proposed approach is not a deep-learning method, it benefits from deep learning based image features through the adoption of a successful pre-trained deep neural networks. This aspect of the work can be further explored by using features from more recent deep NNs such as ResNet [8]. Secondly, image hashing techniques try to enhance the search time of the retrieval algorithms. Usually, the search operation is linear in size of the underlying image set. Hence, the amount of required time could be a bottleneck in real time scenarios. Image hashing techniques try to limit the search to small subsets of the images. Results of the recent works are quite promising in terms of time complexity but their accuracy has more ways to go [9].

The experimental results over three public domain datasets confirm that fusion of multiple types of visual words significantly improves the retrieval performance. Furthermore, the similar trend of improvements over the datasets suggests that the improvement should be mainly attributed to the fusion procedure.

The remainder of the paper is arranged as follows: In the next section, the works related to various levels have been introduced and reviewed. The proposed method is explained in detail in the third section. Then, in the fourth section, the adopted datasets, the experiment and implementation details, and the results are presented. Section 5 concludes the paper with a short discussion and a few future directions.

2 Related work

In this section, the related work is studied in four groups corresponding to four mentioned feature types. Furthermore few works that address the fusion of multiple feature groups are also introduced. Finally, the major strategies of information fusion in learning scenarios are also briefly reviewed.

The tasks performed at the pixel level include extracting the global features from all of the image pixels and the local features from small patches around key points. In [10], *directional local extrema patterns* are used in image retrieval. This feature is inspired by the local binary patterns (LBP) and is used to extract edge information in multiple orientations. The histogram of the oriented edges

constitutes the global feature vector for the image. In [11], histograms of HSV color and an LBP inspired texture feature named *local texon xor patterns* are used for retrieval.

In [12] the image has been partitioned into 4×4 blocks and then texture and color features are extracted from the blocks. The blocks are then clustered and the cluster centers are reported as the region features. Similarly, color and texture features in a small neighborhood around pixels are used in [13]. They used Gaussian mixture models for grouping and segmentation of image pixels. In [14] features of salient regions are extracted and used for retrieval. A salient region is an important part of the image that draws attention from other parts of the image.

As an approach that exploits features of the objects in an image, [15] builds a statistical model of an object out of a small set of samples provided by the users. The system finds images containing the object by matching against this model. [16] follows a similar approach by first segmenting an image into regions; hopefully containing a single object. The set of descriptors are then normalized into a fixed length vector and then is classified by an MLP network. The notion of object classifiers is also adopted in [17]. A set of 200 objects, frequently appearing in web images, is selected for classification. A template of each object slides over the image and the response for all the objects are collected and fed to the classifiers. A nice point in this work is using two distinct classifiers for foreground and background objects.

The problem with tagging images using the correspondence between the visual features of the image and the keywords is that a large number of the irrelevant keywords is generated. The purpose of the research in [18] is to prune the unrelated keywords using word similarities extracted from the WordNet. The purpose of the study at [19] is also refining inappropriate labels of images. The tags assigned to images are ranked and tags below a predefined cutoff are ignored. In [20] the final similarity is a combination of multiple decisions based on the specific feature groups extracted from curvelet transform, wavelet transform, and dominant color descriptor. The color and texture features are extracted in HSV space. After getting the similarity value for each group, a linear combination that is learned by particle swarm optimization (PSO) combines the results.

Several recent works use multiple sources of information for the retrieval task. A recent review studies the available works in the common framework of decision fusion. The differences between the works are mainly the type of features used and their fusion function. In [21] a hybrid feature named *correlated primary visual texon histogram feature (CPV-THF)* has been introduced that combines color, texture, spatial, and structural features of the images using their cross-correlations. Finally L_1 distance is used for

comparison and retrieval. Srivastava et al. also combine SURF and LBP features for image classification [22]. Similarly, Mehmood et. al combine SURF and HOG features to improve retrieval performance [23].

In summary, existing works adopt several sources of information for the retrieval task. However, the incremental effect of the sources is mainly studied in terms of the decision fusion. This approach is limited in the sense that it requires multiple passes over the database. On the other hand, a combined feature vector induces a single description for the image such that its contents convey information from multiple sources but needs to search the database once. Another point that is not explored in the existing research is the use of labels coming from image classification. The text adopted in the studied works mainly come from captions and comments along with the image in the web pages. However, the object names and class labels of the images assigned by the intelligent analysis of the image content need more exploration.

3 The proposed approach

We propose a feature fusion approach that combines features at various granularity levels for efficient image retrieval. The general scheme of this method is shown in Fig. 1. The approach is composed of two main modules. The *Index* module extracts several types of the features from the images and induces a unique representation through the application of autoencoder neural networks.

Fig. 1 Block diagram of the proposed approach

The representations are stored in a database for the later references by the search module.

The *Search* module gets a query image and retrieves similar images from the database. It follows the same procedure for feature extraction and fusion as in the index module. Several metrics can be used to retrieve the most similar images. We have selected *cosine similarity*, since it performed better in the early experiments. In the following subsections, details of the steps are presented.

3.1 Feature extraction and normalization

As depicted in Fig. 2, the features of an image are extracted in 4 levels of pixels, regions, objects, and concepts. Each type of the extracted features are then independently normalized to [0, 1]. The features are concatenated to constitute a single vector. An auto-encoder neural network

non-linearly maps the vector into a lower dimensional space, resulting in the final feature vector of the image. The auto-encoders have been proved to be more efficient than classic methods [24]; making them an ideal tool for feature extraction. The auto-encoder is trained using the training images. The number of new dimensions, i.e. the size of the auto-encoder's hidden layer, is empirically determined.

3.1.1 Pixel level

In this level, initial moments of color, Gabor filter responses and SIFT features are extracted as local descriptors. A code-book is built using k-means clustering algorithm; in which cluster centers are stored as the visual words. The image is then represented as a bag of these words.

Following the approach of [25], to extract color moments, the image is partitioned into 16×16 blocks and

Fig. 2 Feature extraction and fusion

the mean, variance, and skewness of the HSV components make the nine element feature vector for each block.

Gabor features [14] are extracted by applying a bank of filters with five scales and eight directions to the gray-scale images. Each response image is partitioned into 16×16 blocks and then the mean and variance of the corresponding blocks in the 40 images are extracted to build up an 80-element feature vector for each block.

To extract SIFT features, the key points of the image are obtained as the extrema of the Gaussian differences. For each key point, its 16×16 neighbourhood, is partitioned into 4×4 blocks. Histogram of the oriented gradients at the blocks constitute the 128-element SIFT descriptor.

The features extracted from all the training images are described by a bag of visual words model. In this model, a vocabulary is generated from each type of the features using the K-means clustering algorithm. The size of the vocabulary, N , is empirically determined, as in [26]. Here we assign $N / 3$ words to each feature type.

3.1.2 Region level

Similar to the pixel level, a BoVW model is used to describe images at the region level. In this level, the Hue statistics and LBP descriptors are extracted from the segments or regions of the image. Therefore, the image must be partitioned to its constituent regions. For this part, we adopt JSEG method presented in [27].

Each region is divided into 10×10 block and the 64-elements hue descriptor is extracted from each block [28]. The mean of the features for each region is used as its final feature vector. The histogram of LBP responses at a region is used as its LBP descriptor [29]. The visual vocabulary at region levels contains V words. It is equally divided between the Hue and LBP features.

3.1.3 Object level

According to a study of the distribution of images on the web, a large number of the images contain a small group of well-known objects [17]. At the object level, a set of 150 binary support vector machine (SVM) classifiers are trained to detect 150 most frequent objects of this study. SVMs are well-known robust classifiers that induce a maximum decision margin among the compared classes [30]. However, tuning their hyper-parameters requires a substantial amount of try and error efforts. Furthermore, it is hard to develop a probabilistic interpretation of their outputs. These drawback limit the application of SVMs, specially, for small training samples. However, in the case of CBIR systems with sufficiently large datasets, SVMs outperform common classifiers such as regression models and decision trees. Neural networks could be considered as

Fig. 3 AlexNet based image mapping and feature extraction for object classification

Fig. 4 Object-level feature extraction

alternatives for SVM classifiers but they are more memory intensive and require a large amount of training time.

Following the approach of [31] to train the classifiers, a set of 100 images containing the desired object is selected as the positive sample and another set of 100 images without the object (e.g. images of the other object types) is selected as the negative sample.

The images are fed to the AlexNet [32] and the output of the second fully-connected layer is selected as the 4096-element feature vector (Fig. 3). For each class, an SVM is trained using the extracted features. Each of the classifiers predicts the respective label, resulting in a 150-element binary vector of the objects. (Fig. 4). <https://www.overleaf.com/project/5b2e058c5b43cf4aa2ba62c4> AlexNet is one of the well-known convolutional neural networks that has been trained using millions of ImageNet images and showed great successes in various occasions.

3.1.4 Concept level

The previous step labels objects within the image with the natural language words; enabling us to represent the image as a short text document. Therefore, the document can be further processed by the text analysis methods. Word2vec is a method to embed words in a numerical vector space [33] in which the semantic similarity of the words is correlated to the distance of their respective

vector representations. As an example of word2vec application, consider the case of two images that one contains a tiger and the other contains a lion (Fig. 5). Based on the visual features, these images are quite different but their word2vec representations are very close to each other.

We use the word2vec version that is trained on Google News dataset [34]. It maps the words to a 300-element vector space. The final feature vector for the concept level is the mean of the vectors corresponding to the detected objects.

3.2 Feature fusion

The fusion step combines the four groups of features and generates a new feature vector using an autoencoder network. First, each feature is linearly mapped to the range $[0, 1]$ using Eq. 1.

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Then, the autoencoder network (Fig. 6) with sigmoidal activation function, maps the input into a lower dimensional space. The new features, y_i 's, are defined using Eq. 2. Here, w 's are network weights.

$$y_i = \frac{1}{1 + \exp(-\sum_j w_{ji} x_j)} \quad (2)$$

The training procedure adjusts the network weights such that $\hat{x}_i$'s are close approximations to the x_i 's. After training, y_i 's are extracted as the embedding of x_i 's.

4 Experimental results

4.1 Data and setup

In order to evaluate the efficiency of the proposed approach, we have conducted experiments using three sets of images. The first one, Wang [12], contains 1000 images in 10 different classes. Corel9C is the set of images that have been used in [14]. It contains 900 images in nine classes. Corel5K [35] contains 5000 images in 50 classes.

Fig. 5 Conceptually similar but visually different images. Visual similarity of the images **a**, **b**, and **a**, **c** is about 20% while the conceptual similarity of their corresponding object names, i.e. words lion and tiger is about 60%

Fig. 6 The autoencoder network

For all the three datasets, each class contains 100 images of which 10 images are randomly selected as query images.

For a given query image, I , assume that S is the set of semantically related images to I and R is the set of retrieved images. Also, assume that the subset Q of R contains retrieved related images. That is, the subset $R - Q$ contains irrelevant images that are falsely retrieved in response to searching for I . precision, p , and recall, r are defined as below. $|\cdot|$ denotes the size of the set.

$$p = \frac{|Q|}{|R|} \quad \text{and} \quad r = \frac{|Q|}{|S|} \quad (3)$$

Usually Precision and recall are estimated using top K results. For robust evaluation, the results are aggregated over several queries as follows:

$$AP(Q) = \frac{1}{M} \sum_{j=1}^M P(q_j) \quad (4)$$

$$AR(Q) = \frac{1}{M} \sum_{j=1}^M R(q_j) \quad (5)$$

where $Q = q_1, q_2, q_3, \dots, q_M$ is the set of query images and M is its size. The mean precision and recall for K are denoted by $AP@K$ and $AR@K$, respectively.

The precision measures the purity of the retrieved set. Recall on the other hand determines which fraction of the relevant images contained in the database are retrieved. The precision-recall, $(P - R)$, curve denotes the balance between the precision and the recall in different

Table 1 The average precision of the candidate similarity functions

Function	Wang	Corel9C	Corel5K
Manhattan	72.2	50.3	40.2
Euclidean	69.6	47.5	40.6
Cosine	75.7	66.8	47.1
Chebyshev	62.4	43.1	35.3
Canberra	68.2	61.5	36.2

Fig. 7 Average retrieval precision using various feature vector length

thresholds. The higher the area under the curve, the higher the precision and recall. A perfect retrieval system is expected to retrieve the images of the same class as the query in the top ranks.

The number of visual words, i.e. the vocabulary size, for each feature group in pixel and region levels is

experimentally selected to be 250. Thus, the pixel level retrieval system includes 750 words (for three groups) and region level contains 500 words (for two groups). For each group, a feature vector is assigned to its three nearest clusters. That is, each feature vector is represented by three visual words. This number is also selected based on the experimental observations

As discussed earlier, different metrics can be used to evaluate the similarity of image descriptors [36]. Table 1 shows the average accuracy obtained for several metrics. According to its higher results, cosine similarity is selected for the experiments.

Another parameter is the size of the fused feature vector, which is equal to the number of hidden layer neurons in the auto-encoder network. Figure 7 reports the effect of this parameter on average retrieval precision. Based on the results the size of the fused feature vector is selected to be 200.

4.2 Fusion results

An immediate observation is the performance of the individual levels and the fused system. Figure 8 shows two precision/recall curves for Wang and Corel5K datasets. While the results on Wang are much better than Corel5K, the relative order of the systems is quite similar. Region level retrieval has the lowest performance which may be attributed to the image segmentation inconsistencies. A segmentation algorithm may falsely partition a large segment or merge small and distinct segments, resulting in significant changes in the statistical features.

Fig. 8 Comparison of the individual levels and the fused system in Wang and Corel5K datasets

Table 2 Incremental effect of multilevel features

	Wang	Corel9C	Corel5K
Pixel	46.71	39.75	21.47
+Region	60.08	44.22	27.37
+Object	71.71	63.02	44.35
+Concept	75.66	66.78	47.05

Fig. 9 Incremental effect of multilevel features

Pixel level retrieval has slightly better results but it also suffers from misleading regional information such as large and uniform backgrounds. Object and concept level retrieval tend to be close to human perception and got better results. Finally, the system that is based on fused features performed considerably better than all the individual features.

To study the incremental effects of different levels, at first, the experiments are conducted by using features of the pixel level and subsequently included region, object and concept levels. Table 2 summarizes the results, where adding each level constantly improves the precision as expected. The nice point is that while rates for the databases are quite different, their improvements follow a similar trends, as depicted in the Fig. 9 with over-plotted trend lines.

in Fig. 10 a sample of retrieved images is shown for 6 query images. The left column contains original query images. First two rows are from the Wang database, the two subsequent rows are selected from Corel9C and the last two rows are selected from Corel5K.

4.3 Comparison with other systems

In this section, the fused system is compared to the selected similar systems. Table 3 shows the average precision and recall for 10, 20, 50, and 90 retrieved images.

Fig. 10 An example of image retrieval using the proposed method. First two rows are from Wang, two middle rows are from Corel9C and two last rows are from the Corel5K dataset. The first image

of each row is the query image and the remaining are the ranked results. Pictures with dark borders are irrelevant to the query

Table 3 Average precision and recall for the fused system on the three datasets

K	Average precision				Average recall			
	10	20	50	90	10	20	50	90
Wang	87.8	86.4	82.2	75.7	9.8	19.2	45.7	75.7
Corel9C	89.9	84.0	78.0	66.8	9.9	18.7	43.3	66.8
Corel5K	68.0	63.2	55.0	47.1	7.6	14.0	30.6	47.1

Table 4 Comparison of the proposed method with representative works

	Wang				Corel5K		Corel9C	
	AP	AP@10	AP@12	AP@20	AP@10	AP@12	AP@10	AP@20
DLEP [10]		74.8			48.8			
LTXP [11]		79.8			60.0			
PSO [20]				76.5				
CPV-THF [21]			80.8		63.9			
Simplicity [12]	47.2							
ESR [14]							88.2	78.7
DLF [7]	71							
Proposed	75.7	87.8	87.4	86.4	68.0	66.6	89.9	84.0

Table 4 shows the performance of the studied systems, as reported in the respective references. The table is sparse, due to the differences in datasets and evaluations. As stated in Section 2, [10] uses directional local extrema patterns (DLEP), and [11] defines and uses local texton XOR patterns (LTXP). Simplicity [12] is one of the fundamental works that is used for comparison in most of the earlier works. In [14], an extension of salient regions (ESR) has been adopted. Decision level fusion (DLF) is our earlier work that combines the ranked lists of individual features. The results confirm that the proposed fusion system outperformed the other systems.

5 Conclusion and future directions

In this paper, an image retrieval system is proposed that combines different types of image features. The pixel level extracts basic visual features of the image and has a less conceptual load. The region level tries to draw closer to the human visual system by segmenting the image and extracting features from its homogeneous regions. The object level tries to reduce the semantic gap by classifying and labeling objects within the image. The vector space modeling of the object names is further exploited to find similar concepts. The fusion system proposed in this paper incorporates the mentioned feature types in a single combined feature vector. The adopted autoencoder network induces a lower dimensional but informative feature vector.

The experimental results show the proper performance of the proposed method on three datasets including 900, 1000, and 5000 images. The results of the proposed method at the concept level indicate that the conceptual correlation of words corresponding to objects and the scenes within the images can be used as a measure of their similarity. The proposed method in this study has acquired an average precision of about 88% and 89% for the images of Wang and Corel9C and 68% for Corel5K, which is competitive to existing works.

In the future, we plan to explore this study in several directions. The first work is to analyze feature correlations and their importance in the constituting final feature vector. Another point is the direct inclusion of AlexNet's features into the fusion system. Finally, the application areas of the proposed system will be explored.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

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. 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
3. Kuncheva LI (2014) Combining pattern classifiers: methods and algorithms, 2nd edn. Wiley, Hoboken
4. Bloch I (2013) Information fusion in signal and image processing: major probabilistic and non-probabilistic numerical approaches. Wiley, Hoboken
5. Piras L, Giacinto G (2017) Information fusion in content based image retrieval: a comprehensive overview. *Inf Fusion* 37:50–60
6. Mansoorizadeh M, Charkari NM (2010) Multimodal information fusion application to human emotion recognition from face and speech. *Multimed Tools Appl* 49(2):277–297
7. Moghimian A, Mansoorizadeh M, Dezfoulian MH (2018) Content based image retrieval using decision fusion of multilevel bag of words model. *Tabriz J Electr Eng* 50(4):1–13
8. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
9. Liu H, Wang R, Shan S, Chen X (2016) Deep supervised hashing for fast image retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p 2064–2072
10. Murala S, Maheshwari RP, Balasubramanian R (2012) Directional local extrema patterns: a new descriptor for content based image retrieval. *Int J Multimed Inf Retr* 1(3):191–203
11. Bala A, Kaur T (2016) Local texton XOR patterns: a new feature descriptor for content-based image retrieval. *Eng Sci Technol Int J* 19(1):101–112
12. Wang JZ, Li J, Wiederhold G (2001) Simplicity: semantics-sensitive integrated matching for picture libraries. *IEEE Trans Pattern Anal Mach Intell* 23(9):947–963
13. Carson C, Belongie S, Greenspan H, Malik J (2002) Blobworld: image segmentation using expectation-maximization and its application to image querying. *IEEE Trans Pattern Anal Mach Intell* 24(8):1026–1038
14. Zhang J, Feng S, Li D, Gao Y, Chen Z, Yuan Y (2017) Image retrieval using the extended salient region. *Inf Sci* 399:154–182
15. Hoiem D, Sukthankar R, Schneiderman H, Huston L (2004) Object-based image retrieval using the statistical structure of images. In: null. IEEE, p 490–497
16. Li Y (2005) Object and concept recognition for content-based image retrieval. Citeseer
17. Li LJ, Su H, Fei-Fei L, Xing EP (2010) Object bank: a high-level image representation for scene classification & semantic feature sparsification. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A (eds) Advances in neural information processing systems, p 1378–1386
18. Jin Y, Khan L, Wang L, Awad M (2005) Image annotations by combining multiple evidence & wordnet. In: Proceedings of the 13th annual ACM international conference on Multimedia. ACM, p 706–715
19. Wang C, Jing F, Zhang L, Zhang HJ (2007) Content-based image annotation refinement. In: IEEE conference on computer vision and pattern recognition, CVPR'07. IEEE, p 1–8
20. Fadaei S, Amirfattahi R, Ahmadzadeh MR (2016) A new content-based image retrieval system based on optimised integration of DCD, wavelet and curvelet features. *IET Image Process* 11(2):89–98
21. Raza A, Dawood H, Dawood H, Shabbir S, Mehboob Rubab, Banjar Ameen (2018) Correlated primary visual texton histogram features for content base image retrieval. *IEEE Access* 6:46595–46616
22. Srivastava D, Bakthula R, Agarwal S (2019) Image classification using SURF and bag of LBP features constructed by clustering with fixed centers. *Multimed Tools Appl* 78(11):14129–14153
23. Mehmood Z, Abbas F, Mahmood T, Javid MA, Rehman Amjad, Nawaz Tabassam (2018) Content-based image retrieval based on visual words fusion versus features fusion of local and global features. *Arab J Sci Eng* 43(12):7265–7284
24. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507
25. Long F, Zhang H, Feng DD (2003) Fundamentals of content-based image retrieval. In: Feng D, Siu WC, Zhang HJ (eds) Multimedia information retrieval and management. Springer, Berlin, pp 1–26
26. Jing Y, Qin Z, Wan T, Zhang X (2013) Feature integration analysis of bag-of-features model for image retrieval. *Neurocomputing* 120:355–364
27. Deng Y, Manjunath BS (2001) Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans Pattern Anal Mach Intell* 23(8):800–810
28. Van De Weijer J, Schmid C (2006) Coloring local feature extraction. In: European conference on computer vision. Springer, Berlin, p 334–348
29. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24(7):971–987
30. Zou Z, Shi Z, Guo Y, Ye J (2016) Object detection in 20 years: a survey. *arXiv preprint [arXiv:1905.05055](https://arxiv.org/abs/1905.05055)*
31. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)*
32. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems. p 1097–1105
33. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems. p 3111–3119
34. Goldberg Y, Levy O (2014) word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint [arXiv:1402.3722](https://arxiv.org/abs/1402.3722)*
35. Liu GH, Yang JY, Li ZY (2015) Content-based image retrieval using computational visual attention model. *Pattern Recognit* 48(8):2554–2566
36. Cha Sung-Hyuk (2007) Comprehensive survey on distance/similarity measures between probability density functions. *City* 1(2):1

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