Cogent Engineering

ISSN: 2331-1916 (Online) Journal homepage: www.tandfonline.com/journals/oaen20

Content-based image retrieval: A review of recent trends

Ibtihaal M. Hameed, Sadiq H. Abdulhussain & Basheera M. Mahmmod |

To cite this article: Ibtihaal M. Hameed, Sadiq H. Abdulhussain & Basheera M. Mahmmod | (2021) Content-based image retrieval: A review of recent trends, Cogent Engineering, 8:1, 1927469, DOI: 10.1080/23311916.2021.1927469

To link to this article: https://doi.org/10.1080/23311916.2021.1927469

Published online: 02 Jun 2021.

Hameed et al., *Cogent Engineering* (2021), 8: 1927469
https://doi.org/10.1080/23311916.2021.1927469

cogent engineering

COMPUTER SCIENCE | REVIEW ARTICLE

Content-based image retrieval: A review of recent trends

Ibtihaal M. Hameed[1], Sadiq H. Abdulhussain[1]* and Basheera M. Mahmmod[1]

Received: 18 January 2021
Accepted: 29 April 2021

*Corresponding author: Sadiq H. Abdulhussain, Department of Computer Engineering, University of Baghdad, Baghdad 10071, Iraq
E-mail: sadiqh76@yahoo.com

Reviewing editor:
D T Pham, School of Mechanical Engineering, University of Birmingham, UNITED KINGDOM

Additional information is available at the end of the article

Abstract: With the availability of internet technology and the low-cost of digital image sensor, enormous amount of image databases have been created in different kind of applications. These image databases increase the demand to develop efficient image retrieval search methods that meet user requirements. Great attention and efforts have been devoted to improve content-based image retrieval method with a particular focus on reducing the semantic gap between low-level features and human visual perceptions. Due to the increasing research in this field, this paper surveys, analyses and compares the current state-of-the-art methodologies over the last six years in the CBIR field. This paper also provides an overview of CBIR framework, recent low-level feature extraction methods, machine learning algorithms, similarity measures, and a performance evaluation to inspire further research efforts.

Ibtihaal M. Hameed

ABOUT THE AUTHORS

Ibtihaal Monim Hameed was born in Baghdad, Iraq, in 1983. She received the B.Sc. in computer engineering from the University of Baghdad, in 2005. She is currently pursuing an M.Sc. degree with the Department of Computer Engineering, University of Baghdad. Her research interests include computer vision and Image processing.

Sadiq H. Abdulhussain was born in Baghdad, Iraq, in 1976. He received the B.Sc. and M.Sc. degrees in electrical engineering from Baghdad University, in 1998 and 2001, respectively. Received the Ph.D. degree from Universiti Putra Malaysia 2018. Since 2005, he has been a staff member with the Computer Engineering Department, Faculty of Engineering, University of Baghdad. His research interests include computer vision, signal processing, as well as speech and image processing.

Basheera M. Mahmmod was born in Baghdad, Iraq, in 1975. She received the B.Sc. degree in electrical engineering from Baghdad University, in 1998, the master's degree in electronics and communication engineering from Baghdad University, in 2012, and the Ph.D. degree in computer and embedded system engineering from Universiti Putra Malaysia, in 2018. Since 2007, she has been a Staff Member with the Department of Computer Engineering, Faculty of Engineering, University of Baghdad. Her research interests include speech enhancement, signal processing, computer vision, RFID, and cryptography.

PUBLIC INTEREST STATEMENT

Content-based image retrieval appears to overcome the disadvantages of the text-based image retrieval methods (searching image databases through the use of Keywords). CBIR is used to search in an image database to return similar visual content images to a specified query image. This method is fully automated. However, it suffers from "semantic gap", which is the gap between the low-level features that describes images and the high-level concepts (perception) contained in the images, leading to irrelevant image retrieval. Over the past three decades, this gap has been the focus of numerous researches. Due to the increasing research in this field, this paper surveys, analyses and compares the current state-of-the-art methodologies over the last six years in the CBIR field. This paper also provides an overview of CBIR framework, recent low-level feature extraction methods, machine learning algorithms, similarity measures, and a performance evaluation to inspire further research efforts.

Subjects: Artificial Intelligence; Computer Science; General Databases;

Keywords: content-based image retrieval; image feature; feature extraction; similarity measure; machine learning

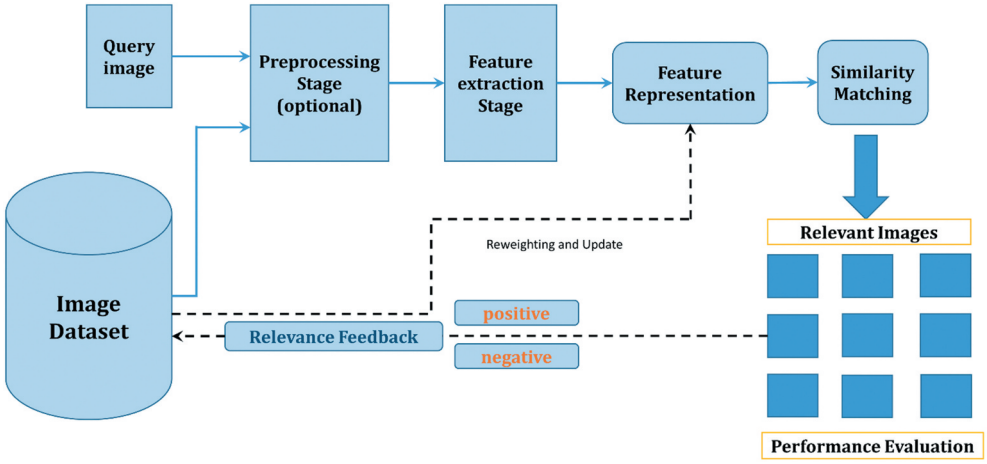
1. Introduction

A massive amount of image databases has been generated by educational, industrial, medical, social, and other life facilities. All these image repositories require a powerful image search mechanism. There are two common search methods. The first method is based on keywords used to annotate images, which is known as text-based image retrieval (Y. Liu et al., 2007). This method suffers from many disadvantages: 1) manually annotating large databases is not feasible, 2) the end user must make annotations, which in turns makes this method subject to human perception, and 3) these annotations are applicable for only one language. The second method is "content-based image retrieval" (CBIR), which highly recommended to overcome the disadvantages of the text-based image retrieval methods (Raghunathan & Acton, 1999).

CBIR is used to search in an image database to retrieve similar visual content images to a specified query image (Jenni et al., 2015) (Ali et al., 2020). This method is fully automated. However, it suffers from "semantic gap", which is the gap between the low-level features that describes images and the high-level concepts (perception) contained in the images (Bai et al., 2018), leading to irrelevant image retrieval. Over the past three decades, this gap has been the focus of numerous studies (Shrivastava & Tyagi, 2017). Many methods have been invented to convert high-level concepts in images to features. These features are the foundation for CBIR. Generally, features are categorized into global features and local features depending on the feature extraction methods. Global features (i.e., color, texture, shape, and spatial information) give a representation for the entire image. They have the advantage of being faster in feature extraction and similarity computations (Datta et al., 2008). On the other hand, they fail to differentiate between the background and the object in the image (different image parts). This makes them unsuitable for retrieval in complex scenes or object recognition (Halawani et al., 2006), but they are appropriate for object classification and detection (Ghrabat et al., 2019). As a comparison to global feature, local features are suitable for image retrieval, matching tasks and recognitions (Halawani et al., 2006). "Object recognition is the task of recognizing the object and labeling the object in an image" (Bansal et al., 2020) while object detection is concerned with the existence of an object that belong to a predefined class in the image and its location (Mittal et al., 2019). Therefore, classification is a sub-task of object detection (Mittal et al., 2019). Local features are defined as the key points or some parts of images for instance, corners, blobs and edges. They are robust to scale, rotation, translation, changes to backgrounds, clutter and partial occlusions (Halawani et al., 2006).

Feature extraction is the first process in CBIR that aims to convert human perception into a numerical description that can be manipulated by machines. The accuracy of the retrieved images is greatly influenced by the features that are extracted (Piras & Giacinto, 2017). However, this selection is based on user requirements. Feeding extracted features to machine learning algorithms (supervised or unsupervised) can improve the CBIR performance (D. Zhang et al., 2012). The trends of recent image retrieval research concentrate on the use of deep learning to improve accuracy at the cost of increasing running time (Markowska-Kaczmar & Kwaśnicka, 2018). Another problem that has a negative effect on CBIR performance (i.e., memory usage, scalability, speed, accuracy) is the high-dimensional features that are usually generated when trying to translate visual image content to a numerical feature form. These high-dimensional feature representations, known as the "curse of dimensionality", usually have a sparsely distributed nature (Zhuo et al., 2014). " Dimensionality reduction" is a solution to this problem (Zhuo et al., 2014). In the literature, a number of comprehensive studies discussed many proposed methods for dimensionality reduction (Zhuo et al., 2014), (Perronnin et al., 2012). Similarity measure is another vital process that has an impact on CBIR performance. Since this measurement

Figure 1. General framework of the CBIR system.



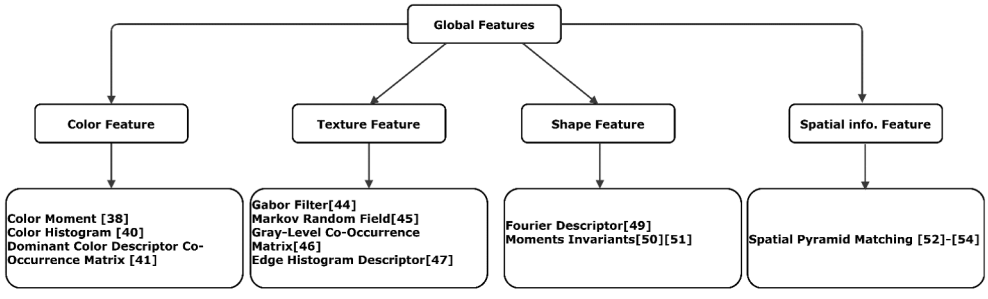
is determined by the layout of the feature vector, choosing an inappropriate measure would result in less identical images being returned, lowering the CBIR system's accuracy. In other words, with the use of suitable similarity measure, high accuracy can be achieved. Since a number of datasets have been used in CBIR frameworks, certain metrics such as precision, recall, and running time are typically used to determine CBIR effectiveness, which are influenced by the image dataset selection (J. Z. Wang et al., 2000; Jain et al., 2011; Oliva and Torralba, 2001; Srivastava and Khare, 2017; Das and Thepade, 2016; Yang et al., 2018; Fei-Fei et al., 2007; Griffin et al., 2007; Ashraf et al., 2020) (more details in Section 7).

An increasing number of studies are conducted in the CBIR domain, which includes many new directions. Many surveys and studies report major challenges and review the existing state-of-the-art (Alzu'bi et al., 2015; Tarawneh et al., 2020; Latif et al., 2019; Ghosh et al., 2018; Tian, 2018). This paper aims to answer the following questions:

- How does feature fusing help in reducing the semantic gap and improve retrieval accuracy?
- How does the usage of latest machine learning algorithms positively affect system accuracy and how do these algorithms affect computational cost and memory usage?
- How does the use of robust measurements and right dataset enhance system performance in terms of specificity, memory usage and computational cost?
- What are the possible future research trends?

Section 2 provides a brief description of the CBIR framework. Section 3 reviews the most recent studies on CBIR in the last six years depending on the image retrieval methods. Section 4 explores the most recent machine learning algorithms deployed in this field, while Section 5 provides

Figure 2. Classification of global features.



a comparison of different CBIR techniques. Similarity measures are the major topic of section 6. Section 7 presents a detailed review of the image datasets used to evaluate performance. Section 8 is devoted to exploring measures used for performance evaluation. Section 9 addresses the most crucial issues in CBIR along with the most important research directions from our point of view. Finally, Section 10 concludes this paper.

2. CBIR framework

The general CBIR framework consists of some mandatory stages and other optional stages, as shown in Figure 1. The first stage in CBIR is the submission of the query image by the user. All the applied processes to the query image will be applied to all of the images in the database and in the same order (Kokare et al., 2002). Usually, these processes are performed on the query image upon user submission and are called online processes; same processes can be applied to dataset images prior to query submission and are called offline processes. An optional preprocessing stage might be included in the architecture of the framework, which could include resizing, segmentation, denoising, and rescaling, etc. This optional stage is followed by the feature extraction stage, which is the most important stage, in which a visual concept is converted to a numerical form. Extracted features could be in the form of low-level features (i.e., color, shape, texture, and spatial information) or local descriptors. Another optional preprocessing stage after feature extraction is normalization or classification. The final stage is the similarity measurement between the extracted features from the query image and all other images in the dataset to retrieve the most relevant images. Relevance feedback is another possible stage that enhances the results through user intervention by deciding relevant and irrelevant returned images. Many techniques have been proposed to apply relevance feedback to enhance the performance of CBIR (Ciocca & Schettini, 1999; Dacheng et al., 2006; Su et al., 2011; Banerjee et al., 2018; Baig et al., 2020).

3. Low-level features

As mentioned earlier, feature extraction and selection that represent the semantic content of images play the main role in CBIR. These features can be divided into global features (i.e., color, texture, shape, and spatial information), which describe the entire image, and local features, which are usually acquired through dividing the image into segments or through the calculation of some key points, such as corners, blobs and edges. Local features are invariant against scale, translation and rotation changes (Low, 2004). Both types will be explained in this section along with a review of recent studies to demonstrate the significance of the selected features and their impact on system performance.

3.1. Global features

Color, texture, shape, and spatial information are the widely used features in image retrieval tasks. These features will be briefly discussed in this subsection. Figure 2 illustrates the classification of global features along with some feature extraction methods.

3.1.1. Color feature

Since the human eye can distinguish visuals based on their colors, the color feature is considered one of the most significant features that are commonly used by researchers. Color features are calculated according to color spaces. The mostly used color spaces in the CBIR domain are HSV (LSV), YCbCr, RGB, and LAB. These color spaces are characterized using color moments (Duanmu, 2010), color correlogram (Huang et al., 1997), color histogram (Flickner et al., 1995), dominant color descriptor color co-occurrence matrix (Qiu, 2003), and many other descriptors. Color features are considered a robust feature because they are invariant against translation, rotation, and scale change (Shrivastava & Tyagi, 2015). However, they have a spatial information constraints, necessitating the use of other descriptors cope with this drawback (Alzu'bi et al., 2015).

3.1.2. Texture

Textures are noticeable patterns that cannot stand alone as a single intensity or color. Texture is considered a pivotal feature in computer vision (CV) because of its existence in many real images,

which is why it is usually used in image retrieval and pattern recognition. The main drawbacks of texture-based image retrieval are its computation complexity and noise sensitivity (Alzu'bi et al., 2015) (D. P. Tian, 2013). Various algorithms are used for texture analysis, such as Gabor filter (Manjunath et al., 2001), Markov random field (Cross & Jain, 1983), wavelet transform, steerable pyramid decomposition, gray-level co-occurrence matrix (GLCM) (Hawlick, 2017), and edge histogram descriptor (EHD) (Won et al., 2002).

3.1.3. Shape

Shape is one of the low-level features that identify objects. Shape extraction can be performed on the basis of a region or a boundary (contour) (D. P. Tian, 2013)(Zhang & Lu, 2004). In the region-based approach, extraction is performed within the entire region, whereas in the boundary-based approach, extraction is performed on the boundary of the region. Many methods such as Fourier descriptor (Zhang & Lu, 2001) and moment invariants (Marxsen et al., 1993), (Suk & Flusser, 2011) are used to extract shape features. In general, shape descriptors are variant to scale and translation. Therefore, they are usually merged with other descriptors to increase accuracy.

3.1.4. Spatial information

Most of the previously discussed methods of extracting low-level features suffer from the lack of spatial information. Spatial feature is basically related to the objects' location in a two-dimensional image. For instance, two different regions with two different spatial contents in the same image may have an equal histogram. Spatial information usually suffers from computational complexity. Spatial pyramid matching is one of the best methods that capture the spatial attributes of images (Lazebnik et al., 2006; Naeem et al., 2017; Mehmood et al., 2016).

Early studies in the CBIR domain used one feature to retrieve images. However, findings were unsatisfactory because images usually contain several pictorial characteristics (Chun et al., 2008). To achieve better retrieval performance and accuracy, researchers used to fuse two or more features in a process that is usually known as feature fusion or feature combination. Shrivastava et al. (Shrivastava & Tyagi, 2015) proposed a CBIR system that has a feed forward architecture with three stages to retrieve images. In the first step, based on color features, appropriate N images were retrieved from images in the dataset, which included M images. The authors used color histogram to calculate the color features. In the next stage, P relevant images are obtained from the N image subset based on texture feature by using Gabor filter to calculate the texture feature. In the final stage, the Fourier descriptor is calculated as shape features to retrieve K relevant images from P. In this manner, relevance feedback can be applied by varying the values of N, K, and P in order to improve precision. The system was tested against two datasets, namely, Corel and CIFAR, with average precision values of 0.7690 and 0.8590, respectively. However, the proposed system does not include a stage for classifying images based on spatial information.

Younus et al. (2015) constructed a new CBIR system that depends on extracting color and texture features. Four extraction methods color moment, color histogram, wavelet moment, and co-occurrence matrices—were used. The authors in Younus et al., 2015 combined k-mean clustering algorithm with particle swarm optimization (PSO), which is a stochastic technique. Tests were conducted by using WANG datasets, which contain 1,000 images divided into 10 classes. Because of incorrect clustering, precision is improved for all classes except architecture and bus as compared to other state-of-the-art techniques. Moreover, the proposed technique did not consider shape feature when computing similarity distance.

Anandh et al. (Ponomarev et al., 2016) presented a novel CBIR system based on the integration of color, texture, and shape. Color auto correlogram, Gabor transform, and wavelet transform were used to extract color, shape, and texture, respectively. The authors used Manhattan distance as a similarity measure between the query image and the dataset images. The achieved average precision values were 0.8300, 0.8800, and 0.7000 for the Corel, Li, and Caltech 101 datasets, respectively. The main drawback of the system is the increased computational complexity because

of the integration of multiple features. Image analysis at a single resolution level may lose some valuable details. Therefore, Srivastava and Khare (Srivastava & Khare, 2017) developed a novel multi-resolution analysis algorithm that analyzes images at multiple levels, with other levels capturing information that one level skipped. This approach is based on the extraction of texture and shape features by using the local binary pattern (LBP) descriptor to extract texture features and Legendre moments to extract shape features from the texture features at multi-resolution levels. Although LBP is used to extract local features, it also creates an influential feature vector when local features are combined with global features. Their technique was tested against five datasets, achieving improved accuracy and sensitivity but with increased computational cost due to the use of multi-resolution analysis.

Sajjad et al (2018) proposed an invariant CBIR system to texture rotation and color change. The proposed system based on concatenating color and texture features to form a feature vector with a size of 360. To extract color features, images are converted to HSV color space and quantized through color histogram. To be invariant to illumination change, only Hue and Saturation channels are utilized. Rotated local binary pattern (RLBP) are used to extract rotation invariant texture features. The proposed system is evaluated through experiments on Zurich Building (ZB), Corel 1 K and Corel 10 K.

In X. Zheng et al., 2016, authors proposed a CBIR approach based on block processing with overlapping. Firstly, images are transformed to HSI color space then are divided to blocks and the main block is selected. Histogram projection is used to extract color features and Roberts Edge detection is used as texture extraction method. The authors used weighted Euclidean distance as a similarity metric to return similar images and the weights were chosen on experimental basis. The proposed approach has a low accuracy value when compared with other state-of-the-art methods (Tadi Bani & Fekri-Ershad, 2019). A novel CBIR approach is presented by combining color, shape and texture features in Z. Zhao et al., 2016. Color distribution entropy (CDE) was used to extract color features while Hue Moments was used to extract shape features. To extract texture features color level co-occurrence matrix (CLCM) was used. For similarity measurements between query image and dataset images, the authors used weighted normalized similarity measure and the weights were decided upon user's experience. Despite the fact that the proposed system achieves high precision value, the performance of the system is affected when the query image contains more objects (complex). This may be because of using Hue Moments to extract shape features which sometimes does not have the ability to recognize images containing more objects or considers different edges as one edge. The authors in Phadikar et al., 2018 proposed a CBIR system in compressed domain (Discrete Cosine Domain). Color moments, color histogram and edge histogram have been extracted directly from compressed domain and GA is employed to assign dissimilar importance to the extracted features to improve image retrieval. Although using GA had great positive impact on system's accuracy, it increased the consumption time; however, extracting features in the compressed domain balanced the total time needed to retrieve images.

A multistage CBIR technique was introduced by Pavithra and Sharmila (Pavithra & Sharmila, 2018). In the first stage, color feature was extracted by using color moment through the calculation of mean and standard deviation for each channel in RGB color space to reduce the search space, which in turn reduce the computational cost. In the second stage, texture and shape (edge) features are extracted from images in the new sub-dataset constructed from the first stage. LBP was used to extract texture information, while Canny edge detector was used to extract edge information. Manhattan distance was used as a metric for the search. Although the proposed multistage system improved performance by increasing precision and decreasing running time, the required running time depends on the number of images in the dataset. If the system is integrated with some machine learning algorithms, then it can be used to search datasets with different sizes and types.

Pavithra and Sharmila (2019) proposed a novel method for the selection of seed points for dominant color-based image retrieval technique. To assess the proposed dominant color descriptor, four image datasets were used in the experiments and improved results were obtained. Nonetheless, the proposed method needs to be fused with other feature extraction methods (shape, texture, and spatial information) to reduce the semantic gap, which still exists because the same color information could be assigned to images in different semantic classes. To reduce the semantic gap, Ashraf et al. (Ashraf et al., 2018) proposed a CBIR system that merges color and edge features to form a feature descriptor. For color extraction, the authors used color histogram. For edge extraction, canny edge histogram was used in the YCbCr color space. To achieve better enhancement, discrete wavelet was calculated, and to accelerate the calculation of discrete wavelet, the authors used Haar wavelet, which is faster in term of calculations (Jacobs et al., 1995). The proposed scheme used artificial neural network (ANN) to learn the semantic class of the images, which needs more time for training and testing purposes. The system used Manhattan distance as its similarity measurement, and the reported results of mean precision and recall proved its efficiency. However, it also suffers from the lack of spatial information, and no information about its computational cost efficiency is available.

Nazir et al. (2018) presented a content-based image retrieval methodology that depends on color and texture feature extraction methods. In the HSV color space, a color histogram is used to extract color features, whereas for texture, discrete wavelet transform (DWT) and EHD are used to include the local distribution of the edges in the image to include local features in the constructed feature vector because EDH is efficient in finding relevant images when used in MPEG-7 (Won et al., 2002). This retrieval system was tested by using the Corel dataset, which is the most common dataset in the CBIR domain. The method has better efficiency in terms of precision and recall than other state-of-the-art techniques, but the algorithm did not use a machine learning algorithm (e.g., ANN), which is considered a significant limitation (Ghrabat et al., 2019).

A novel CBIR system is presented by Tadi Bani et al. (Tadi Bani & Fekri-Ershad, 2019), which is based on extracting global and local texture features in both frequency and spatial domains and color features in spatial domain. To reduce noise effects, images are first filtered by Gaussian filter, then global texture features are extracted in spatial domain by GLCM. Quantized color histogram in RGB color space is used to extract color features. To enhance the retrieval performance, local texture features are extracted through Gabor filter. The proposed system showed high precision values when evaluated against Simplicity dataset and compared to other state-of-the-art methods. Also, it was reported as invariant to rotation and low sensitive to noise, but it had a high run time due to the use of different features. Rana et al. (Rana et al., 2019) presented a methodology for CBIR on the basis of the integration of nonparametric features (texture) and parametric features (color and shape). To extract parametric features, color moments and moment invariants were used, and ranklet transformation was used to extract nonparametric features. The constructed feature vector has a length of 247, which increases the running time and is considered a major limitation of the presented algorithm. Five datasets were used to evaluate the algorithm. Depending on the fusion of the extracted information from the color moment in the HSV color space with GLCM in eight directions, FIF-IRS was proposed by Bella and Vasyki (Thusnavis Bella & Vasuki, 2019). FIF-IRS performance is assessed with three versions of the Corel dataset. Precision, retrieval time, and error rate were used as assessment metrics. The proposed FIF-IRS produced good results, but if a suitable optimization algorithm is used, the retrieval time can be reduced.

Ashraf et al. (2020) developed a subjective methodology for the CBIR system on the basis of the fusion of low-level features (texture and color). Color moments in the HSV color space were used to extract color features, and DWT and Gabor wavelet were used to extract texture features. For further enhancement, the color and edge directivity descriptor were calculated and included in the feature vector, with dimensions of 1 ×250. The larger feature vector dimension gives more accurate retrieval results, but it takes a longer time for searching and comparing. The proposed

Table 1. Summary of the literature on global feature-based methods

Ref.	Features	Feature extraction method	Dataset	Accuracy	Issues	Limitations
(Shrivastava & Tyagi, 2015)	Color Texture Shape	Color Histogram Gabor Filter Fourier Descriptor	Corel CIFAR	0.7690 0.8590	-Avoid linear dataset search.	-Color histogram does not provide any spatial information(Younus et al., 2015). - Fourier descriptor requires high computational cost.
(Younus et al., 2015)	Color Texture	Color Moment Color Histogram Wavelet Moment Co-occurrence Matrix	WANG	0.7352	- comparing query image with sub-set of dataset, because of employing clustering in offline stage. - clustering dataset using PSO and K-mean in offline stage.	- miss clustering cause irrelevant image retrieval. - Not including shape features in similarity measurement.
(Ponomarev et al., 2016)	Color Texture Shape	Color auto correlogram Gabor Transform Wavelet Transform	Corel Li Caltech101	0.8300 0.8800 0.7000	-considering color, texture and shape features in similarity measurement and acquiring high accuracy.	-high computational cost due to using Multiple features.
(Srivastava & Khare, 2017)	Texture Shape	Wavelet Transform/LBP Legendre Moments	Corel 1k Corel 5k Corel 10k Olivia 2688 GHIM10k	0.9995(7^{th}level) 0.5676(7^{th}level) 0.3537(7^{th}level) 0.9999(7^{th}level) 0.9172(7^{th}level)	- combining local and global features. - analyze images at multiple levels to capture missed details from other levels.	-Increased computational cost.
(Sajjad et al., 2018)	Color Texture	Quantized Color Histogram RLBP	Corel 1k Corel 10k ZB building	–0.8777	-Illumination and rotation invariant.	-No clear information about Results of Corel dataset.
(X. Zheng et al., 2016)	Color Texture	Projection Histogram Roberts Edge Detector	-	-	-Using main block to extract color features.	-No information about acquired results.

(*Continued*)

cogent engineering

Table 1. (*Continued*)

Ref.	Features	Feature extraction method	Dataset	Accuracy	Issues	Limitations
(Z. Zhao et al., 2016)	Color Texture Shape	CDE CLCM Hue Moments	Wang	0.9050	-Considering color, texture and shape features in similarity measurement. - High accuracy.	-accuracy depends on the nature of query image.
(Phadikar et al., 2018)	Color Texture	Color Histogram Color Moment MPEG-7 Edge descriptor	Corel 1k GHIM 10 K NUSWIND	0.9850	-Considering color and texture features in similarity measurement. - Using GA increased system accuracy. - Excellent accuracy.	-Using GA had an impact on computational cost.
(Pavithra & Sharmila, 2018)	Color Texture Shape	Color Moment LBP Canny Edge Detector	WANG Corel 5k Corel 10k	0.83225 0.6860 0.5998	-Avoid linear dataset search.	- The required running time depends on the number of images in the new sub-dataset.
(Pavithra & Sree Sharmila, 2019)	Color	Dominant Color Descriptor	Wang Corel-10 K OT-scene OxfordFlower	0.7534 0.4136 0.4352 0.3186	-introduce seed point selection to reduce the drawback of DCD.	-Do not bridge Semantic gap.
(Ashraf et al., 2018)	Color Shape	Color Histogram Canny Edge Histogram	WANG	0.7350	- Using artificial neural network (ANN) to learn the semantic class of the images	- Lack of spatial information. - No information about running time.
(Nazir et al., 2018)	Color Texture	Color Histogram DWT EDH	Corel	0.7350	- combining local and global features. - EHD are used to include the local distribution of the edges.	-Did not use any machine learning algorithm.

(*Continued*)

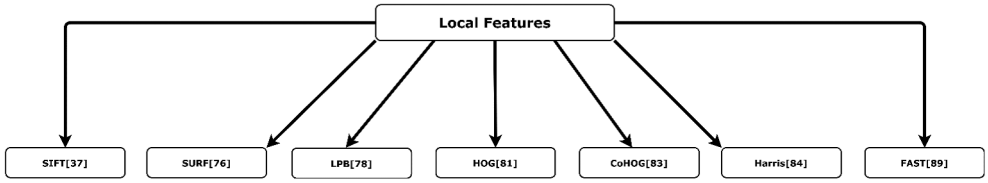
cogent engineering

Table 1. (*Continued*)

Ref.	Features	Feature extraction method	Dataset	Accuracy	Issues	Limitations
(Tadi Bani & Fekri-Ershad, 2019)	Color Texture	Quantized Color Histogram Gabor Filter GLCM	Simplicity	0.8294	-Extracting features in spatial and frequency domains. - Invariant to rotation and low sensitive to noise	-High run time.
(Rana et al., 2019)	Color Texture Shape	Color Moment Ranklet Transformation Invariant Moment	Simplicity Corel 5k Corel 10k Caltech101 MSR	0.6760 0.6744 0.6796 0.6450 0.6480	-Integration of non-parametric and parametric features.	-High computational cost due to high feature vector dimension.
(Thusnavis Bella & Vasuki, 2019)	Color Texture Shape	Color Moment GLCM Geometric Shape Feature	Corel 1k Corel 5k Corel 10k	0.8330 0.6690 0.5640	-Considering color, texture and shape features in similarity measurement.	- The retrieval time can be reduced if a suitable optimization algorithm is applied.
(Ashraf et al., 2020)	Color Texture	Color Moment DWT/Gabor Filter/CEDD	Corel1k Corel1.5k Corel5k GHIM20k	0.8750 0.8633 0.7983 0.7650	-High accuracy values.	-High computational cost due to high feature vector dimension.
(Alsmadi, 2020)	Color Texture Shape	GLCM DWT Canny Edge Histogram	Corel	0.9015	-Intensive and robust number of features.	-Criticality of cooling process. -Increase in calculation time.

cogent engineering

Figure 3. Classification of local feature descriptors.



system was tested against different datasets (Corel 1000, Corel 15,000, Corel 5000, and GHIM-10 K) and achieved high average precision and recall. However, the proposed scheme lacks texture and spatial information, as many other works in the literature.

Alsmadi et al. (2020) introduced a novel content-based image retrieval technique that has an advantage from combining color, shape, and texture. Canny edge histogram and DWT transform in the YCbCr color space were used to extract color features, while GLCM was used to extract texture features. The canny edge method in the RGB color space was used to extract shape features. The proposed technique applied genetic algorithm (GA) with simulated annealing (SA), which increases the fitness number, thus enhancing the solution quality. The proposed CBIR system outperforms other state-of-the-art systems in terms of average precision and recall (0.9015 and 0.1803, respectively). However, SA suffers from the cooling process's criticality and the need for numerous iterations, which slows down the calculation time. Table 1 summarizes the literature on global feature-based methods.

From our point of view, the feature extraction methods used in (Srivastava & Khare, 2017) are preferred when the accuracy is demanded; however, it is considered computationally cost due to the nature of the features extracted. On the other hand, when the computational cost play the major role in system performance, then the methods in (Z. Zhao et al., 2016) can be considered.

3.2. Local features

While many researchers in CBIR use global image features and achieve good accuracy, local image features are gaining popularity because they are superior to global features in terms of being invariant to scale and rotation, and they provide reliable matching in a range of conditions (Low, 2004). The most common local detectors and descriptors that are widely used in the CBIR domain are listed, as shown in Figure 3.

3.2.1. Scale-invariant feature transform

Scale-invariant feature transform (SIFT) is one of the most widely used local descriptors introduced by David Lowe (Low, 2004), which contains a detector and a descriptor for key points. SIFT is robust against image rotation and image scaling, but it performs poorly in matching at high dimensions and need a fixed-size vector for encoding to perform image similarity checking. In image retrieval, SIFT has two drawbacks: it uses a large amount of memory and has a high computational cost (Montazer & Giveki, 2015). Soltanshahi et al. proposed (Montazer & Giveki, 2015) a method for CBIR based on the use of SIFT and local derivative pattern (LDP) to construct the feature descriptor. To overcome its limitation of high memory usage and computation cost, the authors proposed two methods to reduce SIFT dimensionality. The proposed system was tested using four datasets and proved its high retrieval performance for images that contain an object. However, it needs to be improved for nature images.

Sharif et al. (2019) proposed a CBIR system that depends on merging the visual words generated from SIFT and binary robust invariant scalable key points (BRISK). The use of BRISK was used to cope with the drawback of SIFT in low illumination and when key points are localized poorly (Leutenegger et al., 2011). The computational cost of the proposed system could be reduced by choosing different image feature percentages. The proposed approach had not

been tested against large-scale unlabeled datasets. Yousuf et al. (Yousuf et al., 2018) conducted a CBIR system based on SFIT and local intensity order pattern (LIOP). LIOP was used to overcome the limitation of SIFT in illumination changes and low-contrast scenes. LIOP suffers from poor performance in change in scale, whereas SIFT is robust in such state. The fusion of the two descriptors will form visual words on the basis of the bag-of-visual-words (BoVW), which is a standard model that represents local features into a fixed length vector (Tsai, 2012). This fusion will reduce the semantic gap between the high-level and low-level features of an image. After this fusion, k-means clustering is applied on the fused visual words followed by the computation for a histogram to the visual word of each image. These histograms are used to train a support vector machine (SVM) to classify the images. The proposed system was tested on three image collections (Caltech 101, Corel A, and Coral B). It is efficient in terms of MAP and average recall and achieved good computational complexity. SIFT is considered a high-dimensional descriptor.

3.2.2. Speeded-up robust features

Herbert et al. (Bay et al., 2008) introduced speeded-up robust features (SURF) for the first time in 2008 as another robust local descriptor that overcomes the high dimensionality limitation of SIFT. Such high dimensionality can be disabled by using a dimensionality reduction technique. This will degrade the system performance in terms of feature computation. The authors in (Bay et al., 2008) were inspired to create SURF from SIFT, but SURF is faster and more robust than SIFT because it requires less time for feature computation and matching by using an indexing scheme based on the Laplacian sign. However, SURF performs poorly in rotation. Jabeen et al. (Jabeen et al., 2018) proposed a new CBIR system based on SURF and fast retina key point (FREAK). FREAK has better classification performance than SURF, yet SURF is more robust during changes in illumination and scale.

The fusion of the two descriptors will form visual words on the bases of BoVW methodology which will reduce the semantic gap between high-level and low-level features of an image. After the fusion process, K-mean clustering is applied to the fused visual words followed by the computation for a histogram to the visual word of each image. These histograms are used to train an SVM to classify the images on the base of its semantic. The proposed system tested upon three image collections (Caltech 256, Corel 1000 and Coral 1500) and proved its efficiency in terms of mean average precision and average recall and achieves a good value in term of computational complexity. FREAK and SURF do not offer any color information.

3.2.3. Local binary pattern

Ojala et al. (Ojala et al., 2002) introduced LBP, which is based on the qualitative level for local patterns. LBP compares the center pixel and its eight surrounding neighbors, with the center pixel considered as a threshold. LBP is robust because it is invariant to any monotonic transformations in the grayscale. Moreover, LBP is computationally simple. Its limitation is that it loses global spatial information.

Guo et al. (Guo et al., 2010) proposed the LBP variance (LBPV), a variation to LBP to overcome its limitation. LBPV executes a global rotation invariant after performing a local variant LBP. To reduce feature size, a method that uses similarity measurement was proposed to speed up the proposed matching. Sarwar et al. (Sarwar et al., 2019) proposed a CBIR system that uses LBPV and LIOP features to improve the CBIR performance by reducing the semantic gap. These two feature descriptors are used to form two small visual dictionaries, which are then concatenated to form one large visual dictionary. To reduce the size, principal component analysis (PCA) was used, the histogram was calculated, and SVM was trained based on LBPV and LIOP. The proposed system was tested against three image datasets (Holidays, WANG-1 K, and WANK-2 K). It is efficient in terms of precision, recall, and computational cost, but it was not tested against large datasets such as ImageNet or ImageCLEF. It also can't be used to construct feature vectors from multi-spectral images directly, resulting in spectral and spatial information loss.

Table 2. Summary of the literature for local feature-based methods

Ref.	Descriptor	Dataset	Accuracy	Issues	Limitations
(Montazer & Giveki, 2015)	SIFT LDP	Wang OT FP LSP	0.8850 0.9407 0.8860 0.8830	-High performance for images with objects.	- Length of feature vector is 3000. -Need to be improved for nature images.
(Sharif et al., 2019)	SIFT BRISK	Corel 1k Corel 1.5k Corel 5k Caltech 256	0.8439 0.7814 0.5737 0.4752	-Reduce semantic gap between high-level and low-level features.	-Not tested against large-scale unlabeled datasets.
(Yousuf et al., 2018)	SFIT LIOP	Corel 1k Corel 1.5k Caltech 256	0.8730 0.8520 0.3030	-Reduce semantic gap between high-level and low-level features. - Invariant to rotation, changing scale, illumination, and performing better in low contrast cases.	-High dimensional descriptor.
(Jabeen et al., 2018)	SURF FREAK	Corel 1k Corel 1.5k Caltech 256	0.8600 0.8320 0.3898	-Reduce semantic gap between high-level and low-level features.	-Did not offer any color information.
(Sarwar et al., 2019)	LBPV LIOP	WANG 1k WANG 1.5k Holiday	0.8958 0.7602 0.6923	-Reduce semantic gap between high-level and low-level features. - Using PCA to reduce dimensionality.	-Not tested against large-scale datasets. -Cannot be directly used for multispectral images.
(Mehmood et al., 2018)	HOG SURF	Corel 1k Corel 1.5k Corel 5k Caltech 256	0.8061 0.7628 0.6060 0.4630	-Offering more spatial information. -performs better in noise and low illumination situations.	-HOG Cannot be directly used for multispectral images.
(Baig et al., 2020)	CoHOG SURF	Corel 1k Corel 1.5k Sense 15 Caltech 256	0.8641 0.7726 0.8123 0.6839	-Using Relevance feedback to enhance accuracy.	-Not tested against large-scale datasets.
(Qin et al., 2019)	Harris SURF	Corel 1k	-	-Reduce search time.	- For large image dataset, the efficiency is not ideal.

(*Continued*)

cogent engineering

Table 2. (*Continued*)

Ref.	Descriptor	Dataset	Accuracy	Issues	Limitations
(Sharmi et al., 2018)	-FAST	Medical dataset	-	-Using watermark-based protocol.	-FAST performs poorly in high level noise. FAST is dependent on threshold values.
(M. Zhao et al., 2016)	MTSD	Corel 1k Corel 5k Corel 10k Caltech 50	0.7928 0.6298 0.5195 0.4387	- Dimension of the proposed descriptor is 137	- Lack in describing the correlation.
(Raza et al., 2019)	STH	Corel 5k Corel 10k	0.6028 0.4803	- Correlating color information with texture orientation.	-Lack of intensity information.
(Raza et al., 2018)	CPV-THF	Corel 1k Corel 5k Corel 10k	0.8079 0.6390 0.5228	-Integrating semantic and content visual information.	- The dimension of the proposed feature vector is 242.
(song et al., 2018)	DTSD	Corel 1k Corel 5k UCID	0.6119 0.2019 0.3367	-Segmenting images to background and foreground.	- The dimension of the proposed feature vector is 692.
(Agarwal et al., 2019)	MCLTP	Corel 1k Corel 10k CMU-PIE STex MIT VisTex	0.8330 0.5690 0.8630 0.9810 0.4090	-Better retrieval performance.	- The dimension of the proposed feature vector is 3072.

cogent engineering

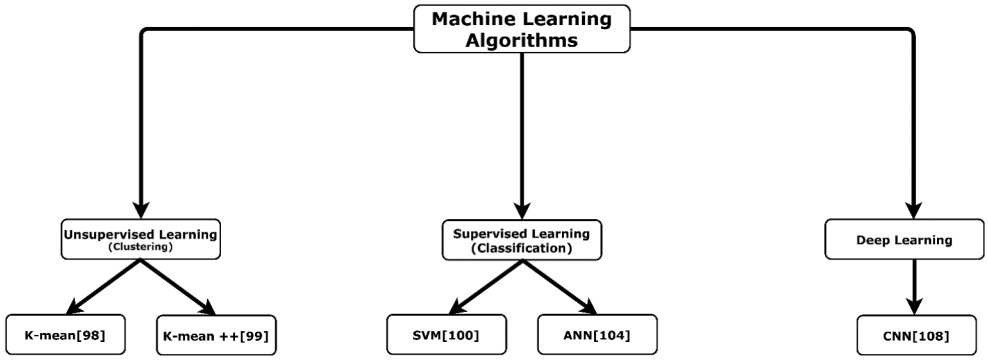


Figure 4. Machine learning algorithms.

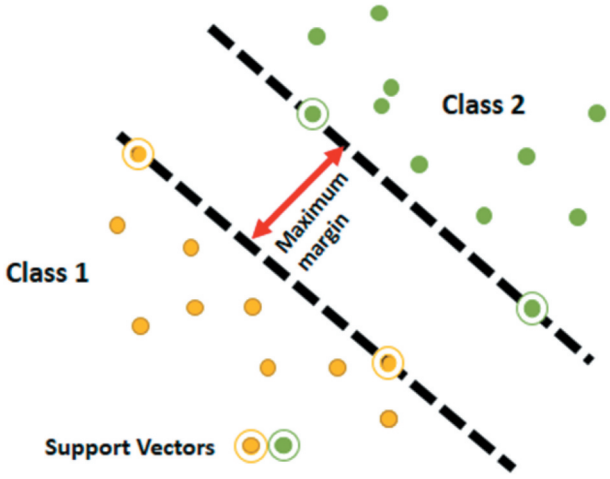


Figure 5. SVM classifier.

3.2.4. Histogram of oriented gradient

Dalal and Friggs (Dalal & Triggs, 2005) proposed histogram of oriented gradient (HOG), a new, locally normalized descriptor that achieves higher performance than other known feature descriptors, including wavelet. Even if no precise information about corresponding gradients or edge positions is available, HOG defines the shape and appearance of the local object based on the directions of the objects' edge or the distribution of local intensity gradients. HOG splits an entire image into small spatial parts known as "cells" or "regions", and edge orientations over the cells' pixels or local 1D histogram of gradient directions are accumulated. The representation of the

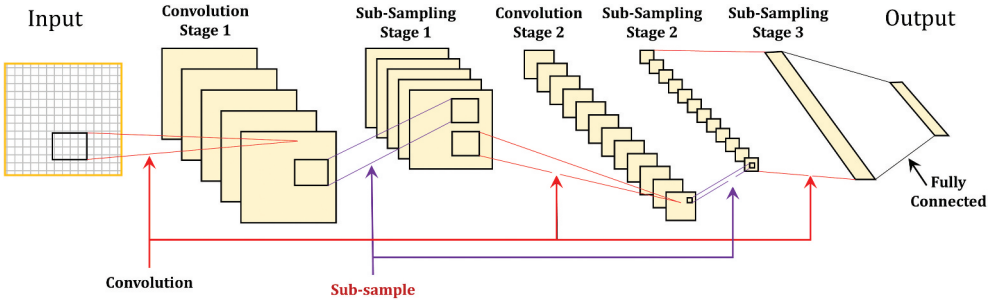


Figure 6. An example of CNN architecture

Table 3. Clarification of the main characteristics, limitations and examples for the main machine learning categories

Methods	Main characteristics	Limitations	Examples
Supervised Learning (Classification)	-requires the presence of labels. -predict and classify data to one of predefined classes.	- the size of the training set has a direct effect on the classification accuracy.	-SVM. -ANN.
Unsupervised Learning (Clustering)	-No need for labels. - learns the similarities and semantics between input data and generalize a model to handle unseen inputs.	-Over fitting. -Scalability. -Number of clusters affects the clustering algorithm performance.	-K-means. -K-means ++.
Deep Learning	-Supervised or Unsupervised. - Generate learning model	-Suffer from high computational cost. - Complex structure.	-CNN. -Deep Neural Network. -Deep Belief Network. -Boltzmann Machine.

image is formed by combining the histogram entries. Moreover, the HOG accumulates energy, which is a measure of the local histogram over larger region "blocks" and normalizes all the block cells by using these results. As a result, HOG is more invariant to illumination and shadowing. For the last decade, HOG has been successfully used in many applications, specifically in object recognition. Mehmood et al. (Mehmood et al., 2018) proposed a CBIR system based on the use of SURF to extract local features and HOG to extract global features. HOG is used because it offers more spatial information, which improves the retrieval performance, while SURF performs better in noisy, low illumination, and clear background images. The two descriptors are used to form two visual vocabularies that are concatenated to form one large vocabulary. The larger vocabulary gives better retrieval results but increases the computational cost. Therefore, the authors took a percentage from the extracted features. K-means++ clustering was applied to the fused feature vector, and the histogram was calculated for each image. SVM was used for classification, and the system was tested against four well-known datasets (Caltech 256, Corel 1 K, Corel 1.5 K, and Corel 5 K). HOG is efficient, but it cannot be directly used for constructing feature vectors from multi-spectral images, causing loss in spectral and spatial information.

3.2.5. Co-occurrence histogram of oriented gradient

HOG does not offer any spatial information about neighboring pixels. It offers only the orientation of the pixel under study. Co-occurrence histogram of oriented gradient (CoHOG) (Watanabe et al., 2010) overcomes this limitation. Baig et al. (Baig et al., 2020) introduced a CBIR system to reduce the semantic gap based on the merging of the benefits of CoHOG and SURF to cope with each other's limitation. CoHOG depends on local discriminant analysis to reduce the dimension of each feature vector of CoHOG and SURF to decrease the computational cost. Relevance feedback is used to enhance the specificity (precision) and sensitivity (recall) with the help of the user by indicating the relevant and irrelevant images retrieved from the search of the database. The proposed system's performance was assessed using Corel 1k, Corel 1.5k, Scene 15, and Caltech 256, and it recorded positive results; however, it has not been tested on large datasets.

3.2.6. Harris corner detector

Harris and Stephens (Harris & Stephens, 1988) introduced Harris detector for the first time in 1988 as a robust corner detector. It is considered a reference technique (Sánchez et al., 2018), which is used for video stabilization, image matching, camera calibration and tracking (Sánchez et al., 2018). Through analyzing the eigenvalues of the autocorrelation matrix (also known as structure tensor) information, points with strong variation in intensity will be located. Harris detector is robust against noise, scale and rotation but it suffers from computational cost. Qin et al. (Qin et al.,

Table 4. Summary of the performance of machine learning algorithms-based approaches for CBIR

Ref.	Machine learning algorithm	Dataset	Accuracy	Issues	Limitations
(Yousuf et al., 2018)	K-mean SVM	Corel 1k Corel 1.5k Caltech 256	0.8730 0.8520 0.3030	-Reduce semantic gap between high-level and low-level features.	High dimensional descriptor.
(Sarwar et al., 2019)	SVM	WANG 1k WANG 1.5k Holiday	0.8958 0.7602 0.6923	-Reduce semantic gap between high-level and low-level features. -Invariant to image rotation and monotonic intensity change. - Using PCA to reduce dimensionality.	-Not tested against large-scale datasets. -Cannot be directly used for multispectral images.
(Mehmood et al., 2018)	K-mean++ SVM	Corel 1k Corel 1.5k Corel 5k Caltech 256	0.8061 0.7628 0.6060 0.4630	-Offering more spatial information. -performs better in noise and low illumination situations.	-HOG Cannot be directly used for multispectral images.
(Ashraf et al., 2015)	ANN	Corel Coil Caltech 101	0.8200	-Retrieval based on the image core (main) object.	-Segmentation slow down the system.
(Alzu'bi et al., 2017)	CNN	Oxford 5k Oxford105k	0.9570 0.8860	-No need for annotations or labels. -Length of feature vector is 16. - Reduce required memory and run time.	- Accuracy decreases when using larger datasets.
(Tzelepi & Tefas, 2018)	CNN/ Unsupervised CNN/Supervised	Paris 6k UKBench	0.9859 0.8347	- Reduce the dimension of the feature descriptor. - Retain spatial information.	- More retrieval time.
(Q. Zheng et al., 2019)	CNN	Caltech 101 Holiday Oxford	0.8850 0.9410 0.9620	-Using VGG-16. -Using similarity score.	-More enhancement is required in terms of speed of the training and testing stages. - More time to construct the gravitational field database
(Sezavar et al., 2019)	CNN	Corel ALOI MPEG-7	0.9559 0.9706 0.7749	-Reduce the computational cost.	- More retrieval time if not using sparse representation.

2019) proposed a CBIR method to retrieve encrypted images from cloud based on improved Harris detector (Li et al., 2018) and SURF detector and descriptor. The authors used Local Sensitive Hash function to create searchable indexes for feature vectors to reduce required time for searching and enhance efficiency. However, Xu et al. (Y. Xu et al., 2019) mentioned that for large image dataset, the efficiency is not ideal.

3.2.7. Features from Accelerated Segment Test (FAST)

FAST (Rosten et al., 2010) overcomes the computational cost limitation of Harris. The authors in (Rosten et al., 2010) reported FAST to have an excellent repeatability. Here, repeatability means under different transformation, interest point can be detected. FAST properties corner over edge detection because a corner has an intensity variation in two dimensions which makes it a point of interest. Sharmi et al. (Sharmi et al., 2018) proposed a CBIR method in cloud computing based on FAST to represent features. These features are encrypted by stream chipper to preserve images' privacy before storing them on the cloud. The Locality Sensitive Hash algorithm is used to provide faster search time. Furthermore, watermark-based protocol is used by cloud server to attach a unique watermark to retrieved images which are in the encrypted form to the query user in an attempt to prevent retrieving images to unauthorized users. The proposed CBIR retrieved similar images with a better efficiency because FAST takes less computation time. But FAST performs poorly in high level noise existence. Another limitation is its dependency on threshold values.

3.2.8. Local structure descriptors developed for CBIR

In CBIR domain, local structure descriptors have been proposed to represent local spatial structure of information which makes these descriptors more semantically. MTSD (M. Zhao et al., 2016) is a novel descriptor for CBIR, which uses local and multi-trend structures. It is based on integrating edge, color and intensity information. It characterizes both local spatial structure information and low-level features (i.e. color, shape and texture) by using multi-trends. The dimension of the proposed descriptor is 137. The proposed descriptor is assessed against Caltech and Corel (1 K, 5 K and 10 K) datasets and obtained results show that the proposed descriptor outperforms many of the state-of-the-art descriptors. The drawback in this descriptor is its lack of describing the correlation between local spatial information, intensity, texture and color (Raza et al., 2018).

Squared Texton Histogram (STH) (Raza et al., 2019) descriptor for CBIR is derived from correlating color information with texture orientation. The performance of the proposed descriptor is tested on Corel datasets (5 K and 10 K). Although the proposed STH depends on texture and color features and achieved good system performance, the absence of intensity information has an effect on the retrieval performance of the system(Raza et al., 2018). Correlated primary visual texton histogram features (CPV-THF) (Raza et al., 2018) is proposed which is based on the correlation between intensity information, edge orientation, color and local spatial structure. The proposed descriptor integrates semantic and content visual information. Corel 1 K, 5 K and 10 K is utilized to evaluate the proposed descriptor. In (Song et al., 2018) a descriptor for CBIR is proposed which is diagonal texture structure descriptor (DTSD). It is based on the segmentation of images to background and foreground according to Otsu algorithm (Otsu, 1979). Texture and color features of these background and foreground are extracted in HSV color space. The dimension of the proposed feature vector is 692, which is considered the drawback of the proposed descriptor (Vimina & Divya, 2020). In Agarwal et al., 2019, a new descriptor called multi-channel local ternary pattern is proposed, which is based on color and texture feature extraction (MCLTP). In this scheme, images are converted from RGB color space to HSV, then local ternary patterns are calculated for two channels simultaneously (i.e. H-S, H-V and V–V). Lower pattern and upper pattern for each channel pair is calculated by considering 3×3 and 5×5 neighborhoods. The proposed feature vector is constructed by concatenating these upper and lower pattern for the three channel pairs with a size of 3072 (Vimina & Divya, 2020). Although, the proposed feature vector is reported to have better retrieval than many other states-of-the-art feature vectors, its high dimensionality is a major disadvantage. Table 2 shows the summary of the literature for local feature-based methods.

Table 5. Comparison of different CBIR techniques

Ref.	Feature extraction method	Accuracy	Strength of used method	Limitation
(Srivastava & Khare, 2017)	global	0.9999	- Provides directional information. - Provides spatial information. - Invariant to scale and rotation changes.	- Increases computation complexity. - Low accuracy when using large datasets. - DWT is sensitive to noise.
(Phadikar et al., 2018)	global	0.9850	-Considering color, texture and shape features in similarity measurement. - Using GA increased system accuracy. - Excellent accuracy.	- Criticality of the cooling process. -Using GA had an impact on computational time.
(Sarwar et al., 2019)	local	0.8958	-Reduce semantic gap between high-level and low-level features. -Invariant to image rotation and monotonic intensity change. - Using PCA to reduce dimensionality.	- Not tested against large datasets. - Loss in spectral and spatial information in multispectral images.
(Yousuf et al., 2018)	local	0.8730	-Reduce semantic gap between high-level and low-level features. - Invariant to rotation, changing scale, illumination, and performing better in low contrast cases.	- High dimensionality of SFIT.
(Tzelepi & Tefas, 2018)	CNN	0.9859	- Reduce the dimension of the feature descriptor. - Retain spatial information.	- More retrieval time.
(Sezavar et al., 2019)	CNN	0.9706	-Reduce the computational cost.	- More retrieval time if not using sparse representation.

From Table 2, the CBIR algorithm presented in (Sarwar et al., 2019) can be used to obtain high accuracy; however, attention must be paid because they cannot be directly used for multispectral images and did not tested against large-scale datasets.

4. Machine learning

Recently, CBIR systems have been shifted toward using machine learning algorithms to obtain a model that can deal with new input data and give correct prediction, which will improve the image search. The most common machine learning algorithms used in the CBIR domain will be

discussed and analyzed, as illustrated in Figure 4 along with the most recent works. This section will be subdivided into three subsections: unsupervised learning (clustering), supervised learning (classification), and deep learning.

4.1. Unsupervised learning (clustering)

After the feature extraction process and feature vector construction, clustering, which is the process of gathering image descriptors into a single group that semantically differs from other groups, is performed. Clustering is considered as unsupervised learning algorithm because it does not previously know which group the images' data should belongs to K-means (Hartigan & Wong, 1979) and k-means++ (Arthur & Vassilvitskii, 2007) clustering are the most widely used clustering algorithms in CBIR, especially when systems depend on local feature extraction methods. These methods are usually followed by a clustering process to decide the semantic group which the image belongs to.

In Yousuf et al., 2018, as described before, k-means was applied on the visual vocabulary constructed from the fusion of the SIFT and LIOP visual vocabularies, which will be larger and will enhance the retrieval process. K-means has the limitation of specifying the number of clusters at the beginning. Moreover, the selection of the initial centroid will affect the performance of the clustering algorithm and make it terminate at the local optimum if no proper initial centroid was chosen. Although a large number of clusters will decrease the error, the risk of overfitting still available. K-means has the drawback of failing in handling outliers and noisy data. In (Mehmood et al., 2018) the authors used k-means++ on the visual dictionary constructed from the fusion of the HOG and SURF visual dictionaries. K-means++ overcomes the limitation of k-means by assigning weights to initial centroids. Although the process of selecting the initial centroid is more complicated and time consuming than k-means, the clustering is more accurate and there are less iterations, lowering the computational cost.

4.2. Supervised learning (classification)

Unlike unsupervised learning, the supervised learning algorithm have prior knowledge of the image groups and labels. Therefore, it is considered a classification task; when a new image is entered, the algorithm knows which predefined group it should consider. Below are the mostly used supervised learning algorithms in the CBIR domain.

4.2.1. Support vector machine

SVM (Cortes & Vapnik, 1995) is one of the most common supervised classifiers used in pattern recognition and image classification. When new data are assigned, SVM will decide the class it should be assigned. There are two types of SVM (Garg & Dhiman, 2020), linear and non-linear. In linear SVM, features can be separated into two classes through the use of straight line while in non-linear SVM, dataset cannot be separated by a straight line, SVM uses kernel functions to enable the separation by adding a new dimension; Figure 5 shows an instance for the SVM classifier (Alzu'bi et al., 2015). Kernel function is considered an essential part that affects the performance of SVM. Many researchers used the SVM classifier to predict the class of an input image (Yousuf et al., 2018), (Sarwar et al., 2019), (Mehmood et al., 2018). They all used SVM with Hellinger kernel(Vedaldi & Zisserman, 2012), which is derived from Additive Kernel that has a low computation cost and better performance than other kernels. Other kernel types are Hyperbolic Tanget, Gaussain Radial Baises, Polynomial and Linear (Fowler, 2000).

4.2.2. Artificial neural networks

ANN is widely used to find an elegant solution for most real-world problems, including image retrieval. The development of these networks is like the behavior of the human neuron system. ANN's excellent information processing characteristics, such as robustness, high parallelism, fault tolerance, noise, and nonlinearity, make it an appealing option for solving a wide range of problems (Basheer & Hajmeer, 2000). In general, ANN consists of neurons and links that interconnect neurons. ANN consists of three layers: input, hidden, and output layers. The input layer consists of n neurons, with every neuron for one independent variable in the network, while the

number of neurons in the hidden layer is experimentally chosen by the user. The output layer has a number of neurons equal to the number of classes, and it is considered a dependent variable. Each connection between the neurons has a weight, which is adjusted in every iteration in the training process of the network. The choice of the network type is associated with the problem that needs a solution. For ANN to act as a classifier, it executes training and testing stages. Ashraf et al. (Ashraf et al., 2015) proposed a CBIR system that retrieves images automatically on the basis of their core (main) object. For feature extraction, the authors applied Bandelet transform on the images' major object. The authors used a back propagation neural network for texture classification in one of four categories (no contour blocks, vertical, horizontal, right/left diagonal). The ANN consisted of 20 neurons in one hidden layer and 4 neurons in the output layer. The authors used Gabor filter for texture features based on the ANN output. To enhance the system performance, color features were extracted in YC_bC_r and RGB color spaces by using color wavelets and color histogram. Another ANN was used to classify the belonging class that query image should be assigned and then compare it with the whole images in the same class. Their system was built on segmentation concept which gives more precise results, but it is slow.

The architecture of ANN has a great impact on the system performance. This architecture is learned by trial and error (Yoon et al., 2013), and it cannot be easily derived. The training error and uncertainty increase if the input data are noisy, and the perception will be incorrect (Alzu'bi et al., 2015), (Wu & Huang, 2013).

4.3. Deep learning

Deep learning is one of the machine learning techniques that has received a great deal of attention in the last decade for solving real-world problems. Deep learning architecture consists of a family of machine learning algorithms, the design is inspired from human brain. These algorithms organize and manipulate information by passing them through stages of representation and transformation. The success of deep learning algorithms in many fields (e.g., object recognition) has inspired their use in the CBIR domain to bridge the semantic gap. The architecture of deep learning makes it capable of mapping data in the input layer to data in the output layer without depending on features provided by humans (Wan, 2014). Deep learning algorithm includes convolutional neural network (CNN), deep neural network (DNN), deep belief network, and Boltzmann machine, with CNN exhibiting outstanding performance in computer vision applications such as face recognition, object detection and semantic segmentation (Voulodimos et al., 2018) and specifically in the CBIR domain (Latif et al., 2019). CNNs consist of three types of layers, convolutional layer, pooling layer and fully connected layer, Figure 6 shows an example of CNN (Chaudhry & Chandra, 2016). Filters are applied to input images through convolutional layer in order to learn features while the function of the intermediate layers (pooling layer) is down sampling the volume of the incoming inputs. The last layer (fully connected layer) predicts the label or the class of the input image. The difference between ANN and CNN is that the last layer of CNN is only fully connected layer while in ANN all neurons are connected to others (Gogul & Kumar, 2017). CNN is also invariant to translation, scale and rotation which made it valuable for specific CV application (Voulodimos et al., 2018). CNN relays on labeled data (Voulodimos et al., 2018), which is considered one of its limitations. Other difference is that CNNs do not demand hand-craft feature extraction (Patil & Rane, 2021). Table 3 clarifies the main characteristics, limitations and examples for the main machine learning categories.

In Wan, 2014, the behavior of CNN in different settings in the CBIR domain is examined to provide feature representation for images, and to perform similarity measurements. The study concluded that CNN could be used to extract features, which will help improves the retrieval results. However, because of the large visual dictionary that is used, the memory storage and training time results degrade retrieval capacity. Alzu'bi et al. (Alzu'bi et al., 2017) proposed a CBIR system based on the use of a bilinear CNN. To the best of the author's knowledge, it was the first study that introduced the bilinear CNN in the CBIR domain. They used CNN to extract features from the image content in an unsupervised manner without depending on the bounding boxes or

Figure 7. Samples from corel, holiday, and brodatz image datasets.

annotation or any class label. The extracted features were characterized by their reduced dimensions because the authors depend on the use of the pooling scheme during the extraction process, thereby reducing memory usage and computational cost. The proposed scheme was assessed against the retrieval of large-scale images and exhibited good retrieval performance.

To improve the retrieval performance in terms of computational cost and memory usage, Tzelepi et al. (Tzelepi & Tefas, 2018) suggested a method for CBIR that utilized CNN for feature representation by using maximum pooling after the convolutional layers rather than using the fully connected layers because fully connected layers discard the spatial information due to the

Table 6. Some widely used datasets in the CBIR domain

Dataset name	Number of images	Version	Number of classes	Number of images in each class
Corel (J. Z. Wang et al., 2000)	1000	Corel 1k	10	100
	1500	Corel 1.5k	15	100
	5000	Corel 5k	50	100
	10,000	Corel 10 K	10	100
WANG (J. Z. Wang et al., 2000)	1000	SIMPLicity	10	100
	10,000	WBIIS	10	1000
Holiday (Jain et al., 2011)	1491	Not- applicable	500	vary
Oliva (Oliva & Torralba, 2001)–(Das et al., 2016)	2688	QT-Scene	8	vary
CIFAR (Yang et al., 2018)	60,000	CIFAR-10	10	6000
Caltech (Fei-Fei et al., 2007)(Griffin et al., 2007)	More than 9000	Caltech 101	101	vary
	30,607	Caltech 256	256	80
GHIM (Srivastava & Khare, 2017) (Ashraf et al., 2020)	10,000	Not- applicable	20	500

Table 7. An example of two class confusion matrix

Predicted Classes			
Actual Classes		A	B
	A	9	1
	B	2	8

connection to the entire input neurons. This architecture will reduce the dimension of the feature descriptor while retaining the spatial information, resulting in high retrieval efficiency while requiring minimal memory storage and processing time. The proposed approach consisted of three schemes depending on the existing information: fully unsupervised retraining, which is used when only the dataset exists; relevance information, which is used in the case of an existing dataset with labels; and relevance feedback-based retraining, which is used when users are present and can give their feedback.

Zheng et al. (Q. Zheng et al., 2019) proposed an end-to-end CBIR based on VGGNet (Simonyan & Zisserman, 2014). To train the CNN in feature extraction, they used a gravitational field dataset and similarity score labels instead of the usual labels. Proposed system achieved accuracies of 0.9620, 0.9410, and 0.8850 in retrieving images when tested against three datasets, namely, Oxford Paris, Holidays, and Caltech 101, respectively. However, the system takes long time to construct the gravitational field database and needs more speed enhancements in the training and testing stages.

Sezavar et al. (Sezavar et al., 2019) proposed a content-based image retrieval approach that uses CNN to extract high-level features. The last layer of Alexnet (Krizhevsky et al., 2017) is used to extract features because the last layer has the smallest feature vector. Sparse representation is used to reduce the computational cost, and it is reportedly efficient in compression. The technique is tested by using the ALOI, Corel, and MPEG-7 datasets, and it has good retrieval speed and accuracy. However, sparse representation achieves less accuracy but ensures faster retrieval. Table 4 summarizes the performance of machine learning algorithm-based approaches for CBIR.

To conclude, the retrieval performance in terms of accuracy can be improved by using machine learning algorithms in different stages of CBIR; however, the processing time required for training and testing phases is high.

5. Comparison among the state-of the art approaches

In this section, six studies that achieved the highest accuracy in global feature extraction will be discussed (Srivastava & Khare, 2017)(Phadikar et al., 2018), local feature extraction (Sarwar et al., 2019)(Yousuf et al., 2018), and machine learning extraction-based approaches (Tzelepi & Tefas, 2018)(Sezavar et al., 2019). All the investigated studies in this section utilize COREL dataset except studies from machines learning approach which utilizes Paris6k and ALOI.

In the global feature approach, Srivastava and Khare (Srivastava & Khare, 2017) achieved the highest accuracy (0.9995) by extracting texture and shape features. This feature integration helps provide details obtained by extraction. To extract texture feature, DWT was used followed by the use of LBP. The use of wavelet transform provides directional information to cope with the limitation of LBP. LBP provides structural information about gray level pixels. The computation of LBP is simple, and LBP is insensitive to gray level scale changes but has sensitivity to noise. Also, LBP does not provide spatial information. Therefore, Legendre moments were used. Legendre moments belong to continuous orthogonal moments. Orthogonal moments are widely used as a shape descriptor because they alleviate information redundancy and can reconstruct signals (Idan et al., 2020). Continuous orthogonal moments are computationally complex, which increases the overall retrieval time. Moreover, using wavelet increases the computation complexity. DWT is

also sensitive to noise, but has a positive effect on accuracy. There is a tradeoff between accuracy and retrieval time because increasing the extracted features will enhance accuracy on the cost of increasing retrieval time. The proposed CBIR approach achieved high precision and recall values for the seventh level of DWT decomposition and there is a noticeable degradation in accuracy as image databases become larger. In other words, the proposed method achieved low accuracy when using large datasets.

The second highest accuracy was achieved by (Phadikar et al., 2018). The authors proposed a CBIR system in compressed domain (Discrete Cosine Domain). Color moments, color histogram and edge histogram have been extracted directly from compressed domain. Color moments and color histogram are widely employed in CBIR domain because of their simplicity in implementation and low computational requirement. However, they do not provide any spatial information. Therefore, the authors used edge histogram to extract texture. The used algorithm describes the local edge distribution which represents the standard semantic of edge histogram in MEPG (Won et al., 2002). These edge distributions may not be satisfactory to describe the edge's global features (Won et al., 2002). To reduce the semantic gap, the authors believed that extracted features had varying importance in similarity measure and should be assigned unequal weights. GA is employed to assign dissimilar importance to the extracted features to improve image retrieval. Although using GA had a great impact on enhancing the accuracy of the system, it increased the consumption time but extracting features in compressed domain made up the total time required to retrieve images.

In the local feature extraction approach, Sarwar et al. (Sarwar et al., 2019) achieved the highest accuracy. The proposed method was based on LBPV and LIOP. LBPV overcomes the limitation of LBP of dropping global spatial information but it is computationally extensive. LIOP is a high discriminative descriptor as for each local patch the global and local intensity order information are encoded. Moreover, this descriptor is invariant to monotonic intensity change, JPEG compression, blur image, image rotation and viewpoint change (Zhenhua Wang et al., 2011). These two feature descriptors (LBPV and LIOP) are used to form two small visual dictionaries, which are then concatenated to form one large visual dictionary. To reduce the size, principal component analysis (PCA) was used, the histogram was calculated. SVM was trained based on LBPV and LIOP. The proposed system was efficient in terms of precision, recall, and computational cost, but it was not tested against large datasets such as ImageNet or ImageCLEF. Moreover, it cannot be directly used for constructing feature vectors from multispectral images, resulting in spectral and spatial information loss. The second highest accuracy in local feature extraction was achieved by Yousuf et al. (Yousuf et al., 2018). The proposed CBIR system was based on SFIT and LIOP. LIOP was used to overcome the limitation of SIFT in illumination changes and low-contrast scenes. LIOP suffers from poor performance in change in scale, whereas SIFT is robust in such state. To reduce the semantic gap between the high-level and low-level features of an image, the two descriptors fused to form visual words based on the bag-of-visual-words (BoVW). After this fusion, k-means clustering is applied on the fused visual words followed by the computation for a histogram to the visual word of each image. These histograms are used to train a support vector machine (SVM) to classify the images. Although, the proposed method was efficient in terms of MAP, SIFT is considered a high-dimensional descriptor.

In the machine learning approach, CNN is efficient at feature representation. The method proposed by Tzelepi et al. (Tzelepi & Tefas, 2018) has the highest accuracy. The strength of their method is the use of CNN for feature representation by using maximum pooling after convolutional layers rather than using fully connected layers because fully connected layers discard the spatial information due to the connection of the entire input neurons. This architecture will reduce the dimension of the feature descriptor while retaining the spatial information, resulting in high retrieval performance and at the same time requiring smaller memory storage. Therefore, the proposed method could be employed in machines such as robots, drones, and smartphones, which have limited computational power and memory.

However, not employing an indexing technique requires more retrieval time. The second most accurate method was proposed in (Sezavar et al., 2019). The proposed CBIR algorithm combine CCN and sparse representation. CNN was used to extract high-level features and as a classifier to specify the class of the query image while the use of sparse representation was to reduce the computational cost. Sparse representation achieves lower accuracy but ensures faster retrieval because computing the Euclidean distance between query image and large databases is computationally intensive. To have an efficient CBIR, accuracy must be high and at same time must have fast retrieval time. AlexNet (Gogul & Kumar, 2017) was used, and features from the last layer were extracted since these features are fed to the softmax (the classifier), and the last layer has the smallest size, reducing computational time. Table 5 shows a comparison of different techniques for CBIR.

6. Similarity/Dissimilarity measures

The performance of image retrieval systems is influenced by the feature extraction process as well as the similarity measurement. The similarity measurement determines which images are considered most relevant to the query image and should be returned from the dataset. Therefore, the similarity measure determines the accuracy of the CBIR indirectly and has an effect on the computational complexity of the system. The selection of the similarity measure is affected by the structure of the constructed feature vector (type and dimensionality of input data). This selection is a major challenging task in the literature. The similarity measure can be divided into distance measure and similarity metric (Sergyan, 2008).

Distance measure usually measures the dissimilarity between two feature vectors. When dissimilarity is measured by using distance measure, the smallest value indicates the most similar images to the query image. Distance measure divided into two main categories: bin-by-bin and cross-bin. Each category has its own advantages, disadvantages and situations where it is recommended to use it. In Bin-by-bin distance metric, the bins from two feature vectors are compared. If $X = (x_1, x_2, x_3, \ldots, x_n)$ and $Y = (y_1, y_2, y_3, \ldots, y_n)$ and $x_i \in X \text{and} y_i \in Y$, then for all, the bin-by-bin distance function examines x_i with only y_i. This category is widely used because of its simplicity in computation and implementation, but it is affected by scaling, quantization, noise, shape deformation and light changes (Tyagi, 2017). On the other hand, cross-bin distance function considers the cross-bin relation between non-corresponding bins in the feature vector. It is better than bin-by-bin distance in being robust and more descriptive, but it has a quite high computational complexity.

Minkowski Family distance is from the bin-by-bin distance function category which is widely used in CBIR as a dissimilarity measure because it is simple in implementation and computation. The mathematical expression is:

$$L_p(X,Y) = \left(\sum_{i=1}^{N} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{1}$$

Where X and Y are two vectors in R^N and N is the total dimensionality of the Euclidean Space. Minkowski distance is also named L_p -Norm. L_1 and L_2distances are from the most widely used distance metrics in CBIR domain and other image processing fields. L_2 is known as Euclidean or Pythagoras distance. It is calculated by set the value of p to 2 in equation (1). It has a special property of being invariant to orthogonal transformation. While L_1 is known as Manhattan or City block or Taxi-cab distance. It is calculated by making the value of p equal to 1 in equation (1). It is variant to coordinate system rotation but robust against reflection and translation. L_∞which is also named Chessboard or Chebyshev distance is another member in the Minkowski family.

$$L_\infty(X,Y) = \max_{i=1}(|x_i - y_i|) = \lim_{p\to\infty}\left(\sum_{i=1}^{N}|x_i - y_i|^p\right)^{\frac{1}{p}} \tag{2}$$

On a chessboard which represents a 2-D space, L_∞ is the minimum number of movements that a king needed to move between two squares. The final member in Minkowski family is the fractional distance, which is the result if p is less than one and $p \in (0,1)$. This distance did not considered a metric because it does not obey triangle inequality condition (Howarth & Rüger, 2005). It is preferred if the dimensionality of the data is high (Tyagi, 2017), the dimensionality of the data is inversely proportional to the order of p . It is the most robust against noise as compared to other members in the Minkowski family.

Chi-square Statistic is another distance metric which is widely used for computing the difference between histogram functions (Pele & Werman, 2010). It is mathematical expression (Cha, 2007) shown below:

$$Chi - square = \sum_{i=1}^{N}\left(\frac{(x_i - y_i)^2}{(x_i + y_i)}\right) \tag{3}$$

Chi-square (Pele & Werman, 2010) showed a success when used for shape classification (Salve & Jondhale, 2010), (Ling & Jacobs, 2007), local descriptors matching (Forssén & Lowe, 2007), boundary detection (Martin et al., 2004), near duplicate image identification (D. Xu et al., 2010) and finally texture and object categories classification (Cula & Dana, 2004; Varma et al., 2009; Zhang et al., 2007).

Histogram intersection distance is another distance metric which is used in image retrieval and other computer vision algorithms such as segmentation, clustering, classification and codebook creation. It is reported to be robust to distraction in object's background, image resolution variation, occlusion and viewpoint variation (Swain & Ballard, 1991). It is proposed by (Swain & Ballard, 1991) and defined as:

$$HistogramIntersectionDistance = \sum_{i=1}^{N}\min(M_i, S_i) \tag{4}$$

where S and M are two histograms with N bins.

Mahalanobis Distance is used to measure the distance between a distribution and a specific feature vector. P. C. Mahalanobis in 1936 proposed Mahalanobis distance, which is defined as:

$$MahalanobisDistance = \sqrt{(v-m)^T(C^{-1})(v-m)} \tag{5}$$

where T represents matrix transpose, v is the feature vector, m is the mean row vector and C is the covariance matrix. The calculation of this distance becomes expansive for high dimensional data because of the computation of the covariance matrix.

G.N. Lance and W. T. Williams in 1966 defined Canberra distance for unsigned numbers (Lance & Williams, 1966). The modified version for signed numbers is proposed in (Lance & Williams, 1967), and is defined as:

$$Canberra\ distance = \sum_{i=1}^{N}\frac{|x_i - y_i|}{|x_i + y_i|} \tag{6}$$

where x and y are vectors with real values. This metric is reported to be sensitive to values near zero while it is appropriate when difference in sign indicates difference in classes(Lance & Williams, 1967).

Squared Chord is another distance metric which is used in image retrieval and it is unsuitable for negative values feature vectors (Rui et al., 2008), it is defined as:

$$Squared\ Chord\ distance = \sum_{i=1}^{N} \left(\sqrt{x_i} - \sqrt{y_i}\right)^2 \tag{7}$$

Unlike the previously described distances, which measure dissimilarity, the cosine distance is used to measure similarity. When measuring the similarity, the larger value indicates the most similar images to the query image. Cosine distance measures the angle between two vectors as defined below(Rui et al., 2008):

$$Cosine\ distance = \frac{(X.Y)}{|X|.|Y|} \tag{8}$$

Choosing the right similarity measure is a challenging task, many researchers made this decision through experiments such as in (Rana et al., 2019), the authors made a comparison between Canberra, Chi-square, Manhattan and Euclidean distances. They found that using Euclidean distance as a similarity measure achieved higher precision. While authors in (Raza et al., 2018) included L1, L2, weighted L1, Chi Square, Square Chord and Extended Canberra in the experiments. They found that using weighted L1 was the best choice to the proposed descriptor because weighted L1 gave better results in terms of computational complexity and accuracy. Although L1 and L2 have lower complexity than Square Chord, Canberra and Chi Square, the retrieval accuracy was also lower because both are noise sensitive and did not consider the neighbor bins. Other researchers believed that features should not be equally treated when measuring the similarity between query image and dataset images. Authors in (Z. Zhao et al., 2016) assigned weights to color, texture and shape features based on their experience.

Although, similarity measure was the topic for many researchers (e.g., (Arevalillo-Herráez et al., 2008),(Rui et al., 2008)) over the last years, choosing the appropriate similarity is still an open research area which needs more exploring. For a detailed investigation of popularly used distance metrics in CBIR, it is recommended to read (Alzu'bi et al., 2015), more details about mathematical expressions could be found in (Cha, 2007).

7. Image datasets

The appropriate dataset is chosen based on the algorithm, the problem to be solved, and the application. In CBIR domain, a wide range of datasets vary in the number of images they may contain, the type of images, and the way they are collected. In general, image datasets usually comprise multiple categories or classes. Each category (class) contains similar semantics, and this categorization is performed by the authors in accordance with the semantic concept. Usually, images of datasets are manually collected and labeled by authors. When it comes to selecting the right dataset, there are a few drawbacks to consider. First, due to variations in perspective and class, some images can be divided into two groups. Second, labelled images (or tagged images) may not be correctly annotated, compromising the system's accuracy. Finally, some images are classified under different categories although they have very similar semantic contexts.

As a result of these reasons, choosing the dataset requires careful attention. For instance, CBIR systems based on machine learning algorithms need a sufficient dataset for training and testing. The Corel dataset has more than one version, namely, Corel 1k, Corel 1.5k, Corel 5k, and Corel 10k. Corel 1k contains 1,000 colored images (with either 256 × 384 or 384 × 256 pixels) divided into 10

classes, and each class contains 100 images. Corel 1.5k contains 15 classes with 100 images in each class, and each image is either 256 × 384 or 384 × 256 pixels. Corel 5k contains 50 semantic classes, and each class has 100 images with a size of 126 × 187 or 187 × 126 pixels, and all images are colored images. Corel 10k contains 10 semantic classes with 100 images with the same size as Corel 1k and Corel 5k. All images in the Corel datasets are natural images (P. Liu et al., 2017). Figure 7(a) shows samples of the Corel dataset. The Holidays dataset (Jain et al., 2011) is also considered as a natural dataset (P. Liu et al., 2017), containing 1,491 images divided into 500 classes with varying numbers of images in each class (ranging from two to thirteen images in each class) (Bibi, Mehmood, Yousaf, Saba, Sardaraz, Rehman et al., 2020a); see Figure 7(b). Grayscale datasets are also available, and Brodatz is a textural dataset (P. Liu et al., 2017)(Bibi, Mehmood, Yousaf, Saba, Sardaraz, Rehman et al., 2020b), the samples of which are shown in Figure 7(c). As mentioned before, many datasets have limited use for a specific field, such as the IRMA dataset (Shamna et al., 2019)(Öztürk, 2020), which is a medical dataset that contains 57 categories with a total of 14,410 images. Table 6 shows the most widely used datasets in the CBIR domain.

8. Performance evaluation

CBIR system evaluation should be conducted by considering predefined system formulas without depending on human intervention because it can be mistaken, long lasting, and subjective. Although a set of well-known criteria are available to assess the performance and accuracy of CBIR systems, choosing the most suitable measurement formula depends on several factors: the used method, the algorithm itself, and the domain of the problem. Below are the most widely used metrics.

(I)

Precision (P) is the number of the relevant retrieved images to the total number of the retrieved images (Haji et al., 2019).

$$P = \frac{No.\ of\ relevant\ images}{No.\ of\ retrieved\ images} \tag{9}$$

(I)

Recall (R) is the number of the relevant retrieved images to the total number of relevant images in the dataset (number of images in each class if the dataset is labeled) (ElAdel et al., 2016)(Bala & Kaur, 2016).

$$R = \frac{No.\ of\ relevant\ images}{No.\ of\ all\ relevant\ images\ in\ dataset} \tag{10}$$

(I)

Average precision (AP) is another globally known performance measure. The AP for one query q is the mean value of the precision values at every relevant image (Latif et al., 2019)

$$APq = \frac{1}{N_{RI}} \sum_{q=1}^{N_{RI}} P_q(R_q) \tag{11}$$

N_{RI} is the number of relevant images in the database for the current query image (q). Rq is a binary pointer that equals to 1 if the q-th retrieved image result is relevant to the present query image and 0 if the q-th retrieved image result is irrelevant to the present query image.

(I)

Mean average precision (MAP) is the mean of AP over all the queries (NQ) (Alzu'bi et al., 2015)

$$\text{MAPq} = \frac{1}{N_Q}\sum_{q=1}^{N_Q} AP_q \tag{12}$$

(I) F1-score (F-measure) is a combination of precision and recall in a single measure, which is the harmonic mean that is defined as (Makhoul et al., 1999)

$$F = \frac{(1+\beta^2)PR}{\beta^2 P+R} \tag{13}$$

β is a non-negative number(Alzu'bi et al., 2015). Generally, $\beta = 1$, which is why the F1-score, or F1-measure is referred to as such. The equation after substituting $\beta = 1$ will be (Baig et al., 2020)

$$F1 = 2 \times \left(\frac{P_k.R_k}{P_k+R_k}\right) \tag{14}$$

(I) Computational cost (running time) is a crucial attribute of CBIR techniques, especially in evaluating real-time applications. Running time can be discussed in terms of feature extraction time only (Sharif et al., 2019), in terms of feature extraction and total retrieval time (Mehmood et al., 2018)(Ahmed et al., 2019), or in terms of total retrieved time only (Yousuf et al., 2018), (ElAdel et al., 2016). However, it is infrequently discussed by researchers.

(II) Confusion Matrix: is a 2-D matrix which is usually used to summarize the classifiers' performance (Sammut & Webb, 2017). One dimension represents the actual class of the object while the other dimension represents the predicted class, Table 7 shows an example of two class confusion matrix.

9. Discussion and future research directions

Text-based image retrieval is the basis for a wide range of image search engines. Although it is not satisfactory, it is a common method. The need for efficient content-based image retrieval has motivated many researchers to develop such systems. These developed systems suffer from many limitations and drawbacks, such as the need for a proper feature selection and extraction method, which is the basis for any CBIR technique. Feature extraction is mainly performed by two methods: either by handcrafted feature extraction which depends on global or local feature extraction or by using machine learning algorithms to provide feature representation. Selecting proper features that reflect semantic and visual perceptions contained in images remains an open research challenge. Feature extraction based on discrete orthogonal polynomial can be investigated to increase the accuracy of CBIR. Examples of discrete orthogonal polynomials are Kravchuk polynomials (Mahmmod et al., 2020), Meixner polynomials (Abdulhussain & Mahmmod, 2021), Charlier polynomials (Abdul-Hadi et al., 2020), Tchebichef polynomials (Abdulhussain et al., 2017), and squared Kravchuk–Tchebichef polynomials (Abdulhussain, Ramli, Mahmmod et al., 2019), as well as the orthogonal polynomial embedded image kernel (Abdulhussain, Ramli, Hussain et al., 2019), which increases accuracy and reduces the computation cost in other computer vision fields (Abdulhussain et al., 2021).

Depending on extracting one feature is not enough to achieve high retrieval performance, and fusing more features has a negative effect of increasing the feature vector dimensions. Therefore, dimensionality reduction methods are another research direction that needs to be investigated more carefully. Moreover, research efforts to improve image representation with low to moderate feature vector dimensions are highly recommended.

In general, information retrieval algorithms in recent years obtained the benefits of using different machine learning algorithms such as deep learning, SVM, and k-means. Therefore, they are predicted to receive more attention in the forthcoming years.

To improve user satisfaction and increase accuracy, many CBIR algorithms in the literature contain a relevance feedback stage, which requires user intervention to specify relevant and irrelevant retrieved images. The system updates or reweights the representation and metrics, and then recalculates the revised results. This stage is a challenging task that needs more attention to achieve user satisfaction with minimal iterations.

CBIR systems in special domains have a number of specific characteristics that should be taken into account when designing such systems from the beginning. One of these characteristics is the use of special datasets such as medical care images. CBIR systems are extensively employed in many precise domains. Another challenging task is to create new image datasets that meet the requirements of these specific domains.

10. Conclusion

The need to find an efficient image retrieval mechanism based on image content is motivated by the large amount of image databases and the absence of an efficient text-based image retrieval method. This paper presented a literature review on different studies in the CBIR domain over the last six years. This paper also discussed the general CBIR framework stages and the most recent techniques used to reduce the semantic gap. Finally, this paper highlighted some of the most important issues that affect the performance of CBIR and concentrated on several directions that may help in creating a novel CBIR. However, designing an algorithm that provides high retrieval accuracy with minimized computational cost is a challenging task.

The challenge of any CBIR algorithm is the semantic gap between the high-level meaning of the image and the visual features because the CBIR algorithms start with simple low-level features. To bridge this gap, many efforts have been performed in this field. The CBIR algorithm has been developed by using different and novel features as well as fusion between features. These types of features are experimentally designed and tested. The literature shows that researchers have made great efforts in this field. CBIR algorithms can be divided into two categories—global feature extraction and local feature extraction—based on the feature extraction methods. Both global and local features are considered low-level features. Feature fusion is necessary to reduce the semantic gap and increase the accuracy. Moreover, fusion between local and global features is one trend that shows good retrieval results and needs more attention. As a development, researchers designed CBIR algorithm using machine learning algorithms because of the limitation of the conventional features and good results were obtained when employing machine learning algorithms. Recently, research in the CBIR domain has focused on employing deep neural networks, which provide good results on many image datasets. These types of algorithm have been developed quickly and efficiently in the recent years.

To accomplish an effective CBIR framework, the framework's components must be chosen in a balanced way; this study helps in investigating these components. Such effective CBIR frameworks will contribute to many real-world applications, such as medical applications, web searches, and social media.

To sum up, an algorithm that elevates the semantic gap is highly demanded. The design of the algorithm should consider the following: first, the algorithm needs to consider the feature extraction as well as similarity measure as they influence the performance of the CBIR. Second, more features can be extracted to enhance the accuracy of the CBIR and maintain the computational cost as it is considered important factor in the real-time applications. Third, merging local and global features will lead to a balanced design because local features are more robust against scale, translation and rotation changes than global features; and global features are faster in feature extraction and similarity measurements. Fourth, machine learning algorithms can be used in different stages of CBIR to increase system accuracy but need more attention to be paid to their computation cost. Finally, there is a tradeoff between system's accuracy and computational cost.

Funding
The authors received no direct funding for this research.

Author details
Ibtihaal M. Hameed[1]
Sadiq H. Abdulhussain[1]
E-mail: sadiqh76@yahoo.com
ORCID ID: http://orcid.org/0000-0002-6439-0082
Basheera M. Mahmmod[1]
ORCID ID: http://orcid.org/0000-0002-4121-0843
[1] Department of Computer Engineering, University of Baghdad, Baghdad 10071, Iraq.

Citation information
Cite this article as: Content-based image retrieval: A review of recent trends, Ibtihaal M. Hameed, Sadiq H. Abdulhussain & Basheera M. Mahmmod, *Cogent Engineering* (2021), 8: 1927469.

References

Abdul-Hadi, A. M., Abdulhussain, S. H., Mahmmod, B. M., & Pham, D. (2020, January). On the computational aspects of Charlier polynomials. *Cogent Engineering*, *7*(1), 1763553. https://doi.org/10.1080/23311916.2020.1763553

Abdulhussain, S. H., & Mahmmod, B. M. (2021, April). Fast and efficient recursive algorithm of Meixner polynomials. *Journal of Real-Time Image Processing*. https://doi.org/10.1007/s11554-021-01093-z

Abdulhussain, S. H., Mahmmod, B. M., Naser, M. A., Alsabah, M. Q., Ali, R., & Al-Haddad, S. A. R. (2021, March). A robust handwritten numeral recognition using hybrid orthogonal polynomials and moments. *Sensors*, *21*(6), 1999. https://doi.org/10.3390/s21061999

Abdulhussain, S. H., Ramli, A. R., Al-Haddad, S. A. R., Mahmmod, B. M., & Jassim, W. A. (2017). On computational aspects of tchebichef polynomials for higher polynomial order. *IEEE Access*, *5*(1), 2470–2478. https://doi.org/10.1109/ACCESS.2017.2669218

Abdulhussain, S. H., Ramli, A. R., Hussain, A. J., Mahmmod, B. M., & Jassim, W. A., "Orthogonal polynomial embedded image kernel," in *Proceedings of the International Conference on Information and Communication Technology - ICICT '19 Baghdad, Iraq* (pp. 215–221).

Abdulhussain, S. H., Ramli, A. R., Mahmmod, B. M., Saripan, M. I., Al-Haddad, S. A. R., & Jassim, W. A. (2019, May). A new hybrid form of Krawtchouk and Tchebichef Polynomials: Design and application. *Journal of Mathematical Imaging and Vision*, *61*(4), 555–570. https://doi.org/10.1007/s10851-018-0863-4

Agarwal, M., Singhal, A., & Lall, B. (2019, November). Multi-channel local ternary pattern for content-based image retrieval. *Pattern Analysis and Applications*, *22*(4), 1585–1596. https://doi.org/10.1007/s10044-019-00787-2

Ahmed, K. T., Ummesafi, S., & Iqbal, A. (2019, November). Content based image retrieval using image features information fusion. *Information Fusion*, *51*, 76–99. https://doi.org/10.1016/j.inffus.2018.11.004

Ali, F., and Hashem, A. (2020, June). Content Based Image Retrieval (CBIR) by statistical methods. *Baghdad Science Journal*, *17* (2(SI)), 694. https://doi.org/10.21123/bsj.2020.17.2(SI).0694

Alsmadi, M. K. (2020, April). Content-based image retrieval using color, shape and texture descriptors and features. *Arabian Journal for Science and Engineering*, *45*(4), 3317–3330. https://doi.org/10.1007/s13369-020-04384-y

Alzu'bi, A., Amira, A., & Ramzan, N. (2015, October). Semantic content-based image retrieval: A comprehensive study. *Journal of Visual Communication and Image Representation*, *32*, 20–54. https://doi.org/10.1016/j.jvcir.2015.07.012

Alzu'bi, A., Amira, A., & Ramzan, N. (2017, August). Content-based image retrieval with compact deep convolutional features. *Neurocomputing*, *249*, 95–105. https://doi.org/10.1016/j.neucom.2017.03.072

Arevalillo-Herráez, M., Domingo, J., & Ferri, F. J. (2008, December). Combining similarity measures in content-based image retrieval. *Pattern Recognition Letters*, *29*(16), 2174–2181. https://doi.org/10.1016/j.patrec.2008.08.003

Arthur, D., & Vassilvitskii, S. (2007, January). K-means++: The advantages of careful seeding. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, *07-09*, 1027–1035. https://dl.acm.org/doi/10.5555/1283383.1283494

Ashraf, R., Ahmedm, M., Jabbar, S., Khalid, S., Ahmad, A., Din, S. & Jeon, G. (2018, March). Content based image retrieval by using color descriptor and discrete Wavelet transform. *Journal of Medical Systems*, *42* (3), 44. https://doi.org/10.1007/s10916-017-0880-7

Ashraf, R., Ahmed, M., Ahmad, U., Habib, M. A., Jabbar, S., & Naseer, K. (2020, April). MDCBIR-MF: Multimedia data for content-based image retrieval by using multiple features. *Multimedia Tools and Applications*, *79*(13–14), 8553–8579. https://doi.org/10.1007/s11042-018-5961-1

Ashraf, R., Bashir, K., Irtaza, A., & Mahmood, M. (2015, May). Content based image retrieval using embedded neural networks with bandletized regions. *Entropy*, *17*(6), 3552–3580. https://doi.org/10.3390/e17063552

Bai, C., Chen, J., Huang, L., Kpalma, K., & Chen, S. (2018, January). Saliency-based multi-feature modeling for semantic image retrieval. *Journal of Visual Communication and Image Representation*, *50*, 199–204. https://doi.org/10.1016/j.jvcir.2017.11.021

Baig, F., Mehmood, Z., Rashid, M., Javid, M. A., Rehman, A., Saba, T., & Adnan, A. (2020, March). Boosting the performance of the BoVW model using SURF–CoHOG-based sparse features with relevance feedback for CBIR. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, *44*(1), 99–118. https://doi.org/10.1007/s40998-019-00237-z

Bala, A., & Kaur, T. (2016, March). Local texton XOR patterns: A new feature descriptor for content-based image retrieval. *International Journal of Engineering Technology and Scientific Innovation*, *19*(1), 101–112. https://doi.org/10.1016/j.jestch.2015.06.008

Banerjee, I., Kurtz, C., Devorah, A. E., Do, B., Rubin, D. L., & Beaulieu, C. F. (2018, August). Relevance feedback for enhancing content based image retrieval and automatic prediction of semantic image features: Application to bone tumor radiographs. *Journal of Biomedical Informatics*, *84*, 123–135. https://doi.org/10.1016/j.jbi.2018.07.002

Bansal, M., Kumar, M., & Kumar, M. (2020, February). 2D object recognition techniques: State-of-the-art work. *Archives of Computational Methods in Engineering, 28* (3), 1147-1161. https://doi.org/10.1007/s11831-020-09409-1

Basheer, I., & Hajmeer, M. (2000, December). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*,

43(1), 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008, June). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359. https://doi.org/10.1016/j.cviu.2007.09.014

Bibi, R., Mehmood, Z., Yousaf, R. M., Saba, T., Sardaraz, M., & Rehman, A. (2020a, April). Query-by-visual-search: Multimodal framework for content-based image retrieval. *Journal of Ambient Intelligence and Humanized Computing, 11*(11), 5629-5648. https://doi.org/10.1007/s12652-020-01923-1

Bibi, R., Mehmood, Z., Yousaf, R. M., Saba, T., Sardaraz, M., & Rehman, A.“*SIPI-USC Brodatz texture image database*.” [Online]. University of Southern California. Available: http://sipi.usc.edu/database/database.php?volume=textures. [Accessed: 01-Jan-2020b].

Cha, S.-H. (2007). Comprehensive Survey on Distance Similarity measures between probability density functions. *International Journal of Mathematics Model Methods Applied Sciiences*, *1*(4). https://www.naun.org/main/NAUN/ijmmas/mmmas-49.pdf

Chaudhry, S., & Chandra, R. (2016, October). Unconstrained face detection from a mobile source using convolutional neural networks. *Lecture Notes in Computer Science*, *9948*, 567–576. *Including sub-series Lecture Notes in Artificial* Intelligence *and* Lecture Notes in Bioinformatics. https://link.springer.com/chapter/10.1007%2F978-3-319-46672-9_63

Chun, Y. D., Kim, N. C., & Jang, I. H. (2008, October). Content-based image retrieval using multiresolution color and texture features. *IEEE Transactions on Multimedia*, *10*(6), 1073–1084. https://doi.org/10.1109/TMM.2008.2001357

Ciocca, G., & Schettini, R. (1999, September). A relevance feedback mechanism for content-based image retrieval. *Information Processing & Management*, *35*(5), 605–632. https://doi.org/10.1016/S0306-4573(99)00021-7

Cortes, C., & Vapnik, V. (1995, September). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cross, G. R., & Jain, A. K. (1983, January). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-5*(1), 25–39. https://doi.org/10.1109/TPAMI.1983.4767341

Cula, O. G., & Dana, K. J. (2004). 3D texture recognition using bidirectional feature histograms. *International Journal of Computer Vision*, *59*(1), 33–60. https://doi.org/10.1023/B:VISI.0000020670.05764.55

Dalal, N., & Triggs, B., “Histograms of oriented gradients for human detection,” In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 2005. San Diego, CA.

Das, R., Thepade, S., Bhattacharya, S., & Ghosh, S. (2016). Retrieval architecture with classified query for content based image recognition. *Applied Computer Intellegent Software Computers*, *2016*, 1–9. doi: 10.1155/2016/1861247

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008, April). Image retrieval. *ACM Computing Surveys*, *40*(2), 1–60. https://doi.org/10.1145/1348246.1348248

Duanmu, X., “Image retrieval using color moment invariant,” in *2010 Seventh International Conference on Information Technology*: New Generations, 2010, pp. 200–203. Las Vegas, Nevada.

ElAdel, A., Ejbali, R., Zaied, M., & Ben Amar, C. (2016, August). A hybrid approach for content-based image retrieval based on fast beta wavelet network and fuzzy decision support system. *Machine Vision and Applications*, *27*(6), 781–799. https://doi.org/10.1007/s00138-016-0789-z

Fei-Fei, L., Fergus, R., & Perona, P. (2007, April). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, *106*(1), 59–70. https://doi.org/10.1016/j.cviu.2005.09.012

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., .. Yanker, P. (1995). Query by image and video content: The QBIC system. *Computer (Long. Beach. Calif)*, *28*(9), 23–32. http://ieeexplore.ieee.org/document/410146/

Forssén, P. E., & Lowe, D. G. (2007). Shape descriptors for maximally stable extremal regions. *Proceedings / IEEE International Conference on Computer Vision*, 0–7. DOI: 10.1109/ICCV.2007.4409025

Fowler, B. (2000, February). A Sociological Analysis of the Satanic Verses Affair. *Theory, Culture & Society*, *17*(1), 39–61. https://doi.org/10.1177/02632760022050997

Garg, M., & Dhiman, G. (2020, June). A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants. *Neural Computing & Applications, 33*(4), 1311-1328. DOI: 10.1007/s00521-020-05017-z

Ghosh, N., Agrawal, S., & Motwani, M. 2018. A survey of feature extraction for content-based image retrieval system. *Lecture notes in networks and systems* (Vol. 34 305–313).Springer. https://doi.org/10.1007/978-981-10-8198-9_32

Ghrabat, M. J. J., Ma, G., Maolood, I. Y., Alresheedi, S. S., & Abduljabbar, Z. A. (2019, December). An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier. *Human-centric Computing and Information Sciences*, *9*(1), 31. https://doi.org/10.1186/s13673-019-0191-8

Gogul, I., & Kumar, V. S., “Flower species recognition system using convolution neural networks and transfer learning,” in *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*, 2017, no. March, pp. 1–6. Chennai, India.

Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset. *Caltech mimeo*, *11*(1), 20. https://authors.library.caltech.edu/7694/

Guo, Z., Zhang, L., & Zhang, D. (2010, March). Rotation invariant texture classification using LBP variance (LBPV) with global matching. *Pattern Recognition*, *43*(3), 706–719. https://doi.org/10.1016/j.patcog.2009.08.017

Haji, M. S., Alkawaz, M. H., Rehman, A., & Saba, T. (2019). Content-based image retrieval: A deep look at features prospectus. *International Journal of Computational Vision and Robotics*, *9*(1), 14. https://doi.org/10.1504/IJCVR.2019.098004

Halawani, A. H., Teynor, A., Setia, L., Brunner, G., & Retrieval, C. I. (2006, January). Fundamentals and Applications of Image Retrieval : An Overview. *Image (Rochester, N.Y.)*, 14–23.

Harris, C., & Stephens, M., “A combined corner and edge detector,” In *Procedings of the Alvey Vision Conference 1988*, 1988, pp. 23.1–23.6. Manchester, UK.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hawlick, R. M. (2017). Statistical and structural approaches to texture. *Advances in Intelligent Systems and Computing*, *459*(5), 1–644. doi: 10.1109/PROC.1979.11328

Howarth, P., & Rüger, S. (2005). Fractional distance measures for content-based image retrieval. 447–456. doi: 10.1007/978-3-540-31865-1_32

Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J., & Zabih, R., "Image indexing using color correlograms," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 762–768. San Juan, PR.

Idan, Z. N., Abdulhussain, S. H., & Al-Haddad, S. A. R. (2020). A new separable moments based on tchebichef-krawtchouk polynomials. *IEEE Access*, *8*, 41013–41025. https://doi.org/10.1109/ACCESS.2020.2977305

Jabeen, S., Mehmood, Z., Mahmood, T., Saba, T., Rehman, A., Mahmood, M. T., & Rubin, D. L. (2018, April). An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model. *PLoS One*, *13*(4), e0194526. https://doi.org/10.1371/journal.pone.0194526

Jacobs, C. E., Finkelstein, A., & Salesin, D. H., "Fast multiresolution image querying," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95, New York, USA*, 1995, pp. 277–286.

Jain, M., Jégou, H., & Gros, P., "Asymmetric hamming embedding," in *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, New York, USA, 2011, no. May, p. 1441.

Jenni, K., Mandala, S., & Sunar, M. S. (2015). Content based image retrieval using colour strings comparison. *Procedia Computer Science*, *50*, 374–379. https://doi.org/10.1016/j.procs.2015.04.032

Kokare, M., Chatterji, B. N., & Biswas, P. K. (2002, May). A survey on current content based image retrieval methods. *IETE Journal of Research*, *48*(3–4), 261–271. https://doi.org/10.1080/03772063.2002.11416285

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, May). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lance, G. N., & Williams, W. T. (1966, May). Computer programs for hierarchical polythetic classification ('Similarity Analyses'). *The Computer Journal*, *9*(1), 60–64. https://doi.org/10.1093/comjnl/9.1.60

Lance, G. N., & Williams, W. T. (1967). Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*, *1*(1), 15–20. https://www.semanticscholar.org/paper/Mixed-Data-Classificatory-Programs-I-Agglomerative-Lance-Williams/4072b6353dd761ba0c219a4b860073f08bb3db27

Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N. I., Zafar, B., Dar, S. H., Sajid, M., & Khalil, T. (2019, August). Content-based image retrieval and feature extraction: a comprehensive review. *Mathematical Problems in Engineering*, 1–21. https://doi.org/10.1155/2019/9658350

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

Leutenegger, S., Chli, M., & Siegwart, R. Y., "BRISK: Binary Robust invariant scalable keypoints," in *2011 International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 2548–2555.

Li, H., Qin, J., Xiang, X., Pan, L., Ma, W., & Xiong, N. N. (2018). An efficient image matching algorithm based on adaptive threshold and RANSAC. *IEEE Access*, *6*, 66963–66971. https://doi.org/10.1109/ACCESS.2018.2878147

Ling, H., & Jacobs, D. W. (2007). Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 286–299. https://doi.org/10.1109/TPAMI.2007.41

Liu, P., Guo, J.-M., Wu, C.-Y., & Cai, D. (2017, December). Fusion of deep learning and compressed domain features for content-based image retrieval. *The Computer Journal*, *26*(12), 5706–5717. https://doi.org/10.1109/TIP.2017.2736343

Liu, Y., Zhang, D., Lu, G., & Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, *40*(1), 262–282. https://doi.org/10.1016/j.patcog.2006.04.045

Low, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Mahmmod, B. M., Abdul-Hadi, A. M., Abdulhussain, S. H., & Hussien, A. (2020, August). On computational aspects of Krawtchouk polynomials for high orders. *Journal of Imaging*, *6*(8), 81. https://doi.org/10.3390/jimaging6080081

Makhoul, J., Kubala, F., Schwartz, R., & Weischedel, R. (1999, February). Performance measures for information extraction. In *Proc. DARPA broadcast news work.* (pp. 249–252).

Manjunath, B. S., Ohm, J.-R., Vasudevan, V. V., & Yamada, A. (2001, June). Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, *11*(6), 703–715. https://doi.org/10.1109/76.927424

Markowska-Kaczmar, U., & Kwaśnicka, H. (2018). Deep learning—a new era in bridging the semantic gap. *Intelligent Systems Reference Library*, *145*, 123–159. doi: 10.1007/978-3-319-73891-8_7

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(5), 530–549. https://doi.org/10.1109/TPAMI.2004.1273918

Marxsen, S., Hilbert, D., Laubenbacher, R., Sturmfels, B., David, H., & Laubenbacher, R. C. (1993). *Theory of algebraic invariants*. Cambridge University Press.

Mehmood, Z., Abbas, F., Mahmood, T., Javid, M. A., Rehman, A., & Nawaz, T. (2018, December). Content-based image retrieval based on visual words fusion versus features fusion of local and global features. *Arabian Journal for Science and Engineering*, *43*(12), 7265–7284. https://doi.org/10.1007/s13369-018-3062-0

Mehmood, Z., Anwar, S. M., Ali, N., Habib, H. A., & Rashid, M. (2016). A novel image retrieval based on a combination of local and global histograms of visual words. *Mathematical Problems in Engineering*, *2016*, 1–12. https://doi.org/10.1155/2016/8217250

Mittal, U., Srivastava, S., & Chawla, P., "Review of different techniques for object detection using deep learning," in *Proceedings of the Third International Conference on Advanced Informatics for Computing Research - ICAICR '19*, New York, USA, 2019, pp. 1–8.

Montazer, G. A., & Giveki, D. (2015, September). Content based image retrieval system using clustered scale invariant feature transforms. *Optik (Stuttg)*, *126*(18), 1695–1699. https://doi.org/10.1016/j.ijleo.2015.05.002

Naeem, M., Ashraf, R., Ali, N., Ahmad, M., & Habib, M. A., "Bottom up approach for better requirements elicitation," in *Proceedings of the International*

Conference on Future Networks and Distributed Systems - ICFNDS '17, New York, USA, 2017, 1305, pp. 1–4.

Nazir, A., Ashraf, R., Hamdani, T., & Ali, N., "Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor," in *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, 2018, January, pp. 1–6.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002, July). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Otsu, N. (1979, January). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, *9*(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Öztürk, Ş. (2020, December). Stacked auto-encoder based tagging with deep features for content-based medical image retrieval. *Expert Systems with Applications*, *161*, 113693. https://doi.org/10.1016/j.eswa.2020.113693

Patil, A., & Rane, M. (2021). Convolutional neural networks: an overview and its applications in pattern recognition. *In Smart Innovation, Systems and Technologies*, *195*(Insights into Imaging), 21–30. DOI: 10.1007/978-981-15-7078-0_3

Pavithra, L. K., & Sharmila, T. S. (2018, August). An efficient framework for image retrieval using color, texture and edge features. *Computers & Electrical Engineering*, *70*, 580–593. https://doi.org/10.1016/j.compeleceng.2017.08.030

Pavithra, L. K., & Sree Sharmila, T. (2019, December). An efficient seed points selection approach in dominant color descriptors (DCD). *Cluster Computing*, *22*(4), 1225–1240. https://doi.org/10.1007/s10586-019-02907-3

Pele, O., & Werman, M. (2010). The quadratic-chi histogram distance family. *Lecture Notes in Computer Science*, *6312* (2), 749–762. *Including subseries Lecture Notes in Artificial* Intelligence *and* Lecture Notes in Bioinformatics. DOI: 10.1007/978-3-642-15552-9_54

Perronnin, F., Douze, M., Pe, P., Schmid, C., & Sa, J. (2012). Aggregating local image descriptors into compact codes. *Analysis*, *34*(9), 1704–1716. https://doi.org/10.1109/TPAMI.2011.235

Phadikar, B. S., Phadikar, A., & Maity, G. K. (2018, May). Content-based image retrieval in DCT compressed domain with MPEG-7 edge descriptor and genetic algorithm. *Pattern Analysis and Applications*, *21*(2), 469–489. https://doi.org/10.1007/s10044-016-0589-0

Piras, L., & Giacinto, G. (2017, September). Information fusion in content based image retrieval: A comprehensive overview. *Information Fusion*, *37*, 50–60. https://doi.org/10.1016/j.inffus.2017.01.003

Ponomarev, A., Nalamwar, H. S., Babakov, I., Parkhi, C. S., & Buddhawar, G. (2016, February). Content-based image retrieval using color, texture and shape features. *Key Engineering Materials*, *685*, 872–876. https://doi.org/10.4028/www.scientific.net/KEM.685.872

Qin, J., Li, H., Xiang, X., Tan, Y., Pan, W., Ma, W., & Xiong, N. N. (2019). An encrypted image retrieval method based on harris corner optimization and LSH in cloud computing. *IEEE Access*, *7*, 24626–24633. https://doi.org/10.1109/ACCESS.2019.2894673

Qiu, G. (2003, January). Color image indexing using BTC. *IEEE Transactions on Image Processing*, *12*(1), 93–101. https://doi.org/10.1109/TIP.2002.807356

Raghunathan, B., & Acton, S. T., "A content based retrieval engine for circuit board inspection," in *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, vol. 1, pp. 104–108 1999. Kobe, Japan.

Rana, S. P., Dey, M., & Siarry, P. (2019, January). Boosting content based image retrieval performance through integration of parametric & nonparametric approaches. *Journal of Visual Communication and Image Representation*, *58*(3), 205–219. https://doi.org/10.1016/j.jvcir.2018.11.015

Raza, A., Dawood, H., Dawood, H., Shabbir, S., Mehboob, R., & Banjar, A. (2018). Correlated primary visual texton histogram features for content base image retrieval. *IEEE Access*, *6*, 46595–46616. https://doi.org/10.1109/ACCESS.2018.2866091

Raza, A., Nawaz, T., Dawood, H., & Dawood, H. (2019). Square texton histogram features for image retrieval. *Multimedia Tools and Applications*, *78*(3), 2719–2746. https://doi.org/10.1007/s11042-018-5795-x

Rosten, E., Porter, R., & Drummond, T., "Faster and better: A machine learning approach to corner detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, January. 2010. Hannover, Germany. https://doi.org/10.1109/TPAMI.2008.275

Rui, H. S., Ruger, D. S., Liu, H., & Huang, Z., "Dissimilarity measures for content-based image retrieval," In *2008 IEEE International Conference on Multimedia and Expo*, Hannover, Germany, 2008, pp. 1365–1368.

Sajjad, M., Ullah, A., Ahmad, J., Abbas, N., Rho, S., & Baik, S. W. (2018, February). Integrating salient colors with rotational invariant texture features for image representation in retrieval systems. *Multimedia Tools and Applications*, *77*(4), 4769–4789. https://doi.org/10.1007/s11042-017-5010-5

Salve, S. G., & Jondhale, K. C. (2010). Shape matching and object recognition using shape contexts. *Proceedings - 2010 3rd IEEE International Conference Computer Science Information Technology ICCSIT 2010*, *9*, 471–474. DOI: 10.1109/ICCSIT.2010.5565098

Sammut, C., & Webb, G. I. (Eds.). (2017). *Encyclopedia of machine learning and data mining*. Springer US.

Sánchez, J., Monzón, N., & Salgado, A. (2018, October). An analysis and implementation of the harris corner detector. *Image Processing Line*, *8*(9), 305–328. DOI: 10.5201/ipol.2018.229

Sarwar, A., Mehmood, Z., Saba, T., Qazi, K. A., Adnan, A., & Jamal, H. (2019, February). A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine. *Journal of Information Science*, *45* (1), 117–135. https://doi.org/10.1177/0165551518782825

Sergyan, S., "Color histogram features based image classification in content-based image retrieval systems," in *2008 6th International Symposium on Applied Machine Intelligence and Informatics*, Herlany, Slovakia, 2008, pp. 221–224.

Sezavar, A., Farsi, H., & Mohamadzadeh, S. (2019, August). Content-based image retrieval by combining convolutional neural networks and sparse representation. *Multimedia Tools and Applications*, *78*(15), 20895–20912. https://doi.org/10.1007/s11042-019-7321-1

Shamna, P., Govindan, V. K., & Abdul Nazeer, K. A. (2019, March). Content based medical image retrieval using topic and location model. *Journal of Biomedical Informatics*, *91*, 103112. https://doi.org/10.1016/j.jbi.2019.103112

Sharif, U., Mehmood, Z., Mahmood, T., Javid, M. A., Rehman, A., & Saba, T. (2019, August). Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artificial Intelligence Review*, *52*(2), 901–925. https://doi.org/10.1007/s10462-018-9636-0

Sharmi, N., Mohamed Shameem, P., & Parvathy, R. (2018). Content-based image retrieval using fast machine learning approach in cloud computing. In *Communications in computer and information science* (Vol. 837, pp. 434–444). Elsevier: Springer Singapore.

Shrivastava, N., & Tyagi, V. (2015, August). An efficient technique for retrieval of color images in large databases. *Computers & Electrical Engineering*, *46*, 314–327. https://doi.org/10.1016/j.compeleceng.2014.11.009

Shrivastava, N., & Tyagi, V. (2017, December). Corrigendum to "Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching" [Information Sciences 259 (2014) 212–224]. *Information Sciences*, *421*, 273. https://doi.org/10.1016/j.ins.2017.09.017

Simonyan, K., & Zisserman, A. (2014, September). Very deep convolutional networks for large-scale image recognition. *3rd International Conference Learning Represent ICLR 2015 - Conference Track Proceedings*, 1–14.

Song, W., Zhang, Y., Liu, F., Chai, Z., Ding, F., Qian, X., & Park, S. C. (2018, April). Taking advantage of multi-regions-based diagonal texture structure descriptor for image retrieval. *Expert Systems with Applications*, *96*, 347–357. https://doi.org/10.1016/j.eswa.2017.12.006

Srivastava, P., & Khare, A. (2017, January). Integration of wavelet transform, local binary patterns and moments for content-based image retrieval. *Journal of Visual Communication and Image Representation*, *42*, 78–103. https://doi.org/10.1016/j.jvcir.2016.11.008

Su, J.-H., Huang, W.-J., Yu, P. S., & Tseng, V. S. (2011, March). Efficient relevance feedback for content-based image retrieval by mining user navigation patterns. *IEEE Transactions on Knowledge and Data Engineering*, *23*(3), 360–372. https://doi.org/10.1109/TKDE.2010.124

Suk, T., & Flusser, J. (2011, September). Affine moment invariants generated by graph method. *Pattern Recognition*, *40*(2), 2047–2056. https://doi.org/10.1016/j.patcog.2010.05.015

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32. https://doi.org/10.1007/BF00130487

Tadi Bani, N., & Fekri-Ershad, S. (2019, August). Content-based image retrieval based on combination of texture and colour information extracted in spatial and frequency domains. *Electronic Library*, *37*(4), 650–666. https://doi.org/10.1108/EL-03-2019-0067

Tao, D., Tang, X., Xuelong, L., & Rui, Y. (2006, August). Direct kernel biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm. *IEEE Transactions on Multimedia*, *8*(4), 716–727. https://doi.org/10.1109/TMM.2005.861375

Tarawneh, A. S., Celik, C., Hassanat, A. B., & Chetverikov, D. (2020, February). Detailed investigation of deep features with sparse representation and dimensionality reduction in CBIR: A comparative study. *Intelligent Data Analysis*, *24*(1), 47–68. https://doi.org/10.3233/IDA-184411

Thusnavis Bella, M. I., & Vasuki, A. (2019, May). An efficient image retrieval framework using fused information feature. *Computers & Electrical Engineering*, *75*, 46–60. https://doi.org/10.1016/j.compeleceng.2019.01.022

Tian, D. (2018). Support vector machine for content-based image retrieval: A comprehensive overview. *Journal Information Hiding Multimedia Signal Processing*, *9*(6), 1464–1478. http://bit.kuas.edu.tw/~jihmsp/2018/vol9/Number6/JIH-MSP-2018-11-011.pdf

Tian, D. P. (2013). A review on image feature extraction and representation techniques. *International Journal Multimedia Ubiquitous Engineering*, *8*(4), 385–395. https://gvpress.com/journals/IJMUE/vol8_no4/39.pdf

Tsai, C.-F. (2012). Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*, *2012*, 1–19. https://doi.org/10.5402/2012/376804

Tyagi, V. (2017). Similarity measures and performance evaluation. In *Content-based image retrieval* (pp. 63–83). Springer Singapore.

Tzelepi, M., & Tefas, A. (2018, January). Deep convolutional learning for content based image retrieval. *Neurocomputing*, *275*, 2467–2478. https://doi.org/10.1016/j.neucom.2017.11.022

Varma, M., & Zisserman, A. (2009). CUReT1. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(11), 2032–2047. https://doi.org/10.1109/TPAMI.2008.182

Vedaldi, A., & Zisserman, A., "Sparse kernel approximations for efficient classification and detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 2320–2327.

Vimina, E. R., & Divya, M. O. (2020, September). Maximal multi-channel local binary pattern with colour information for CBIR. *Multimedia Tools and Applications*, *79*(35–36), 25357–25377. https://doi.org/10.1007/s11042-020-09207-8

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, *2018*, 1–13. DOI: 10.1155/2018/7068349

Wan, J., "Deep learning for content-based image retrieval," In *Proceedings of the ACM International Conference on Multimedia - MM '14*, New York, USA, 2014, pp. 157–166.

Wang, J. Z., Li, J., & Wiederholdy, G. (2000). SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. *Lecture Notes in Computer Science*, *1929* (9), 360–371. *(including subseries Lecture Notes in Artificial* Intelligence *and* Lecture Notes in Bioinformatics). DOI: 10.1007/3-540-40053-2_32

Wang, Z., Fan, B., & Wu, F., "Local intensity order pattern for feature description," in *2011 International Conference on Computer Vision*, 2011, no. May, pp. 603–610. New York, USA.

Watanabe, T., Ito, S., & Yokoi, K. (2010). Co-occurrence histograms of oriented gradients for human detection. *IPSJ Transactions on Computer Vision and Applications*, *2*, 39–47. https://doi.org/10.2197/ipsjtcva.2.39

Won, C. S. W., Park, D. K. P., & Park, S.-J. P. (2002, February). Efficient use of MPEG-7 edge histogram descriptor. *ETRI Journal*, *24*(1), 23–30. https://doi.org/10.4218/etrij.02.0102.0103

Wu, G.-D., & Huang, P.-H. (2013, February). A Vectorization-optimization-method-based type-2 fuzzy neural network for noisy data classification.

IEEE Transactions on Fuzzy Systems, *21*(1), 1–15. https://doi.org/10.1109/TFUZZ.2012.2197754

Xu, D., Cham, T. J., Yan, S., Duan, L., & Chang, S. (2010). Aligned pyramid matching. *IEEE Transactions on Circuits and Systems*, *20*(8), 1068–1079. https://doi.org/10.1109/TCSVT.2010.2051286

Xu, Y., Zhao, X., & Gong, J. (2019). A large-scale secure image retrieval method in cloud environment. *IEEE Access*, *7*, 160082–160090. https://doi.org/10.1109/ACCESS.2019.2951175

Yang, H.-F., Lin, K., & Chen, C.-S. (2018, February). Supervised learning of semantics-preserving hash via deep convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(2), 437–451. https://doi.org/10.1109/TPAMI.2017.2666812

Yoon, H., Park, C.-S., Kim, J. S., & Baek, J.-G. (2013, January). Algorithm learning based neural network integrating feature selection and classification. *Expert Systems with Applications*, *40*(1), 231–241. https://doi.org/10.1016/j.eswa.2012.07.018

Younus, Z. S., Mohamad, D., Saba, T., Alkawaz, H. M., Rehman, A., Al-Rodhaan, M., Al-Dhelaan, A. (2015, August). Content-based image retrieval using PSO and k-means clustering algorithm. *Arabic Journal Geoscience*, *8*(8), 6211–6224. doi: 10.1007/s12517-014-1584-7

Yousuf, M., Mehmood, Z., Habib, H. A., Mahmood, T., Saba, T., Rehman, A., & Rashid, M. (2018). A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval. *Mathematical Problems in Engineering*, *2018*, 1–13. https://doi.org/10.1155/2018/2134395

Zhang, D., Islam, M. M., & Lu, G. (2012, January). A review on automatic image annotation techniques. *Pattern Recognition*, *45*(1), 346–362. https://doi.org/10.1016/j.patcog.2011.05.013

Zhang, D., & Lu, G., "A comparative study on shape retrieval using Fourier descriptiors with different shape signatures," in *Intelligent Multimedia, Computing and Communications: Technologies and Applications of the Future: Proceedings of the International Conference on Intellient Multimedia and Distance Education*, 2001, pp. 1–9.

Zhang, D., & Lu, G. (2004, January). Review of shape representation and description techniques. *Pattern Recognition*, *37*(1), 1–19. https://doi.org/10.1016/j.patcog.2003.07.008

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computational Vision and Robotics*, *73*(2), 213–238. https://doi.org/10.1007/s11263-006-9794-4

Zhao, M., Zhang, H., & Sun, J. (2016, July). A novel image retrieval method based on multi-trend structure descriptor. *Journal of Visual Communication and Image Representation*, *38*, 73–81. https://doi.org/10.1016/j.jvcir.2016.02.016

Zhao, Z., Tian, Q., Sun, H., Jin, X., & Guo, J. (2016, January). Content based image retrieval scheme using color, texture and shape features. *International Journal of Signal Processing Image Processing Pattern Recognition*, *9*(1), 203–212. doi: 10.14257/ijsip.2016.9.1.19

Zheng, Q., Tian, X., Yang, M., & Wang, H. (2019). Differential learning: A powerful tool for interactive content-based image retrieval. *Engineering Letters*, *27*(1), 202–215. http://www.engineeringletters.com/issues_v27/issue_1/EL_27_1_23.pdf

Zheng, X., Tang, B., Gao, Z., Liu, E., & Luo, W. (2016, November). Study on image retrieval based on image texture and color statistical projection. *Neurocomputing*, *215*, 217–224. https://doi.org/10.1016/j.neucom.2015.07.157

Zhuo, L., Cheng, B., & Zhang, J. (2014). A comparative study of dimensionality reduction methods for large-scale image retrieval. *Neurocomputing*, *141*, 202–210. https://doi.org/10.1016/j.neucom.2014.03.014

Cogent Engineering **(ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com