Multimedia Tools and Applications (2021) 80:29561–29583
https://doi.org/10.1007/s11042-021-11045-1

State of the art content based image retrieval techniques using deep learning: a survey

Rajiv Kapoor[1] · Deepak Sharma[2] · Tarun Gulati[2]

Received: 26 January 2020 / Revised: 20 February 2021 / Accepted: 5 May 2021 /
Published online: 2 July 2021

Abstract
In the recent years the rapid growth of multimedia content makes the image retrieval a challenging research task. Content Based Image Retrieval (CBIR) is a technique which uses features of image to search user required image from large image dataset according to the user's request in the form of query image. Effective feature representation and similarity measures are very crucial to the retrieval performance of CBIR. The key challenge has been attributed to the well known semantic gap issue. The machine learning has been actively investigated as possible solution to bridge the semantic gap. The recent success of deep learning inspires as a hope for bridging the semantic gap in CBIR. In this paper, we investigate deep learning approach used for CBIR tasks under varied settings from our empirical studies; we find some encouraging conclusions and insights for future research.

Keywords Bits-scalable deep binary codes · Spatial maximal activator pooling · Semantic assisted visual hashing · Cosine distance · Deep hashing based on classification and quantization error (DHCQ) · Dot-diffused block truncation coding

1 Introduction

The multimedia researchers have broadly studied about feature representations and similarity measurement approaches which imperatively affect the retrieval pursuance of a Content Based Image Retrieval (CBIR) system. In spite of numerous researches carried out in the field of Content Based Image Retrieval (CBIR) system, the bridging of the semantic gap between low level features captured by machines and high level feature anticipated by human remains a vivid area of researcher's interest.

✉ Rajiv Kapoor
rajivkapoor.dtu@gmail.com

[1] Department of ECE, Delhi Technological University, Delhi, India

[2] Department of ECE, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, India

In last two decades, a number of Content Based Image Retrieval approaches were proposed to handle the challenge of semantic gap. The CBIR methods based on low level features like color, shape and texture are extensively explored and discussed in detail [6, 11, 37, 41, 63]. But CBIR based on color, shape and texture feature can limitedly address efficient retrieval as there is a considerable gap that exists between the analyzing ability using color, texture and shape features compared to the inability at the semantic level. Kokare et al. [28] explored the use of multidimensional indexing techniques for the Content Based Image Retrieval applications. As the database has grown with time, multidimensional indexing was used to speed up the queries. But the multidimensional indexing has proved to be of limited use for most image retrieval applications, due to its loss of efficiency as the number of dimensions increases. To further improve the retrieval efficiency, the focus of research shifted from low level feature extraction to the semantic gap reduction with high level semantics. Liu et al. [34] surveyed the distinct facets of research in this area like the use of object ontology to define high level concepts, the use of machine learning association methods, the introduction of relevance feedback in retrieval loop, generation of semantic templates and making use of text information. But an efficient complete CBIR system is not possible with only high level semantics. It requires the combination of high level semantics with low level features and efficient indexing tools. The research in the field of CBIR has focused on the traditional attributes. Another effective but not much explored method of CBIR, is browsing. Heesch [22] presented a survey to provide an organized overview of distinct browsing models that has been investigated over last one to two decades. Some of the browsing models solve retrieval problem better than others but there are some important issues with the browsing models like creating a browsing structure for the large collection can be computationally costly. The browsing models either use a pre computed structure that often are too rigid or initialize new query at every step that make the system slow for large collection. Another challenge with browsing models is the requirement of long-term learning. Rafiee et al. [50] presented a survey on CBIR using the techniques such as supervised, unsupervised learning and relevance feedback approaches to reduce the semantic gap. The supervised approach does not reasonably achieve desired level of generalization and accuracy, whereas in unsupervised approach, extra information has to be employed to seek desired retrievals. In case of relevance feedback, the efficient retrieval can be achieved by its integration with supervised / unsupervised learning techniques. Dharani et al. [10] also explored the application of various machine learning techniques in CBIR but in the survey they discussed Content Based Image Retrieval based on unlabeled images only. Another survey on CBIR based on different indexing techniques has been carried out by Mukherjee et al. [36]. They explored various bag of words based indexing approaches and concluded that the retrieval efficiency of L1 norm based similarity search is better for small size database. But the other methods like SR tree and Local Search Hashing approaches take less computation time than L1 or L2 norm based approach. One of the key challenges in CBIR is an accurate and fast indexing technique. Patel et al. [44] explored the various Content Based Image Retrieval system using hashing based indexing techniques. They illustrated that albeit hashing based indexing approaches performs better than tree based indexing methods still an efficient indexing method in CBIR requires of future research. There are different attributes that affects the retrieval performance of the CBIR system. Similarity techniques are one of the important attributes. Patel et al. [43] reviewed the various CBIR systems on the basis of distinct similarity measure techniques. The discussed methods have less computation cost but the performance of similarity approach depends on the number of relevant images related to the concept. Wan et al. [64] exploits a study on Content Based Image Retrieval system based

on deep learning. In this survey, CBIR systems using basic deep learning approaches are discussed. The basic deep learning performs better than shallow approaches but it would be very interesting to review CBIR system based on advanced deep learning approaches.

Even though the success of deep learning inspired researchers to explore deep learning based CBIR approaches, most applications that use deep learning are in the field of image recognition and classification and there is still limited focus on application of deep learning in the field of CBIR. This motivates to address questions like, whether deep learning is effective in CBIR work or not and how these methods perform in different domain image databases.

In recent years, the main focus of researchers was on efficient feature extraction methods and similarity metrics. It is very interesting to investigate its application in CBIR system; so this paper limits its focus to the description and analysis of the recent deep learning based CBIR methods. The deep binary codes performed better than other state of art methods in terms of accuracy by using late fusion approach. In case of large surveillance data the high retrieval performance achieved by using prioritized and segmented data with Spatial maximal activators (SMA) pooling. The semantic-assisted visual hashing approach shows superior performance by effectively integrating the visual and text semantics using offline learning. The bilinear CNN model is better in terms of learning more complex features by reducing feature maps to compact size. The VAE-GAN based deep hashing approach generates deep hash codes which contribute to high retrieval performance than other hashing approaches. The R-MAC based method performs better when used with the combination of query expansion (QE) and database side feature augmentation (DBA). Hopefully, this comprehensive study on CBIR systems using deep learning will provide lucrative knowledge and direction to the research in the field of CBIR. The remaining paper is organized as follows.

In Section 2, distinct deep learning based CBIR methods are reviewed and tabulated. The different databases used in CBIR systems based on deep learning are reviewed in Section 3. Section 4 & 5 summarizes the various similarity measures and evaluation parameters respectively used in CBIR. Section 6 contains the concluding remarks and future directions in the field of deep learning based CBIR systems.

2 Deep learning based CBIR approaches

In this Section, the content based image retrieval techniques using deep leaning methods are reviewed. The distinct content based image retrieval (CBIR) approaches based on deep learning are illustrated in Fig. 1.

2.1 Binary multi-scale multi-pooling approach

In this approach, deep binary codes [67] based on deep convolutional features are generated for image retrieval. The response of each feature map is compared to the average response of all features on the deep convolutional layer to create deep binary codes. The bit scalable binary codes are directly produced by the proposed spatial cross-summary strategy. When the image is passed through Convolutional Neural Network (CNN), the binary codes are produced on different deep layers. The deep binary codes generated at different layers contribute differently to search accuracy. The retrieval precision is further enhanced using late fusion approach. The deep learning codes generation is based on generic model and the late fusion approach is query adaptive.

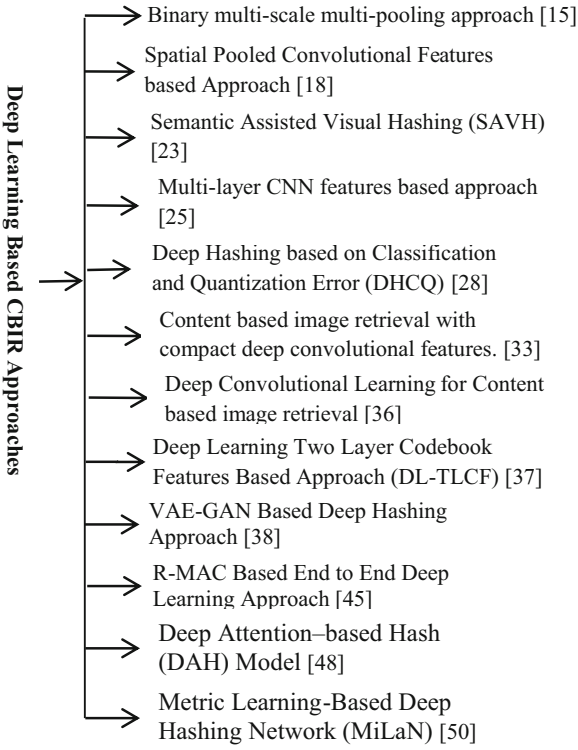

Fig. 1 Overview of Deep Learning based CBIR Approaches

2.1.1 Deep binary codes generation

To generate deep binary codes [67], a pre-trained CNN with 'L' deep convolutional layers is considered. When the input image 'I' is passed through the pre-trained network, the feature maps = $\{F_{i,j}\}$ are obtained for i = 1, 2, 3, −—— C_i. The feature map $F_{i,j}$ is the j^{th} feature map of i^{th} deep convolutional layer and C_i is number of convolutional kernels of deep convolutional layer. The size of feature map $F_{i,j}$ is $W_i \; x \; H_i \; x \; C_i$ where W_i and H_i are the width and height of each channel.

The feature maps generated by max-pooling or sum pooling are only translation invariant. The multi-scale pooling presents the scale-invariant feature maps. The features extracted from multiple scale regions are added and l2-normalized to represent the image. The feature maps are converted into binary codes. Since binary codes have less memory requirements and consume less retrieval time, theses codes are more suitable for large scale image search.

2.1.2 Spatial cross-summing

The different lengths of the feature maps result in a tradeoff between accuracy and efficiency. Low dimensional binary code have less requirement of computational and storage resources, while high dimensional binary codes attain more accuracy. The bit scalable seep binary codes are created by using a spatial cross summing strategy.

2.1.3 Dynamic late fusion

To further enhance the retrieval precision; the top-N search scores from deep binary codes [67] are fused. The hamming distance is used to measure the similarity between the images. The higher similarity means lower value of hamming distance. The modified hamming distance H_I=K-HI, where K in the size of deep learning codes.

2.2 Spatial pooled convolutional features based approach

Selective and low dimensional features are required for an efficient and accurate search of surveillance of data. Multi Trend Structure Descriptor (MTSD) [72] and Multi Scale Local Feature Pattern (MS-LSP) [2] apply pooling principle to local image features. Each sensor node of the modern surveillance network capture, prioritize and transmit visual data to central repositories. The transmission, storage and retrieval of a casual data become a difficult task due to the exponential growth of visual data from surveillance networks [1]. The most common but challenging task is the search of a particular visual object from the data. The proposed work is based on the extraction of optional kernels from the first layer of CNN that is pre trained on ImageNet dataset. The feature maps are generated by convolutional of input image with set of kernels. Then these feature maps are unified into single feature map using spatial maximal activator pooling technique. The particular surveillance objects are represented by computing a low dimensional feature vector. There is an improvement in performance over other state of art approaches for surveillance dataset. The potent features are required for the efficient retrieval of surveillance object. Before feature extraction the prioritization and segmentation of desired objects is necessary.

2.2.1 Image prioritization

All the data apprehended by the visual sensor is not useful. The important data is picked by using some efficient algorithms. Image prioritization scheme is used in this work. This approach is based on the concept of motion saliency.

2.2.2 Object segmentation

The prioritized data is further reduced by removing the unnecessary portion of the image using segmentation. Thus fewer amounts of data are used to extract efficient features transmitted to the server.

2.2.3 Feature extraction

The subnets of selected kernels are used to extract the features. The suitable kernels are selected on the basis of highest frequency usage frequency rate. Then, spatial maximal activator pooling strategy [1] is used to consolidate all the feature maps into a single feature map. The maximum activator value for each pixel position and each color channel produce by consolidated feature map have unique id that includes data, time sensor id and object identification.

2.2.4 Feature matching

Various similarity measures are used for image matching like Euclidean, Manhattan, Hellinger and Canberra distance. The performance of the histogram intersection approach and the Manhattan distance is very close. Therefore the Manhattan distance [1] is preferred due to less computational requirements.

2.3 Semantic assisted visual hashing (SAVH)

Hashing has become one of the active domains of research in the field of content based image retrieval. Single feature visual hashing can be categorized in two categories: data independent and data dependent hashing. Manifold based hashing [55] and deep learning based hashing [13] are in trend to extract binary hash codes. Multi View Latent Hashing (MVLH) [56] formulates binary representation learning of multimodal features. Multi View Alignment Hashing (MVAH) [33] incorporates regularized kernel non-negative matrix factorization in hash codes. The multi feature visual hashing provides comprehensive interpretation of visual contents and optional performance. In this Semantic Assisted Visual Hashing (SAVH) [73] method, the hash codes are integrated with the semantic assistance of auxiliary texts to characterize the correlation between images and shared topics. This unified structure enhances the effectiveness of the visual hashing approach. This method demonstrates superior performance over several state of the art approaches using standard dataset. The structure of the SAVH based Content Based Image Retrieval (CBIR) system has the phrases: offline learning and offline hashing.

2.3.1 Offline learning

The hash codes of database images and hash function of query images are generated in this phase. In the first step, the image pixels using visual and text features of image are identified. Then, with the help of topic hyper graph, a text-enhanced visual graph is constructed while text information guides the latent semantic topic detection. The framework having correlation of images and semantic topics is utilized for hash code of database images.

2.3.2 Online hashing

In online hashing [73], the extracted virtual features of the query image are mapped into binary codes with hash functions. The hashing space is used to calculate to similarity between query and database images. The most similar has minimum distance.

2.4 Multi-layer CNN features based approach

For image retrieval task, the CNN features are firstly used in Overfeat approach. All convolutional layers of GoogleNet [57] are evaluated to extend good results. Convolutional neural network features contain different information at different layers. Lower layer feature have more detailed information but contains semantic ambiguity also. Whereas higher layer features have less but more semantic information. The proposed work effectively attained the complementary strengths of the different layers of CNN [71]. The effectiveness of lower layer similarity is focused on a mapping function, where similarity between query image and its

nearest neighbor is measured in terms of similar semantics. This work outperforms renowned retrievals benchmark datasets. The fully connected layer features are the output of their corresponding layer whereas the aggregated features are the features of convolutional layers. The problem in direct concatenation is the inference of high layer features or low layer features.

The threshold t and exponent P is fixed to a certain value get the best performance from the system. The threshold t is threshold of similarity of the selected layers of the images for similarity matching and exponent P controls the degree of weakening in the mapping function of this approach. The value of P ranges between 0 and 1. The two layers (fc2 and conv 4) are used to evaluate the impact of t and P. The CNN is trained on ImageNet ILSVRC 2012. The framework of CNN has five convolutional layers and three fc [71]. The experiments are conducted using three convolutional layers (3, 4) and two fully connected layer (fc and fc2) to calculate similarity between images. The fully connected layer output of convolutional layer is the extracted set of local features. The Bag of Words (BoW) is used to encode the output of the convolutional layer in a universal representation of the image.

2.5 Deep hashing based on classification and quantization error (DHCQ)

The Supervised Discrete Hashing (SDH) [54] methods are handcrafted visual feature based methods. The deep hashing methods are also very important for image retrieval. Erin binary et al. [31] proposed a DH approach by minimizing the quantization error whereas Yang et al. [68] presented a DH method by minimizing classification error. Due to the fast retrieval speed and low storage cost, hashing approach has been very popular for Approx Nearest Neighbor search. A novel supervised approach is proposed for scalable face image retrieval by concurrently learning feature of images, hash codes and classifier in [59]: the Deep Hashing based on Classification and Quantization error (DHCQ). The deep convolutional network framework extracts discriminative features to generate hash codes and image labels. The learning process of the deep network is guided by the quantization error and the prediction error. Experiments are carried out to demonstrate the effectiveness of the presented work in comparison with the state of the art of existing hashing methods.

2.5.1 Deep binary function

A CNN with powerful feature extraction capabilities is used as a deep hashing learning framework [59]. The deep network has four convolutional layers with max pooling and two fully connected layers: deep feature layer and latent layer. The multi scale features are extracted by connecting third and fourth convolutional layers with deep featured fully connected layer. The features of the fourth layer are more global than third layer features. To learn hashing function, the output of the latent layer is quantized to discrete values. The number of nodes of the latent layer μ is set equal to the length of the desired hashing bits. The sigmoid function is used as activation function of the latent layer. The activation function maps the data into the range of (0, 1). The ReLU function is used as activation function to reset the layers.

2.6 Content based image retrieval with compact deep convolutional features

VLAD and Fisher Vector (FV) are popularly used due to their effectiveness in image retrieval. The VLAD and its variants are applied especially at high layer; in case of CNN image retrieval

applications. Ng et al. [39] encoded the multiple scale and layer features into single VLAD vector. Paulin et al. [45] utilize the patch level descriptor by using a convolutional kernel network. Lin et al. [30] also proposed a bilinear CNN-loss framework for image categorization using two features extractor. Gao et al. [12] recommended a compact pooling approach which is important in terms of computation and memory usage. The various CNN based approaches are utilized for image retrieval task. The image features are extracted at the last layer of the single CNN framework using order less quantization approaches. But this work presents bilinear CNN based structure using two parallel CNNs as feature extractors in context of CBIR. The image features are extracted at different locations and scales using activation of convolutional layers. The deep CNN framework [3] is initially pre-trained on a large generic image dataset and fine tune for CBIR application. The compact and highly discrimination features are produced by applying bilinear root pooling approach to low dimensional pooling layer. The fine tune of network is performed by end to end back propagation to tune parameters for image retrieval work. The work comprises of three steps: 1. Initialization of deep CNN network pre rained on million of images. 2. Fine tuning of bilinear CNN framework on image retrieval dataset. 3. Feature extractor of query and dataset image.

2.6.1 Framework of retrieval and deep learning

The bilinear CNN framework is based on two CNN variant: VGG-m and VGG-16. Both CNN has convolutional layers, pooling layers and fully connected layers [3]. Image size 224 × 224 is initialized as an input. The features are extracted from activation of convolutional feature maps. The fully connected layers are discarded from the fine tune bilinear framework for the sake of simplification. The CNNs are truncated at the last convolutional layer 14th layer of VGG-m and 30th layer from VGG-16. Then three additional layers are added at the end of resulting CNN framework. These three layers are: (1) root compact bilinear pooling to transfer the data in small size. (2) SQRT (3) L2 NORMALIZATION. The resulting network is fine tuned using specific data set. In the pre trained network image features are extracted in the last convolutional C14 and C30 in VGG-m and VGG-16 respectively. The default output is 512 for each of two CNNs. The bilinear model can be represented as $B = (f_A, f_B, P)$. Where f_A and f_B are features set and 'P' is the pooling function. The features f_A, f_B are combined at each location '*l*' using bilinear function.

$$Bilinear(l, I, f_A, f_B) = f_A(l, I)^T . f_B(l, I) \tag{1}$$

A low dimensional projection I applied to the extracted features using the root compared bilinear pooling. The low dimensional descriptor then passed through next layer: SQRT and L2 NORMALIZATION respectively. The final fine tuned structure is used to extract image vector of quantization data sets to calculate rank of retrieved images.

This bilinear CNN-Based architecture uses two parallel CNN as feature extractor whereas most of the other related retrieval methods used single CNN architecture. This bilinear model is used to extract image features at various image locations and scales but the other methods extract image features at last layer of CNN. Therefore the bilinear architecture has more discriminative features due to local image pattern extraction from intermediate convolution layers. This bilinear model has also an efficient bilinear root pooling than other related methods, this pooling approach reduces the image features to compact size. The network is then fine-tuned on the domain specific dataset using end to end training. Hence this model works better in terms of accuracy.

2.7 Deep convolutional learning for content based image retrieval

Feature aggregation approach using spatial convolutional layer activation produced compact feature vectors. E. Mohedano et al. [35] proposed an approach having convolutional CNN features and bag of word aggregation scheme. W. Yu et al. [70] exploited the strength of CNN features at different layers which outperforms the concatenation of multiple layers. The features are extracted from the convolutional layer of deep CNN. The most efficient and compact image descriptors are produced in term of retrieval accuracy and memory requirements. This work [62] presents three different images of retraining the network: fully unsupervised retraining when only dataset is available; the retraining with relevance information if label in the dataset are also available and last is relevance feedback based retraining if user feedback is available. The proposed approach is effective as compare to the hand crafted feature based technique and other CNN based approaches. The training of deep networks is not suitable due to huge set of required parameters and the training procedure is also performed offline. This work [62] proposed smaller and faster framework which enhance the performance with less memory requirement.

2.8 Deep learning two layer codebook features based approach (DL-TLCF)

This paper [32] presents an effective image retrieval method by combining high-level features from convolutional neural network (CNN) model and low-level features from dot-diffused block truncation coding (DDBTC). The low-level features, e.g., texture and color, are constructed by vector quantization -indexed histogram from DDBTC bitmap, maximum, and minimum quantizers. Conversely, high-level features from CNN can effectively capture human perception. With the fusion of the DDBTC and CNN features, the extended deep learning two-layer codebook features is generated using the proposed two layer codebook, dimension reduction, and similarity reweighting to improve the overall retrieval rate. Two metrics, average precision rate and average recall rate (ARR), are employed to examine various data sets. As documented in the experimental results, the proposed schemes can achieve superior performance compared with the state-of-the-art methods with either lower high-level feature in terms of the retrieval rate. Thus, it can be a strong candidate for various image retrieval related applications. In the proposed flowchart, to begin with, OpenMP is adopted to extract both GL-FCF and two-layer codebook feature (TLCF) [32] with different threads. As the threads are completed, both features are combined into several fusion features of various dimensions, including low-dimensional deep learning two-layer codebook features (LD(DL-TLCF)), mid-dimensional DL-TLCF (MD(DL-TLCF)) and high dimensional DL-TLCF (HD(DL-TLCF)). In GL-FCF, the released Convolutional Neural Network framework [32] is adopted to train the GoogLeNet model, and extract the features from the last fully-connection layer. In TLCF, DDBTC features, bitmap, color maximum quantizer and minimum quantizer, are extracted from color images to form VQ-indexed histogram by the proposed two-layer codebook, namely color two-layer combination histogram feature (CTLCHF) and bit two-layer combination histogram feature (BTLCHF).

2.9 VAE-GAN based deep hashing approach

Albeit many deep hashing methods are suggested to learn similarity preserving binary codes but usually they suffer from the problems of meager labeled training data or unreliable

semantic constraints. These limitations can be tackled by VGH network [20] as auxiliary network to learn discriminative image representation and compact binary codes concurrently. This method does not need any manual annotation information like image labels or text DSH-GAN and ACMR approaches. The DSH-GAN [48] uses semi supervised generative adversarial network to produce synthetic hashing training data and ACMR [65] learns discriminative and modality in variant representations for retrieval. This architecture fuses variational auto encoder (VAE) [27] with generative adversarial network (GAN) [16] to produce content preserving image for pair-wise hash learning. A semantic preserving feature mapping model is learned by using a real and a synthesized image in pair-wise form under an adversarial generative process. The unsupervised learning is implemented by the network that takes as input pair-wise images *{I, I′, s}* and refines them by using a deep hashing pipeline. In the pairwise image description, I and I' are the input and synthesized images respectively and s is the class similarity of pair-wise images. If I and I' belong to same class then $s = 1$ otherwise $s = 0$. The binary hash codes are learned by the hashing network using k-dimensional representation of the hash layer. Once the deep architecture is trained, q-bit hash code can be generated for an input image. Then the final query binary code is obtained by simple quantization using the sign function.

2.10 R-MAC based end to end deep learning approach

Most of the time networks are used as local feature extractors in deep retrieval approaches. They concentrate on designing image representation for image retrieval. Some contributions allow a deep architecture to accurately represent input images of distinct sizes and aspect ratios [26, 60] and address the deficit of geometric invariance of CNN based features. These deep architectures are restricted for retrieval applications due to the lack of supervised learning for the particular instance level retrieval task. In [18], the regional maximum activation of computation (R-MAC) descriptor is used for representation. CNN based descriptors are computed for distinct image region at different scales. These descriptors are further aggregated into a compact feature vector of fixed length and are moderately robust to scale and translation. This approach can encode images at high resolution without distorting their aspect ratio. Originally the R-MAC descriptor uses a pre-trained CNN on ImageNet which may not be an optimal solution for retrieval tasks. The pre-trained CNN is further improved by fine tuning on external set of images. The research work in [17, 49] illustrates that fine-tuned models can bring significant improvement. Siamese network fuses three streams (i.e. query, relevant, irrelevant) with triplet loss to optimize weights for generation of well suited representation of retrieval tasks. Originally the R-MAC is used to find the location of region which is pooled to produce the final image level descriptor. For retrieval task, the network learns to choose regions using Region Proposal Network (RPN). Finally this architecture encodes an image into compact fixed length vector in a single forward pass. Then the dot product is used to compare the representation of different images. Hence this approach encodes information at different resolutions that significantly improves retrieval results.

2.11 Deep attention–based hash (DAH) model

In this work, a Deep Attention–based Hash (DAH) [29] retrieval model was proposed. It is a combination of an attention module and a convolution neural network that is used to obtain strongly representable hash codes. This method significantly improves the image retrieval

precision by effectively extracting the semantic features of the data. The paired batches of images are organized as input that contains classified one-hot tags and similarity information. These strong hash codes achieved the improved level of classification due to interconnection and optimization of semantic similarity representation and classification of recognition ability. The original sigmoid activation function is improved and logarithmic hash loss function is designed. These functions cause a smooth gradient change between the hash layer and the classification layer and also the trained model is allowed to produce similarity features. In the forward structure of DAH network, the core structure of 30 layers of the ResNet50 [21] convolution structure is replaced by the depth-wise separable convolution kernel, from Xception [8]. The back structure finds the direction of change of model. It calculates partial derivatives of the loss function in the opposite direction of the forward structure. The initial experiments in this study prove the positive effect of the design on the model parameter training.

2.12 Metric learning-based deep hashing network (MiLaN)

Recent deep neural network (DNNs) have shown success in optimize feature learning for the retrieval tasks in end to end fashion. But in case of training of DNNs with small labeled dataset can cause problem of over fitting. The metric learning-based deep hashing network (MiLaN) [52] used a pre-trained DNN as a intermediate feature representation step to reduce the problem of over fitting on small remote sensing labeled datasets without the retraining a fine tuning requirements. The big CNN (Inception Net [58]) pre-trained on Imagenet is used to extract in intermediate features of remote sensing images. Then, using these intermediate features a smaller network is trained in order to learn final hash function. Each image is fed to the Inception Net [58], where the last layer (pool 3) is applied just before the softmax layer. This last layer consist of 2049 neurons, whose activation shown the result of average pooling on the complete feature maps of the layer. It optimize the features by learning a semantic based metric space then compact binary hash codes are computed for fast retrieval task. The hamming distance is calculated between the query and each database image.

The overview of various state of the art content based image retrieval techniques using deep learning is summarized in Table 1. In this study, the mean average precision (mAP) is the main performance evaluation parameter. The bilinear root compact pooling [3] has higher mAP for Oxford 5 k database (0.886) and holiday dataset (0.951) than other approaches. The spatial pooling of deep convolutional feature approach [1] is used for the person retrieval applications. From large database, the supervised deep hashing method [59] performed better than other methods used for information retrieval for face images database in youtube (0.7786) and face-scrap (0278) database. The superior values of the mAP for large databases like MIRFlickr (0.6704) and NUS-Wide (0.5281) are obtained by using semantic assisted visual hashing [73] for retrieval applications. In terms of top N-score, the R-MAC based end to end deep learning based approach [18] has best score for UKBench (3.84) database.

3 Dataset

The performance of deep binary codes and the dynamic late fusion scheme are evaluated on four datasets: INRIA Holidays, Oxford, UKBench and MIRFLICKR 1 M [24]. The INRIA Holidays dataset [25] has been created for the ANR RAFFUT project i.e. the joint project of

Table 1 State of the Art Content Based Image Retrieval Techniques using Deep Learning

Author	Algorithm	Deep Network Used	Database	Similarity Measure	Evaluation Parameters	Description
Song Wu et al. [67]	Deep Binary Codes	VGGNet	INRIA Holidays, Oxford 5 k, UKBench, MIRFLICKR	Hamming distance	mean Average Precision (mAP), **0.7719** (INRIA Holidays+ FLICKr) **0.5204** (Oxford 5 k +FLICKr), **0.9156** (UKBench + MIRFLICKR)	Significantly reduces the memory requirements and computational cost compared to other state of the art methods and the dynamic late fusion scheme provides consistent improvement in accuracy.
Jamil Ahmad et al. [1]	Spatial Pooling of Deep Convolutional Features	AlexNet	Weizmann, ETHZ, CUHK, CAVIAR	Euclidean, Manhattan, Hellinger & Canberra distances, Histogram Intersection	mean Average Precision (mAP) **0.8150** (Weizmann) **0.6790** (ETHZ) **0.6810** (CUHK) **0.4760** (CAVIAR)	Effective mechanism to deal with surveillance data but a more improved feature extraction process can be used to extract more sophisticated features.
Lei Zhu et al. [73]	Semantic Assisted Visual Hashing	Proposed SAVH	WiKi, MIR Flickr, NUS-WIDE	Hamming space	mean Average Precision (mAP) **0.1991** (WiKi) **0.6704** (MIRFlickr) **0.5281** (NUS-WIDE)	Superior performance than other visual hashing methods. Its effectiveness can be further validated by involving more associated modalities.
Wei Yu et al. [71]	Multi-layer CNN	AlexNet, VGG-16	Oxford 5 K, Holiday, UKB	Cosine distance	mean Average Precision (mAP) **0.615** (Oxford 5 K) **0.912** (Holiday) top –N Score **3.69** (UKB)	Outperforms the methods using handcrafted features and single layer CNN but can be further improved using complementary information
		Proposed DHCQ		Hamming Ranking		

Table 1 (continued)

Author	Algorithm	Deep Network Used	Database	Similarity Measure	Evaluation Parameters	Description
Jinhui Tang et al. [59]	Supervised Deep Hashing		Youtube, FaceScrub & Cifar-10		mean Average Precision (mAP) **0.7786** (Youtube) **0.2728** (FaceScrub) **0.7094** (Cifar-10)	Retrieval Performance is enhanced using optimized objective function defined over classification and quantization error.
Ahmad Alzu'bi et al. [3]	Bilinear Root Compact Pooling of Deep Convolutional Features	VGG-16	Holidays, Oxford, UKBench	Euclidean distance, Manhattan distance	mean Average Precision (mAP) **0.886** (Oxford 5 K) **0.951** (Holiday) top –N Score **3.56** (UKB)	High retrieval performance in terms of extraction & search time and storage cost.
Maria Tzelepi et al. [62]	Deep convolutional learning using model retraining	CaffeNet	Paris 6 k, UKBench,	Euclidean distance, Cosine distance	mean Average Precision (mAP), **0.9859** (Paris 6 k) top-N score **3.9710** (UKB)	Outperforms others CNN-based retrieval techniques as well as conventional handcrafted feature based approaches.
Peizhong et al. [32]	Fusion of Deep features and compressed features	GoogLeNet	Corel 1000, Brodatz, Holidays, UKBench	Proposed similarity assessment algorithm	Average Precision, **0.7509** (Corel 1000) **0.8205** (Holidays) Average Recall **0.9225** (Brodatz) NS- Score **3.636** (UKBench)	Achieve the best performance compared to the other state of the art methods on both natural and textural databases and is a very robust and effective method for retrieval applications.
Jin Guoqiang et al. [20]	VAE-GAN based deep hashing approach	Variational Auto-encoder (VAE), Generative Adversarial Network (GAN)	CIFAR-10, Flickr, NUS-WIDE	Hamming Ranking	mean Average Precision (mAP), **0.129** (CIFAR-10), **0.529** (Flickr),	Attains state of the art retrieval performance by yielding accurate deep hash codes.

Table 1 (continued)

Author	Algorithm	Deep Network Used	Database	Similarity Measure	Evaluation Parameters	Description
Albert Gordo et al. [18]	R-MAC based end to end deep learning based approach	VGG-16, ResNet101	Oxford 5 k, Paris 6 k, Holidays, UKBench	Dot Product	**0.475** (NUS-WIDE) mean Average Precision (mAP), **0.861** (Oxford 5 k), **0.946** (Paris 6 k), **0.948** (Holidays) Recall@4 Score **3.84** (UKBench)	Faster and more memory efficient. It outperforms other state of the art approaches.
Xinlu Li et al. [29]	Deep Attention–based Hash (DAH) Model	ResNet50	CIFAR-10,	Hamming distance	mean Average Precision (mAP), **0.9267** (CIFAR-10),	The inclusion of the attention model in DAH improves the performance of this method in term of precision.
Subhankar Roy et al. [52]	Metric Learning-Based Deep Hashing Network (Mi-LaN)	Inception Net	UCMD, AID	Hamming distance	mean Average Precision (mAP), **0.991** (UCMD), **0.926** (AID),	The MiLaN method is more fast and accurate than other CBIR used for remote sensing data retrieval applications.

INRIA and Advestigo Company. This has 1491 personal holidays photos divided in 500 categories. This dataset is mainly used for evaluation of image search approaches. It is collection of personal holiday images of natural, man-made and water scenes etc. The Oxford building dataset [46] contains 5062 of 11 different Oxford landmarks that are collected from Flickr. This dataset has 55 queries in all having 5 queries for each landmark category. Henrik Stewenius and David Nister created the UK bench dataset [40] of 10,200 images of 2550 groups having 4 images each of size 640 × 480. The dataset blurred and rotated images for different evaluation purposes. This dataset is typically used of image retrieval applications using one query image from each group. MIRFLICKR 1 M dataset [24] has 1 million Flickr images introduced as an ACM Sponsored image retrieval evaluation. These datasets are popularly used for benchmarking. The different types of datasets are used to evaluate the performance of the spatial pooled convolutional feature based approach. The dataset 3DPeS (3D People Surveillance dataset) [5] mainly created for people re-identification multicar system having many videos capture from real surveillance set up of 8 cameras. This is capture by monitoring an area of the University of Modena and Reggio Emilia campus. The images are uncompressed with resolution of 704 × 576 pixels. VIPER (Visual Person Detection made Reliable) [19], data was acquired at Brugge railway station, Belgium with cooperation of the thermal imaging company FLIR. There are 25,000 images that are extracted as frames rom 28 separated videos. Multi-view car dataset [42] have images of cars and rotated by 360 degree. This dataset is captured by Nikon D70 camera a tripod at the Geneva international motor show'08. These comprehensive car dataset [69] are used for kernel selection. The vehicle image dataset [4] have 4000 vehicle image under real road environment. The Weizmam [14] dataset contains 1464 image of 22 different vehicles from different viewpoints. The ETHZ public surveillance image dataset [53] has 8500 object patches from 149 people included the background clutter and degree of occlusion. The most challenging dataset CAVIAR [7] and CUHK [1] person dataset are also utilized for the evaluation of proposed work. The CAVIAR dataset contains 10 images captured from different view of 72 different individuals in shopping mall using surveillance cameras. The CUHK dataset consists of 3884 images of 971 different people captured with surveillance camera from different views. The performance of the SAVH is validated by conducting experiments on the publically available databases: Wiki [51], MIR flicker [23] and NUS-WIDE [9]. All the datasets include image and text. The Wiki dataset [51] has 2866 Images of 10 different semantic categories whereas NUS-WIDE dataset [9] has 269,648 images categorized in 81 different concepts. It was developed by the lab for media search in National University of Singapore. The three well known dataset Oxford 5 K [46], Holiday [25] and UKB [41] are used to evaluate the performance of the multilayer CNN feature based approach. The DHCQ method [59] is evaluated by conducting experiments on dataset such as: Youtube face dataset [66], face scrub face datasets [38] and cifar-10 [61]. The Youtube dataset contains 1595 face images of people captured from 3425 videos. The face scrub dataset comprised of 107,818 face images of 530 different celebrities with about 200 images per person whereas the Cifar-10 contains 60,000 images of size32 ×32 each that are classified in 10 different categories. Cifar-10 was created by Alex Krizhevsky, Vinod Jain and Geoffrey Hinton. The standard datasets like the Holiday dataset [] and Oxford building dataset and UK-bench dataset are used to evaluate the performance of the proposed work in comparison to other state of the art approaches. The performance of deep learning approach [15] for CBIR has been tested using dataset Paris 6 K [47] and UK-bench [42]. The Paris 6 K dataset contains 6392 images collected from Flickr by search of Paris landmarks. University of California Merced [52] (UCMD) dataset contains 2100 aerial images from 21 different land-

cover categories, where each category includes 100 images. Another benchmark dataset is the aerial image data set [52] (AID) contains 10,000 images from 30 different categories, and the number of images in each category varies between 220 and 420. The accomplishments of DL-TLCF approach [32] has been accessed using various types of datasets like COREL 1000 with 1000 images divided in 10 categories of 100 images each, the Holiday, the UK-bench and the Brodatz-1856 dataset which has 1856 images divided in 116 categories of 16 images each etc. The comparison of various datasets used for the evaluation of CBIR techniques that use deep learning based on the dataset size and of the number of categories is shown in Figs. 2 and 3 respectively.

4 Similarity measures

In content based image retrieval approaches, the similarity measures play a very crucial role in retrieving relevant images from databases. In CBIR, the image contents (features) like color, shape or texture extracted from images are compared with the contents of the query image with the help of similarity measures. The accuracy of the retrieval results highly depends on the performance of the retrieval results similarity measures. The Manhattan distance [1, 3] is one of the simple and robust measures used to find similarities in image retrieval. The Manhattan distance is given by Eq. (2).

$$d_{MH}(a,b) = \sum_{j=1}^{m} a_j - b_j \tag{2}$$

Where a & b are the two set of image feature vectors and j denotes the j^{th} feature vector. The Euclidean distance [1, 3, 62] is another very popular similarity metric used in many applications particularly best suited for the measurement of image similarity in terms of best match. Equation (3) shows the formula for Euclidean distance as:

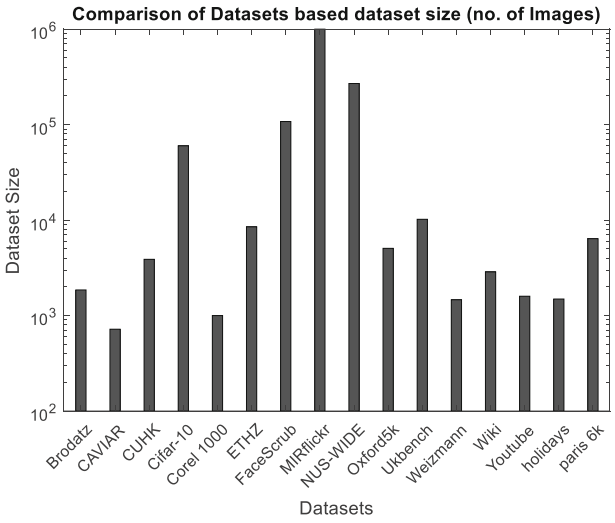

Fig. 2 Comparison of the dataset size of datasets that use Deep Learnig for CBIR

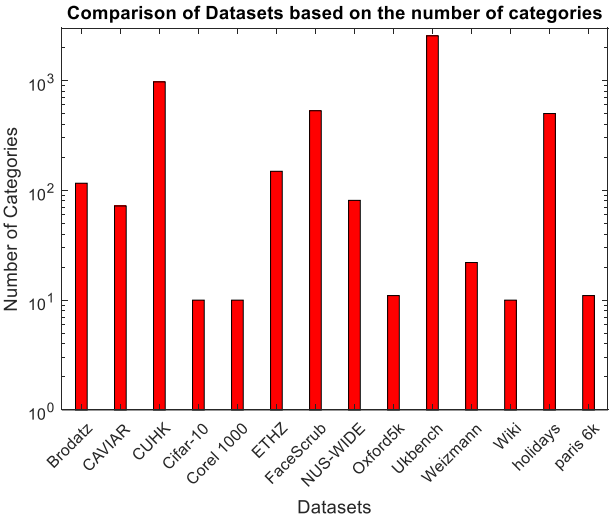

Fig. 3 Comparison based on the number of categories for the datasets that use Deep Learning for CBIR

$$d_E(a,b) = \sqrt{\sum_{j=1}^{m} \left(a_j - b_j\right)^2} \tag{3}$$

The histogram intersection distance [1] is a similarity metric proposed for the color image retrieval applications. This similarity metric is defined for color histograms as given in Eq. (4) where |a| and |b| are the respective histogram magnitudes.

$$d_{HI}(a,b) = \frac{\sum_x \sum_y \sum_z \min(a(x,y,z), b(x,y,z))}{\min(|a|,|b|)} \tag{4}$$

Where a & b are the RGB images, (x, y) are the pixel coordinates and z represents the colour channel of RGB image. Initially, however the Canberra distance [1] was not widely used but due to its significant performance in retrieval application it has become popular in CBIR applications in these days. The formula of the Canberra distance is shown in Eq. (5).

$$d_{CAN}(a,b) = \sum_{j=1}^{m} \frac{|a_j - b_j|}{|a_j| + |b_j|} \tag{5}$$

The cosine distance d_c [62, 71] is often used as similarity metric for image retrieval applications. It is calculated by using Eq. (6):

$$d_C(a,b) = 1 - CS(a,b) \tag{6}$$

CS is the cosine similarity and calculated by Eq. (7).

$$CS(a,b) = \frac{a.b}{\|a\|\|b\|} = \frac{\sum_{j=1}^{m} a_j b_j}{\sqrt{\sum_{j=1}^{m} a_j^2}\sqrt{\sum_{j=1}^{m} b_j^2}} \tag{7}$$

with a_j and b_j are the components of vector a and b respectively. Liu et al. [32] proposed a similarity assessment and a normalization algorithm for finding similarity between images with improved overall performance. The hamming distance is used to evaluate the similarity between two images if features of the images are represented as deep binary codes. It evaluates the number of bits difference between two binary code vectors: the lower the number of the dissimilar bits, the higher the similarity. This metric is used when features are in the form of binary code vectors.

The Hellinger distance [1] is also used as similarity measure in image retrieval applications where the image features are represented in the form of probability distribution function. Given two probability distribution functions A & B, the Hellinger distance between them is shown in Eq. (8).

$$d_H(A,B) = \frac{1}{\sqrt{2}} \left\| \sqrt{A} - \sqrt{B} \right\|_2 \tag{8}$$

5 Evaluation parameters

The performance of a content based image retrieval (CBIR) system can be evaluated by various evaluation parameters (performance metrics) such as precision, recall, average precision rate (APR), average recall rate (ARR), mean average precision (mAP) and top-N score. The precision [62] is defined as the ratio of relevant retrieved images to the total number of retrieved images.

$$P = \frac{A_R}{B_T} \tag{9}$$

where P is the precision of B_T retrieved images, A_R is the number of retrieved images and B_T is the total number of retrieved images for a query and the average precision [32] is the average of the precision scores of all the relevant images of a query image.

$$P_{avg} = \frac{\sum_{l=1}^{L} P(l) \times R(l)}{B_T} \tag{10}$$

where P_{avg} is the average precision of G retrieved images. P(g) is the precision of top g retrieved images and R(g) indicates the relevance of the image of rank g to the query image. The value of R(g) =1 for relevant image at rank g and 0 for irrelevant image. The mean average precision (mAP) [1, 3, 59, 67, 71, 73] is the mean of all the average precisions of different query images applied to the system.

$$mAP = \sum_{j=1}^{J} \frac{P_{avg}(j)}{J} \tag{11}$$

where mAP is the mean average precision score of J different queries and $P_{avg}(j)$ represents the average power of j^{th} query. The recall [32, 62] is the ratio of total number of relevant images retrieved and the total number of relevant images present in the database. The recall is given by the Eq. (12).

$$R = \frac{A_R}{C_T} \tag{12}$$

where A_R is the number of relevant retrieved images and C_T is the total number of relevant images in the database. The parameter, top-N score [62] refers to average number of relevant images, within the top-N ranked images.

6 Discussion

In recent past, in order to exploit potential advantage of different deep learning approaches, content based image retrieval system has been explored. The deep binary codes have many advantages over deep convolutional features as they do not need any additional training. The experiments carried out using standard dataset demonstrate that the deep binary codes are competitive to other state of art approaches. The dynamic late fusion approach presents considerable enhancement in accuracy. The prioritized and segmented surveillance image dataset is very helpful to extract compact and discriminative features. The features are utilized by selected kernel with SMA to achieve high retrieval performance. This technique is very effective when dealing with large collection of surveillance data. The SFVH and MFVH schemes utilize only visual features and do not take advantage of valuable text segments for performance enhancement. The UCMH technique can treat both visual and text equally but fails to achieve good performance. The SAVH approach effectively integrates the semantics of text by offline learning using image as input. The SAVH shows performance superiority over other techniques. The objective function based on classification and quantization error is optimized. The obtained learned hashing codes boost the retrieval performance. The bilinear CNN model shows higher efficiency in learning more complex image features. This model extracts more discriminative features as it uses local image pattern taken from intermediate convolutional layers. This model reduces the features map to compact size which enhances the performance in terms of retrieval time and storage. This shows high retrieval performance for large scale image retrieval. The VAE-GAN based deep hashing approach attains state-of-the-art retrieval performance by generating deep hash codes. The R-MAC based approach with its global descriptors performs well and the performance is improved by combining some complex methods like query expansion (QE) and database-side feature augmentation (DBA). These methods made the system faster and memory efficient but use costly matching and verification. The DAH learns supervised information end to end by developing a loss function for learning the correlation information between images. MiLaN avoids the over fitting problem in small – size labeled remote sensing dataset. This is more accurate and time efficient than the state of the art methods.

The different dataset for deep learning based CBIR system are summarized in Figs. 2 and 3 with different parameters. The datasets are generated under distinct conditions. In general, the research of deep learning based CBIR system is performed using standard dataset so that results can be reused for further study and meaningful comparison of performance of various CBIR methods can be performed. In this review, the various deep learning based methods are tested for different standard dataset or different domain like human face and surveillance datasets.

In this survey, the distinct similarity measures like Euclidean, Hamming, Manhattan, Canberra, Hellinger distance and histogram intersection for CBIR are discussed. Evaluation

parameters like mean average precision (mAP), top-N score, average precision and average recall are mostly considered for performance evaluation of deep learning based CBIR systems.

7 Conclusion and future direction

In present day content based image retrieval approaches, deep learning based CBIR methods have reached miraculous attainment due to its potent learning competence. The numerous latest advancements of deep neural network based CBIR techniques have been collected in this paper. The deep binary codes with late fusion have shown improved accuracy for large scale image retrieval. The convolutional neural network pre-trained on ImageNet database has shown better performance and accuracy in a surveillance dataset. Due to hashing methods, an unsupervised visual hashing approach, called semantic-assisted visual hashing, has been used in CBIR application. This method attains better performance than other state of the art approaches. The use of multilayer CNN features for the CBIR performs better than single layer features and also presents a competitive performance on various popular CBIR benchmarks. The hashing based CBIR approach, with a deep convolution network, is used to produce hash codes and predicts labels of images. The CBIR method based on the bilinear root pooling, utilizes the compact and highly discriminative image feature set. The deep CNN based CBIR system has efficient retrieval performance and memory efficiency due to compact image feature descriptor. The CBIR model based on fusion of deep learning and compressed feature also achieve superior performance in terms of retrieval rate for both low and high features. The inclusion of the attention model in DAH improves the performance of this method in term of precision. The different model structures can be considered and their effect on retrieval performance can be analyzed as a future work. The MiLaN method is more fast and accurate than other CBIR used for remote sensing data retrieval applications. The survey is concluded with the following remarks:

1. Deep learning approaches can handle CBIR problems more efficiently than other state of the art method.
2. The deep learning based CBIR methods attains advancement in real application.
3. The dataset with more associated modality, need further research for validation in CBIR.
4. Features extracted using deep learning methods are more compact and more discriminative for different retrieval scenarios.
5. The resolution between retrieval accuracy and storage space can be further exploited and improved.

Finally, in this survey, some of the key deep learning based CBIR approaches have been investigated. The most recent developments have been compared and displayed in tabular form. This review will clarify future directions of research in the ever growing domain of deep learning based CBIR systems.

References

1. Ahmad J, Mehmood I, Baik SW (2017) Efficient Object-Based Surveillance Image Search using Spatial Pooling of Convolutional Features. J Vis Commun Image R 45:62–76

2. Ahmad J, Sajjad M, Rho S, Baik SW (2016) Multi-scale local structure patterns histogram for describing visual contents in social image retrieval systems. Multimed Tools Appl 75(20):12669–12692
3. Alzu'bi A, Amira A, Ramzan N (2017) Content based image retrieval with compact deep convolutional features. Neurocomputing 249:95–105
4. Arrospide J, Salgado L, Nieto M (2012) Video Analysis Based Vehicle Detection and Tracking using an MCMC sampling Framework, EURASIP J Adv Signal Process, pp. 1–20
5. Baltieri D, Vezzani R, Cucchiara R (2011) 3dpes: 3d People Dataset for Surveillance and Forensics, In Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, ACM
6. Bhagyalaksluni A, Vijayachamundeeswari V (2014) A survey on content based image retrieval using various operators IEEE international conference on computer communication and systems (ICCCS '14)
7. Cheng DS et al. (2011) Custom pictorial structures for re-identification, In BMVC
8. Chollet F (2017) Xception: Deep Learning with Depth-wise Separable Convolutions, In IEEE Computer Vision and Pattern Recognition, Hawaii, USApp. 1251–1258
9. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nuswide: A Real-world Web Image Database from National University of Singapore In Proc ACM Int Conf Image Video Retrieval, pp. 48:1–48:9
10. Dharani T, Aroquiaraj IL (2013) A survey on content based image retrieval. In: Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, pp 485–490
11. Felci Rajam I, Valli S (2013) A survey on content based image retrieval. Life Sci J 10(2):2475–2487
12. Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact Bilinear Pooling In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 317–326
13. Gao L, Song J, Zou F, Zhang D, Shao J (2015) Scalable multimedia retrieval by deep learning hashing with relative similarity learning. In: Proc ACM Int Conf Multimedia, pp 903–906
14. Glasner D, Galun M, Alpert S, Basri R, Shakhnarovich G (2012) Viewpoint aware object detection and continuous pose estimation. Image Vis Comput 30(12):923–933
15. Goel R, Sharma A, Kapoor R (2019) Object recognition using deep learning. in J Comput Theor Nanosci 16(9):4044–4052
16. Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2014) Generative adversarial nets, Proc of International Conference on Neural Information Processing Systems, pp.2672–2680
17. Gordo A, Almazán J, Revaud J, Larlus D (2016) Deep image retrieval: learning global representations for image search, In Proceedings European conference on computer vision (ECCV)
18. Gordo A, Almazan J, Revaud J, Larlus D (2017) End-to-end Learning of Deep Visual Representations for Image Retrieval. Int J Comput Vis 124(2):237–254
19. Gray D, Brennan S, Tao H (2007) Evaluating Appearance Models for Recognition, Reacquisition, and Tracking, In Proc IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Citeseer
20. Guoqing J, Yongdong Z, Ke L (2019) Deep hashing based on VAE-GAN for efficient similarity retrieval. Chi J Electron 28(6):1191–1197
21. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. IEEE Comput Vis Pattern Recognit, Las Vegas, NV, USA, pp 770–778
22. Heesch D (2008) A survey of browsing models for content based image retrieval. J Multimed Tools Appl (Springer) 40:261–284
23. Huiskes MJ, Lew MS (2008) The Mir Flickr retrieval evaluation. In: Proc ACM Int. Multimedia Inf Retrieval, pp 39–43
24. Huiskes MJ, Thomee B, Lew MS (2010) New trends and ideas in visual concept detection: the MssIRFLICKR retrieval evaluation initiative. In: Proceedings of the International Conference on Multimedia Information Retrieval, pp 527–536
25. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search, In Computer Vision–ECCV, pp. 304–317
26. Kalantidis Y, Mellina C, Osindero S (2016) Cross-dimensional weighting for aggregated deep convolutional features. In: Workshop on Web Scale Vision and Social Media (VSM), European Conference on Computer Vision (ECCV), pp 685–701
27. Kingma D, Welling M (2013) Auto-encoding Variational Bayes, arXiv: 1312.6114
28. Kokare M, Chatterji BN, Biswas PK (2002) A survey on current content based image retrieval methods. IETE J Res 48(3–4):261–271
29. Li X, Xu M, Xu J, Weise T, Zou L, Sun F, Wu Z (2020) Image retrieval using a deep attention-based hash. IEEE Access 8:142229–142242
30. Lin TY, Roy Chowdhury A, Maji S (2015) Bilinear CNN Models for Fine Grained Visual Recognition, In Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1457
31. Liong VE, Lu J, Wang G, Moulin P, Zhou J (2015) Deep Hashing for Compact Binary Codes Learning, In IEEE Comput Vis Pattern Recognit, pp. 2475–2483

32. Liu P, Guo JM, Wu CY, Cai D (2017) Fusion of Deep Learning and Compressed Domain Features for Content Based Image Retrieval, IEEE transactions on image processing, vol. 26, no. 12
33. Liu L, Yu M, Shao L (2015) Multi-view alignment hashing for efficient image search. IEEE Trans Image Process 24(3):956–966
34. Liu Y, Zhang D, Lu G, Ma W-Y (2007) A Survey of Content Based Image Retrieval with High-Level Semantics. J Pattern Recognit (Elsevier) 40:262–282
35. Mohedano E, McGuinness K, O'Connor NE, Salvador A, Marques F, Giroi Nieto X (2016) Bags of Local Convolutional Features for Scalable Instance Search, In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ACM, pp. 327–331
36. Mukherjee J, Mukhopadhyay J, Mitra P (2014) A survey on image retrieval performance of different bag of visual words indexing techniques. In: Proceedings of the IEEE Students' Technology Symposium, pp 99–104
37. Muller H, Michoux N, Bandon D, Geissbuhler A (2004) A Review of Content Based Image Retrieval Systems in Medical Applications—Clinical Benefits and Future Directions. Int J Med Inform 73:1–23
38. Ng HW, Winkler S (2014) A Data-driven Approach to Cleaning Large Face Datasets, In IEEE International Conference on Image Processing pp. 343–347
39. Ng J, Yang F, Davis L (2015) Exploiting Local Features from Deep Networks for Image Retrieval, In Proceedings of the IEEE Conf Comput Vis Pattern Recognit Workshops, pp. 53–61
40. Nister D, Stewenius H (2006) Scalable Recognition with a Vocabulary Tree. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2:2161–2168
41. Oussalah M (2008) Content based image retrieval: review of state of art and future directions. In: First Workshops on Image Processing Theory, Tools and Applications (IEEE), pp 1–10
42. Ozuysal M, Lepetit V, Fua P (2009) Pose Estimation for Category Specific Multi-view Object Localization, In IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, IEEE
43. Patel T, Gandhi S (2017) A survey on content based similarity techniques for image retrieval. In: International Conference on Innovative Mechanisms for Industry Applications (IEEE), pp 219–223
44. Patel F, Kasat D (2017) Hashing based indexing techniques for content based image retrieval: a survey. In: International Conference on Innovative Mechanisms for Industry Applications (IEEE), pp 279–283
45. Paulin M, Douze M, Harchaoui Z, Mairal J, Perronin F, Schmid C (2015) Local Convolutional Features with Unsupervised Training for Image Retrieval, In Proceedings of the IEEE International Conference on Computer Vision, pp. 91–99
46. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: Computer Vision and Pattern Recognition (CVPR), pp. 1–8
47. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2008) Lost in quantization: improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8
48. Qiu Z, Pan Y, Yao T, et al. (2017) Deep Semantic Hashing with Generative Adversarial Networks, Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan, pp. 225–234
49. Radenovic F, Tolias G, Chum O (2016) CNN image retrieval learns from BoW: unsupervised fine-tuning with hard examples, In Proceedings European conference on computer vision (ECCV)
50. Rafiee G, Dlay SS, Woo WL (2010) A review of content-based image retrieval. In: 7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010), pp 775–779
51. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A New Approach to Cross-modal Multimedia Retrieval, In Proc ACMInt Conf Multimedia, pp. 251–260
52. Roy S, Sangineto E, Demir B, Sebe N (2020) Metric learning based deep hashing network for content-based retrieval of remote sensing images. IEEE Geosci Remote Sens Lett 18(2):226–230
53. Schwartz WR, Davis LS (2009) Learning discriminative appearance based models using partial least squares. In: 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing. IEEE, pp 322–329
54. Shen F, Shen C, Liu W, Shen HT (2015) Supervised discrete hashing, In IEEE Computer Vision and Pattern Recognition, pp. 37–45
55. Shen F, Shen C, Shi Q, van den Hengel A, Tang Z, Shen HT (2015) Hashing on nonlinear manifolds. IEEE Trans Image Process 24(6):1839–1851
56. Shen X, Shen F, Sun QS, Yuan YH (2015) Multi-view latent hashing for efficient multimedia search. In: Proc ACM Int Conf Multimedia, pp 831–834
57. Szegedy C, Liu W, Jia Y, Sermanet P (2015) Going deeper with convolutions. In: IEEE Computer Vision and Pattern Recognition (CVPR), pp 1–9

58. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision, In Proc IEEE Conference Computer Vision Pattern Recognition (CVPR), pp. 2818–2826
59. Tang J, Li Z, Zhu X (2018) Supervised deep hashing for scalable face image retrieval. Pattern Recogn 75: 25–32
60. Tolias G, Sicre R, Jegou H (2016) Particular object retrieval with integral max pooling of CNN activations. In: International Conference on Learning Representations (ICLR), pp 1–12
61. Torralba A, Fergus R, Freeman W (2008) 80 million tiny images: a large dataset for nonparametric object and scene recognition. IEEE Trans Pattern Anal Mach Intel 30(11):1958–1970
62. Tzelepi M, Tefas A (2018) Deep convolutional learning for content based image retrieval. Neurocomputing 275:2467–2478
63. Veltkamp RC, Tanase M (2002) Content-based image retrieval systems: A survey. Multimed Sys Appl Ser (Springer) 21:47–101
64. Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content based image retrieval: a comprehensive study. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp 157–166
65. Wang B, Yang Y, Xing X, Alan H, Heng TS (2017) Adversarial cross-modal retrieval. In: Proc of ACM on Multimedia Conference, pp 154–162
66. Wolf L, Hassner T, Maoz I (2011) Face Recognition in Unconstrained Videos with Matched Background Similarity, In IEEE Computer Vision and Pattern Recognition, pp. 529–534
67. Wu S, Oerleman A, Bakker EM, Lew MS (2017) Deep binary codes for large scale image retrieval. Neurocomputing 257:5–15
68. Yang H, Lin K, Chen C (2017) Supervised learning of semantics-preserving hash via deep convolutional neural networks. IEEE Trans Pattern Anal Mach Intell 40(2):437–451
69. Yang L et al. (2015) A large-scale Car dataset for fine grained categorization and verification, In Proceedings of the IEEE conference on computer vision and pattern recognition
70. Yu W, Yang K, Yao H, Sun X, Xu P (2017) Exploiting the complementary strengths of multilayer CNN features for image retrieval. Neurocomputing 237:235–241
71. Yua W, Yangb K, Yaoa H, Suna X, Xub P (2017) Exploiting the complementary strengths of multilayer CNN features for image retrieval. Neurocomputing 237:235–241
72. Zhao M, Zhang H, Sun J (2016) A novel image retrieval method based on multi trend structure descriptor. J Vis Commun Image Represent 38:73–81
73. Zhu L, Shen J, Xie L, Cheng Z (2017) Unsupervised visual hashing with semantic assistant for content based image retrieval. IEEE Trans Knowl Data Eng 29(2):472–486