TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 19, No. 3, June 2021, pp. 858~871
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i3.18157

A comparative analysis of automatic deep neural networks for image retrieval

Hanan A. Al-Jubouri, Sawsan M. Mahmmod
Computer Engineering Dept. College of Engineering, Mustansiriyah University, Baghdad, Iraq

Article Info

Article history:

Received Jul 10, 2020
Revised Sep 9, 2020
Accepted Sep 20, 2020

Keywords:

Content-based image retrieval
Convolutional neural networks
Image classification
Image retrieval deep learning

ABSTRACT

Feature descriptor and similarity measures are the two core components in content-based image retrieval and crucial issues due to "semantic gap" between human conceptual meaning and a machine low-level feature. Recently, deep learning techniques have shown a great interest in image recognition especially in extracting features information about the images. In this paper, we investigated, compared, and evaluated different deep convolutional neural networks and their applications for image classification and automatic image retrieval. The approaches are: simple convolutional neural network, AlexNet, GoogleNet, ResNet-50, Vgg-16, and Vgg-19. We compared the performance of the different approaches to prior works in this domain by using known accuracy metrics and analyzed the differences between the approaches. The performances of these approaches are investigated using public image datasets corel 1K, corel 10K, and Caltech 256. Hence, we deduced that GoogleNet approach yields the best overall results. In addition, we investigated and compared different similarity measures. Based on exhausted mentioned investigations, we developed a novel algorithm for image retrieval.

Corresponding Author:

Hanan A. Al-Jubouri
Department of Computer Engineering
Mustansiriyah University
Palestine Street, Baghdad, Iraq
Email: hananaljubouri@uomustansiriyah.edu.iq

1. INTRODUCTION

Today, digital photographic devices are widely used resulting large volumes of digital images have being acquired and stored in databases in different fields such as scientific research, medical, forensic analysis, and social networking. So, the retrieval of these images should be done effectively and fast. Information retrieval (IR) attempts to find material such as images or texts (documents) which have unstructured form to get information from large volume of these materials [1, 2]. In early image retrieval systems, images are indexed in a database using textual annotation such as keywords or phrases. A user asks the system to find similar images by entering the textual annotation and the system retrieves images in order according to the degree of match to the annotation. However, some limitations face such a method. For instance, it is time consuming to annotate images in a large-scale database manually and the text may not available during image capturing respectively. Consequently, content-based image retrieval (CBIR) is a process that extract image feature (visual content) to represent images automatically and index them in a database [3].

Figure 1 illustrates a typical diagram of CBIR system that stores images in the database by extracting image features at off-line phase [4]. Meanwhile, the system extracts a feature vector from a query image in the

same way and compares it with the image features in the database using a similarity measure. The most similar images are ordered in ranked list and returned to the user at on-line phase. Hence, some irrelevant images are retrieved in the ranked list due to a challenge so-called "semantic gap" which is the gap between high and low level features in meaning [4]. Therefore, the aim of researchers in CBIR is how to develop a system or algorithm that can bridge the semantic gap between human conceptual meaning for images and machines such as a computer. In other words, how the CBIR system can extract effective features that represent the image in the database and retrieve in terms of relevant images.

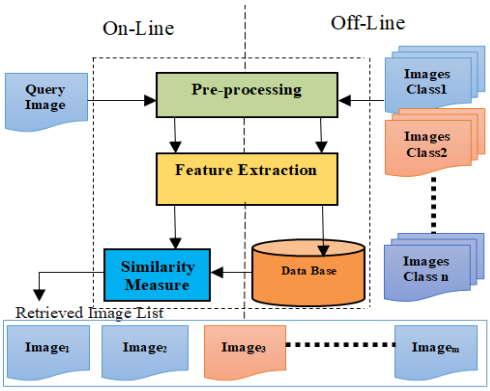

Figure 1. Typical diagram of CBIR system

The main contributions of this paper are as follows: first, convolution neural networks (CNNs) are investigated to classify huge amount of images. In our investigation, different deep learning approaches are used in classification such images. Second, the CNNs approaches are exploited to learn features of images for image retrieval. Third, different distance functions are tested for similarity measures. The aim is to judge which deep learning approach can produce effective features and which distance function is more accurate to reduce the semantic gap issue in CBIR. Consequently, a novel algorithm for image retrieval is developed. The remainder of this paper is structured as follows. The relevant literatures are presented in section 2. The proposed CNNs used in this paper are described in section 3 while section 4 describes the images datasets used in the investigation and presents the experimental results analysis of our evaluation system. Finally, section 5 draws the finding of our paper and gives a recommendation for further works.

2. RELETATED WORK

Numerous studies of literature have investigated CNNs in image retrieval. In this section, we will present some of the literatures using CNNs in these studies. For example, in [5] three CNN features for IR are proposed by fusing the product rule and the weighted average of features similarity. The authors extract the features of images using three kinds of CNNs. After that, by using product rule, the weighted feature similarities between the query and database image are calculated. Finally, the retrieval result is found by returning the images with the highest top-n scores. Also, in [6], the features of the images are extracted by analyzing the classical CNN and then the results are compared with three classical algorithms. The performance of the retrieval system is improved by combining a cosine similarity measurement approach.

A deep CNN model is utilized in [7] to extract the feature representation from the activations of the convolutional layers in a large image dataset for applications such as remote sensing and plant biology. Then database indexing structure and recursive density estimation are established to retrieve the images in a fast and efficient way. Also, to improve the accuracy of the image retrieval and prevent the overfitting of training a CNN, the authors in [8] propose a deep CNN with L1 regularization and an activation function named PRelu. The deep network is successfully used to simulate the brain of human by receiving and transferring information and it contains a convolution operation which is appropriate in image processing.

In [1], deep belief network is investigated and trained to learn large scale representations from the images for application where CBIR jobs are used. In that work, similarity measures are applied for CBIR tasks. The authors in [9] investigate the using of CNN for CBIR jobs as well where different setting are implemented and tested. A hybrid of CNN and support vector machine (SVM) model is proposed in [2] using the minimum number of materials and time resources. The last output layer of the proposed CNN is changed with a classifier based on SVM. There are two parts used in that work, convolutional part and recognition part. In the

convolutional part, the images are passed through a sequence of several filters where new images are forming named convolution maps. In the recognition part, a SVM classifier is trained to automatically extract features on testing images and take the final decisions. A kind of deep learning is applied to classify images in [10]. AlexNet deep learning network is effectively used on images selected from ImageNet database. The experiments are conducted on the images after cropping images for different areas. In [11], the semantic features of the images are extracted using CNN model. Then, a distance function is computed to find the similarity between the semantic features of the images.

In [12], a CNN called ConvNet are trained to classify medical images. The medical images are acquired using computed tomography of an organ or body part-specific anatomical. The performance of the classification is improved using data augmentation. Also, deep CNNs are proposed in [13] for content based medical image retrieval. For retrieval process, two approaches are proposed. The first approach, the network is trained to get the prediction of the query image class and then the specific class is searched for relevant images. In the second approach, the whole dataset is searched for the relevant images without including information related to the query image class.

A CBIR system is built using a combination of deep features generated by CNN and SVM to train a linear hyperplane in [14]. The authors use CNN for feature extraction while SVM is applied to find the similarity between image pairs. A deep representation for image retrieval called regional-maximum activations of convolutions (R-MAC) is built in [15]. Using R-MAC, a number of image regions are aggregated into a small and fixed length feature vector robust hence it is robust to scale and translation. This deep CNN gives high accuracy since it can deal with images have high resolution of different ratios. In [16], a CNN model is trained on ImageNet-2012. Then, for CBIR task, the four layers, which are extracted as the feature representation of the data, are evaluated using the retrieval performance. Finally, the original features are compared with the binarized feature representation.

Different CNNs with application to CBIR tasks are examined and compared using varied settings in [9]. The features representation of the images and the similarity measures between image pairs are learnt to process the tasks of CBIR. The authors attempts to approve if CNNs are effective in learning the features of images when applied to CBIR tasks. A deep CNN model is proposed in [17] to learn the features representation from the activations of the convolutional layers. The authors suggest three retraining methods in order to improve the performance of the retrieval process and the amount of the required memory. These are: fully unsupervised retraining when no information is available but only from the dataset itself, retraining with relevance information when the labels of the training data are exists, and relevance feedback-based retraining when there are feedbacks from users.

3. DEEP CONVOLUTION NEURAL NETWORKS

Over the past years there have been extensive studies using deep learning networks (DLNs), for example, deep belief network, Boltzmann machines, restricted Boltzmann machines, deep Boltzmann machine, and deep neural networks (DNN) [9]. In this study, we have investigated, compared, and evaluated some common DLNs and their applications for image classification and automatic image retrieval. These are: AlexNet, VGG-16 and VGG-19 networks, GoogleNet, ResNet. We also have compared the performance of these networks to prior works in this domain by using known accuracy metrics and analyzed the differences between the approaches. In the following subsections, we will explain these DLNs.

3.1. AlexNet

AlexNet is a kind of DLNs introduced by Alex Krizhevsky [18]. The architecture of AlexNet convolutional network is illustrated in Figure 2. As shown in this figure, convolution and max pooling operations are implemented at the first convolutional layer with local response normalization (LRN). The convolutional layer parameters consist of a set of learnable filters. These filters can be used to calculate the features of the images in classification. The filters of the convolutional layers are updated by performing the full convolutional operation on the feature maps between the convolutional layer and its immediate previous layer. In this layer, about 96 different receptive filters are used where the sizes of these filters are 11*11. Also, a stride size of 2 and 3*3 filters are used to perform the max pooling operation. The job of pooling layer is to reduce the computational complexity when nonlinear down sampling is performed. The same operations are implemented but with 5*5 filters in the second layer, 3*3 filters with 384, 384 and 296 features maps in the third, fourth and fifth convolutional layers. More image details and local feature images are extracted since the size of convolutional layer and stride is small. Two layers, which are fully connected (FC), are used with dropout. In AlexNet network, the problems of training time consuming and over-fitting problems are solved by dropout operation. Finally, a softmax layer is used. AlexNet has been used in a wide range of applications such as object detection, video classification and image segmentation [6, 12, 19-22].

3.2. VGG-E Net

VGG-E net has been proposed by Simoyan *et al.* to simulate the relation of depth of the network with its capacity, VGG-E net made 19 deep layers comparing with AlexNet. Figure 3 shows the architecture of the VGG net. It consists of ReLU activation function which is used by two convolutional layers. ReLU is also used by a single max pooling layer and some fully connected layers. The purpose behind putting max pooling after the convolutional layer is to tune the network and the padding is done to preserve the spatial resolution. The last layer is a softmax layer which is used for classification. The size of the convolution filter is 3x3 and has a stride of 2. By using small size of filters, it provides low computational complexity and reduces the number of parameters. There are different kinds of VGG-E models were proposed. These are: VGG-11, VGG-16, and VGG-19 where these models have 11, 16, and 19 layers respectively. Although, the three models of VGG-E have three fully connected at the end, VGG-11 contain 8 convolution layers, VGG-16 has 13 convolution layers and VGG-19 contain 138M weights and 15.5M MACS [21, 23].

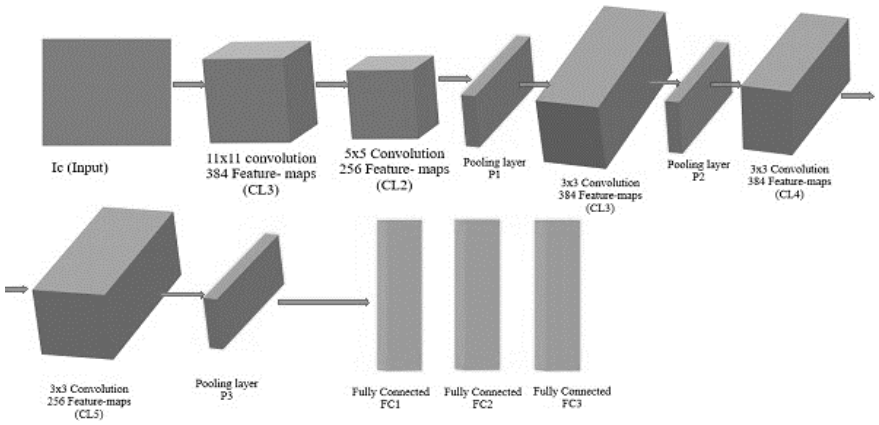

Figure 2. AlexNet Layout with its convolution and connected layers

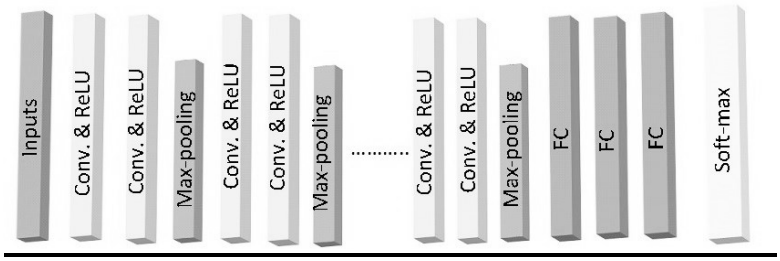

Figure 3. VGG network layout where Conv is the convolution layer and FC is full connected

3.3. GoogleNet

GoogleNet DLN is proposed by Christian Szegedy *et al.* [22]. GoogleNet network has been especially designed to reduce the computational cost and achieve high accuracy compared with traditional CNNs. It presents the concept of inception block. It helps in combining multi scale convolutional transformations by exploiting the idea of split merge and transform operations. Thus, different types of variations in the same category images with diverse resolutions are learnt. Inception blocks are used in replacing the conventional layer. They hide filters of different sizes (1*1 and 3*3) to capture spatial information [23, 21].

The architecture of GoogleNet is illustrated in Figure 4. In this network, nine inception modules are used consists of 22 layers. Although, GoogleNet has many layers compared to other networks before it, the number of the parameters is much lower than AlexNet and VGG networks. It has 7M parameters while AlexNet and VGG have 60M and 138M parameters respectively. Also, GoogleNet network has four max pooling layers and one average pooling layer i.e. only layers with parameters. The average pooling layer has a filter with a size of 5*5 and has three strides which is used before the classifier. It also uses dropout layer which has a ratio of 70% from dropped outputs. All convolutional layers and inception modules use ReLu [21, 22].

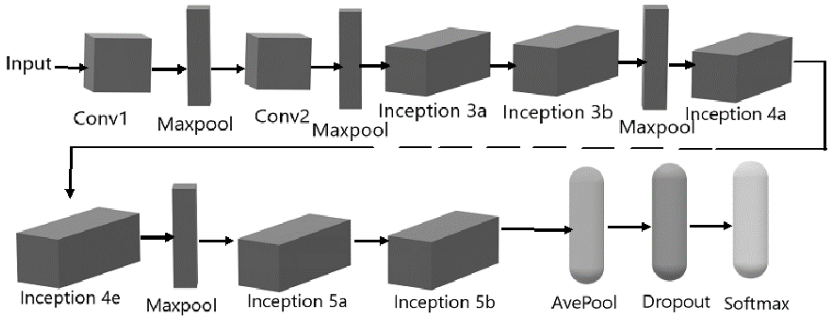

Figure 4. GoogleNet architecture

3.3. ResNet

Deep residual networks or called ResNet is proposed by Kaiming He *et al*. [24]. It is one of the states of art and greatest CNNs used for image recognition. In ImageNet Large Scale Visual Recognition Challenges in 2015(ILSVRC-15), ResNet won that challenge with a top 5 error of 3.57%. For instance, ResNet-50 has reached an average of 5.25% of top-5 error when it is trained on 1.28 million training images in 1000 classes. It has shown a high accuracy in computer vision. Figure 5 shows the architecture of ResNet-50. In this study, ResNet-50 has been used for image classification. In this network, 5 convolutional layers are used and the input images are of size 224*224*3. ResNet-50, which has 50-layer CNN architecture, is considered to be the first deep CNN that applied residual learning [24, 25].

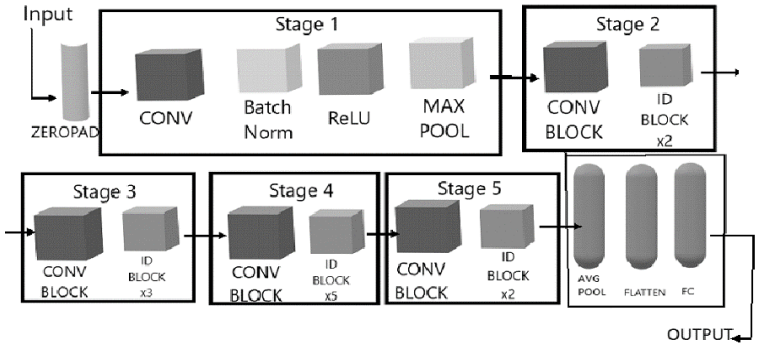

Figure 5. ResNet architecture

4. PROPOSED METHOD

In this work, two scenarios are followed: image classification and image retrieval. Figure 6 shows the stages of the framework, training, CNN model training, image classification, feature extraction, similarity measure, and image retrieval. For image classification, CNNs are investigated to classify huge amount of images. In our investigation, different deep learning approaches are used in classification such images. The CNNs approaches are exploited to learn features of images. Image classification is achieved by two stages. First, a set of training images that associated with class label are used to train a classifier. Second, the trained classifier is used to predict the class label of a query image based on its trained knowledge about the class. Hence, the accuracy of the classifier is evaluated according to correct prediction. Image retrieval is implemented using features that are learned by the CNNs approaches and then results are compared. Based on outcomes and analyses a new algorithm for image retrieval is developed (see subsection 4.2.3).

4.1. Data sets

Different datasets have been used for testing algorithms or approaches in CBIR. The datasets used in this paper to evaluate the performance of CNNs are datasets with a high quality where the images are non-labeled and compressed. Datasets corel 1K [26], corel 50K [26] and Caltech 256 [27] are used in this work to validate the proposed system.

4.1.1. Corel 1K

Corel 1K dataset [26] consists of 1000 images with 100 for each class. The size of images is (256x384) or (384x256) each image may be one of the ten class labels (African peapole, beach, buidings, buses, dinosaurs, elephants, flowers, hourses, mountains, and foods). These labels are annotated manually using an Excel file. A sample of 20 images is shown in Figure 7 with their labels.

4.1.2. Corel 10K

Corel 10K dataset [26] consists of 10000 images with 100 for each class. The size of images is (126x187) or (187x126). We slected 50 classes, art, weman, dog, cloud, machroom, castle, glass, bear, fighting people, and fruit. Figure 8 shows a sample of images.

4.1.3. Caltech 256

Caltech 256 dataset [27] consists of 30,607 images of objects with different sizes. Images are divided into 256 classs. Researchers select some classes to evaluate their approaches or algorithms. In our experiment, we chose 50 classes with 100 for each class. Figure 9 shows sample of some images.

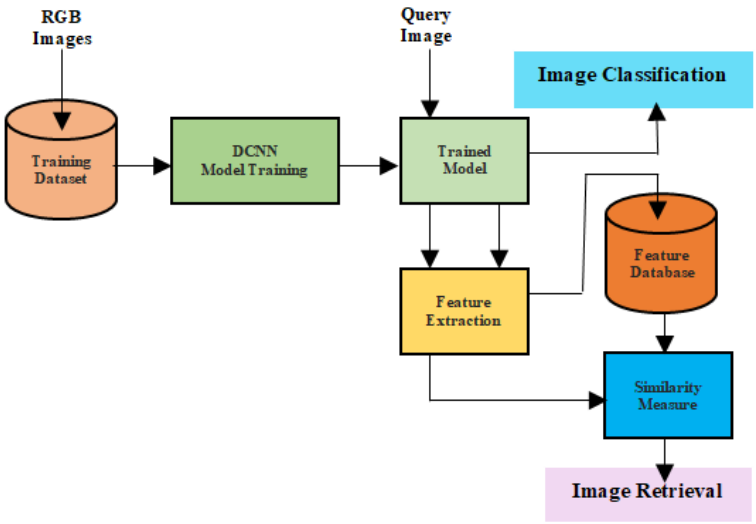

Figure 6. Framework of image classification and retrieval

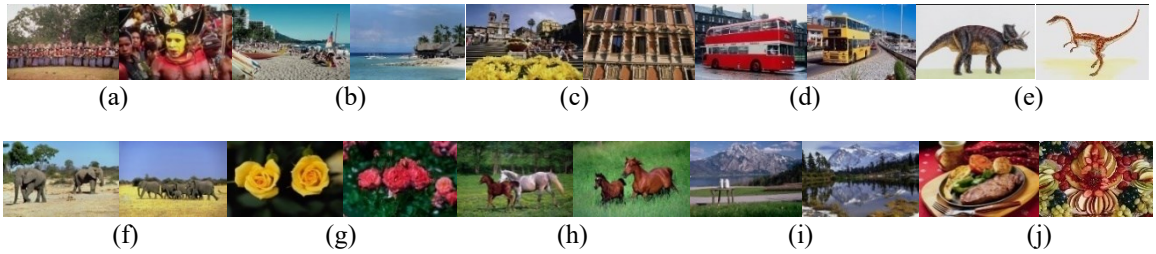

Figure 7. Sample of corel 1K images; (a) African People 1, (b) Beach 2, (c) Buildings 3, (d) Buses 4, (e) Dinosaurs 5, (f) Elephants 6, (g) Flowers 7, (h) Horses 8, (i) Mountains 9, and (j) Foods 10

Figure 8. Sample of corel 10K images

Figure 9. Sample of Caltch256 images

4.2. Experimental results and analysis

In this section, we present the results of the experiments conducted to evaluate the accuracy of IR and computational efficiency based on prposed CNNs in terms of image classification and image rerieval. Image classification is achieved by two stages. First, a set of training images that associated with class label are used to train a classifier. Second, the trained classifier is used to predict the class label of a query image based on its trained knowledge about the class. Hence, the accuracy of the classifier is evaluated according to correct prediction. IR returns top T images as a ranked list from database images that are most similar to a query image by using a similarity measure without using class labels. The accuracy is evaluated according to how many correct images out of the T images in the ranked list. All experiments are performed using MATLAB 2018a, on a computer with a processor Intel core i7 CPU 2.5 GHz 2.6 GHz and 8 GB RAM.

4.2.1. Evaluation of the performance

In image classification, a confusion matrix is usually used to evaluate the performance of a classifier. Table 1 shows a confusion matrix for two classes and it can be extended into m classes (i.e. m x m). True positive (TP), true negative (TN), false negative (FN), and false positive (FP) are the terms given to an image classification test [28]. Precision or accuracy is calculated as follows:

$$AC_i = C(i,i)/\sum_{j=1}^{n} C(j,i) \quad (1)$$

$$AC = \sum_{i=1}^{n} AC_i/_n \quad (2)$$

where, AC is the precision or accuracy.

Table 1. Confusion Matrix

		Predicted class	
		C_1	C_2
Actual class	C_1	TP	FN
	C_2	FP	TN

In image retrieval, a mean average precision (MAP) is used for evaluation based on precision (P) and average precision (AP) [28].

$$P = \frac{IM}{TIM} \quad (3)$$

where, P is the precision of image retrieval, IM is number of relevant retrieved images and TIM is total number of retrieved images,

$$AP = \frac{\sum_{i=1}^{n} P_i}{n} \quad (4)$$

where, AP is average precision of image retrieval, P_i is precision of i image in the class, and n is total number of images in the class.

$$MAP = \frac{\sum_{j=1}^{m} AP_j}{m} \quad (5)$$

where, MAP is mean average precision of image retrieval, AP_j is average precision of j class image, and m is total number of classes in the database.

4.2.2. Image classification

Many experiments are conducted on the image datasets. The training models of the networks are set up as follows: the datasets are divided into 70% for training, 15% for validation and 15% for testing data. In addition, the training parameters for the CNNs are set as follows: the learning rate is set to 0.00001; the maximum epoch number is 435. Also, the weight of the learning rate factor and bias learning rate factor are set to 20 for the layer of fully connected.

The most common CNNs used in the paper as mentioned in the previous section are: simple CNN, AlexNet, GoogleNet, ResNet-50, Vgg-16 and Vgg-19. These models are compared with the conventional methods used for IR such as the hue saturation value (HSV) colour feature, gray level co-occurrence matrix (GLCM) features and scale invariant feature transform (SIFT) [6]. The accuracy of the results of the testing and validation data sets is used on image data to evaluate the performance of these methods. The results of the conventional methods and CNNs models as feature extractors based on corel 1K dataset are shown in Table 2 with data augmentation. As can be seen from this table, the best accuracy are 99%, 97% and 95% achieved by CNNs models when the training, validating and testing data are augmented compared with the conventional approaches.

Table 2. Accuracy of CNNs models as feature extractors based on corel 1K data augmentation dataset

Method	Accuracy (%)	Training Time
HSV	0.43	Not Available
GLCM	0.39	Not Available
SIFT	0.53	Not Available
CNN [5]	0.79	Not Available
AlexNet	0.99	10 min.
GoogleNet	0.97	7 min.
ResNet-50	0.95	89 min.
Vgg-16	0.98	43 min.
Vgg-19	0.98	131 min.

For the corel 1K datasets, the models based on the CNNs models did converge to excellent accuracy and demonstrate high performance in training stage with the least number of epochs. Although, there are no significant differences in the convergences of the models (more than 95%), they took more training time for convergence as the complexity of the CNNs are increased. On the other hand, the convergence accuracy results for the same datasets without data augmentation have not given good accuracy. For example, the testing accuracy is 10%, 46% and 68% for the simple CNN, AlexNet and GoogleNet respectively. It is shown that to improve the performance of CNNs, data augmentation can successfully be used.

A sample of 30 class probabilities results for both AlexNet and GoogleNet convolutional neural network as feature extractors with augmentation is shown in Figure 10. From the results, it is observed that most classes have high accuracy, the classification is almost successful. Also, it is shown that the CNNs models results are superior to the known three methods. On the otherhand, the results of the CNNs models as feature extractors based on corel 50K and Caltech 256 datasets are shown in Table 3 and Table 4 with data augmentation. From the experiments, it is apparent that the corel 1K images data are classified correctly using CNNs models with high accuracy while the Caltech 256 data with 50 classes has low accuracy. It is concluded that the features of the images are learnt from the pre-trained models and it does not need to search features manually.

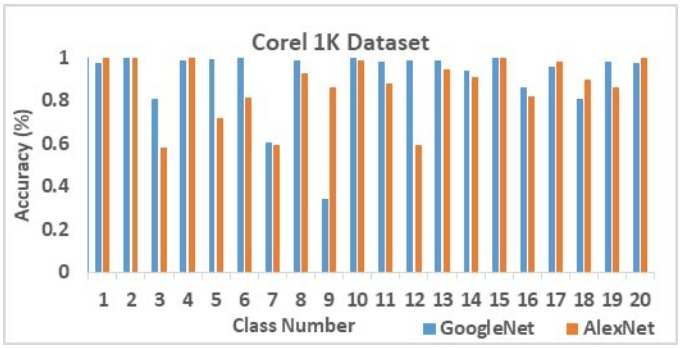

Figure 10. A sample of class probabilities of AlexNet and GoogleNet feature extraction of corel 1K dataset

Table 3. Accuracy of CNNs models as feature extractors based on corel 50K data augmentation dataset

Method	Accuracy (%)	Training Time
AlexNet	0.96	198 min.
GoogleNet	0.97	324 min.
ResNet-50	0.93	306 min.
Vgg-16	0.97	564 min.
Vgg-19	0.98	591 min.

Table 4. Accuracy of CNNs models as feature extractors based on Caltech 256 data augmentation dataset

Method	Accuracy (%)	Training Time
AlexNet	0.43	48 min.
GoogleNet	0.44	48 min.
ResNet-50	0.45	889 min.
Vgg-16	0.42	1759 min.
Vgg-19	0.44	555 min.

4.2.3. Image retrieval

As mentioned earlier, feature representation is one challenge of semintac gap in CBIR. Recently, CNN has been used to learn features to be more accurate. Hence, our aim in these experiments is using above CNNs approaches to learn features and handle image retrieval without using class labels by using them. Resulted features from five CNNs (AlexNet, GoogleNet, ResNet-50, Vgg-16, and Vgg-19) are seperately tested according to the framework in Figure 6.

Firstly, experiments of image retrieval are conducted on corel 1K standard database to judge which deep learning approach can produce effective feature than others. Leave-one-out manner is used to calculate Precisions (P) for images and then MAPs are computed. City-block (L_1) distance function is used to compute the similarity between a query image vector (feature) and database images vectors. Resulted similarity values are ranked in ascending order. Top (5-100) retrieved images in terms of MAPs for CNN approaches are calculated and illustrated in Figure 11.

It is clear that the performance of using feature (10D) that is produced from GoogleNet with 22-layers is more effective and robust than others as long as Top (5-100) retrieved images. Meanwhile, AlexNet with 20-layers extracted feature (4096D) that has lowest achievement. Vgg-16 and -19 produced features which are the same as that of AlexNet in length but they performed higher. ResNet-50 extracted a smaller dimension of feature which is 2048 compared to the AlexNet, Vgg-16 and -19 approches but the features are more robust especially at Top30-100 ranked list of images. Therefore, it is interest to analys individual class images between Googlenet and Resenet-50 at Top100 retrieved images. Hence, APs are clarified in Figure 12.

At the first view, there is a big difference between two approaches where the performance of GoogleNet is higher than ResNet-50 over all classes except for the bus class, the rate is equal. In order to judge how the difference is significant, a t-test statistical method is used that can be calculated as [29]:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)S_1^2+(N_2-1)S_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}} \quad (6)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample precision rates (P), S_1 and S_2 are standards deviations, and N_1 and N_2 are the sample sizes. Two hypotheses are regarded and determined based on t-test, the null hypothesis (H_0) where $\bar{X}_1 - \bar{X}_2 = 0$ and alternative hypothesis (H_A) where $\bar{X}_1 - \bar{X}_2 \neq 0$. P-value of the test is the probability of observing a test. Small values of p refers to that the null hypothesis is rejected at significance level 0.05.

For each class in the corel 1K database, the test was computed. This means the size of each sample is 100 elements (i.e. precision values). Hence, the first sample (S_1) and second sample (S_2) have precision rates of Top100 retrieved images from using Googlenet feature and ResNet-50 feature respectively. The t-test proved that all diffrences between precesion values are signifigant even for Buses class. Figure 13 shows the two samples where the most values of S_1 99% compared to S_2. We can conclude that GoogleNet learened a feature with low dimension (10) means less computation and high accuracy due to the inception block that exploits split, merge and transform operations to combine multi scale convolutional transformations. Therefore, different types of variations in the same category images with diverse resolutions are learnt. In other words, Googlenet has ability to extract more discriminative information about interested objects than Resnet-50 at layer 22.

We conducted other retrieval experiments to investigate the second issue in CBIR which is similarity measures. In the literature, different measures have been used to compute the similarity between a query image and database images depending on image descriptor. For instance, the descriptor is represented as a single vector or a set of vector, in linear space or non-linear manifold. [29, 30]. Hence, correlation (D_1), cosine (D_2),

and Euclidean (D_3) were applied rather than city-block (D_4) in our system separately. Suppose x_{QI} and x_{DI} refere to query image and database image feature vectors respectively with the n dimension, then D_1, D_2, D_3, and D_4 are defined as follows [31].

$$D_1 = 1 - \frac{(x_{QI}-\bar{x}_{QI})(x_{DI}-\bar{x}_{DI})^T}{\sqrt{(x_{QI}-\bar{x}_{QI})(x_{QI}-\bar{x}_{QI})^T}\sqrt{(x_{DI}-\bar{x}_{DI})(x_{DI}-\bar{x}_{DI})^T}} \quad (7)$$

where $\bar{x}_{QI} = \frac{1}{n}\sum_j x_{QI_j}$ and $\bar{x}_{DI} = \frac{1}{n}\sum_j x_{DI_j}$

$$D_2 = 1 - \frac{x_{QI}x_{DI}^T}{\sqrt{(x_{QI}x_{QI}^T)(x_{DI}x_{DI}^T)}} \quad (8)$$

$$D_3 = \sqrt{\sum_{j=1}^{n}(x_{QI_j} - x_{DI_j})^2} \quad (9)$$

$$D_4 = \sum_{j=1}^{n}\left|x_{QI_j} - x_{DI_j}\right| \quad (10)$$

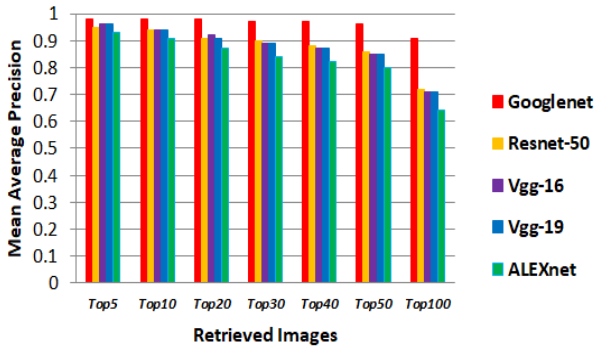

Figure 11. MAPs for CNN approaches using corel 1K

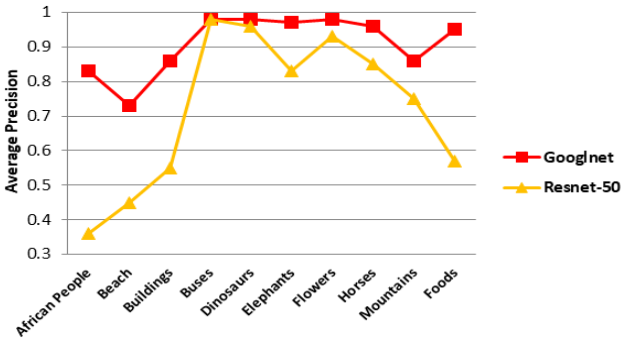

Figure 12. APs of Top100 retrieved images for corel 1K database image classes

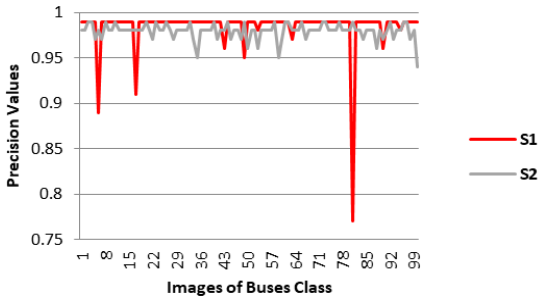

Figure 13. Precision values along bus class images

Table 5 shows APs for individual corel 1K classes for Top20 retrieved images compared to recent work in 2020. It is clear that the ability of the Googlenet approach to learn feature with low dimension (10D) led to reduce the semantic gap across all classes using above four distances. Hence, our proposed method achieved remarkable rates comparing with recent methods [32, 33] which are more complicated. Where the method in [32] combines two features in terms of fusion, the first one was produced from using detects salient objects, spatial color and texture features and the second one from using ResNet CNN approach. Our experience referred to that fusing normal and CNN features degrade or do not affect rates of image retrieval when we fused the learned feature from Googlenet CNN and global local binary patterns (LBP) colour texture feature (177D) from YCbCr images. Meanwhile, the method in [33] used the fusion between two normal features. The first one can detect shapes, objects, and texture by locating interest points and the second one is color features extracted from the spatially arranged L2 normalized coefficients. This evidence supports that the learned feature from CNN approachs is more effective.

Table 6 illustrates APs of image retrieval for Top100 retrieved images using above four distance functions to calculate the similarity between the query image and database feature vectors. As we can see that correlation and cosine perform equally and are higher than city-block overall classes because the cosine similarity between two images is the cosine of the angle formed by two vectors relative to visual content of images and the correlation similarity between two vectors is a mean centered cosine similarity. Both similarity measures are subtracted from 1 as in (7) and (8). Meanwhile, Euclidean approaches the correlation and cosine distances. To judge the significant differences betweeen D_1 and D_4, *t*-test was used by taking samples of precision values that were achived from D_1 and D_4 for each class as shown Figure 14. Then the *t*-test proved that alternative hypothesis (H_A) is not equal to zero and values of *p* are small means the null hypothesis is rejected at significance level 0.05 for all classes as shown Table 7.

Table 5. APs comparision for Top20 retrieved images with recent works

	People	Beach	Buildings	Buses	Dinosaurs	Elephants	Flowers	Horses	Mountains	Foods	MAP
D_1	0.97	0.93	0.99	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.98
D_2	0.97	0.92	0.99	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.98
D_3	0.96	0.90	0.97	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98
D_4	0.96	0.88	0.96	1.00	1.00	1.00	1.00	1.00	0.95	1.00	0.98
[32]	0.79	0.85	0.78	0.97	1.00	0.77	1.00	1.00	0.95	0.81	0.89
[33]	0.89	0.70	0.79	0.70	1.00	0.72	0.82	1.00	0.65	0.74	0.80

Table 6. AP comparision for Top100 retrieved images using D_1, D_2, D_3, and D_4 distance functions

	People	Beach	Buildings	Buses	Dinosaurs	Elephants	Flowers	Horses	Mountains	Foods	MAP
D_1	0.88	0.82	0.95	0.99	0.99	0.99	0.99	0.98	0.92	0.98	0.95
D_2	0.88	0.82	0.95	0.99	0.99	0.99	0.99	0.98	0.92	0.98	0.95
D_3	0.88	0.79	0.91	0.98	0.99	0.98	0.99	0.97	0.89	0.97	0.94
D_4	0.83	0.73	0.86	0.98	0.98	0.97	0.98	0.96	0.86	0.95	0.91

Table 7. Alternative hypothesis and *p*-values

	People	Beach	Buildings	Buses	Dinosaurs	Elephants	Flowers	Horses	Mountains	Foods
H_A	1	1	1	1	1	1	1	1	1	1
p-value	0.0015	3.86E-05	1.20E-07	0.0376	0.0284	1.66E-06	1.68E-06	0.0119	0.0006	0.0001

We expanded the experiment to corel 10K and Clatech250 databases with 50 classes using the best approach (i.e. Googlenet). The approach produced a learned feature within 50 dimentional in length for both databases. Consequently, the process of image retrieval was applied using above four distances. Results showed that D_1 and D_2 performe better than D_3 and D_4 about 6% more for Top100 retrieved images as shown in Table 8. Hence, we ended up with a novel algorithm that uses the Googlenet CNN approach to learn image feature and correlation or cosine distance function to compute the similarity between query and database images as shown in Figure 15.

Table 8. MAPs of Top100 retrieved images using D_1, D_2 and D_4 for corel 10K and Caltech 250

Distance Function	Corel 10K	Caltech 256
D_1 and D_2	0.46	0.26
D_4	0.40	0.20

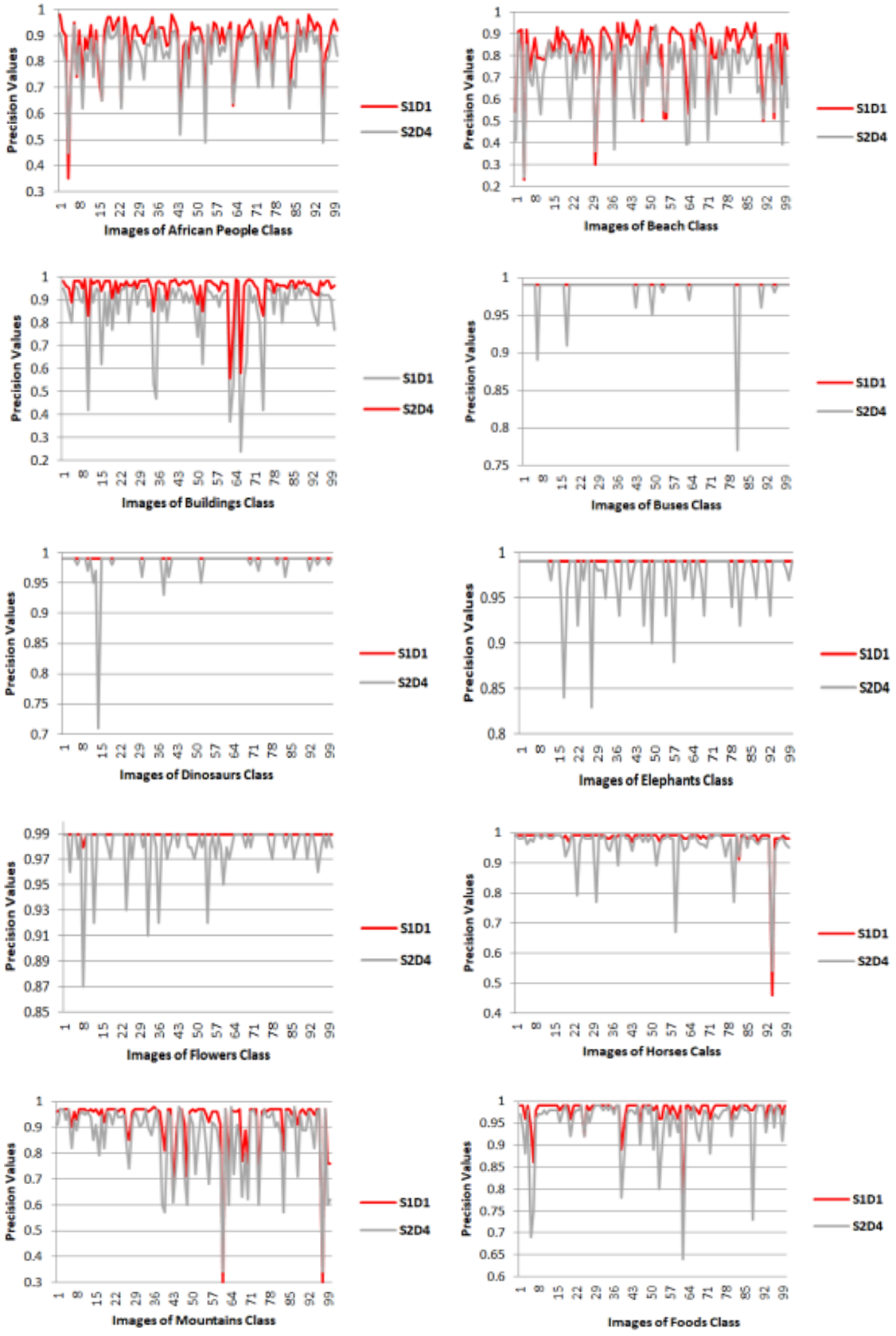

Figure 14. Precision values along corel 1K classes

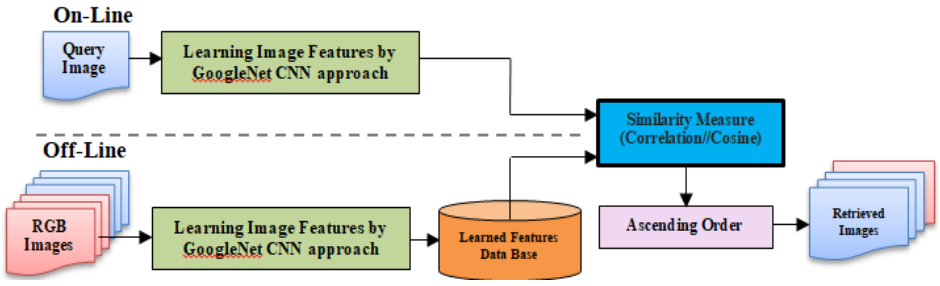

Figure 15. A diagram of developed algorithm

5. CONCLUSION

In this paper, a novel algorithm for image retrieval using a deep neural networks learning was developed based on experience from exhausted experiments in terms of image classification and retrieval when class label is available and unavailable respectively. Different CNNs are used and compared with the other conventional IR methods. The developed algorithm used Googlenet CNN approach to learn feature and correlation/cosine distance function to compare two images. Hence, remarkable rates were achieved comparing with recent methods due to the effective learned feature and accurate distance function. The semantic gap challenge was consequently reduced. We plan to evaluate this algorithm on faces and medical database images. Also, our future investigation is to implement CNNs approaches using different colour spaces such as YCbCr and HSV to see the impact on accuracy.

ACKNOWLEDGEMENTS

We commend the efforts made to make corel 1K, corel 10K and Caltech 256 dataset available, making it easier to do this study. Also, we would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad, Iraq for its support of this work.

REFERENCES

[1] R. R. Saritha *et al.*, "Content based image retrieval using deep learning process," *Cluster Computing*, vol. 16, no. 3, pp. 1-14, 2018.
[2] O. Mohamed *et al.*, "Content-based image retrieval using convolutional neural networks," In *First International Conference on Real Time Intelligent Systems*, Springer, Cham, October 2017, pp. 463-476.
[3] D. D. Feng *et al.*, "Multimedia informaton retrieval and management technolgical fundamentals and applications," *Springer-Verlag Berlin Heidelberg,* 2003.
[4] A. Smeulders *et al.*, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 1349-1380, 2000.
[5] F. ZHANG, and Z. Bao-jiang, "Image Retrieval Based on Fused CNN Features," *DEStech Transactions on Computer Science and Engineering aics*, 2016.
[6] WQ. Huang, and Q. Wu, "Image retrieval algorithm based on convolutional neural network," In *Current Trends in Computer Science and Mechanical Automation*, *Sciendo Migration*, vol. 1, pp. 304-314, Dec 31, 2017.
[7] P. Sadeghi-Tehran *et al.*, "Scalable Database Indexing and Fast Image Retrieval Based on Deep Learning and Hierarchically Nested Structure Applied to Remote Sensing and Plant Biology," *J. Imaging,* vol. 5, no. 33, pp. 1-21, 2019.
[8] W. Qing Jie, and WB. Wang, "Research on image retrieval using deep convolutional neural network combining L1 regularization and PRelu activation function," *IOP Conference Series: Earth and Environmental Science*, vol. 69, no. 1, 2017.
[9] J. WAN *et al.*, "Deep learning for content-based image retrieval: A comprehensive study," *Conference: Proceedings of the ACM International Conference on Multimedia*, 2014.
[10] K. Manoj *et al.*, "Image classification using Deep learning, " *International Journal of Engineering & Technology*, vol. 7, no. 2.7, pp. 614-617, 2018.
[11] H. Liu *et al.*, "Image Retrieval Algorithm Based on Convolutional Neural Network," *2nd International Conference on Artificial Intelligence and Industrial Engineering*, *Atlantis Press*, 2016.
[12] H. Roth *et al.*, "Anatomy-specific classification of medical images using deep convolutional nets," *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015.
[13] A. Qayyum, *et al.*, "Medical image retrieval using deep convolutional neural network," *Neurocomputing,* vol. 266, pp. 8-20, 2017.
[14] R. Fu *et al.*,"Content-based image retrieval based on CNN and SVM," *2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 638-642.

[15] A. Gordo *et al*., "Deep image retrieval: Learning global representations for image search," In *European conference on computer vision*, Springer, Cham, pp. 241-257, 2016.
[16] H. Wang *et al*., "Deep learning for image retrieval: What works and what doesn't," In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 1576-1583.
[17] M. Tzelepi, and T. Anastasios, "Deep convolutional learning for content based image retrieval," *Neurocomputing,* vol. 275, pp. 2467-2478, 2018.
[18] SH. Wang, *et al.*, "Alcoholism identification based on an AlexNet transfer learning model," *Frontiers in psychiatry*, 2019.
[19] W. Yu *et al.*, "Visualizing and comparing AlexNet and VGG using deconvolutional layers," *Proceedings of the 33 rd International Conference on Machine Learning*, 2016.
[20] W. Nawaz, *et al*., "Classification of Breast Cancer Histology Images Using ALEXNET," *International Conference Image Analysis and Recognition*, *Springer*, *Cham*, 2018.
[21] M. Z. Alom *et al.*, "A State-of-the-Art Survey on Deep Learning Theory and Architectures," *Electronics*, vol. 8, vol. 3, pp. 1-67, 2019.
[22] R. Sustika *et al.*, "Evaluation of Deep Convolutional Neural Network Architectures for Strawberry Quality Inspection," *International Journal of Engineering & Technology*, vol. 7, no. 4.40, pp. 75-80, 2018.
[23] A. Khan *et al.*, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review,* vol. 53, pp. 5455-5516, 2020.
[24] M. Sankupellay and D.A. Konovalov, " Bird call recognition using deep convolutional neural network, ResNet-50," *In Proceedings of ACOUSTICS*, vol. 7, no. 9, Nov, 2018.
[25] S. Almabdy and E. Lamiaa, "Deep Convolutional Neural Network-Based Approaches for Face Recognition," *Applied Sciences,* vol. 9, no. 20, pp. 1-21, 2019.
[26] Z. James *et al.*,"SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947-963, 2001.
[27] G. Griffin *et al.*,"Caltech-256 object category dataset," 2007.
[28] H. Müller *et al.*,"Performance Evaluation in Content-based Image Retrieval: Overview and Proposals," *Pattern Recogn. Lett.,* vol. 22, no. 5, pp. 593-601, 2001.
[29] A. M. Graziano and M. L., "Research methods: A Process of Inquiry," *8th ed., New York, HarperCollins College Publishers*, 2012.
[30] H. Du *et al.*, "Effectiveness of image features and similarity measures in cluster-based approaches for content-based image retrieval," *Proceedings of SPIE - The International Society for Optical Engineering,* 2014.
[31] Stephen L. France, *et al.*,"Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization," *Information Sciences,* vol. 184, no. 1, pp. 92-110 2012.
[32] K. Kanwal *et al.*," Deep Learning Using Symmetry, FAST Scores, Shape-Based Filtering and Spatial Mapping Integrated with CNN for Large Scale Image Retrieval," *Symmetry*, vol. 12, no. 4, 2020.
[33] KT. Ahmed *et al.*,"Deep Image Sensing and Retrieval Using Suppression, Scale Spacing and Division, Interpolation and Spatial Color Coordinates with Bag of Words for Large and Complex Datasets," *IEEE Access*, vol. xx, pp. 1-32, 2020.

BIOGRAPHIES OF AUTHORS

Hanan A. Al-Jubouri received her B.Sc. in Computer Science from University of Technology/Baghdad, Iraq in 1994. She obtained her M.Sc. from University of Technology in 2001. Her Ph.D. degree in Information Systems, Buckingham University, Buckingham, UK, 2015. Hanan joined Mustansiriyah University/ Engineering College in 1994 as a member of the academic staff. Her research interests are mainly Content-Based Image Retrieval and Data mining.

Sawsan M. Mahmoud received her B.Sc. in Computer Science from University of Technology/Baghdad, Iraq in 1994. She obtained her M.Sc. from University of Baghdad in 1998. Her Ph.D. degree in Computational Intelligence is obtained from Nottingham Trent University, Nottingham, UK in 2012. Sawsan joined Mustansiriyah University/ Engineering College in 1994 as a member of the academic staff. Her research interests include, but not limited to, Computational Intelligence, Ambient Intelligence (Smart Home and Intelligent Environment), Wireless Sensor Network, Data Mining, and Health Monitoring.