

Combining convolutional neural networks and slantlet transform for an effective image retrieval scheme

Mohammed Sabbih Hamoud Al-Tamimi

Department of Computer Science, College of Science, University of Baghdad, Iraq

Article Info

Article history:

Received Des 26, 2018

Revised Apr 10, 2019

Accepted Apr 25, 2019

Keywords:

Content base image retrieval
wavelet transforms
Convolutional neural networks
Deep learning
Information retrieval
Slantlet transform

ABSTRACT

In the latest years there has been a profound evolution in computer science and technology, which incorporated several fields. Under this evolution, Content Base Image Retrieval (CBIR) is among the image processing field. There are several image retrieval methods that can easily extract feature as a result of the image retrieval methods' progresses. To the researchers, finding resourceful image retrieval devices has therefore become an extensive area of concern. Image retrieval technique refers to a system used to search and retrieve images from digital images' huge database. In this paper, the author focuses on recommendation of a fresh method for retrieving image. For multi presentation of image in Convolutional Neural Network (CNN), Convolutional Neural Network - Slantlet Transform (CNN-SLT) model uses Slantlet Transform (SLT). The CBIR system was therefore inspected and the outcomes benchmarked. The results clearly illustrate that generally, the recommended technique outdid the rest with accuracy of 89 percent out of the three datasets that were applied in our experiments. This remarkable performance clearly illustrated that the CNN-SLT method worked well for all three datasets, where the previous phase (CNN) and the successive phase (CNN-SLT) harmoniously worked together.

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Mohammed Sabbih Hamoud Al-Tamimi,
Department of Computer Science, College of Science,
University of Baghdad,
Baghdad, Iraq.
Email: m_altamimi75@yahoo.com

1. INTRODUCTION

Of late, there has been an upsurge in the consumption of digital images with the developing accessibility of the internet and computers, particularly because digital image media creation is speedily increasing. The availability of inexpensive storage devices and the user demand together with high quality printers give room for public consumers to print and collect digital images with ease from the Internet. In addition, the prompt development of network technologies has stimulated the application of digital images as one of the most essential communication media for everyday life.

CBIR refers to the retrieval of pertinent images from an image database in accordance with spontaneously resultant features, for instance shape, color and texture representing the image's information content. In several application areas, the need for efficient content-based image retrieval has immensely increased, for instance, entertainment, biomedicine, education, crime prevention, commerce, military and culture. In general, retrieval through documents or images founded on textual description is very easy though the process necessities to manually tag the images, which is time consuming and laborious apart from being highly susceptible to error. The manual process is reliant on the human knowledge and this clues the process to uncertainty since dissimilar individuals have diverse image understanding.

Over the traditional text-based retrieval, CBIR has several benefits. CBIR is capable of overcoming the pointed out challenges by automatically tackling them or via machines, which is very efficient and

precise devoid of human interference as a result of the use of the visual contents of the query image in CBIR [1]. CBIR has previously been accomplished via several methods and research is still underway for progressive upgrading. In addition, with the current digital technology, as well as the enormous number of images all over the world, an automated system for the retrieval via images is therefore compulsory. It is a more effective and efficient means of discovering pertinent images as compared to searching according to text annotations. Moreover, CBIR does not use the wasted time in the process of text-based method's manual annotation. In CBIR field, these benefits have encouraged this study to design image retrieval technique. This challenge has resulted in the increase of development and research in the CBIR field. Images are retrieved according to features that are extracted automatically from images in CBIR.

Amongst the image features that are low-level, for instance, shape, color, spatial location and texture, texture used to be objective and effective in retrieval of content base image. Diverse methods advanced for extraction of texture features, largely categorized into the spectral (similarly referred to as frequency) and the spatial techniques. Generally, the spatial methods are reliant on statistical computations on the image. These statistic methods are however sensitive to image noise and lack adequate number of features [2]. Spectral techniques of texture analysis for image retrieval are on the other hand strong to noise. The spectral techniques comprise using multi-resolution texture extract approaches such as Wavelet Transform (WT) [3, 4] for texture illustration, discrete cosine transform [5] and Multi Resolution (MR) techniques for instance, Gabor filters [6, 7]. The disadvantage with these spectral techniques is that they do not effectively capture the image's edge information. This is the rationale behind looking for a better resolution to conglomerate the finest features from spectral approach and spatial approach, which can be strong to noise with simple statistical computation.

The problem background was discussed earlier. It indicated that the problems related to the CBIR process need further investigation. One of the major challenges of the CBIR was that the images must be represented using effective and accurate extraction techniques. The higher dimensionality of the colour feature vector along with the extracted features does not display any spatial information and consists of semantic gaps. These issues must be resolved for improving the precision of the retrieval performance [3, 4]. Along with developing techniques for combining the texture, shape or colour-based similarities, the drawbacks noted in the earlier studies were investigated for determining the questions that need answers. Here, the researchers have proposed and developed a novel automatic features extraction process that was based on the SLT and CNN method, which was considered as one of the best deep learning methods in the field of machine learning algorithms.

2. RELATED WORK AND DEFINITION

The finger veins are not visible to the naked human eyes under normal illuminating conditions. On the other hand, they could be viewed using the Near-InfraRed (NIR) light between the wavelengths of 700 and 1000 nm. The human tissues are seen to absorb the NIR light waves, however, these waves get blocked by the deoxidised Haemoglobin (HbO) molecule, which is present in higher concentrations in the human veins, which make these veins darker in the acquired images [1]. The vein-scanners generally support the NIR light waves, which are generated from the Light Emitting Diodes (LEDs), and the Charge Coupled Device (CCD) cameras or the Complementary Metal-Oxide Semiconductor (CMOS) cameras. These devices capture the images of the veins from a particular region. Also, these devices comprise of several optical filters which scatter the NIR-emitted beams and also, increase the contrast of the captured raw images, as described in Figure 1.

There are a number of factors that determine the CBIR; these aspects include feature extraction technique, using suitable features in CBIR, resemblance measurement technique and selected mathematical convert to compute operational characteristics, reaction procedure. In CBIR, all these aspects are very important. An efficient retrieval mechanism can be accomplished through improvement of some of the prompting aspects. We first avail a short review of the aspects that could influence CBIR for this purpose. From the discussion on CBIR in introduction, it is assumed that using low-level image features like color, shape and texture is to amass information from an image to restoration. A variety of spectral techniques of extracting texture features as well as all of the existing techniques of validating image texture characteristics in contemporary literature have been talked over with the intention of attaining the objective of the research, which is to ascertain the most sufficient characteristics in CBIR. We attempt to handle the weaknesses of a spectral method as well as the way a different method could offer resolutions, and which is the most operational among them in representation of texture characteristics in this discussion. The vital concerns of content based image retrieval system include the following: (1) Similarity measurement, (2) Low-level image, (3) features extraction, (4) Selection of image database, and (5) Performance evaluation of the retrieval process.

2.1. Deep learning

In numerous artificial intelligent tasks, for instance, machine translation, object detection and speech recognition, deep learning has radically enhanced the state-of-the-art [8]. Its deep architecture nature provides deep learning with the likelihood of resolving numerous problematical artificial intelligent tasks [9]. Therefore, researchers are prolonging deep learning to diverse contemporary domains and tasks together with traditional tasks such as face recognition, language models, or object detection, [8] employs the insistent neural network to denoising speech signals, [10] applies stacked auto encoders in discovering gene expressions' clustering patterns. [11] Applies a neural model in creating images with diverse styles [12]. Make use of deep learning to enable instantaneous sentiment analysis from numerous modalities, [13] Put into use deep learning for classification of biological image.

This is an era of witnessing the deep learning research's flourishing. Deep learning performs better as compared to other machine learning algorithms the way it is proposed by the empirical outcomes. There are a number who have recommended that it is due to the fact that it sloppily imitates the functions of a brain, neural networks' numerous layers stacked one after another one such as the classical brain model. Up to date on the other hand, there is no strong hypothetical foundation for deep learning, [14] then deep learning machines regularly work better as compared to traditional ML devices since they learn the part of feature extraction as well. The aim of deep learning techniques is to learn feature hierarchies with characteristics from higher levels of the hierarchy designed by aligning features from lower level. Spontaneously learning characteristics at numerous levels of abstraction enable a system to learn intricate functions that directly map the input to the output from data, devoid of depending completely on human crafted characteristics [14]. A point in case is image recognition, where the traditional system is to remove handcrafted characteristics before feeding a Support Vector Machine (SVM). In contrast, deep learning schemes optimize the extracted features that largely enlighten on the rationale behind their better performance.

The salient variance between traditional machine learning and deep learning is its performance as data increases' scale. Deep learning algorithms do not perform well in a situation where the data is small. This is due to the fact that deep learning algorithms necessitate a huge amount of data for it to be perfectly understood [12].

2.2. Convolutional neural network (CNN)

One specific kind of deep feedforward network that was trained with much ease generalized much better as compared to networks having full connectivity between layers that are nearby. This was the CNN [15, 16], which accomplished several practical successes at a time when neural networks were beyond good turn and it had currently been widely adopted by the community of computer vision.

CNN are aimed at processing data coming in the form of numerous arrays, and a point in case is a grayscale image comprising three Two Dimension (2D) arrays that contain pixel intensities. Numerous data modalities are in the form of manifold arrays: Three Dimensions (3D) for volumetric or video images; One Dimension (1D) for sequences and signals, comprising language; and 2D for audio or images spectrograms. Behind CNN, there are four crucial ideas taking advantage of natural signals' properties: shared weights, the application of several layers, pooling and local connections [15-18].

The typical CNN's architecture as shown in Figure 1 is structured as a sequence of phases. The few initial stages comprise two kinds of layers: pooling layers and convolutional layers. In a convolutional layer, units are prearranged in feature maps, whereby every unit is linked to local patches in the preceding layer's feature maps via a set of weights known as a filter bank. The outcome of this local weighted sum is therefore passed via a non-linearity, for instance a Rectified Linear Units (ReLU) [19]. In a feature map, each and every unit shares a similar filter bank. In a layer, diverse feature maps employ dissimilar filter banks. The rationale behind this architecture is double. Foremost, in array data, for instance images, local groups of values are usually greatly interrelated, as a result forms a unique local motifs that could be detected with ease. Secondly, the images' local statistics as well as other signals are invariant to position. To simplify, in a situation where a motif could appear in the image's one part, it can appear at anyplace, as a result the idea of units at diverse locations that share identical weights and detect similar pattern in the array's diverse parts. Arithmetically, the filtering operation carried out by a feature map is a distinct convolution, and therefore the name.

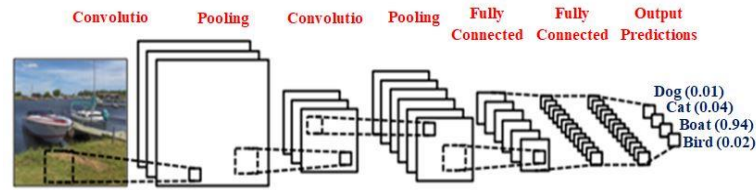


Figure 1. Effects of selecting different switching under dynamic condition

Despite the fact that the convolutional layer’s role is detection of local permutations of features from the preceding layer, the pooling layer’s role is integrating into one, the semantically comparable features. Because the relative positions of the features that form a motif could vary to some extent, consistent detection of the motif could be done by coarse-graining every feature’s position. In one feature map, a classic pooling unit calculates the supreme of a local patch of units.

For instance, it is clearly demonstrated in Figure 2 that a CNN could learn to distinguish edges from raw pixels in the first layer in image taxonomy, and then apply the edges in detecting simple shapes in the second layer, and then apply these shapes in preventing higher-level characteristics, for instance shapes of faces in higher layers. Therefore, the final layer is a classifier that applies these characteristics of high level.

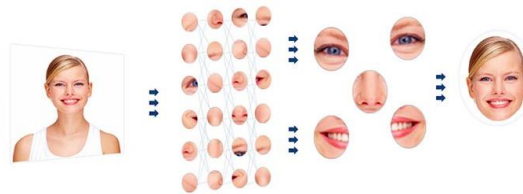


Figure 2. Eyeris’ deep learning based facial feature extraction using convolutional neural network [20]

2.3. Convolutional neural network architecture

As clearly demonstrated in Figure 1 [15, 21, 22], a CNN comprises an output and an input layer, together with numerous hidden layers. The hidden layers are either fully connected, convolutional or pooling.

2.3.1. Convolutional layer

Convolutional layers use a convolution procedure to the input, conveying the outcome to the subsequent layer. The convolution imitates a person’s response neuron to visual stimuli [23]. A convolutional layer comprises manifold neurons’ maps, referred to as feature filters or maps, with their size equivalent to the input image’s dimension. Two perceptions enable reduction of the number of model strictures: stricture sharing and local connectivity. Initially, not like in a network that is fully connected, every neuron in a feature map is simply linked to a local patch of neurons in the preceding layer, also known as receptive field. Secondly, each and every neuron in a certain feature map shares similar strictures. Therefore, each and every neuron in a feature map scans for equivalent feature in the preceding layer, though at diverse localities. Diverse feature maps could, for instance, perceive edges of diverse orientation in an image, or series motifs in a genomic series. The acquisition of neuron’s activity is via computation of a distinct convolution of its approachable field, whereby it is computation of the subjective sum of input neurons, as well as application of an activation function. Figure 3 clearly illustrate the Discreet Convolution.

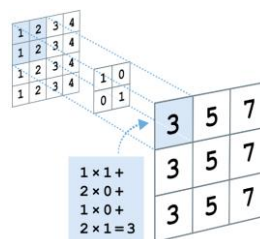


Figure 3. The CNN’s first layer is discreet convolution

2.3.2. Rectified linear units (ReLU) layers

It is a resolution to use an activation layer (or a layer that is nonlinear) instantaneously afterward following every convolutional layer [19]. This layer's purpose is presentation of nonlinearity to a system that mainly has just been calculating linear operations in the course of the convolutional layers. In the past, nonlinear functions like sigmoid and tanh were applied, though researchers discovered that ReLU layers are more enhanced in the way they operate since the network is capable of training a lot quicker (due to the computational effectiveness) devoid of making a substantial difference to the precision. In addition, it assist in relieving the challenge of vanishing gradient, which is the concern, whereby the network's lower layers train sluggishly since the gradient exponentially declines via the layers. To all of the values in the input volume, the ReLU layer make use of the function $f(x) = \max(0, x)$. All the negative activations are changed to zero by this layer in simple terms [16]. The nonlinear properties of the model and the overall network are increased by this layer devoid of interfering the conv layer's receptive fields. Figure 4 illustrates the ReLU activation function.

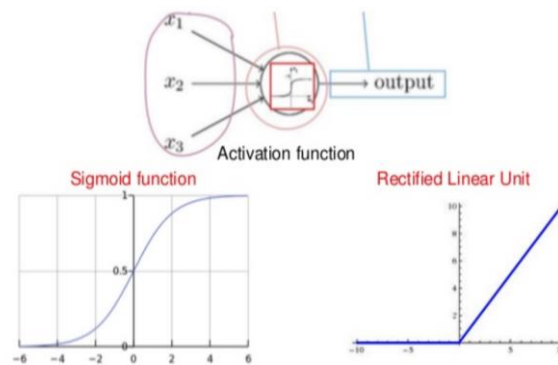


Figure 4. The relu activation function

2.3.3. Pooling layer

Pooling layer lessens their input's size and gives room for analysis of multi-scale. The most popular pooling operators are average-pooling and max-pooling. Within a small spatial block, these operators calculate the average or the maximum value. Max pooling operation with (2×2) filters is illustrated in Figure 5.

In a number of applications, the features' exact frequency and position is not pertinent for the last expectation, for instance ability to recognize objects in an image [24]. With the use of this assumption, the pooling layer outlines adjoining neurons through computation, for instance, the average or maximum over their activity, leading to feature activities' representation that is smoother. Through application of similar pooling operation to small image patches that are moved by pixel beyond one, the input image is efficiently down-sampled, and as a result reducing further the model parameters' number. The size of the output could be regulated by three hyper strictures which are the zero-padding, depth and stride.

- Stride: amount of pixels the filter jumps as they slide over the image.
- Depth: in order to the input image, it is basically the amount of filters that is employed. These filters are capable of detecting structure for instance, blobs, edges and corners.
- Zero-Padding: padding zeros around the input's borders for its size to be preserved.

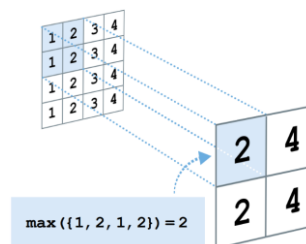


Figure 5. Max pooling operation with (2×2) filters

2.3.4. Fully-connected layer

Usually, a CNN comprises numerous pooling and convolutional layers, which gives room for increasingly learning abstract characteristics at snowballing scales including object parts, entire objects and small edges. It is possible for one or more completely connected layers to follow the final pooling layer. Model hyper-structures for instance the size of receptive fields, the number of feature maps and the number of convolutional layers refer to application-dependent and should be firmly chosen on a validation data set [23].

Fully-Connected layer attaches to each and every neuron of the preceding layer. Typically, fully connected layers are applied as the network’s final layer and carry out the taxonomy. Figure 6 depicts a CNN’s sample, illustrating all the three formerly revealed layers.

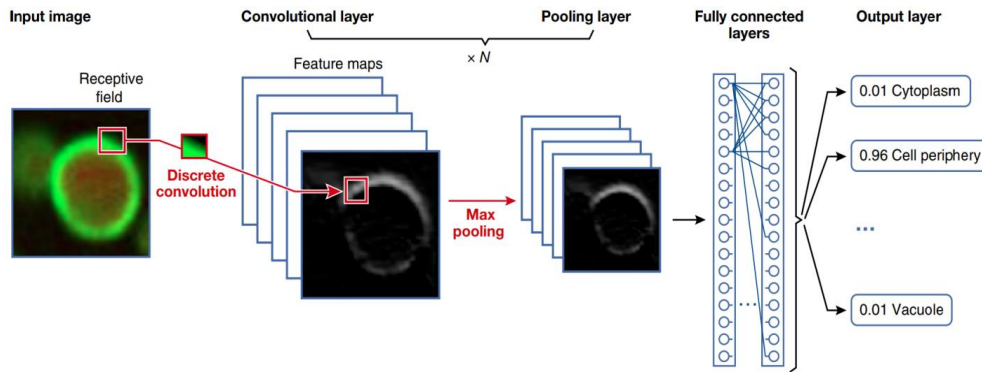


Figure 6. A Sample of CNN architecture

2.4. Slantlet transform (SLT)

The SLT refers to a orthogonal Discrete Wavelet Transform (DWT) with two zero moments, having enhanced time localization. SLT maintains typical filter bank implementation’s features having a scale dilation factor of two. Its foundation is not on reiterated filter bank such as DWT; as an alternative, diverse filters are applied for every scale. This paper recommends a new way of applying SLT in image retrieval through conversion of the image from spatial domain to convert domain with the intention of ordering them and selecting the most relevant and informative part of image for the retrieval model to be improved. In multi CNN, a novel technique of image retrieval CNN-SLT model applies merging. As image representation, the researchers have therefore applied the transform domain. We will as well associate this novel technique with image retrieval’s present techniques.

In a 2D SLT decomposition, there is usually an image that is divided into four parts, High-High (HH), Low-Low (LL), High-Low (HL) and Low-High (LH), as Figure 7 illustrates, where H and L represent the high and low frequency band, correspondingly. Each is carrying diverse image information. The low-frequency band component of the image, which is marked as LL, retains the inventive image information. On the contrary, the high - and medium -frequency bands, HH, LH and HL carry the information associated with the contour, edge, as well as the image’s other details. In the image, the important information is characterized by high coefficients. In the meantime, the small (insignificant) coefficients are deliberated as worthless information or noise. These small coefficients therefore ought to be ignored for the best outcomes in succeeding operations to be attained.

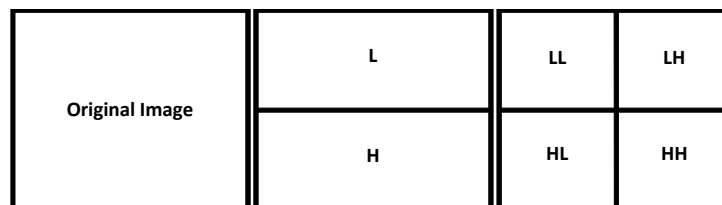


Figure 7. The conventional 2D SLT decomposition schemes for dividing an image

As a multi-resolution method, the SLT [17] is well-matched for piecewise linear data. The SLT refers to an orthogonal DWT, having enhanced time localization characteristics and two zeros moments. It is founded on the principle of designing diverse filters for diverse scales not like iterated filters method with the use of DWT. Formerly, SLT was applied in a various applications, for instance, compression, de-noising of various input images, estimation and fast algorithms. SLT is executed as a filter-bank having parallel structures and used in parallel processing, where diverse filters are configured for every scale rather than filter reiteration at personal level. According to Selesnick [17], the filters' coefficients are computed with the use of the SLT equations.

3. METHODS AND MATERIALS

Primarily, the utilized datasets are high pointed with the intention of evaluating the recommended framework. After that, the construction of Network architecture is described accompanied by the deep convolutional network's design. Subsequently, the merging input representation to the system is defined. To test the system, various experimental benchmarks are finally applied.

3.1. Data sets

For the proposed image retrieval technique to be validated, three standard datasets are applied. The initial two datasets are acquired through Wang V2.0 and Wang V1.0 – the Wang refers to the Corel database's subclass [18]. The third dataset is the Caltech 101, which consists of objects' pictures, which belong to 101 classifications - approximately 40 to 800 images for each category, though majority of categories have approximately 50 images. Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato established the database in September 2003. Every image's size is approximately 300×200 pixels. The most widely used and popular database in numerous latest studies is WANG V1.0 database [25]. It comprises one thousand images of ten classes. Each class consists of 100 images that nearly resemble each other. These classes are beach, buses, elephant's, horse, dinosaur, mountain, monuments, roses, food and Africa, illustrated in Figure 8.

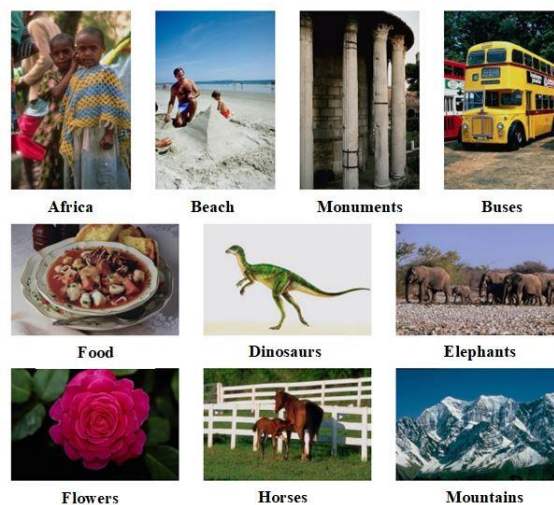


Figure 8. An example of image from each of the 10 classes of the WANG database as well as their class labels

WANG_10000 Database V2.0 As compared to its predecessor, this dataset is ten times larger, Wang V1.0. Majority of the images have low resolution, which adversely influences any image retrieval systems' performance. Figures 9 illustrate the images according to the categories that are established. Database V2.0 is however more comprehensive and challenging than database V1.0.

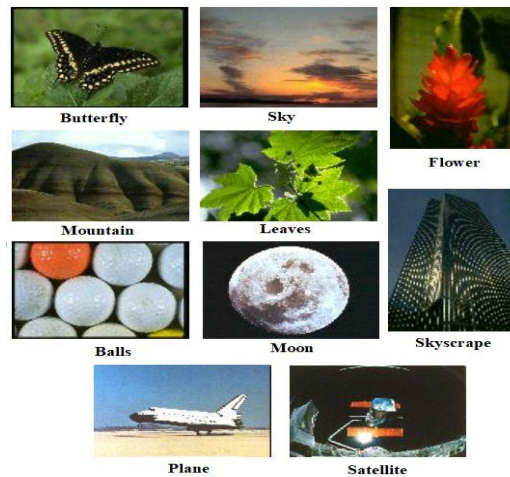


Figure 9. An example image from each of the 10 classes of the WANG V2.0 database together with their class labels

Caltech 101 refers to a data set of digital images generated in September 2003 [26]. It aimed at facilitating image recognition, classification, and computer vision. The dataset consists of a total of 9,146 images, divided between 101 distinct object classes (Pianos, Faces, Ants, Watches and others) as well as a background classification. For every category, it is approximately 40 to 800 images. Majority of categories have approximately 50 images. The size of every image is about 300 x 200 pixels. Figures 10 illustrate the images' samples. Description of used datasets as shown in Table 1.



Figure 10. An example image from each of the 10 classes of the caltech 101 database together with their class labels

Table 1. Description of used datasets

Dataset Name	Number of Images	Class	Number of Images In Each Class	Image Size
WANG 1	1000	10	100	256×384
WANG 2	10000	100	100	256×128
Caltech 101	9146	101	40-800	300×200

3.2. Benchmarking

The benchmarking refers to the most imperative steps that ought to be used in most of the image processing research with the intention of determining the developed techniques' reliability and effectiveness compared to the current one. Typically, the benchmarking is attained either by using similar dataset or using algorithms utilized in no different problem domain. The benchmarking is also performed with the use of the best and famous methods since in the literature, there was retrieval of image. A number of benchmarking techniques that use similar standard dataset are enlisted in Table 2.

Table 2. Existing methods for benchmarking

Dataset Name	Benchmarking
WANG 1	ElAlami [28], Lin [29], Wang [25]
WANG 2	Shrivastava & Tyagi [30]
Caltech 101	Bosch [31]

3.3. CNN network architecture

Following the collection of all the data, the convolutional architecture, having completely connected layers was deliberated as the avoidance architecture. In designing the recommended model configuration, the Krizhevsky principles [24] were applied, in which the source code might be seen [27]. The aforementioned generic design followed the configuration [24]. Figure 11 illustrates the suggested configuration, in which the images were passed via a stack of 4 convolutional (conv.) layers, where a (3×3) feature map size was applied for the conv. layers (this is a virtuous size for center, right/left, down/up). There were diverse number of pooling layers and conv. The initial two conv. layers applied 32 kernel filters; whereas the final two conv. layers applied 64 kernel filters. Max-pooling's two layers separated every convolution's step. This combination helped the models to mutually benefit from and upgraded the proposed configuration's performance, leading to the retrieval of image.

The conv. Layer was followed by the flattened layer (having architectures with diverse depths) and this assisted to transform into a vector the 2D matrix data. This gave room for the processing of the output with the completely-connected layers, referred to as dense layers. The initial completely-connected layer contained 256 nodes, whereas the second fully-connected layer was built using 128 nodes. The regularization layer applied dropouts and was configured to randomly omit 50 percent of the neurons with the intention of reducing overfitting. The final layer was designed by the softmax layer [14, 24, 23].

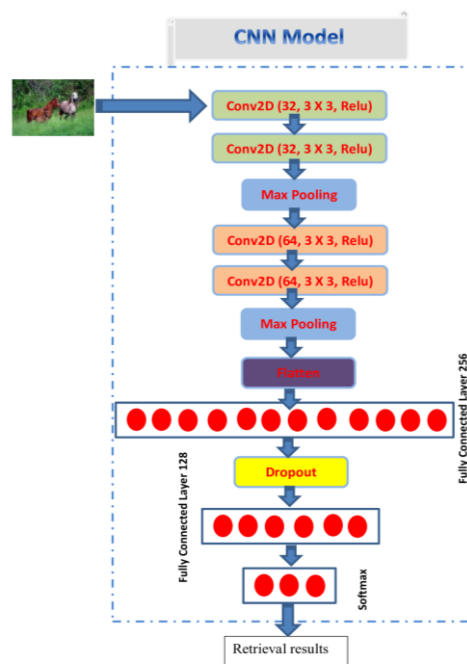


Figure 11. The recommended CNN configuration

a. CBIR based of merging and combine multi CNN:

CBIR refers to a powerful device for retrieval of image. During translating an image into mathematical variables, major components of the technique are the illustrations that are applied. A number of image presentation and descriptors are corresponding to other descriptors; therefore if the combination is applied, better results could be produced. This opinion implies that diverse descriptors could create diverse outcomes for retrieval of image, and integrate several image presentations to combine and merge multi convolutional neural network model, which could enhance the CBIR model's performance. The fundamental perception is accumulating in one CNN a variety of knowledge sources and features.

Deep learning does perform better than other machine learning algorithms as the empirical results suggest. Some has suggested that is because it loosely mimic the brain functions, multiple layers of neural networks stacked one after another like the classical brain model. However, until now there is no robust theoretical background for deep learning [8, 9, 11, 13], otherwise Deep Learning machines usually work better than traditional ML tools because they also learn the feature extraction part. Deep learning methods aim at learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower level features. Automatically learning features at multiple levels of abstraction allow a system to learn complex functions mapping the input to the output directly from data, without depending completely on human-crafted features [13, 14]. In image recognition, for example, the traditional setup is to extract handcrafted features and then feed a SVM. On the contrary, deep learning CNN schemes also optimize the features that are extracted which largely explains why they perform better. The most important difference between deep learning and traditional machine learning is its performance as the scale of data increases. When the data is small, deep learning algorithms don't perform that well. This is because deep learning algorithms need a large amount of data to understand it perfectly. On the other hand, traditional machine learning algorithms with their handcrafted rules prevail in this scenario.

By understanding the problem Statements which has been discussed earlier, the CBIR try to measure the similarities of images. Since the traditional CBIR systems still suffer from their poor retrieval accuracy and sensitivity, more works are still required to develop new approaches for the area of similarities of images measurement. Therefore, this research raises several challenges, such as improving the retrieval accuracy and enhancing the image descriptors and features extraction step.

The concepts of combining and merging multi convolutional neural network model have employed to develop a novel CNN-SLT-CBIR model according to SLT presentation combination with CNN model with the intention of enhancing and improving the recommended CBIR model's performance. The general CNN-SLT-CBIR model's framework based on combination of SLT with CNN is illustrated in Figure 12.

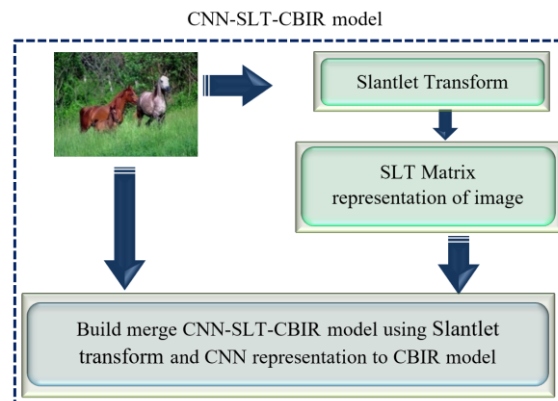


Figure 12. The general framework of CNN-SLT-CBIR model

Development of CNN-SLT-CBIR model with the use of SLT image presentation combined with CNN model with the intention of developing novel CBIR model. The researchers introduce a new CNN architecture that joins information from two presentation of an image into a compact and single image descriptor that offers even better retrieval of image.

4. EXPERIMENTAL OUTCOMES

In the Theano, the recommended code has been executed [28], which is a public deep learning software, according to the Keras [29]. In the neural networks, the weights were initialized according to the settings of Keras. In the deep network, all layers were simultaneously initialized using ADADELTA [30]. The entire network was trained with the use of the Dell Precision T1700 CPU system using the professional-grade NVIDIA-Quadro discrete graphics and a 14GB memory.

4.1. Evaluation measures

The performance of a retrieval system is assessed according to a number of criteria. Some of the commonly applied performance measures are average Accuracy and average precision. All these strictures are calculated with the use of precision and Accuracy values calculated for every query image. The retrieval precision is described as the fraction of the retrieved images that are definitely pertinent for the query. The Accuracy and precision measures are applied in assessing the image retrieval's performance recommended system. They produce:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \text{ and } precision = \frac{TP}{TP+FP}$$

Whereby, False Positive (FP), True Negative (TN), True Positive (TP) and False Negative (FN) have their own implication.

4.2. Result

The researcher in this study recommended new CBIR model according to the image and merging with presentation of SLT, which was a new CBIR. CNN-SLT-CBIR refers to a convolutional neural network, with a novel image representation, and is applied for retrieval of image. In addition, it is a deep learning system integrating the information of image. The recommended CNN-SLT-CBIR method was therefore associated with 5 other CBIR model existing in the literature.

This subsection introduces discussions and outcomes of performance measures of the recommended CNN-SLT-CBIR method. Concerning that, experiments' series were carried out with the use of three standard datasets, Caltech 101, Wang V1.0 and Wang V2.0. In accordance with standard practices in the retrieval of image domain, this study puts more emphasis on precision and accuracy with the intention of measuring the value of having high consistency and accuracy of the recommended technique.

For a fast grasp of the performance assessment, the outcomes are given in the form of tables and Figures, benchmark, and lastly discussions and analyses. With regard to that, the entire outcomes are given in Tables 3, 4 and 5 together with its boxplot diagram in Figure 13 were carried out with the use of three standard datasets, Caltech 101 Wang V2.0 and Wang V1.0.

Table 3. Accuracy and precision of retrieved images acquired by the CNN and CNN-SLT methods for the above ten classes of WANG V1.0 dataset

Classes	CNN		CNN-SLT	
	Accuracy	Precision	Accuracy	Precision
Africa	0.9331	0.716	0.9401	0.753
Beach	0.9235	0.6443	0.9335	0.6743
Monuments	0.9515	0.7837	0.9555	0.8057
Buses	0.9736	0.896	0.9826	0.918
Dinosaur	1	1	1	1
Elephant's	0.9532	0.8003	0.9622	0.8303
Roses	1	1	1	1
Horse	1	1	1	1
Mountains	0.9349	0.7117	0.9379	0.7387
Food	0.926	0.6613	0.93	0.6883
Average	0.95958	0.82133	0.96158	0.84833

Table 4. Accuracy and precision of retrieved images attained by the CNN and CNN-SLT techniques for the above ten classes of WANG V2.0 dataset

Classes	CNN		CNN-SLT	
	Accuracy	Precision	Accuracy	Precision
Accordion	1	1	1	1
Airplanes	1	1	1	1
Beaver	0.9919	0.9257	0.9979	0.9507
Binocular	0.9907	0.9263	0.9947	0.9473
Bonsai	0.994	0.957	1	1
Brontosaurus	0.9905	0.9067	0.9985	0.9307
Buddha	0.9928	0.922	0.9958	0.945
Camera	0.9952	0.9653	1	1
Cannon	0.9855	0.8603	0.9915	0.8873
Cellphone	1	1	1	1
Chair	0.9885	0.914	0.9975	0.951
Chandelier	0.9932	0.9317	0.9972	0.9547
Cougar_Body	0.9754	0.7513	0.9824	0.7693
Cougar_Face	1	1	1	1
Crab	0.9945	0.946	0.9965	0.962
Crayfish	0.9797	0.799	0.9807	0.824
Crocodile	0.9853	0.8523	0.9913	0.8763
Crocodile_Head	0.9838	0.8373	0.9908	0.8653
ad				
Cup	1	1	1	1
Dalmatian	0.9924	0.9367	1	1
Dollar_Bill	0.9931	0.9367	0.9961	0.9567
Dolphin	0.9877	0.8737	0.9897	0.8847
Average	0.995	0.9202	0.998	0.9482

Table 5. Accuracy and precision of retrieved images attained by the CNN and CNN-SLT methods for the above ten classes of caltech dataset

Classes	CNN		CNN-SLT	
	Accuracy	Precision	Accuracy	Precision
Accordion	1	1	1	1
Airplanes	1	1	1	1
Beaver	0.9919	0.9257	0.9979	0.9507
Binocular	0.9907	0.9263	0.9947	0.9473
Bonsai	0.994	0.957	1	1
Brontosaurus	0.9905	0.9067	0.9985	0.9307
Buddha	0.9928	0.922	0.9958	0.945
Camera	0.9952	0.9653	1	1
Cannon	0.9855	0.8603	0.9915	0.8873
Cellphone	1	1	1	1
Chair	0.9885	0.914	0.9975	0.951
Chandelier	0.9932	0.9317	0.9972	0.9547
Cougar_Body	0.9754	0.7513	0.9824	0.7693
Cougar_Face	1	1	1	1
Crab	0.9945	0.946	0.9965	0.962
Crayfish	0.9797	0.799	0.9807	0.824
Crocodile	0.9853	0.8523	0.9913	0.8763
Crocodile_Head	0.9838	0.8373	0.9908	0.8653
Cup	1	1	1	1
Dalmatian	0.9924	0.9367	1	1
Dollar_Bill	0.9931	0.9367	0.9961	0.9567
Dolphin	0.9877	0.8737	0.9897	0.8847
Average	0.995	0.9202	0.998	0.9482

In this research, according to the experiment, it is witnessed that a number of the retrieved images are discrete from their query images. This is anticipated since image texture computes roughness or smoothness of the image in contradiction of physical appearance, defining features or traits, which is the first thing someone realizes when he looks at the image, (which could be shape, object or color).

Beginning with outcomes of the models CNN-SLT and CNN are illustrated in Tables 3, 4 and 5. Figures 13 show comparison of precision values and the precision correspondingly; for the CNN-SLT and CNN models that use boxplot diagram for the three datasets applied in this research.

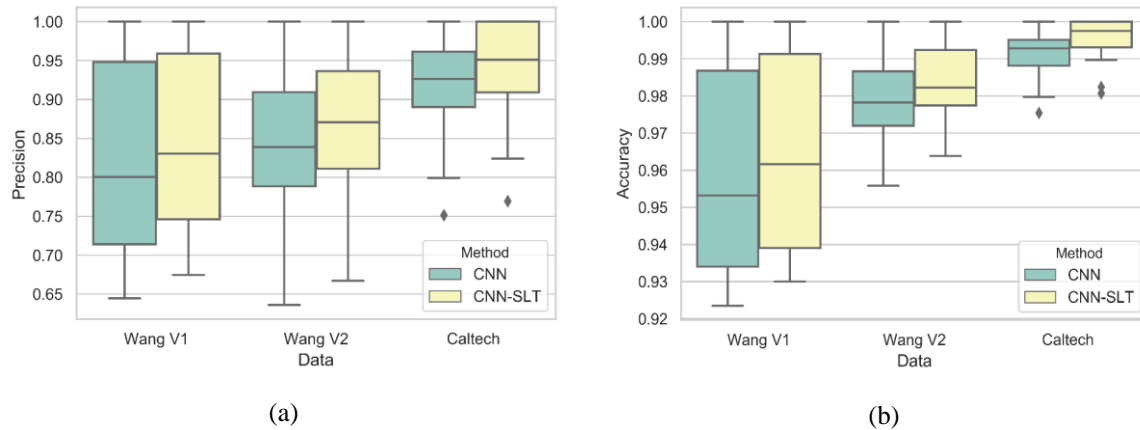


Figure 13. Comparison of the values for the CNN and CNN-SLT models using boxplot diagram for the three datasets used in this research: (a) Precision, and (b) Accuracy

Generally, Figure 13 and Tables 3, 4 and 5 clearly illustrate that the accuracy attained by the CNN-SLT is far better as compared to the CNN, as high accuracy is maintained whereby; the Caltech dataset scored impeccable outcomes in both precision and accuracy. The substantial enhancement of the precision is primarily attributed to the fine texture features revealed by the Slantlet convert joined with CNN deep learning technique.

Figure 13 described the Boxplot chart results after comparing the Precision and Accuracy values of the CNN and CNN-SLT Models using the Three Datasets Used in this Research. The CNN algorithm showed a higher variance, compared to the CNN-SLT for the three datasets. This diversity in all Precision and Accuracy values was especially seen in the WANG V1 data that displayed a variance of 7.5×10^{-3} . Furthermore, the CNN-SLT algorithm showed a high Precision value of 0.99. The results showed a small p-value of 8.2×10^{-2} , which highlighted a significant difference between both models.

Associated to the preceding study several authors for instance, Wang and Elhami design system according to special domain and feature extraction, as well as combining it with frequency domain methods to enrich the results of retrieval of image and attain high average of precision. In our recommended system however, that based on deep learning CNN, joined with Slantlet transfer to enrich the extracted feature from color channel attained highest precision. The precision of the proposed system as well as other system is illustrated in Table 6 below.

Table 6. The proposed method versus the State-of-the-Art techniques in terms of precision

Dataset Name	CNN-SLT	CNN	Shrivastava & Tyagi [30]	ElAlami [28]	Lin [29]	Wang [25]	Bosch [31]
WANG V1.0	0.84833	0.82133	n/a	0.739	0.727	0.59	n/a
WANG V2.0	0.8733	0.8413	0.769	n/a	n/a	n/a	n/a
Caltech 1	0.9482	0.9202	n/a	n/a	n/a	n/a	0.627

The outcomes of the benchmark, given in Table 6 above, have evidently illustrated that the recommended technique outdid the others with precision gaps fluctuating between 25.83% and 10.93 percent for Wang V1.0 dataset. For Wang V2.0 however, the recommended technique has up surged by 10.43 percent. On the other hand, for Caltech 101 dataset the recommended technique has substantially outperformed Bosch work with larger percentage, 32.12 percent. This impressive performance clearly revealed that the CNN-SLT method operated well for all the three datasets, whereby the CNN (preceding phase) and the CNN-SLT (subsequent phase) worked together amicably.

5. CONCLUSION

This part resolves the accomplishments of the recommended research technique in accomplishing the objectives and handling the research gaps. We assessed deep learning convolutional networks for image retrieval in this work. For the retrieval accuracy, it was established that the representation multi image presentation is not rewarding. The precision's substantial improvement is primarily accredited to the fine

texture characteristics unveiled by the Slantlet transform joined with CNN deep learning technique. There are still lots of complicated matters that have not been resolved, for instance, low precision and false positive (false matching). Against the backdrop therefore, this study has recommended a CNN methodology. In enhancing the CNN model, the former applied Slantlet transform-based characteristics. The two worked together amicably in order to achieve high accuracy and high precision. In addition, this study's findings showed that texture characteristics are trustworthy and proficient in production of outstanding outcomes and not vulnerable to low resolution as illustrated in the outcomes of Wang V2.0 dataset. It is well verified that spatial-based texture only could not attain the desired accuracy. This study has therefore demonstrated that the perfect mate is the SLT-based texture feature.

REFERENCES

- [1] Fuhui Long, Hongjiang Zhang and David Dagan Feng, "Fundamentals of Content-Based Image Retrieval," *Multimed. Inf. Retr. Manag. Technol. Fundam.*, pp. 1-6, 2003.
- [2] I. Sumana and M. Islam, "Content based image retrieval using curvelet transform," *Signal Process.* 2008, pp. 11-16, 2008.
- [3] S. Bhagavathy and K. Chhabra, "A Wavelet-based Image Retrieval System," *Comput. Eng.*, pp. 1-7, 2007.
- [4] N. Suematsu, Y. Ishida, A. Hayashi, T. Kanbara, "Region-Based Image Retrieval using Wavelet Transform," 2002.
- [5] Huang, "A fast method for textural analysis of DCT-based image," *J. Inf. Sci. Eng.*, vol. 21, no. 1, pp. 181-194, 2005.
- [6] L. C. L. Chen, G. L. G. Lu, and D. Z. D. Zhang, "Effects of different Gabor filters parameters on image retrieval by texture," *10th Int. Multimed. Model. Conf., Proceedings.*, pp. 1-5, 2004.
- [7] J. Wang, L. Jia, and G. Wiederhold, "SIMPLiCity: Semantics-sensitive Integrated Matching for Picture Libraries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 9, pp. 947-963, 2001.
- [8] Y. LeCun, B. Yoshua, and H. Geoffrey, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
- [9] Y. Bengio, "Learning Deep Architectures for AI," *Journal Foundations and Trends in Machine Learning*, vol. 2, no. 1, 2009.
- [10] A. Gupta, H. Wang, and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering," *2015 IEEE Int. Conf. Bioinforma. Biomed.*, pp. 1328-1335, 2015.
- [11] L. a. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *arXiv Prepr.*, pp. 1-16, 2015.
- [12] H. Wang, A. Meghawati, L.-P. Morency, and E. P. Xing, "Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis," vol. 1, 2016.
- [13] C. Affonso, A. L. D. Rossi, F. H. A. Vieira, and A. C. P. D. L. F. de Carvalho, "Deep learning for biological image classification," *Expert Syst. Appl.*, vol. 85, pp. 114-122, 2017.
- [14] H. Wang and B. Raj, "On the Origin of Deep Learning," *Arxiv*, pp. 1-72, 2017.
- [15] S. J. Lee and S. W. Kim, "Localization of the slab information in factory scenes using deep convolutional neural networks," *Expert Syst. Appl.*, vol. 77, pp. 34-43, 2017.
- [16] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving Deep Neural Networks for {LVCSR} Using Rectified Linear Units and Dropout," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 8609-8613, 2013.
- [17] I. W. Selesnick, "The Slantlet Transform," *IEEE Trans. Signal Process.*, vol. 47, no. 5, pp. 1304-1313, 1999.
- [18] X. Wang, Y. Yu, and H. Yang, "Computer Standards & Interfaces An effective image retrieval scheme using color, texture and shape features," *Comput. Stand. Interfaces*, vol. 33, no. 1, pp. 59-68, 2011.
- [19] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proc. 27th Int. Conf. Mach. Learn.*, no. 3, pp. 807-814, 2010.
- [20] Chen Yu-Hsin, Krishna Tushar, Emer Joel and Sze Vivienne, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," in *IEEE International Solid-State Circuits Conference, ISSCC 2016, Digest of Technical Papers*, pp. 262-263, 2016.
- [21] M. E. Elalami, "A novel image retrieval model based on the most relevant features," *Knowledge-Based Syst.*, vol. 24, no. 1, pp. 23-32, 2011.
- [22] B. L. Tseng, F. Hills, and N. Y. Us, "(12) United States Patent," 2009.
- [23] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Mol. Syst. Biol.*, vol. 12, no. 7, pp. 1-16, 2016.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1-9, 2012.
- [25] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1-60, 2008.
- [26] Li Fe-Fei, Fergus, and Perona, "A Bayesian approach to unsupervised one-shot learning of object categories," *Proc. Ninth IEEE Int. Conf. Comput. Vis.*, pp. 1134-1141 vol.2, 2003.
- [27] V. GUPTA, "Image Classification using Convolutional Neural Networks in Keras," 2017.
- [28] F. Bastien, *et al.*, "Theano: new features and speed improvements," pp. 1-10, 2012.
- [29] F. Chollet, "Keras Documentation," *Keras.Io*, 2015.
- [30] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," 2012.
- [31] A. Bosch, A. Zisserman, X. Mu, and X. Munoz, "Image Classification Using Random Forests and Ferns," *Comput. Vis. (ICCV), IEEE 11th Int. Conf.*, pp. 1-8, 2007.