A Novel Re-Ranking Based Image Retrieval Using Convolutional Neural Network and Support Vector Machine in Cloud Environment

K. Nithya[1*], V. Rajamani[2]
[1]Research Scholar, Anna University, Chennai, 600025,India
[2]Professor, Department of Electronics and Communication Engineering, VeltechMultitechDr. Rangarajan Dr.Sakunthala Engineering College, Chennai, 600062, India
[*]Corresponding Author: K. Nithya. Email: nithyakmaha@gmail.com

Article Info
Page Number: 665 – 675
Publication Issue:
Vol. 71 No. 3 (2022)

Abstract:
Content Based Image Retrieval (CBIR) is the prominent research area now-a-days. In spite of massive researches exist in the field of CBIR, the retrieval of relevant images from the large set of image database in the cloud, remains more challenging task. The real problem lies in the feature representation of images and semantic gap exists between the image representation as a pixel in the machine and the concepts viewed by the human. Among all the techniques, the Convolutional Neural Network (CNN) stands top in bridging this gap. In this paper, a deep CNN ResNet50 model is used to represent the features of the image. During training the CNN, the Support Vector Machine (SVM) is used in the place of Softmax layer in the CNN to predict the class label of the training image. To further improve the accuracy of the proposed model, a Novel re-ranking method based on K-Nearest Neighbourhood (KNN) is used. The proposed system exhibits better performance and accuracy than the existing systems.

Keywords: Content Based Image retrieval; Feature representation; Semantic gap; Convolutional Neural Networks; Support Vector Machine and KNN.

Article History
Article Received: *12 January 2022*
Revised: *25 February 2022*
Accepted: *20 April 2022*
Publication: *09 June 2022*

1. Introduction

The query image and the relevant images in the data base are compared using features present in both the images. These comparisons are done in 3 ways[1]. First method is to compare shape, texture andcolor. These features are called as low level features which are extracted directlyfrom the images and are compared. But these features are not able to explain the semantic meaning of the images. The second method uses the visual vocabulary spaces using feature encoding techniques. This method is capable of handling illumination variation, image truncation and transformation thus decreases the semantic gap. The third method uses the high level features which are directly obtained from the CNN. This method further narrows down the semantic gap in a remarkable manner[2]. Content Based Image retrieval (CBIR) is used to extract the stored images by queuing with an image instead of text based image retrieval from a large database. The real challenge in CBIR exists in reducing the semantic gap between the internal representation of the images as pixel and high level concepts viewed by human as an image.

In recent years, Machine Learning (ML) is prominent technologies that can be used with CBIR. It is used to represent the higher level features from the raw images directly. The ML can be further deeply evolved as Deep Learning which tries to synchronize with the human brain. With the introduction of CNN in machine learning, image feature extraction and classification is made much easier. The general architecture of the CNN is shown in Fig. 1, it consists of one input layer and more Convolutional layers followed by pooling layers and one or more Fully Connected (FC) layers. The output of the FC layer can be given to SoftMax layer for classification. The Convolutional layer is used to convolve the input from the previous layer with the number of filters (normally called as kernels) and apply any one of the activation functions. The changes in the parameters of one layer will be propagated to subsequent layers will slow down the learning rate. Here, Batch normalization is used to improve the learning rates. Either Max or Average Pooling is used in the pooling layers. The Last layer of the CNN is Fully Connected Layer. This layer can be connected to SoftMax layer for classification in which softmax pooling is used.

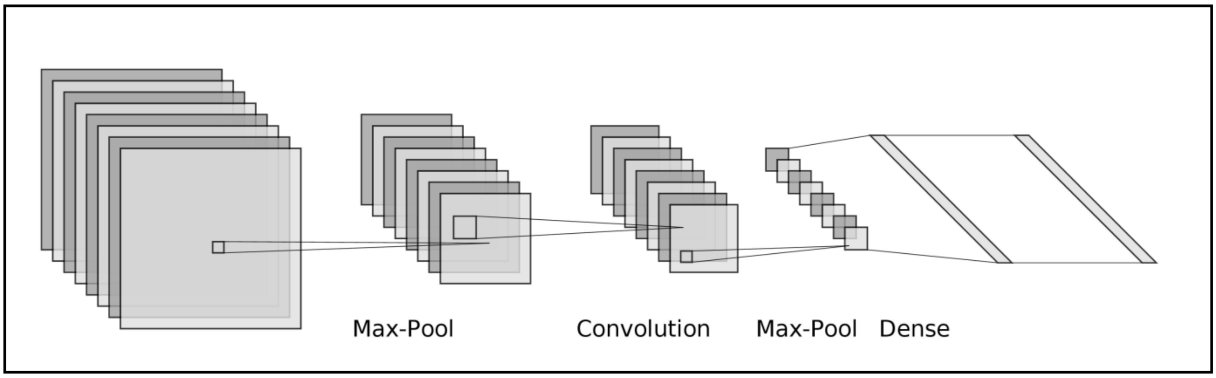

Figure 1: General architecture of a CNN

The Support Vector Machine is a popular classifier which provides better classification result even for the noisy input. It takes longer time for training, but provides accurate classification than the other classifiers. In the retrieval process, usually the Euclidean distance is used for calculating the relevancy between the query image and the retrieved images. The relevancy also can be improved further with the introduction of re-ranking mechanism which can be applied over the retrieved images. The remaining section of the paper is structured as follows. The related works are reviewed in section 2. The proposed model is elaborated in section 3 and the results are discussed and compared in section 4. The section 5 elaborates conclusion and future scope.

2 Related works

This section brief about the related works which were carried out in the above domains explained in the previous section. Last decades, Content based image retrieval (CBIR) techniques is used to retrieve relevant images by using visual contents. The CBIR is becoming a challenging job. As there is a large gap between the machine representations of an image a pixels and the concept that are viewed by the human. Numerous techniques for CBIR has been categorized and evaluated in [3]. Traditional system uses only one feature for retrieval. The proposed system[4], combined color, texture and shape.The machine learning has been a prominent field of study to bridge the semantic gap between the feature representation in computer (machine) as pixels and concept viewed by [5] attempted to address how deep learning techniques are used for feature representation and they examined the deep learning method under variable settings. [6] discussed deep learning methods are also used in the field of big data analysis. The various practical applications of deep learning are discussed with appropriate architectures.The CNN can be used for classification and retrieval in [7]. [8] proposed a new architecture based on unsupervised object recognition is presented to address the problems associated with backprobagation neural networks, a combination of pretrained CNN model and Hopfield Network based Associative memory bank are used. The back propagation is eradicated by Associative memory bank.

Deep learning based CNN is used to detect the object. Various deep learning frameworks are services so far has been discussed in [9]. Also in various domains these techniques can be applied.[10] presented a deep learning frame work based on CNN in combination with a new classifiers called support vector Machine(SVM). In this method, the SVM classifier is used instead of soft max classifier. [11] proposed a model retrieving method to improve the efficient of the learning. They used deep CNN model. From the Convolutional layers the features are obtained and they used maxpooling for reducing the number of parameters. They concentrated on Convolutional layers instead of fully connected layers to improve the retrieval time by producing smaller number of descriptors.

The input images for the previous CNNs are constrained with fixed size. This constraint reduced the accuracy of different sized images with other size. [12], introduced the concept of SPP-net to produce the fixed length. Using this method, the features maps are computed only once, thus avoiding the repeated computation of features.For generating the augmented images, [13] used Generative adversial network, whereas [14], generated sequences that specified the strategies of the data augmentation by using GANs. CNN model scaling is studied in [15] and identified that even better performance is achieved by carefully balancing of network depth, width and resolution. The family of models called EfficientNets are obtained which exhibited better efficiency and accuracy.

A new framework using CNN with hash-coding is proposed by [16]. The images present within the same class have same features are learnt by the contrastive loss function and weight sharing. The dimensionality of the feature vector by hash mapping is represented. [17], [18] and [19] proposed a weighted distance model to retrieve images from the IR database. Here a pre-trained CNN is fine tuned with some sample training data and is used for feature extraction. A new re-ranking mechanism was proposed in [20]. In this, the relevant images are calculated with the help of 2 image to class distances. One is image to training class distance and the second is query image to query class distances. They take the average of these 2 distances for relevancy calculation.

3. Proposed system

This section explains about the architecture of the proposed system. The entire process is carried out in 2 phases. (i) Training phase and (ii) Retrieval phase. Here the ResNet50 a pretrained CNN is used. Fig. 2 and Fig. 3 portray the proposed system architecture in training and retrieval phases.

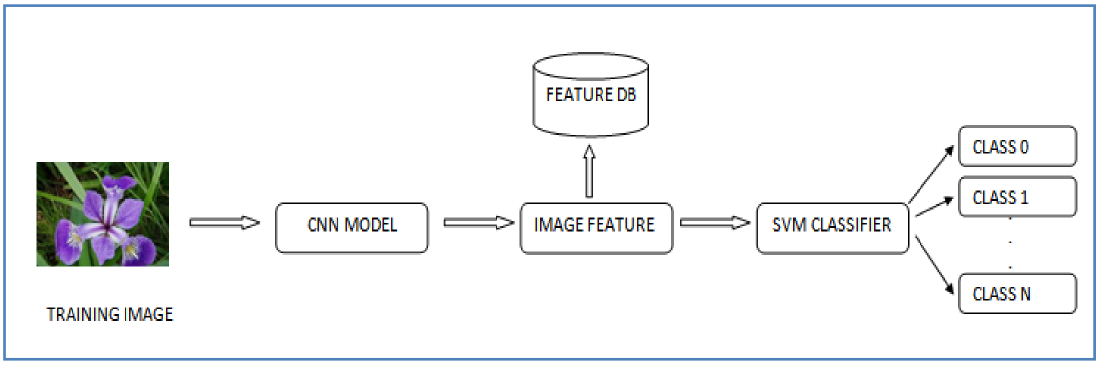

Figure 2: Proposed System architecture - During training phase

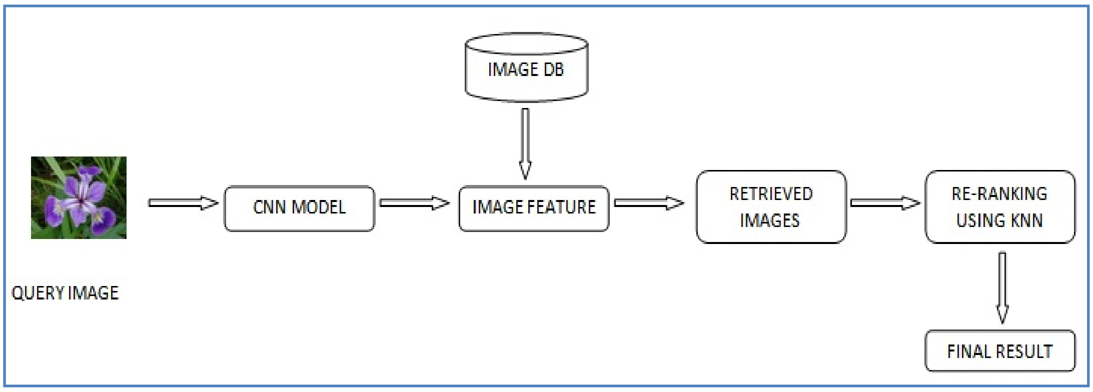

Figure 3: Proposed System architecture - During Retrieval phase.

3.1 Training phase

The data set in the entire collection is partitioned into 2 independent sets called training and testing set. The model is given training data to learn. To improve the accuracy of the model, it requires thorough training. The ResNet50, a pretrained model is used as a feature extractor in this phase which achieved better results in ILSVRC-2015 competition with top-5 error rate of 3.57%.

3.1.1 Network Setup

Overall trainable parameters used in the model are 23 million. Batch normalization is used in every Convolutional layer’s output to improve the learning. Learning rate of 0.01 is used. Each neuron in the network will produce the output as per the ReLU activation function that is modelled in Equ., 1.

$$g(x) = \max(0,x\ x). \qquad (1)$$

3.1.2 Performance Tuning

As the network uses large number of parameters, there is more possibilities for overfitting. In order to reduce this, data augmentation is used. In this, the original images are transformed to produce other images and these transformed images need not be stored in the database. To do this, in the RGB values of the image, the Principal Component Analysis is used and each pixels in the images are transformed as per Equ.2.

$[q_1,q_2,q_3][\ \alpha_1\ \lambda_1,\ \alpha_2\ \lambda_2,\ \alpha_3\ \lambda_3]^T$ (2)

Where q_i is ith Eigen vector and λ_i is the ith Eigen value of RGB pixel. α is the random variable. As per "dropout" technique, the neuron's output is set to "zero" with the probability of 0.5. This will make the neuron to learn independently of other neurons. The second and third layer in the network has dropout technique introduced.

3.1.3 The convolution operation

For a given query image I, Filter F and the number of channels C, the output of each neuron in the forward pass is performed as follows.

$$(I * F) = \sum_{x=0}^{f1-1} \sum_{y=0}^{f2-1} \sum_{c=1}^{cn} I_{i+x,j+y,c} . F_{x,y,c} + b \qquad (3)$$

During back probagation, the mean squared error is calculated.

$$E = ½ \sum (ot - oe)^2 \qquad (4)$$

Where o_t and o_e are targeted and expected outcomes of the network. Based on this calculated error value, the weights and δs must be updated as per the Equ.5 and Equ.6 respectively.

$$\frac{\partial E}{\partial w^m} = \sum\sum \frac{\partial E}{\partial o^m} * \frac{\partial o^m}{\partial w^m} \qquad (5)$$

$$\frac{\partial E}{\partial o^m} = \sum\sum \frac{\partial E}{\partial o^{m+1}} * \frac{\partial o^{m+1}}{\partial o^m} \qquad (6)$$

3.1.4 The retrieval process

The last fully connected layer outputs the feature vector of the training image. These features are stored in the feature database and passed through Support Vector Machine for classification instead of softmax layer. As the SVM is well known for its accurate classification, it is combined with CNN in the proposed system.

The SVM finds a hyperplane which has higher margin to classify the images accurately. The hyperplane is found by the following.

A separating hyperplane can be written as,

$W.X + b = 0$ (7)

Where W is the weight vector, X is the training vector and b is the bias value. For example, if X has 2 attributes, say x_1 and x_2, then, the Equ.7 can be rewritten as,

$w_1x_1+w_2x_2+b=0$ (8)

2 hyperplanes say H_1 and H_2 that can be used to separate the input data into 2 classes, can be written as,

$H_1 : w_1x_1+w_2x_2+b >=1$ (9)

$H_2 : w_1x_1+w_2x_2+b <=1$ (10)

Any data falls on or above H_1 can be classified in to class 1 and, data falls on or below H_2, can be classified into class 2. As the data falls on the hyperplanes H_1 and H_2 support the classification process, these data are called as support vectors.

Now, the accuracy of the SVM prediction lies in the finding of a highest margin between these 2 hyperplanes. The maximal margin is calculated as follows,

$$\frac{2}{||W||}.$$

(11)

As per the Equ.11, the margin can be maximized by fixing the value of W as minimum as possible.

3.1.5. Retrieval phase

The query image is given input to the ResNet50 model. The features of the image are extracted from the last fully connected layer. According to the features of the query image, the relevant images from the image database are retrieved. The relevancy is calculated by the Euclidean distance between the imput image and the output image by the Equ. 12.

$$\mathrm{Ed}(I_q, I_r) = \sqrt{\sum_{i=1}^{m} fi(\mathrm{Iq}) - fi(Ir)} \qquad (12)$$

Where Iq and Ir are the query image and retrieved image with m features respectively. $fi(I_q)$ and $fi(I_r)$ are the i^{th} feature vector of the input image and output image respectively.

In order to retrieve the relevant images, the Euclidean distance is made as minimum as possible. The less is the distance the more is the relevancy. But, only considering the image to image distance is not the optimum solution for finding the relevancy. In order to improve the relevancy, a novel re-ranking mechanism is proposed based on KNN algorithm.

3.1.6 The steps involved in the novel re-ranking mechanism are as follows.

A. Consider the retrieved image dataset as DS_R, with "n" images returned as a result of query image. Calculate the KNN set for the input image, by calculating distance between input image to all the images in the retrieved dataset and sort the images based on the distance in ascending order. Now retrieve "k" images from the ordered data set as KNN set. This can be written as,

$$\mathrm{KNN(q,r)} = \{\hat{x} \mid D(x,\grave{x}) \le d(k) \ \land \ \hat{x} \subset DSR\} \qquad (13)$$

Where d(k) is the k^{th} element's distance and the elements in the DSR are ordered according to the distance in ascending order. Now the query class formed by query image concatenated with the KNN of the query image.

Now the modified distance between the input image and the output image can be calculated by combining the Euclidean distance and the average distance of the output image and knn of the query image as follows.

$$\mathrm{MAD(q,r)} = \frac{1}{2}(\ (D(q,r)\ +\ \frac{1}{k}\sum_{r\subset KNN(q)} D(n,r)) \qquad (14)$$

The modified distance calculated from equ.,14 can be used to re-rank the retrieval database. The orders of the images are rearranged based on the modified distance and returned as a new retrieval database.

The results show that the proposed system returns more relevant images than the existing systems and the accuracy of the overall retrieval is improved.

4. Experimental Results

The proposed system is tested on CalTech256 datasets. Various parameters of the proposed system were analysed under different conditions and the comparisons were made with the existing systems. The experiment results show that our system exhibited better performance over the existing systems.

Evaluation metrics:

Precision (P)

P is defined as the ratio of relevant images retrieved to total images retrieved.

$$P = \frac{RIR}{TIR}$$

(15)

Where RIR is the number of relevant images retrieved and TIR is the total number of images retrieved.

mean Average Precision(mAP)

mAP is defined as the ratio between the average precision for a given query and the total number of queries. This is used to assess the quality of the retrieval algorithm. mAP is calculated as follows.

$$mAP = \frac{1}{N}\sum_{k=0}^{n} AP(n)$$

Where n denotes number of queries and the AP is the average precision value.

Recall (R)

R is the fraction of relevant documents that are extracted successfully.

Table 1: The effect of number of epochs on Accuracy

No. of Epochs	10	20	30	40	50	60	70	80	90	100
proposed	50.09	91.13	96.48	97.79	96.68	98.56	98.01	97.93	94.12	93.26
Rui et al. [2020]	37.89	92.55	98.49	98.11	96.89	94.62	93.15	93.39	93.88	94.37

The result of the number of epochs with accuracy is shown in Table 1. This shows that the accuracy is improved linearly by increasing the number of epochs in the network and at one point, the accuracy becomes decreasing, and this effect is shown in Figure 4. The accuracy of our method is compared with the existing system and the proposed system outperforms the existing system at the point where the number of epochs is 60.

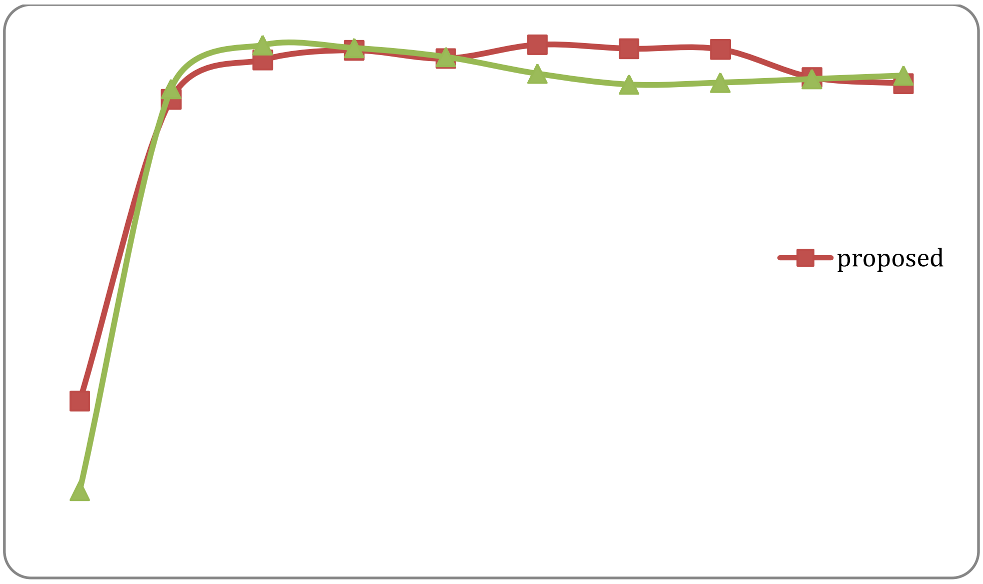

Figure 4: The accuracy of the system over number of Epochs.

Table 2: Precision and recall value of the proposed system.

Class	Number of images in the DB	Number of images retrieved	Number of relevant images	Precision	Recall
Class 1	50	48	47	0.98	0.94
Class 2	50	48	46	0.96	0.92
Class 3	50	49	47	0.96	0.94
Class 4	50	48	47	0.98	0.94

The table 2 shows, the performance of the proposed system with precision and recall value. The proposed system is trained with 50 images in each class. For a given query, our model retrieved the relevant images with the average precision of 0.97 and average recall of 0.94. This is because the relevancy is calculated by taking the average of Euclidean distance and the k-relevant images retrieved which gives more accurate matching. From the above value, it is clearly inferred that our proposed system shows high accuracy.

Our system is compared with the existing systems and exhibited better performance than the existing systems. Value of mean Average Precision (mAP) with the varying number of training images is shown in the figure 5. It shows that our system shows better mAP than the existing system from which it clearly understood that the proposed system consistently gives higher accuracy.

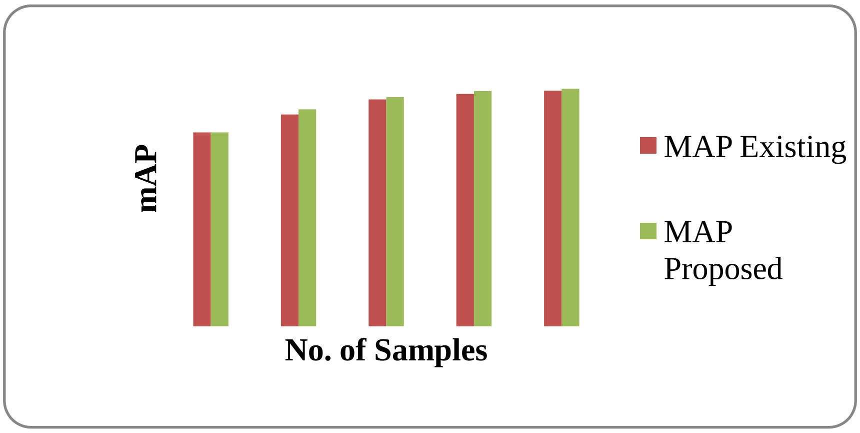

Figure 5: Number of training images compared with mean average precision (mAP).

The error rates of the test set with and without augmentation are shown in the table 3. Our proposed method is compared with various methods that used the augmentation. In this comparison, our results show that with augmented data, the error rate is reduced. The LSTM provided 1.6 reductions in the error rate whereas the MF provided 2.1 error reduction rates. As our proposed method used the average of Euclidean distance of class average and the k most relevant images, the error rate has been reduced significantly. With the results, the proposed method outperforms all the other existing system with the error rate reduction of 2.4. Thus our method exhibits better accuracy.

Table 3: Comparison of error rate reduction

Method	With original data (%)	With augmented data (%)	Reduction in error rate (%)
LSTM	7.7	6.0	1.6
MF	7.7	5.6	2.1
Proposed method	6.6	4.2	2.0

The real challenge lies in selecting k value in KNN algorithm for re-ranking. The table 4 shows some k value effects on the precision.

Table 4: K-value in KNN with Precision

KNN-k values	Precision-Existing	Precision-Proposed
5	98.99	99.5
10	98.87	99.1
50	98.7	98.9
100	98.66	98.81

The small k value gives the higher precision because it will only retrieve the most relevant images but, it leaves some of the relevant images from the data set. If the value of k increases, the precision value decreases but it includes all the relevant images in its re-ranked dataset. So, selecting k-value is purely user specific as well as application dependent.

Table 5: Comparisons of retrieval time with existing methods

Methods	Retrieval time(in sec)
Color based Method	20.85
Texture based Method	22.97
Adaptive Method	40.54
Wavelet optimization	34.44
Differential Learning	30.27
Proposed Method	19.03

The table 5 portrays the retrieval time of the proposed system compared with the other existing models. The color based and texture based methods are faster than the other methods like Adaptive and wavelet methods. Our proposed system exhibited faster retrieval time thus provided better performance.

5. Conclusion

The introduction of CNN made the Content Based Image Retrieval very effective and attracted many researchers towards this domain. In the proposed method, the CNN (ResNet50) is used to retrieve the features of the images and SVM is used for better classification of images while training. Next, a novel re-ranking mechanism based on KNN is also proposed to improve the precision of the system. From the experimental result, it is found that our system exhibited higher accuracy in classification and better precision value in retrieval. The future scope of the proposed system can be extended by adding Pre-processing to the input images or query images for better classification or retrieval. Instead of SVM classifier, some other classifier can be used. The different techniques in data augmentation can be added to improve further improve the accuracy. It can be better utilized for other applications if proper introduction of transfer learning. The retrieval process may be added with security according to the need of applications.

References

[1] Zhou, Wen-gang et al. “Recent Advance in Content-based Image Retrieval: A Literature Survey.” *ArXiv* abs/1706.06064 (2017): n. pag.

[2] L. Zheng, Y. Yang, and Q. Tian, “SIFT meets CNN: A decade survey of instance retrieval,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1224–1244, 2018.

[3] Zhou W, Newsam, S Li C and Shao Z, “PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” J. Photogramm. Remote Sens., vol. 145, pp. 197–209, 2018.

[4] Duan, Guoyong et al. “Content-Based Image Retrieval Research.” *Physics Procedia* 22, pp. 471-477, 2017.

[5] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu *et al.*, “Deep Learning for Content-Based Image Retrieval: A Comprehensive Study,” In Proceedings of the 22nd ACM international conference on Multimedia (MM '14). Association for Computing Machinery, New York, NY, USA, pp. 157–166, 2014.

[6] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu*et al.*, ”A survey of deep neural network architectures and their applications”, Neurocomputing, vol. 234, pp. 11-26, 2017.

[7] Z. Rian, V. Christanti and J. Hendryli, “Content-Based Image Retrieval using Convolutional Neural Networks,” IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, pp. 1-7, 2019.

[8] Q. Liu andS. Mukhopadhyay “Unsupervised Learning using Pretrained CNN and Associative Memory Bank,” In Proceedings of IEEE International Joint Conference on Neural NetworksIJCNN, Rio de Janeiro, Brazil, pp. 1-8, 2018.

[9] Pathak, Ajeet Ram et al. “Application of Deep Learning for Object Detection.” *Procedia Computer Science* 132 (2018): 1706-1717.

[10] O. Mohamed,E.A.Khalid, O. Mohammed andA.Brahim, “Content-Based Image Retrieval Using Convolutional Neural Networks,” In: Mizera-Pietraszko J., Pichappan P., Mohamed L. (eds) Lecture Notes in Real-Time Intelligent Systems. RTIS 2017.

[11] Maria Tzelepi and Anastasios Tefas, ”Deep convolutional learning for Content Based Image Retrieval,” Neuro computing, vol. 275, pp. 2467-2478, 2018.

[12] He K,Zhang X, Ren S andSun J “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8691. Springer, Cham.

[13] Perez, Luis and Jason Wang. “The Effectiveness of Data Augmentation in Image Classification using Deep Learning.” *ArXiv* abs/1712.04621 (2017): n. pag.

[14] J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Re, “Learning to compose domain-specific transformations for data augmentation,” In Advances in Neural Information Processing Systems, pp. 3239–3249, 2017.

[15] Mingxing Tan and Quoc V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” arXiv preprint arXiv:1905.11946, 2019.

[16] Y. Cai, Y. Li, C. Qiu, J. Ma and X. Gao, "Medical Image Retrieval Based on Convolutional Neural Network and Supervised Hashing," in IEEE Access, vol. 7, pp. 51877-51885, 2019.

[17] Jun Chen, Yong Wang, Linbo Luo, Jin-Gang Yu andJiayi Ma, “Image retrieval based on image-to-class similarity” , Pattern Recognition Letters, vol. 83, pp- 379-387, 2016.

[18] Ye F, Xiao H, Zhao X,Dong M,LuoW*et al.*, “Remote Sensing Image Retrieval Using Convolutional Neural Network Features and Weighted Distance,” in IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 10, pp. 1535-1539, 2018.

[19] Ye F, Dong M,Luo W, Chen X and Min W, “A New Re-Ranking Method Based on Convolutional Neural Network and Two Image-to-Class Distances for Remote Sensing Image Retrieval,” in IEEE Access, vol. 7, pp. 141498-141507, 2019.