I.J. Image, Graphics and Signal Processing, **2023, 2, 1-12**
Published Online on April 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2023.02.01

A Novel Technique for Image Retrieval based on Concatenated Features Extracted from Big Dataset Pre-Trained CNNs

Chandra Mohan Bhuma
Bapatla Engineering College, Department of ECE, Bapatla, Andhra Pradesh, India
E-mail: chandrabhuma@gmail.com
ORCID iD: https://orcid.org/0000-0002-7566-4739

Ramanjaneyulu Kongara*
PVP Siddhartha Institute of Technology, Department of ECE, Vijayawada, Andhra Pradesh, India
E-mail: kongara.raman@gmail.com
ORCID iD: https://orcid.org/0000-0003-0711-0547
*Corresponding Author

Received: 08 April, 2022; Revised: 14 June, 2022; Accepted: 10 February, 2023; Published: 08 April, 2023

Abstract: Accessing semantically relevant data from a database is not only essential in commercial applications but also in medical imaging diagnosis. Representation of the query image by its features and subsequently the dataset are the key factors in Content Based Image Retrieval (CBIR). Texture, shape and color are commonly used features for this purpose. Features extracted from the pre-trained Convolutional Neural Networks (CNNs) are used to improve the performance of CBIR methods. In this work, we explore a recent state of the art big dataset pre-trained CNNs which are known as Big Transfer Networks. Features extracted from Big Transfer Network have higher discriminative power compared to features of many other pre-trained CNNs. The idea behind the proposed work is to demonstrate the effectiveness of using features of big transfer networks for image retrieval. Further, features extracted from big transfer networks are concatenated to improve the performance of the proposed method. Feature diversity supplemented with network diversity should ensure good discriminative power for image retrieval. This idea is supported by performing simulations on four datasets with varying sizes in terms of number of images and classes. As feature size increases with the concatenation, we applied a dimensionality reduction algorithm i.e., Principal Component Analysis. Several distance metrics are explored in this work. By properly choosing the pre-trained CNNs and distance metric, it is possible to achieve higher mean average precisions. ImageNet-21K pre-trained CNN and Instagram pre-trained CNN are chosen in this work. Further, a pre-trained network trained on Imagenet-21K dataset is superior compared to the networks trained on ImageNet-1K dataset as there are more classes and presence of wide variety of images. This is demonstrated by applying our algorithm on four datasets i.e., COREL-100, CALTECH-101, FLOWER-17 and COIL-100. Simulations are presented for various precisions (scopes), and distance metrics. Results are compared with the existing algorithms and superiority of the proposed method in terms of mean Average Precision is shown.

Index Terms: Content Based Image Retrieval, Big Transfer Network, Pre-trained Convolutional Neural Network, ImageNet-21K, COREL-100, CALTECH-101, FLOWER-17, COIL-100 datasets.

1. Introduction

Based on the content of the query image, extracting relevant images from the large databases is known as Content Based Image Retrieval (CBIR) [1]. Image search engines utilize this for displaying images when the user poses a query. Finger print recognition, iris recognition, crime detection and abnormality assessment from the medical images are some applications of CBIR. Two steps in general CBIR systems are feature extraction and similarity measure. Feature extraction refers to representation of the images in a compact manner without any loss of image content. Several features i.e., texture, color, and shape [2] are used for this purpose. In recent times, Convolutional Neural Networks (CNNs) are gaining popularity due to their powerful representation of the images with minimal preprocessing. Several architectures have been attempted using CNNs and excellent performance metrics have been reported on various

datasets. Most of the architectures were trained on ImageNet-1K [3] which is a 1000 class and 1.2 Million image dataset. A larger version is Imagenet-21K [4] dataset comprising 14.19 Millions of images and 21,843 classes. However, recently Google team has trained their networks on a much larger dataset known as JFT-300M [5] which is bigger than ImageNet-1K and ImageNet-21K and has 18291 classes and 300 Millions of images. A pre-trained network is a network which is trained on some dataset. This pre-trained network can be used as a feature extractor. Usually features are extracted from the last pooling layer of the pre-trained CNN.

It is expected that the features extracted from a pre-trained CNN which was trained on large and similar dataset (similar to target dataset) performs well for the target dataset. ImageNet-21K comprises more classes and images than ImageNet-1K. In this work, features extracted from a Big Transfer (BiT) Network [6] are used for CBIR. To improve the mean average precision (mAP), the features from BiT are concatenated with features derived from an Instagram trained ResNeXt [7]. The dimensionality of the resulting concatenated feature vector is reduced by using Principal Component Analysis (PCA) [8].

Once features are extracted for the query and dataset images, a distance or similarity metric is used to obtain the semantically closest image. Top 20 or 10 images (given as scope) are retrieved and the number of relevant images is computed. This is known as precision. Mean Average Precision (mAP) is calculated for all the classes of the dataset.

The major research objectives of this work are to propose an algorithm which offers higher mAP and faster retrieval from the database. Most of the existing solutions in CBIR consider the features from the images for retrieval. These features could be based on color, texture and shape. Usage of features i.e, image moments, Histogram of Oriented Gradients (HOG) features, wavelet features, Haralick features, Speeded up Robust Features (SURF) features are common ones under this category. Most of the times, it is unsure that what sort of features to use for a given dataset. Understanding the nature and characteristic of the images in the database is essential for applying these features. However, it is not essential when pre trained CNNs are used. This work addresses this problem by extracting the features from a pre-trained CNN which was trained on a large dataset. A large dataset comprises many classes and many images per class. Hence, wide variety of features can be obtained with this pre-trained CNN. For this, we used a pre trained CNN trained on ImageNet-21K. ImageNet-21K dataset has 14.19 Millions of images with 21843 classes. To impart more discriminating power to the feature vector, a feature vector derived from another pre-trained CNN trained on Instagram dataset is concatenated. Instagram dataset has billions of Instagram images with distinct hashtags as labels. In CBIR, there are two major constraints. One is mAP and the other one is the feature vector size which in turn, leads to computational complexity. The increase in mAP is obtained with concatenation. However, concatenation increases the feature vector size. Hence, PCA is used to reduce the dimension of the feature vector size. The size of the BiT is large in terms of number of layers. When the feature vector is to be computed for the query image, it may take some time as the image passes through all the layers. This may take some time in real time applications. This limitation can be overcome by utilizing mobile architectures which are lightweight in terms of number of parameters and layers.

This paper is organized as follows. A brief review of existing works in CBIR is presented in section 2. Details of the BiT network and Instagram trained ResNeXt are given in section 3. Section 4 describes the details of the datasets used in this work. Proposed algorithm is given in section 5. Simulations and experimental results are given in section 6. Concluding remarks are given in section 7.

2. Existing Works in CBIR

Searching for relevant images from the databases is common in search engines. Due to rapid usage of internet, this requirement is growing at a fast pace. Relevant images have several meanings in the context of image processing terminology. When the retrieval depends on the content of the image, it is known as content based image retrieval. Images are represented using features. Pixel level features are not powerful for describing the content. Features describing the color, texture and shape are useful in this regard.

A CBIR system known as QBIC [9] was developed by IBM in 1995. By using sketches and drawings as query images, one can obtain relevant images from the datasets. QBIC used the combination of all the features i.e., color, shape, and texture. Later many CBIR systems used Color Co-Occurrence matrix (CCM).

For identifying the prominent objects, a Bandlet transform has been used by Ashraf et al., [10] for detecting geometric boundaries. A Gabor filter is applied to find the textural content from the boundaries. By using back propagation neural network, accurate geometric classification is achieved. They have demonstrated their results on COREL100, COIL100 and CALTECH101 datasets. Khokar [11] have used the color, texture and shape features in cohesion. For color features, they have used color histogram, and texture content is represented by Gray Level Co-occurrence Matrix (GLCM). Zernike moments are used to describe the shape. Feature reduction is done using ReliefF algorithm. The final system is trained with a back propagation neural network. They have applied the algorithm on COREL-100 dataset. Uzma Sharif et al., [12] have claimed that shape, color and texture descriptors alone cannot represent the image content. They have used visual words of a Scale Invariant Feature Transform (SIFT). Fusing these features with Binary Robust Invariant Scalable Key (BRISK) points an improvement in CBIR system is observed. The algorithm has demonstrated superior performance on COREL-1K, COREL-1.5K, COREL-5K, and CALTECH-256 datasets.

Ahmed et al., [13] used signatures of the object and color for retrieving images. By utilizing the Gaussian variance and convolution, features are extracted. PCA is used for reducing the dimensionality. Results are shown for CALTECH-101 and CORELl-1000 datasets. In complex and occluding scenarios also their method performed well.

Subhadip Maji and Samarjit Bose [14] used deep features from the pre-trained CNNs for features. They have used DenseNet, InceptionResNetV2, InceptionV3, MobileNetV2, NasNet Large, ResNet50, VGG19, and Xception networks for feature extraction. As the networks are already pre-trained on ImageNet1K there is no separate training needed. Higher level features can be extracted within less time. PCA was used to reduce the dimensions of the feature vector. Simulations were done on DB-2000, COREL-1000, and CALTECH-101 datasets.

Samarjit Bose et al., [15] used relevance feedback by employing feature reweighing in the CBIR system. They have applied their algorithm on CALTECH, COREL, and COREL datasets. Soumya et al., [16] have used three types of features Color Moments, Ranklet Transformation, and Moment Invariants implicitly including color, texture and shape features respectively. Concatenation of all the three features is used to retrieve the relevant images. They have demonstrated superior results on SIMPLICITY, COREL-5K, COREL-10K and CALTECH-101 and MSR datasets. On COIL-100 dataset, various techniques described in [17-19] are available. For FLOWER-17 dataset comparison, references [20-22] are used.

Obulesu et al., [23] have presented an approach for CBIR, where in the image is divided into 2x2 grids. Two images are obtained with an initiation from top left most pixel and bottom right most pixel. Two different Peano scan motif indices are formulated for each grid. Co-occurrence matrices are calculated for these two images. They called these two matrices as Motif Co-occurrence Matrix from Bottom Right and Top Left. These two features are concatenated. Simulations are carried on Corel-1k, Corel-10k, MIT-VisTex, Brodtaz, and CMU-PIE.

There are several feature extraction methods and similarity metrics available in the literature. A systematic approach for selecting these features and similarity metrics is suggested by S.M. Mohidul Islam and Rameswar Debnath [24]. With exhaustive simulations, they identified that color moments and wavelet packet entropy features fare well compared to color autocorrelogram and wavelet moments. Cosine and correlation measures did good job for majority of the datasets. City block distance measure is also able to retrieve similar images well. However, L1 and Mahalanobis distance metrics are not suitable for the majority of the datasets. The work presented a comparative report of 6 features and 12 distance measures.

Nitin Arora et al., [25] have used Haar wavelet features and Gabor features. They have shown that by using hybrid features some improvement in the retrieval accuracy can be obtained. A. Anbarasa Pandian and R. Balasubramanian [26] have applied the CBIR technique for the MRI brain tumor images dataset. They used a shape feature for this. By using Zernike Moments, they got an accuracy of 99%. They compared the results with Scale invariant feature transform (SIFT) and Harris corner detection techniques. Further, they classified the images as normal and tumor images by using a Deep Neural Network and Extreme learning machine (ELM). They proved that ELM is better than DNN for the MRI image dataset.

3. Big Transfer Network

In general, the deeper the Convolutional Neural Netwsorks (CNN) the higher is the classification accuracy. This is an empirical one. However, training the CNN from the scratch is computationally expensive. Hence, transfer learning is usually employed. Network trained on one dataset is utilized for other dataset which may be completely different from the trained dataset. Freezing some layers and training it on the new dataset is a common apporach in transfer learning. Another alternative is using a pre-traind CNN as a feature extractor. After extracting the features, they are given to a traditional classifier i.e., Support Vector Machine (SVM) or K Nearest Neighbour (KNN) classifier if the problem is that of a classification one. But in this work, we utilize the features for retrieving the relevant images from the dataset with an appropriate similarity metric.

In Big Transfer (BiT) Networks, large networks are trained on ‘Big’ datasets and these weights are used for subsequent application on small datasets by using transfer learning. There are three categories of datasets used in this work. BiT-L (Big Transfer Network-Large) uses JFT-300M dataset, BiT-M (Big Transfer Network-Medium) uses ImageNet-21K and BiT-S (Big Transfer Network-Small) uses ImageNet-1K. Number of images in JFT-300M, ImageNet-21K and ImageNet-1K are 300 Millions, 14 Millions and 1.3 Million respectively.

In this work, a content based image retrieval algorithm using BiT network trained on ImageNet-21K is proposed. BiT network is available as BiT (Small), BiT (Medium) and BiT (Large). However, BiT (Medium) is publicly available and trained on a public dataset ImageNet-21K. BiT (Small) was trained on ImageNet-1K and BiT (Large) was trained on JFT-300M dataset having 300 Millions of images. Simply training on large dataset like JFT-300M by architecture like ResNet50 does not show any improvement. However, when larger architectures like ResNet152x4 were trained on JFT-300M a significant improvement was demonstrated. Conceptually there are two steps involved in Big transfer networks i.e., upstream training and downstream training. In upstream training, large architectures are trained on larger datasets i.e., JFT-300M and ImageNet-21K. The inference from BiT networks is, larger models trained on larger datasets only will offer significant improvements. When handling larger datasets, training time has to be longer in order to have good accuracy. Number of epochs has to be increased while training on larger datasets. For ImageNet-1K usually 90 epochs training is standard. However, this is not sufficient for ImageNet-21K or JFT-300M datasets. Group

normalization with weight standardization is used instead of batch normalization as batch normalization performance is poor when the number of images fit in a batch is low on a 'CPU' or 'GPU'. In downstream training, for the hyper parameter BiT- Hyper Rule is proposed. In BiT-Hyper Rule, stochastic gradient descent optimizer, initial learning rate of 0.003 with a momentum of 0.9 is used. With a batch size of 512 and a decay learning rate with a factor 10 at 30%, 60% and 90% training steps the model can be trained on any small dataset. Resizing, random cropping, and random horizontal flip are the pre-processing steps used in the downstream. Out of the three models, only two models BiT-S and BiT-M were released which were trained on ImageNet-1K and ImageNet-21K datasets. The basic architecture selected for BiT training was ResNet model. In addition, variants of ResNet architectures like ResNet50x1, ResNet50x3, ResNet101x1, ResNet101x3, ResNet152x2 and ResNet152x4 were also used on all the three datasets. We have used ResNet152x2 trained on ImageNet21K as the base network for feature extraction.

To improve the mAP further, the features from the BiT network is concatenated with another weakly supervised network i.e., IGResNeXt. ResNeXt models were pre-trained on 940 Millions of images with 1.5K hashtags which are publicly available. Training was done in a weakly supervised method. Later they were fine-tuned on ImageNet-1K dataset. In this work, four variants of IGResNeXt are used. They are IGResNeXt101-8d, IGResNeXt101-16d, IGResNeXt101-32d and IGResNeXt101-48d. BiT models were released by Google and IGResNeXt models were released by Facebook AI.

4. Dataset

Datasets used for conducting simulations are COREL-100, CALTECH-101 and FLOWER-17. There are 10 categories of images in COREL-100 dataset. In each category, 100 images are there. African People, Beach, Building, Bus, Dinosaurs, Elephant, Flower, Horse, Mountain and Food are the categories of COREL-100 dataset. Sample images from each category are shown in Fig. 1. There are 102 classes in CALTECH-101 dataset. The number of images in each category is not constant and it varies from 34 to 800. There are 9144 images in CALTECH-101 dataset and sample images are shown in Fig. 2. Fig. 3 displays the 17 classes of Oxford FLOWER-17 dataset. There are 80 images per class.

Fig. 1. Sample images from COREL-100 dataset'

Fig. 2. Sample images from CALTECH-101 dataset

Total images in FLOWER-17 dataset are 1360. There are 100 classes and 72 images per class in COIL-100 dataset. Sample images per class are shown in Fig. 4. The idea in selecting these databases is to demonstrate the validity of the proposed algorithm on datasets with varying number of classes and varying number of images per class. Further, the performance of the algorithm is tested on both balanced and imbalanced datasets.

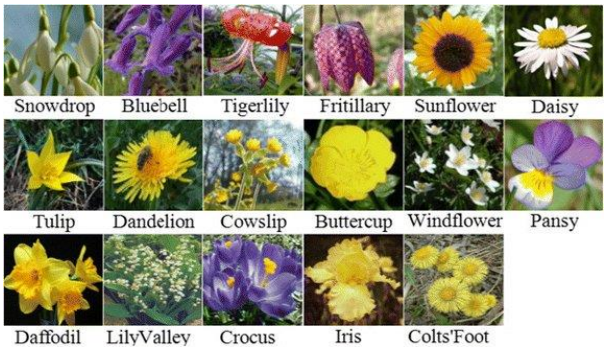


Fig. 3. Sample images from FLOWER-17 dataset

Fig. 4. Sample images from COIL-100 dataset

5. Proposed CBIR Algorithm

Detailed algorithm is as follows:

1. All the images in the dataset are resized and normalized with appropriate mean and standard deviation as per the chosen pre-trained network configuration.
2. All the images in the dataset except query image are passed through the chosen pre-trained networks and features are extracted from the last pooling layer.
3. Features from the two pre-trained networks are concatenated.
4. A dimensionality reduction technique, Principal Component Analysis (PCA), is applied and size of the feature vector is reduced.
5. Steps 1 to 4 are repeated for the query image and the distances between the reduced feature vector of the query image and the images of the database are computed.
6. The distances are sorted and the top 10 or 20 (as per the scope chosen) with minimum distance are retrieved.
7. For each query image in the dataset, steps 5 and 6 are repeated and the mean Average Precision (mAP) is calculated.

Steps of the proposed algorithm are illustrated in Fig. 5. For feature concatenation two pre-trained networks are used. One network was trained on ImageNet-21K and the other network was trained on Instagram 1.5K hashtag images and fine-tuned on ImageNet1k dataset. Various similarity measures (distance metrics) i.e., Euclidean distance, City block distance, Chebychev distance, Cosine distance, Correlation distance, Hamming distance and Spearman distances were used to find the relevant images from the dataset.

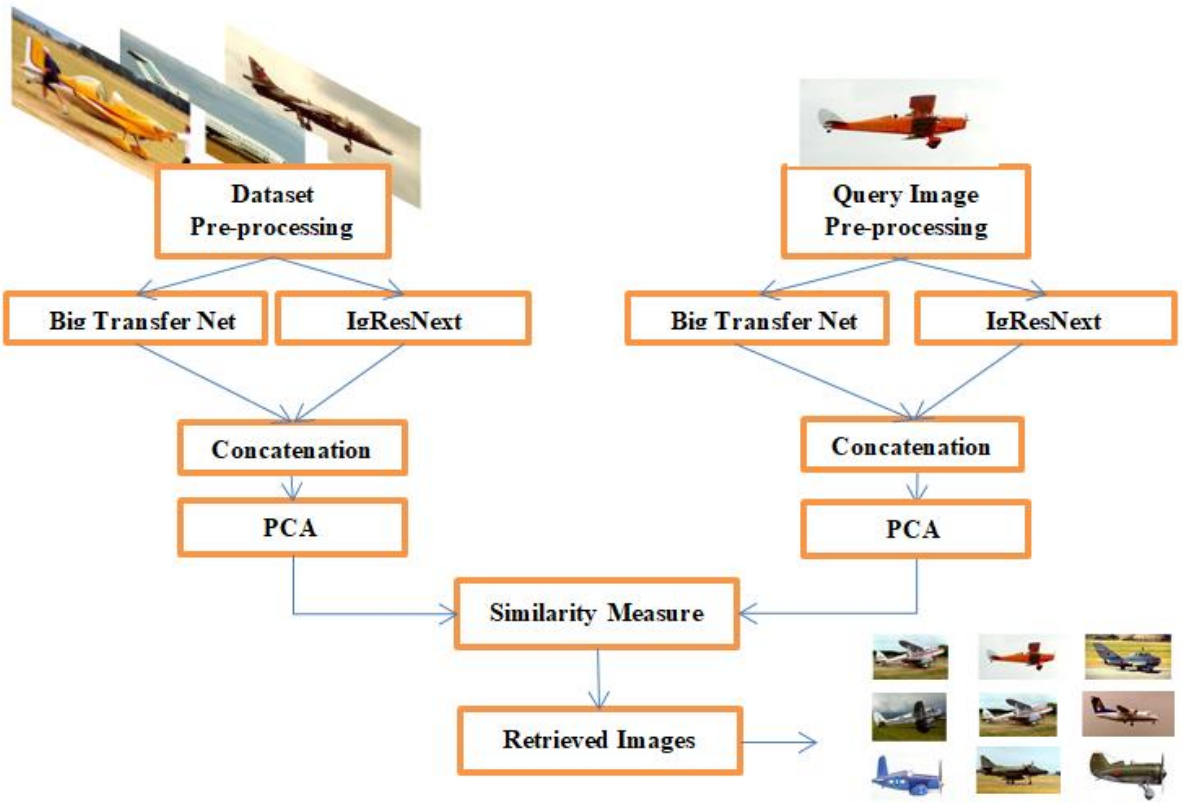


Fig. 5. Block diagram of the proposed algorithm

6. Simulations and Results

Simulations are carried in Google Colab Pro environment with PyTorch framework. All the images in the datasets were resized to 224x224 resolutions. Six models in BiT-L networks and six models in BiT-M networks were used for experimentation. Further, four models under IGResNeXt are selected for concatenation. After rigorous experimentation with various distance metrics, the best model and metric is identified and results are presented. In Table 1, the models and the size of the feature vector is presented.

Table 1. Models and the Feature Vector Size

Category	Model Name	Feature Vector Size
BiT-L and BiT-M	ResNetV2_50x1	2048
	ResNetV2_50x3	6144
	ResNetV2_100x1	2048
	ResNetV2_100x3	6144
	ResNetV2_152x2	4096
	ResNetV2_152x4	8192
IGResNeXt	IGResNeXt101_32x8	2048
	IGResNeXt101_32x16	2048
	IGResNeXt101_32x32	2048
	IGResNeXt101_32x48	2048

Overall accuracy of the CBIR systems depends on the number of retrieved images. This is known as the ‘Scope’ of the CBIR. Precision tells us the number of relevant images retrieved compared to the retrieved images. Precision varies from one class to the other. Mean Average Precision (mAP) is chosen as a general metric to assess the performance of the CBIR system. This is nothing but the average precision taken over all the classes. The base model used for computing the mAP for all the datasets is ResNetV2_152_2. For COREL-100 dataset, mAP for various scopes is shown in Fig. 6.

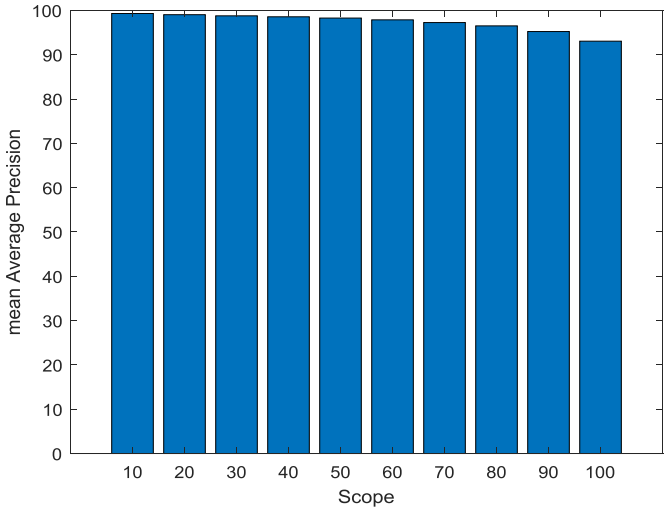


Fig. 6. mAP vs Scope for COREL-100 dataset (Correlation Distance)

At a scope of 20, the mAP for COREL-100 dataset is 99.03% with Correlation distance as a similarity metric. Per class precision for various scopes is given in Table 2. As shown in Table 2, for a scope of 20, 100% precision is achieved for five classes, i.e., Bus, Dinosaurs, Elephant, Flower, Horse, Mountain and Food. Even for African people the precision is 95% which is much higher than the work of [14]. All the classes have a precision well above 95% which reflects the discriminative power of the selected pre-trained network. CALTECH-101 dataset is slightly complex due to the variations in the intra class and inter class. The dataset is highly imbalanced one. Hence the maximum scope is restricted to 30 only. For both city block and cosine distances, at a scope of 20, the mAP value obtained is 87.29%. This is also higher than the results presented in the work of [14]. Mean Average Precision achieved for CALTECH-101 by the authors of [14] is 82.54%. Performance of the proposed algorithm on CALTECH-101 dataset for various distance metrics and scopes is presented in Table 3. Performance of the correlation distance metric is inferior for the cases of African people, Beach, Food and Building.

Table 2. Scope vs per class Mean Accuracy on COREL-101 dataset (Correlation Distance Metric)

Class Scope	African People	Beach	Building	Bus	Dinosaurs	Elephant	Flower	Horse	Mountain	Food
10	96.70	98.50	98.60	100.00	100.00	100.00	100.00	100.00	100.00	99.20
20	95.00	98.25	98.15	100.00	100.00	100.00	100.00	100.00	100.00	98.90
30	94.37	97.67	97.43	100.00	100.00	100.00	100.00	99.83	100.00	98.43
40	93.95	97.33	96.68	100.00	100.00	100.00	100.00	99.70	100.00	97.90
50	93.64	96.54	95.66	100.00	100.00	100.00	100.00	99.54	100.00	97.34
60	93.22	95.83	93.63	100.00	100.00	100.00	100.00	99.40	100.00	96.58
70	92.69	94.63	90.40	100.00	100.00	100.00	100.00	99.29	100.00	95.51
80	91.79	93.13	86.95	100.00	100.00	100.00	100.00	99.08	100.00	94.11
90	88.80	90.76	82.67	100.00	100.00	100.00	100.00	98.54	100.00	91.51
100	82.87	86.69	77.64	99.99	100.00	99.97	99.77	96.05	99.97	87.60

Table 3. mAP vs distance metric for CALTECH-101 for various scopes

Distance Metric	mAP scope=10	mAP scope=20	mAP scope=30
Euclidean	90.10	87.06	83.88
Cityblock	90.28	87.29	84.19
Chebychev	74.21	66.03	59.60
Cosine	90.32	87.29	84.49
Correlation	90.34	87.27	84.45
Hamming	60.75	51.54	45.41
Spearman	89.94	86.71	83.48

Table 4 depicts the results obtained on FlOWER-17 dataset for Spearman distance. There are 80 images per class at the most. Hence, the scope is restricted to 80 only. Out of 17 classes, 6 classes were retrieved with a mAP of 100%. For a scope of 20, the mAP achieved is 99.56% for Spearman distance. Even for a scope of 80, mAP achieved with Spearman distance is 95.51%. For the cases of Buttercup, Dandelion, Windflower, Snowdrop, Lilyvalley, Bluebell, Crocus and Tulip, 100% mAP is not achieved even with a scope of 10.

As expected, as scope increases, the mAP decreases and is reflected in Table 5. The results obtained with City block distance are comparable with spearman distance for FLOWER-17 dataset. However, for the case of COREL-100 dataset, the performance of correlation distance is good offering mAP of 99.03% which is shown in Table 6.

Even for a scope of 100, the Correlation distance did good job in retrieving the relevant images from COREL-100 dataset. Performance of hamming distance and Chebychev distance is inferior compared to the other distances for COREL-100, CALTECH-101 and FLOWER-17 datasets. Good interclass dissimilarity and intra class similarity can be seen with the images of COIL-100 dataset. Dataset of COIL-100 comprises the images rotated by certain angle. Hence, similarity in each class is good. All the similarity measures are able to retrieve the images with a minimum mAP of 96.21%.

However, a highest mAP of 99.47% is achieved for the Euclidean, Cosine and Correlation distances. Results obtained on COIL-100 dataset are given in Table 7.

Table 4. Scope vs per class Mean Precision on FLOWER-17 dataset (Spearman distance)

Class Name	Scope 10	Scope 20	Scope 30	Scope 40	Scope 50	Scope 60	Scope70	Scope80
Buttercup	99.75	99.56	99.50	99.34	99.30	99.33	99.25	97.53
Colts' Foot	100.0	100.00	100.00	100.00	100.00	100.00	100.00	99.98
Daffodil	100.0	100.00	100.00	100.00	100.00	100.00	100.00	99.41
Daisy	100.0	99.81	99.75	99.69	99.68	99.65	99.61	98.17
Dandelion	99.88	99.75	99.79	99.78	99.73	99.67	99.64	97.75
Fritillary	100.0	100.00	100.00	100.00	100.00	100.00	100.00	99.50
Iris	100.0	100.00	100.00	100.00	100.00	100.00	99.86	97.91
Pansy	100.0	99.88	99.75	99.66	99.65	99.50	99.43	97.05
Sunflower	100.0	100.00	100.00	100.00	100.00	100.00	99.98	99.73
Windflower	99.88	99.81	99.79	99.69	99.60	99.46	98.95	95.66
Snowdrop	98.75	98.56	98.21	98.13	97.60	96.90	95.29	88.94
Lily Valley	99.38	99.25	99.13	99.06	98.90	98.60	98.04	91.95
Bluebell	99.25	98.56	97.75	97.22	96.55	95.25	92.34	86.09
Crocus	99.38	98.81	98.79	98.69	98.63	98.50	98.09	96.06
Tigerlily	100.0	99.69	99.67	99.50	99.50	99.42	99.21	95.77
Tulip	98.75	98.06	97.42	96.16	94.48	92.46	89.57	84.73
Cowslip	100.0	100.00	99.92	99.88	99.73	99.63	99.48	98.33

Table 5. mAP with various distances and scopes for FLOWER-17 dataset

Distance	Scope 10	Scope 20	Scope 30	Scope 40	Scope 50	Scope 60	Scope 70	Scope 80
Euclidean	99.57	99.32	99.11	98.85	98.56	97.97	96.81	93.14
Cityblock	99.65	99.45	99.27	99.06	98.78	98.34	97.40	94.02
Chebychev	91.98	87.79	83.96	80.06	75.77	70.79	65.41	60.11
Cosine	99.57	99.42	99.32	99.16	98.94	98.64	97.96	95.37
Correlation	99.57	99.42	99.33	99.17	98.94	98.63	97.97	95.39
Hamming	94.74	92.79	90.32	86.90	82.66	77.98	72.70	66.59
Spearman	99.71	99.51	99.38	99.22	99.02	98.73	98.16	95.56

Improvement in the mAP can be achieved by concatenating the features from the other pre-trained networks. In this work, an Instagram trained ResNeXt network which is fine tuned on ImageNet1K is chosen for concatenation with varying widths. This network was trained by Facebook and a top-1 accuracy of 85.4% was obtained with ResNeXt-101_32x48d. They have released four versions of this ResNeXt network. ResNeXt-101_32x48d has 101 layers, 32

groups, and a group width of 48. It has 829 Millions of parameters and 153 Billion FLOPS of multiplication and additions. In fact it took almost twenty two days to train the ResNeXt -101_32x16d on 3.5 Billion of images. Improved results after concatenating the ResNetV2_152_2 with various variants of IGResNeXts are given in Table 8. After the concatenation, size of the feature vector increases and hence a dimensionality reduction technique like PCA is applied. 300 principal components are chosen for COREL-100, CALTECH-101, and FLOWER-17 datasets. For COIL-100, 21 principal components were enough to obtain mAP of 99.23% with no improvement with dimensionality reduction. Highest mAPs obtained for COREL-100, CALTECH-101, FLOWER-17 and COIL-100 are 99.13%, 93.37%, 99.51%, and 99.51%. The results are compared with the existing works in the literature and are shown in Fig.7 (a), (b), (c), and (d).

Table 6. mAP with various distances and scope for COREL-100 dataset

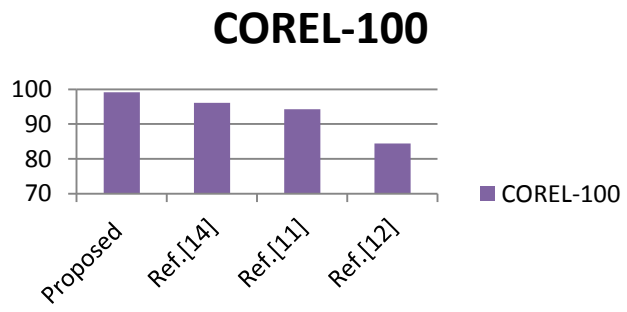
Distance	Scope 10	Scope 20	Scope 30	Scope 40	Scope 50	Scope 60	Scope 70	Scope 80	Scope 90	Scope 100
Euclidean	99.24	98.85	98.51	98.13	97.61	96.95	96.10	94.88	93.10	90.41
Cityblock	99.27	98.90	98.58	98.25	97.71	97.03	96.19	95.01	93.25	90.64
Chebychev	95.67	92.53	89.73	86.79	83.63	80.30	77.01	73.70	70.26	66.12
Cosine	99.33	98.99	98.75	98.51	98.24	97.79	97.15	96.36	95.05	92.83
Correlation	99.30	99.03	98.77	98.56	98.27	97.87	97.25	96.51	95.23	93.06
Hamming	90.83	87.67	85.31	83.04	80.50	77.71	74.93	72.07	69.25	65.99
Spearman	99.27	98.91	98.58	98.27	97.87	97.29	96.49	95.37	93.75	91.28

Table 7. mAP with various distances and scope for COIL-100 dataset

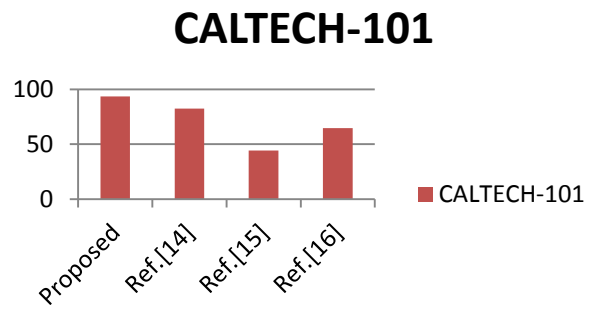
Distance	Scope 10	Scope 20	Scope 30	Scope 40	Scope 50	Scope 60	Scope 70
Euclidean	99.87	99.47	98.81	97.66	95.97	93.51	90.39
Cityblock	99.88	99.42	98.67	97.28	95.38	92.75	89.46
Chebychev	99.49	98.41	96.94	94.78	91.83	88.11	83.77
Cosine	99.86	99.47	98.80	97.58	95.84	93.38	90.29
Correlation	99.86	99.47	98.79	97.55	95.82	93.35	90.25
Hamming	99.05	96.21	92.99	89.49	85.61	81.46	76.77
Spearman	99.87	99.40	98.63	97.22	95.26	92.67	89.42

Table 8. mAP for all the four datasets with concatenation for scope=20

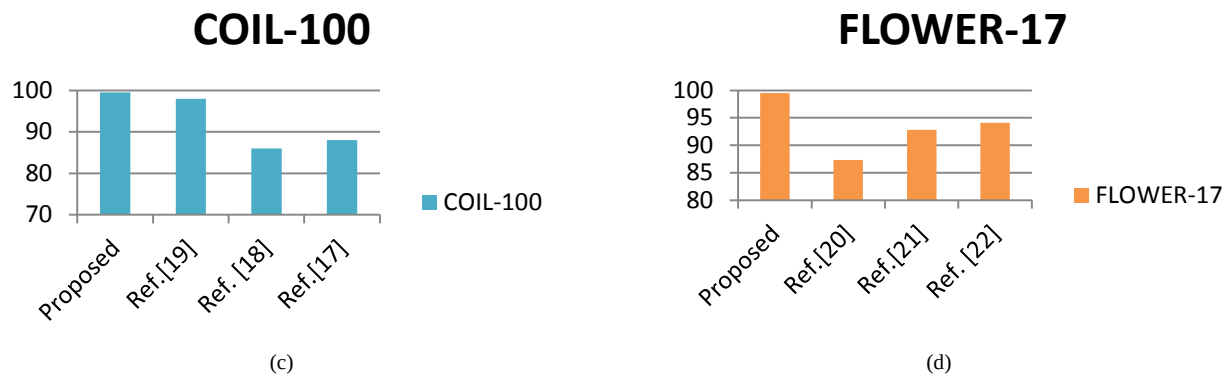
Dataset	Metric	Model Concatenated	Without Concatenation	With Concatenation	PCA Components	mAP with PCA
COREL-100	Correlation	ig_resnext101_32x32	99.03	99.09	300	99.13
CALTECH-101	Spearman	ig_resnext101_32x8	86.71	93.37	300	91.76
FLOWER-17	Spearman	ig_resnext101_32x48	99.51	99.49	300	99.40
COIL-100	Spearman	ig_resnext101_32x8	99.47	99.51	21	99.23



(a)



(b)



(c) (d)

Fig. 7. (a) COREL-100 (b) CALTECH-101 (c) COIL-100 and (d) FLOWER-17 comparison with the existing works (for scope=20)

In a typical simulation environment i.e, the proposed algorithm, the query images are taken from the database itself. Features are extracted in an offline mode by utilizing the power of GPU (Graphical Processing Unit). Feature extraction from the networks i.e, BiT and IGResNeXt is time consuming and may demand more computational resources. This may pose some restriction in scenarios where mobile devices are used.

7. Conclusion

This In this work, a CBIR algorithm using BiT network is proposed. Pre-trained BiT network trained on ImageNet-21K is used as a feature extractor. Improved performance in terms of mAP is achieved for the four datasets i.e., COREL-100, CALTECH-101, FLOWER-17 and COIL-100. By concatenating with another Iinstagram' trained and fine-tuned on ImageNet-1K network, the performance of COREL-100, CALTECH-101 and COIL-100 is improved further. After the concatenation, the feature vector dimension increases hence a dimensionality reduction technique like PCA with 300 components is used and for COREL-100, CALTECH-101 and FLOWER-17 datasets. For COIL-100 only 21 principal components are chosen. Similarity measures i.e., Euclidean, Cityblock, Chebychev, Correlation, Hamming and Spearman are used to compute the mAP for all the datasets. Spearman distance offered higher accuracy for CALTECH-101, FLOWER-17 and COIL-100. Whereas, Correlation distance metric for COREL-100 offered a higher mAP. The present state of the art work uses either traditional features utilizing color, texture and shape or using CNNs trained with small datasets, the proposed algorithm paves the way for effective retrieval based on the query image using CNNs trained on large and diversified datasets. Image based search engines i.e., Google Images, Yahoo Image Search, Bing Image Search, Openverse, and Getty Images are quite popular. The proposed algorithm can be used in these search engines for faster and accurate image retrieval. Hence, it may be concluded that by properly choosing the pre-trained network, similarity metric (distance metric), and suitable dimensionality reduction technique, it is possible to achieve higher mean Average Precisions (mAP) for the image datasets in a CBIR system. By applying a meta-heuristic algorithms i.e., Genetic Algorithm, Particle Swarm Optimization and others, selecting a combination of model, similarity metric and dimensionality reduction technique can be chosen with an objective of improving mAP.

Acknowledgment

The authors wish to thank the Bapatla Engineering College and PVP Siddhartha Institute of Technology for their support in carrying out this research work.

References

[1] Liu Y, Zhang D, Lu G, Ma WY. A survey of content-based image retrieval with high-level semantics. Pattern recognition. 2007 Jan 1; 40(1):262-82.

[2] Manjunath BS, Ohm JR, Vasudevan VV, Yamada A. Color and texture descriptors. IEEE Transactions on circuits and systems for video technology. 2001 Jun; 11(6):703-15.

[3] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec; 115(3):211-52.

[4] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In2009 IEEE conference on computer vision and pattern recognition 2009 Jun 20 (pp. 248-255). Ieee.

[5] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 843-852). https://ai.googleblog.com/2017/07/revisiting-unreasonable-effectiveness.html

[6] Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, Houlsby N. Big transfer (bit): General visual representation learning. InEuropean conference on computer vision 2020 Aug 23 (pp. 491-507). Springer, Cham. https://arxiv.org/pdf/1912.11370.pdf

[7] Mahajan D, Girshick R, Ramanathan V, He K, Paluri M, Li Y, Bharambe A, Van Der Maaten L. Exploring the limits of weakly supervised pretraining. In Proceedings of the European conference on computer vision (ECCV) 2018 (pp. 181-196). arXiv. 1805.00932.
[8] Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science. 1901 Nov 1; 2(11):559-72.
[9] Niblack CW, Barber R, Equitz W, Flickner MD, Glasman EH, Petkovic D, Yanker P, Faloutsos C, Taubin G. QBIC project: querying images by content, using color, texture, and shape. InStorage and retrieval for image and video databases 1993 Apr 14 (Vol. 1908, pp. 173-187). Spie.
[10] Ashraf R, Bashir K, Irtaza A, Mahmood MT. Content based image retrieval using embedded neural networks with bandletized regions. Entropy. 2015 May 29;17(6):3552-80.
[11] Khokhar S, Verma S. Content based image retrieval with multi-feature classification by back-propagation neural network. Int. J. Comput. Appl. Technol. Res. 2017 Jul;6:278-84.
[12] Sharif U, Mehmood Z, Mahmood T, Javid MA, Rehman A, Saba T. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. Artificial Intelligence Review. 2019 Aug; 52(2):901-25.
[13] Ahmed KT, Naqvi SA, Rehman A, Saba T. Convolution, approximation and spatial information based object and color signatures for content based image retrieval. In2019 International Conference on Computer and Information Sciences (ICCIS) 2019 Apr 3 (pp. 1-6). IEEE.
[14] Maji S, Bose S. CBIR using features derived by deep learning. ACM/IMS Transactions on Data Science (TDS). 2021 Aug 31; 2(3):1-24.
[15] Bose S, Pal A, Chakrabarti D, Mukherjee T. Improved content-based image retrieval via discriminant analysis. International Journal of Machine Learning and Computing. 2017 Jun; 7(3):44-48
[16] Rana SP, Dey M, Siarry P. Boosting content based image retrieval performance through integration of parametric & nonparametric approaches. Journal of Visual Communication and Image Representation. 2019 Jan 1; 58:205-19.
[17] Fadaei S, Amirfattahi R, Ahmadzadeh MR. New content‐based image retrieval system based on optimised integration of DCD, wavelet and curvelet features. IET Image Processing. 2017 Feb;11(2):89-98.
[18] Cui C, Lin P, Nie X, Yin Y, Zhu Q. Hybrid textual-visual relevance learning for content-based image retrieval. Journal of Visual Communication and Image Representation. 2017 Oct 1; 48:367-74.
[19] Dhingra S, Bansal P. A novel & efficient fusion based image retrieval model for speedy image recovery. EAI Endorsed Transactions on Scalable Information Systems. 2020; 7(27).
[20] Putri RD, Prabawa HW, Wihardi Y. Color and texture features extraction on content-based image retrieval. In2017 3rd International Conference on Science in Information Technology (ICSITech) 2017 Oct 25 (pp. 711-715). IEEE.
[21] Kumar V, Tripathi V, Pant B. Content based fine-grained image retrieval using convolutional neural network. In2020 7th International Conference on Signal Processing and Integrated Networks (SPIN) 2020 Feb 27 (pp. 1120-1125). IEEE.
[22] Kumar A, Choudhary S, Khokhar VS, Meena V, Chattopadhyay C. Automatic feature weight determination using indexing and pseudo-relevance feedback for multi-feature content-based image retrieval. arXiv preprint arXiv:1812.04215. 2018 Dec 11.
[23] Obulesu A, Kumar VV, Sumalatha L. Content based image retrieval using multi motif co-occurrence matrix. International Journal of Image, Graphics and Signal Processing. 2018 Apr 1; 10(4):59-72
[24] Islam SM, Debnath R. A comparative evaluation of feature extraction and similarity measurement methods for content-based image retrieval. International Journal of Image, Graphics and Signal Processing (IJIGSP). 2020; 12(6):19-32.
[25] Arora N, Ashok A, Tiwari S. Efficient Image Retrieval through Hybrid Feature Set and Neural Network. International Journal of Image, Graphics & Signal Processing. 2019 Jan 1; 11(1):44-53
[26] Pandian AA, Balasubramanian R. Analysis on Shape Image Retrieval Using DNN and ELM Classifiers for MRI Brain Tumor Images. International Journal of Information Engineering & Electronic Business. 2016 Jul 1; 8(4):63-72

Authors' Profiles

Chandra Mohan Bhuma received his B. Tech in Electronics and Communication Engineering (ECE), M.Tech in Microwave & Radar and Doctoral Degree in Image Watermarking JNTU, Hyderabad. He is currently working as a Professor in the Bapatla Engineering College. His research interest includes applications of machine learning and Advanced Deep Learning.

Ramanjaneyulu Kongara received his B. Tech in Electronics and Communication Engineering (ECE) and completed M. Tech in Electronics and Communication Engineering (ECE) from Pondicherry Engineering College. He received Doctoral Degree in domain of Digital Image Watermarking from Andhra University in 2012.Currently working as Professor in PVP Siddhartha Institute of Technology, Vijayawada. Area of Research includes Image processing, Wireless Communications and Machine Learning Applications.

How to cite this paper: Chandra Mohan Bhuma, Ramanjaneyulu Kongara, "A Novel Technique for Image Retrieval based on Concatenated Features Extracted from Big Dataset Pre-Trained CNNs", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.15, No.2, pp. 1-12, 2023. DOI:10.5815/ijigsp.2023.02.01