Application of CBIR in E-commerce

Shweta Sadwani, Vaibhavi Sangawar, Rushabh Sanap, Akanksha Kakade, Minakshi Vharkate

Abstract*: The rise of technology and the rapidly increasing inventions in Science have completely changed many aspects of the world today. Many sectors such as communication, banking, media, etc. have gained momentum because of the internet. Online shopping is one such sector that has flourished in recent times because of the internet. This paper presents a method which employs the system of Content Based Image Retrieval (CBIR) in online shopping. Using this system, the time required to shop online will be reduced. CBIR is the activity of fetching images from the database which have some similarity to the given query image. Traditionally customers would have to search from different categories and apply various filters to buy the product that they want. But in this system, they will be provided with an option to directly upload the image of the product that they wish to buy. If similar products are available, it will be displayed to the customer immediately. Thus, the time required for a customer to buy a product reduces considerably thereby making the shopping experience fun, easy and convenient. The system works in a way such that when an image is uploaded, the features of this image are extracted by using the deep learning method of Convolutional Neural Network (CNN). These extracted features are compared with the features of the available images stored in the database. Then, the similarity measure is calculated and images that are akin to the query image are found and are set out as result. This method significantly helps in reducing the time required to search for a particular product.

***Keywords*: classification and indexing, Content based Image Retrieval, Convolutional Neural Network, deep learning, image processing .**

I. INTRODUCTION

Latest trends in internet technology have eased many processes such as communications, social media, banking and online shopping too. The internet is solely responsible for making a very powerful impact on various sections of marketing and created a completely different type of retail transaction known as e-commerce which is used for online shopping. Electronic commerce or e-commerce is an activity that has recently started gaining momentum and is used for selling and buying different goods and products over the internet.

Revised Manuscript Received on June 22, 2020.

***** Correspondence Author

Shweta Sadwani, B. Tech. in Computer Science, MIT Academy of Engineering, Pune, India. Email: shweta.sadwani.6@gmail.com

Vaibhavi Sangawar, B. Tech. in Computer Science, MIT Academy of Engineering, Pune, India. Email: vmsangawar03@gmail.com

Rushabh Sanap, B. Tech. in Computer Science, MIT Academy of Engineering, Pune, India. Email: sanaprushabh2016@gmail.com

Akanksha kakade, B. Tech. in Computer Science, MIT Academy of Engineering, Pune, India. Email: akanksharkakade @gmail.com

Minakshi N.Vharkate, Sr. Assistant Professor, MIT Academy of Engineering, Pune, India. Email: mnvharkate@comp.maepune.ac.in

In the previous few years, it has been observed that there is an exponential rise in the users indulging in shopping online and it is predicted that this trend will only keep on increasing in the near future. Also, each and every person nowadays is equipped with a smart phone which has the latest state-of-art multi-media and camera technologies and also supports fast internet access

In this paper, we have designed and discussed a system that extracts and stores the features of the images present in the database. This is done by CBIR. CBIR is an approach which uses visual contents and features to search for analogous images from large image databases according to the customers or users interests. In essence, CBIR consists of retrieving the most visually analogous images to a given query image from a database of images. Now-a-days there is a huge rise in the number of images generated on social media platforms this huge database is to be used for an effective and robust retrieval and search technique is required.

The traditional approach for image searching and retrieval was performed by explaining every image with a text annotation and retrieving images by searching the keywords. To search the entire database for a product that a customer is looking for gets very tedious. Hence, the content-based image retrieval technique enhanced the searching functionality of the customers by narrowing down their searching bracket considerably. This technique makes use of the machine learning concept known as Convolution Neural Network (CNN). CNN algorithm which is responsible for the feature extraction, similarity matching and classification. Then, during run-time, whenever a user or customer has to search for a particular product, they have to simply upload the image of that product. This image is treated as the query image. When the query image is uploaded, CNN will then extract the features of this image by carefully selecting the accurate area of the image and eliminating unwanted backgrounds of the image. The features extracted are a fusion of colour, size, shape, edge, texture, etc. using architecture of CNN (In this paper, we have used Visual Geometry Group Net). Once the features of the query image are extracted, it is then matched to the features of the images available in the database. If similar images are present, they are retrieved and returned to the user as a result. [1]

The idea of this system is to employ the method of image processing and image retrieval in e-commerce and to help in online shopping. The system gives the buyer an option to upload a photo of any product that they want and to return similar products using the image retrieval system with better accuracy.

II. RELATED WORK

After searching on the internet, going through various topics with an amalgamation of machine learning and data science domain, we came across the problem of image processing and image retrieval. The application of image retrieval in e- commerce and online shopping is very crucial in today's world. The increased number of users indulging in online shopping day by day has risen to an all new high. For this purpose, it is very crucial to understand the importance of image retrieval techniques in online shopping. This field is an emerging field in which some of the pioneering works by researchers have been carried out. The listed below papers are some of the references we took into consideration while making the project.

The paper [2] proposes a system that is based on CBIR which utilizes the methods of CNN along with Support Vector Machine (SVM). Here, CNN method is used for visual feature extraction of the images that are given and then SVM is used for the purpose of classification of these images as it is used for binary classification. The architecture of CNN that is used in this paper is VGGnet. The images that are tested, are ranked using SVM on the basis of the distance between the paired features of two or more images and also the hyperplane in SVM. This paper lays emphasis on the good results obtained by using the original CNN features and moreover employs SVM as a method for learning similarity measures.

The paper [3] focuses on primarily using only CNN for both feature extraction as well as to classify images as similar and dissimilar. This is done to obtain better retrieval results for the proposed system. Here, the architecture of CNN that is being used is the Alex Net architecture. The Alex Net is architecture of CNN that is very useful in returning the most similar results that may be contained in the system. Set of database images and the given query image are evaluated by the use of a distance formula which helps in ranking all the relevant searches in the descending order which makes it possible to attain the most relevant results right at the top.

This paper [4] describes user relevance feedback using deep neural networks. CNN and Siamese net are used in this paper. CNN is used to draw out the features of the images, Siamese net is a neural net architecture that gives similarity measures between images using Euclidean distance formula and it is also used to label images as similar or dissimilar. CNN learns a new distance measure by user feedback i.e. the relevant set of k images are displayed to the user out of which user randomly selects images the images that are selected by the user are considered as relevant.

This paper [5] represents the pros of the CBIR system and also the key technologies. The authors of the paper suggested a technique that groups features together. Colour, texture and shape along with GrabCut method are used for retrieval of images. Then the paper concentrates on the representation, extraction of features and user feedback.

Different Methods compared in paper are as follows:

a. Shape Feature Extraction
b. Texture Feature Abstraction
c. Colour Feature Abstraction

Comparison between outcomes of different methods as Precision (P), Recall(R) and Response Time (RT) is given in paper based on the images provided by authors as shown in fig.1.

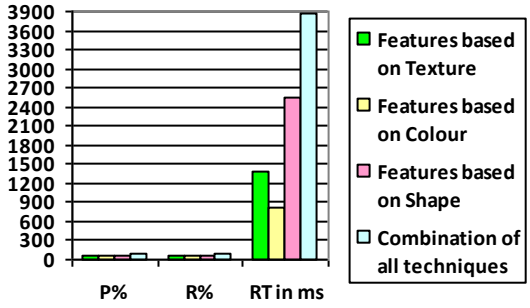


Fig.1. Comparison of different methods

Here, we can conclude from the above table that the grouping of all techniques along with Grab Cut strongly improves precision and recall but on other hand response time increases as a result of more complexity.

This paper [6] shows the comparison between CNN (on Graphical Processing Unit) and CNN-SVM (on Central Processing Unit) i.e. using deep learning methods for fast feature extraction and classification of images. The CNN on GPU used binary hash codes for Speedy image retrieval. The CNN-SVM on CPU - here, CNN is combined with linear SVM where CNN is used for extracting features of images and SVM performs recognition step. The performance of the two is evaluated on the basis of precision and recall based on relevant images and retrieved images. Accuracy of CNN on GPU is more than CNN-SVM on CPU.

This paper describes [7] online shoe shopping using Multi-Task View-invariant Convolutional Neural Network method. In this they have prepared their own data set of various different kinds of shoes by crawling on different online sites. They have used Hierarchical grouping (tree-structured) semantic shoe attributes. They have grouped the shoes in 3 groups mainly based on colour, toe shape, heel shape; these groups have various attributes. Multi-Task View-invariant Convolutional Neural Network is used for attribute prediction, style identification and view point invariance. Above CNN deep triplet ranking loss function is used as similarity comparison metric used to categorize images.

The paper depicts [8] performance of CBIR system using colour features using CN – colour descriptor. It provides relevant feedback to the user based on retrieved images. In this method RGB query images are converted to grey scale, their edge map is calculated and similarity measure is done based on minimum distance metrics i.e. the Canberra Distance Measure. The model proposed in this paper lacked user's expectations.

III. SYSTEM MODEL

Online shopping websites host around 500 million products of various categories addressed for specific audiences. These products are sold faster than ever because of the tremendous use of e-shopping.

Previously, the customer had to search for particular item in different categories by applying filters. Now using CNN algorithm the process of searching similar image is simplified and hence will reduce online shopping time.

Convolutional Neural Network (CNN) also known as ConvNet is an emerging deep learning algorithm which takes image as input also known as the query image, assigns importance values i.e. biases and learnable weights to different aspects or objects present in the image which makes it possible to differentiate one image from the other. A CNN requires certain pre-processing methods but the best feature of it is that it requires much lower pre-processing with respect to other classification algorithms. While in simple methods, filters are manually applied and are hand-engineered, but with enough training, CNNs acquire the ability to learn from these filters. The main aim of the convolution operation is to withdraw high-level and multi-dimensional features such as edges, colours, shapes, etc. from the query image. In a particular application of a CNN, there might be one or more number of convolutional layers depending on the complexity of that application. Classically, the first layer is responsible for capturing the lower-level features. As the number-of-layers keep on adding, the architecture to various high-level features giving rise to a dense network. Similar in working to the convolutional layer, the pooling layer is accountable for reducing the spatial size of the convo feature. This is done in order to lower the computational power that is required to process the data obtained through dimensional lowering. The convolutional layer and the pooling layer together, mainly from the layers of a CNN. After going through the above process, the system is now efficient enough to deduce and understand the features of the images. [1]

There are various architectures of CNNs present which can be used in varying fields. These architectures act as key features in building effective systems and applications as they employ many features together. Some of the architectures are below:

1.) LeNet 2.) AlexNet 3.) VGGNet
4.) ZFNet 5.) GoogLeNet 6.)ResNet

The proposed work uses CNN as a feature extractor instead of any normal feature extractor. CNN has become an important research topic in the field of machine learning and computer vision. So the basic purpose of using CNN as a feature extractor is to match CNN based CBIR system with conventional CBIR system in order to find if it's effective any way.

There are three different layers in all CNN architectures they are [5, 9, 10]:

A) Convolutional Layer

Images are represented as matrices of pixels. In this layer arithmetic operation between image matrix and filter (kernel) matrix is done. Consider image has size n*n and filter has size m*m. This filter contains weights for calculation, by this calculation information is extracted from images matrix. Weights in filter combining might be extracting edges, whereas another might focus on colour or they may remove the noise. This filter(m*m) passes through every pixel in image matrix(n*n) generating convolutional output matrix. One operation between the image matrix and filter produces one output value which is stored in a convolutional output matrix. There may be many convolutional layers possessing filters for extracting different features. Initial layer extracts basic features, as the network connectivity increases i.e. becomes deeper the next convolutional layer extracts complex and extremely fine features. More clear and proper features from the image and help in better prediction of results.

Suppose that the filter matrix moves across the input image matrix by one pixel at a time to cover whole image, then that parameter value one by which filter moves across the image matrix is called stride.

While moving filter by stride value over the image matrix the information present at the border of image matrix is not extracted properly as filter moves very less times through it. The output convolution matrix dimensions are less than image matrix dimensions as it depends on filter size. To overcome this both drawbacks extra layers at the border are added to image matrix. This process is called padding. Usually the same values added in padding layers. If zero is taken as values of pixels then it is referred as zero padding. Due to padding feature information at border is extracted properly also the dimension of output convolution matrix is kept same as the input image matrix.

B) Pooling Layer

This is an optional layer used for reducing the size of number of parameters when input image size is too large. This layer is added after the convolutional layer and usually added periodically in model. Only motive behind the pooling layer is to reduce the spatial image. Depth of image remains same after the pooling process. There are different ways for pooling and max pooling is a generally used technique.

The output dimensions of images depend on three variable components such as number and size of filter, Size of stride and size of pad added during padding. The output size can be calculated using a simple formula deduced by these variables. The formula is represented as

$$[(i + f + 2p)/s] + 1 \quad (1)$$

Where, i is input image size, f is size of filter, p is padding size and s is value of stride by which filter is moving.

C) Fully Connected Layer

The previous two layers that are convolutional and pooling layers are only useful for feature extraction and reduction of number of parameters present in image. For generating final output we required a fully connected layer. The training of models based on the features extracted by Convolutional layer is performed by fully connected layers. This layer is same as a normal neural network and possesses loss functions to reduce error in prediction. Similar to normal neural network backpropagation is done for updating weights and biases and for error reduction. After the execution of fully connected layer classification of images based on features extracted is done for application purpose.

The process can be explained in detail, flow of the working of the CNN algorithm for feature extraction (refer Fig. 1):

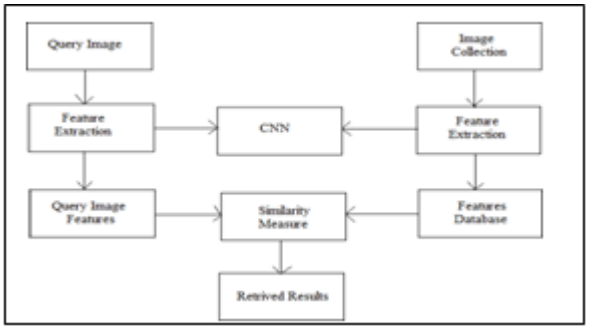


Fig.2. Workflow of system

The query image uploaded by the customer will be scanned, and the products having similar features to the query image will be extracted and shown as a result to the customer.

The major functions that product must perform are:

- Enable searching for similar products
- Enable uploading of query image of the product that is desired
- Be able to crop the image as and when necessary

Features of the images database are extracted by using CNN and stored in a file. Like wisely CNN extracts the features of Query images (QI) Features of QI at run time and these features are then compared one by one with the database images similarity is measured and higher the similarity higher the indexing rank.

IV. METHODOLOGY

In this paper we have used Visual Geometry Group 16 (VGG16) architecture of CNN as it has smaller network of layers and more desirable results. It is also known as Oxford net. It has of 16 layers convolutional layers, 5 pooling layers and 2 fully connected layers. Max Pool 2*2

VGG16 Architecture is represented as:

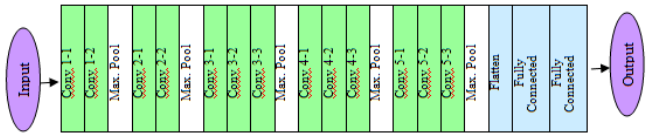


Fig.3:VGG16 net Architecture

We have used a pre-trained VGG16 net for the project which is trained on Image net data set. VGG takes image as an input. Input image is given to the 1st layer of convolutional layer of VGG net. Images are represented as matrices of pixels of height weight as 224 x 224 and accept 3 channels that are RGB – Red, Green and blue coloured images as input. These convolutional layers are used as filters whose parameters are to be acquired though learning. Convolutional layers along with Rectified Linear Unit (ReLu) activation function are used to learn new features with less similarity and error rate is also minimized. ReLu function is represented as:

$$f(y) = 0 \; if \; x < 0 \; || \; = x \; if \; x \geq 0 \qquad (2)$$

Here, x is the input. The convolutional layers extracts fine, complex and clear features which help in better prediction. The output convolutional matrix has size less than then image matrix which depends on the filter the paddings(extra layers added to the image matrix) are added to extract features from the borders properly. The output of convolutional layers is then passed to max pooling layers.

Max pooling reduces the number of parameters and takes only prominent features and is used to scale in variant features and helps reducing over-fitting. This reduces the size of matrix by dividing the matrix in parts and taking the maximum value.

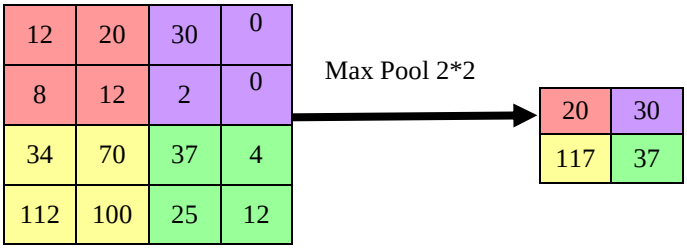


Fig.4: Max Pooling on a matrix

The output from the convolutional and pooling layers is in the form of matrix which is then converted to vector by flatten layer and is given to fully connected layer. Fully connected layer is a feed forward neural network in which information moves only in forward direction and is used for classification. These layers are similar to artificial neural neural network and they perform similar computational operations. The model processes these images and gives output as a vector. This vector consists of classification probabilities. The softmax function is used at the end so that all these probabilities add up-to 1. Output vector can be represented by:

$$Softmax(x_i) = \frac{\exp(x_i)}{\sum^{j=N} \exp(x_j)} \qquad (3)$$

Here N is the number of classes, x is the input vector. The features of the dataset are calculated and stored in a file then at run time the features of query image are calculated.

Dot product or inner product is used for measuring the similarity between the query image vector and image vectors from the database. This helps to find the set of images similar to the query image. The different product obtained from different images from the database are then sorted and ranked. This helps to view top 5 similar images.

Efficiency is calculated using mean average precision. It is given by

$$MeanAveragePrecision(mAP) = \frac{\sum^{q=Q} AP(q)}{Q} \qquad (4)$$

$$AveragePrecision(AP) = \sum^{i=N} \frac{P(i)}{R(i)} \qquad (5)$$

Here mAP is average precision of query q; Q is the total number of queries. Average Precision is sum of product of precision (P) and recall (R)

V. RESULT ANALYSIS AND DISCUSSION

After referring various papers we get to know about different CNN architectures like LeNet, Alex net, VGG net, Google Net, ResNet etc.

All these architectures require Graphical Processing Unit (GPU) for better processing and efficiency of results. The accuracy for these architectures differs depending on the dataset – number of images, parameters, classes to be classified. Alex net it is a deep CNN it takes input as RGB image and gives output as a probability vector. It consists of 8 convolutional layers and works with 60 million parameters. It consists of convolution layers, ReLu activation function, max pooling, and dropout/data augmentation. It has a top5 Accuracy of 84% and error rate is around 16%.

VGG net i.e. visual graphic group architecture it's steady in parameters while depth increases. It ranges from 11 to 19 convolutional layers. It consists of 3 by 3 convolution layers with pooling and filters. Decision function becomes complicated because of non-linearity. It takes image as input and returns a layer array. It works with 138 million parameters. This architecture has fixed kernel sizes. It has top5 Accuracy of 90% and has an error rate of 10%.

Google Net is also known as inception. It has 20 layers and 25 million parameters. Its basic focus is to reduce computer complexity. Inception block is an ensemble type of method; it forms clusters of images with high correlation. This architecture uses parallel kernels. It has top5 Accuracy of 93% and has an error rate of 7%.

ResNet or residual net is an artificial neural network that forms pyramidal network and performs jumps or skips layers i.e. shortcuts; this helps to increase the speed. Multiple parallel jumps are referred as dense nets. It has a top5 Accuracy of 95% and error rate of 5% and works with 60 million parameters.

In case of Alex net and VGG net accuracy can be improved by increasing the no of layers. As this cannot be done always we need to take in consideration the new methods like Google Net and ResNet but training them requires a lot of time and powerful GPUs. [11, 12, 13]

Table II. Comparison between accuracy and error of CNN architecture

CNN Architecture	Top 5 Accuracy (%)	Error Rate (%)
Alex Net	84	16
VGG Net	90	10
Google Net	93	7
ResNet	95	5

Note: All the values of accuracy and error are approximate; they may change depending on the dataset.

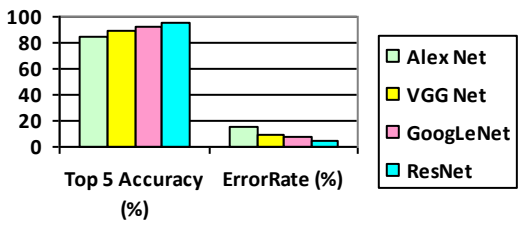


Fig.5. Comparison between accuracy and error of CNN architecture

For this system, we have taken 30 various categories in order to implement online shopping using the Caltech256 dataset. Deep learning method of CNN that we have used to model is VGG16 which consists of 16 layers, which is pre-trained on ImageNet dataset.

The Top 5 accuracy of the model for our dataset is 81.4%. Time taken to display the results is 516 ms.

Set of similar images were displayed based on the query image as shown below:

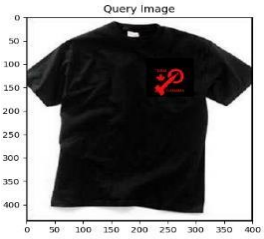


Fig.6. Query Image

Fig.7. Top 5 Retrieved Images

Table III. Results

Top 5 Accuracy (%)	82.4
Time (ms)	516

The suggested method in this paper makes it easy to find the product present on online shopping websites as well as provides retrieved products, material and objects which are very similar to query image. The speed and accuracy of this search is considerably increased by focusing only on the region of interest and particular category of product.

VI. CONCLUSION AND FUTURE WORK

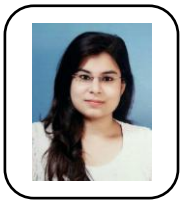
The system present in the paper is an efficient way of implementing online shopping by using an efficient CBIR technique in which for feature extraction the CNNs VGG net architecture is used. Extraction of features is quite accurate in CNN.

Then functions are used to match the similarities present in query image and database images then they are ranked accordingly. The proposed system initiates a very suitable way of employing the CBIR technique to the e- commerce domain. The system study in this CBIR system is done by CNN for characteristic extraction and it can become a base for extended advanced approaches to CBIR in future. To improve the efficiency of this CBIR system techniques like classification and clustering provide prominent direction for research. Further, the efficiency of the system can be improved by adding the user relevance and object detection for better results.

REFERENCES

1. Mutasem Alsmadi, "An efficient similarity measure for Content Based Image Retrieval using memetic algorithm", Taylor and Francis, 2019
2. Ruigang Fu, Biao Li, Yinghui, Ping Wang - ATRKey Lab, "Content-Based Image Retrieval Based on CNN and SVM" - National University of Defense Technology, 2nd IEEE International Conference on Computer and Communications, 2016
3. Amjad Shah, Rashid Naseem, Sadia, Shahid Iqbal, and Muhammad Arif Shah,"Improving CBIR Accuracy using Convolutional Neural Network for Feature Extraction"- Department Of Computer Science, City University of Science and Information Technology
4. Joel Pyykko and Dorota G, "Interactive Content-Based Image Retrieval with Deep Neural Networks", LNCS 9961, pp. 77–88, 2017
5. GuoyongDuana, Jing Yanga, Yilong Yanga,"Content-Based Image Retrieval Research", International Conference on Physics Science and Technology (ICPST 2011)
6. OuhdaMohamed, El Asnaoui Khalid, "Content-Based Image Retrieval Using Convolutional Neural Networks", Springer- 2019
7. HuijingZhan ,Boxin Shi, Ling-Yu Duan, Alex C. Kot, "DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval", Science Direct, Computer Vision and Image understanding 180 (2019) 23-33
8. S.Rubini, R.Divya, G.Divyalakshmi, Mr T.M. Senthil Ganesan, "Content Based Image Retrieval", International Research Journal of Engineering and Technology, Volume: 05 Issue: 03, Mar-2018
9. Y. Jing, D. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel,"Visual search at pinterest," in Proceedings of the 21th ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining.ACM, 2015
10. K. Lin, H.-F. Yang, K.-H. Liu, J.-H. Hsiao, and C.-S. Chen, "Rapid clothing retrieval via deep learning of binary codes and hierarchical search," in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, 2015
11. "Alex net VGG and Inception Architectures, Deep Learning in Computer Vision", National Research University Higher School of Economics, Coursera
12. Fei-Fei Li & Justin Johnson & Serena Yeung , "Lecture 9 – CNN Architecture", Stanford University, 30 April 2019
13. Aqeel Anwar, "Difference between AlexNet, VGGNet, ResNet and Inception", medium- Towards Data Science, Jun 7, 2019

AUTHORS PROFILE



Shweta Sadwani is a student in the final semester of Bachelor of Technology in Computer Science Engineering at MIT Academy of Engineering, Pune. Her areas of interest include Data Science and Analytics and Machine Learning. She has done various projects and research in Data Science and Machine Learning

Vaibhavi Sangawar is a student in the final semester of Bachelor of Technology in Computer Science Engineering at MIT Academy of Engineering, Pune Maharashtra India. Areas of interest include Data Analytics, Business Intelligence, Image Recognition and Content Based Image Retrieval. Member of Computer Society of India committee(CSI). She has done project and research in Data Science and Machine Learning. Recruited by Amdocs as Software Engineer Associate.
.

Rushabh Sanap is a student in the final year of Bachelor of Technology in Computer Science Engineering at MIT Academy of Engineering Pune. Areas of interest are Computer Vision, Image Recognition and Processing and Content Based Image Retrieval. He is a Member of Computer Society of India committee. Have done research in Deep Learning and Machine Learning. He is selected as Business Technical Analyst at Deloitte USI.
Email sanaprushabh2016@gmail.com.

Akanksha kakade is a student in the final semester of Bachelor of Technology in Computer Science Engineering at MIT Academy of Engineering, Pune. Her areas of interest include Image Recognition, Cyber Security and Content based Image Retrieval. She is a member of Computer Society of India committee (CSI).
E-mail: akanksharkakade@gmail.com

Minakshi N.Vharkate received the BE degree in Computer Science and Engineering College of Engineering Osmanabad, affiliated to Dr. B.A.M.U Aurangabad University, India in 1999 and M.E. in Computer Science and Engineering from Walchand College of Engineering, Sangali affiliated to Shivaji University, Kolhapur, India in 2007. She is pursuing a degree in Computer science and Engineering from Dr. B.A.M.U Aurangabad. She is currently working as a Sr. Assistant Professor at MIT Academy of Engineering, Alandi(D), Pune. Her research the area is Remote sensing image Processing, Machine Learning, and Data Science.