Content-Based Image Retrieval System

Anjali MT20082 | Shradha Sabhlok MT20069 | Akhil Mahajan MT20107 | Prabal Jain MT20115

1. PROBLEM FORMULATION

Image Data is growing at an exponential rate in today's world, where we are surrounded by platforms like Instagram, Facebook, and Twitter, to name a few. The traditional search systems mainly rely on the user entering a text query, which can differ from the tags/labels attached to the images. This creates not so relevant results many times. We are targeting to bridge this gap using a Content-Based Image Retrieval System where the input query will also be in the form of an image that the user wants to obtain outputs similar to. This involves extracting relevant information from the image itself, like colors, edges, shapes, texture, etc.

A few CBIR systems have already been in place, but the major challenge involved is the trade-off between the system's efficiency in terms of time and accuracy. The current CBIR systems involve a higher computation cost, and hence there is a need to work on its performance improvement. So, in our proposed method, we are targeting to improve the performance of the Image Retrieval system to not only get a high accuracy but also to decrease the retrieval speed of the system.

2. LITERATURE REVIEW

There are various papers written in the area of Content-Based Image Retrieval systems. Some of the important literature which we covered to get better insight into the field are as below:

Pattanaik and Bhalke [1] dealt with retrieving top matching images for a given query image and retrieval is done using a large database. This paper has used both low level (color, texture, shape, etc.) and high level (dependent on human perception) features of images to get the desired results. Multiple features are utilized by Author to reduce the semantic gap while feature extraction. CBIR is the main focus area of the Author, which is a two-step process: Feature Extraction and Similarity Matching. For similarity matching, the Author used the Euclidean Distance method.The Author experimented with different features and their combinations like Gray Histogram (representing the distribution of intensity of color in the image), Color Histogram, Color Mean (mean of pixel colors), using color and texture, etc. It was concluded that combined features give better results.

Krizhevsky et al. [2] in his paper has tried to use deep Convolutional Neural Network (CNN) architecture for Computer Vision. It was the first break-through in the ImageNet classification challenge (LSVRC-2010, having 1000 classes). ReLU was a key aspect of reducing training time. Multi-GPU training, local response normalization, and overlapping clustering are applied across the network for better performance. To combat overfitting, the Author used two data augmentation methods (namely, generating image translations, altering the intensity of RGB channels) and a Dropout method. CNN has a large learning capacity. Through this paper, it was deduced that depth is critical to get desired results. Removing a single convolution layer degrades the network's performance. The results still have many orders of magnitude to go in a direction to match the human visual system.

Babenko and Lempitsky [3] in their paper have considered descriptors based on activations of pre-trained deep CNNs. They have also compared the distribution properties of deep convolutional features and SIFTs and proposed a new global image descriptor that avoids the embedding step necessary for SIFTs. They have described the SpoC descriptor which was based on the aggregation of raw deep convolutional features without embedding, which starts with Sum Pooling of deep Features and then Centering prior, as the object of interest is tend to be located close to the geometrical center of an Image, SPoC descriptor can be modified by using Simple weight Heuristic. They have performed image retrieval on different datasets like INRIA Holidays dataset, Oxford Building dataset. SpoC with Center Prioring performs better than Fisher Vector method, Triangular Embedding method, and Max Pooling method.They proposed SPoC features provide a considerable improvement over previous state-of-the-art for compact descriptors.

Wang et al. [4] in his paper proposed a new image retrieval algorithm based on SIFT feature matching. In this a fraction of the image was given as an input to the Algorithm, in which Height and width of that region are defined by the user and then extracting features of each image in the database from training images and ROI (Fraction of an Image) by using SIFT to gain feature key points and then finding candidate matching features based on Euclidean distance of their feature vectors. Also, they used a Dynamic Probability function instead of fixed value feature matching threshold to judge the matching. To identify whether feature matching is successfully done or not, they used Dynamic Ratio of distance in which feature keypoint of ROI and the first and second nearest neighbor keypoints by Euclidean distance such that, if the value, which is the nearest distance divides the second-nearest distance, is less than a certain ratio of distance, feature matching is achieved.

Jain and Dhar [5] in their paper, have explored the applications of CNNs towards solving classification and retrieval problems. The authors investigated an architecture of deep learning for CBIR systems by applying an advanced deep learning system, that is, CNNs for studying feature representations from picture data. Overall, their approach is to retrain the pre-trained CNN model, that is, Inception-v3 model of GoogleNet deep architecture on our dataset. Then, the trained network is used to perform two tasks: classify objects into its appropriate classes, and perform a nearest-neighbors analysis to return the most similar and most relevant images to the input image. For retrieval of similar images, they used transfer learning to apply the GoogleNet deep architecture to the problem. Extracting the last-but-one fully connected layer from the retraining of GoogleNet CNN model served as the feature vectors for each image, computing Euclidean distances between these feature vectors and that of the query image to return the closest matches in the dataset.

Chen et al. [6] in his paper has developed a novel scheme based on one-class SVM, which fits a tight hyper-sphere in the nonlinearly transformed feature space to include most of the target images based on the positive examples. The use of kernel provides an elegant way to deal with nonlinearity in the distribution of the target images, while the regularization term in SVM provides good generalization ability. To validate the efficacy of the proposed approach, they test it on both synthesized data and real-world images. Further, statistical learning method is used to attack the problems in content-based image retrieval. They developed a common framework to deal with the problem of training with small samples. Kernel machines provide us a way to deal with non-linearity in an elegant way. Their strategy is to map the data into the feature space and then try to use a hyper-sphere to describe the data in feature space and put most of the data into the hyper sphere. This can be formulated into an optimization problem. The aim is to get the ball to be as small as possible while at the same time, including most of the training data.

Choudhary et al. [7] in his paper has implemented CBIR system by using Color Moment (CM) and LBP for feature extraction and Euclidean distance to compare database images and query image. In this paper, Wang database has been used for analysis. A query image is taken as an input. CM and LBP are applied on both query image and database images and both these features are combined to get a combined feature vector for each image. Then, Euclidean distance is calculated between the feature vector of query image and feature vectors of all the database images. The images with the least distance from the query image are retrieved based on a specified threshold.

Chadha et al. [8] in his paper has implemented CBIR by using Average RGB, Color Moments, Co-occurrence, Local Color Histogram, Global Color Histogram, Geometric Moments as feature extraction techniques and Euclidean distance to find the similarity between images. The place where this paper outshines other papers is the way it tries to optimize the above process resulting in much better accuracy. It uses Wang Database for its analysis and defines three parameters for analysis namely: Time, Accuracy and Redundancy Factor. Redundancy Factor is calculated by subtracting total images in a class from the total number of images retrieved and then dividing them by the total number of images in a class. The ideal RF should be 0 but can range from -1 to 9.The authors got below 50% accuracy if work was done by using individual extraction techniques. So, the authors combined these features into a single feature vector which resulted in 91.51% accuracy. The authors then cropped the images and saw a significant jump in the accuracy as cropping an image reduced the unwanted information of an image and thus helped in increasing accuracy for the desired result.

3. DATASETS

We will be working with three main datasets in our project which are mentioned below:

- **CBIR-50:** It consists of 10,000 images, which are clustered into 50 categories namely Mobiles, India Gate, Kangaroo, Jeans etc. , each category has 200 images of varying sizes.
- **Oxford:** The Oxford Buildings Dataset consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries.
- **Paris:** The Paris Dataset consists of 6412 images collected from Flickr by searching for particular Paris landmarks in 12 categories.The Paris Dataset consists of images provided by Flickr.

Figure 1: CBIR50 Dataset

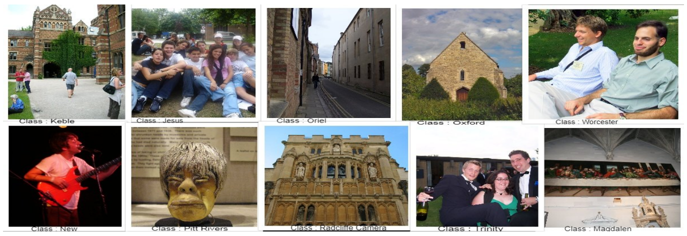

Figure 2: Oxford Dataset

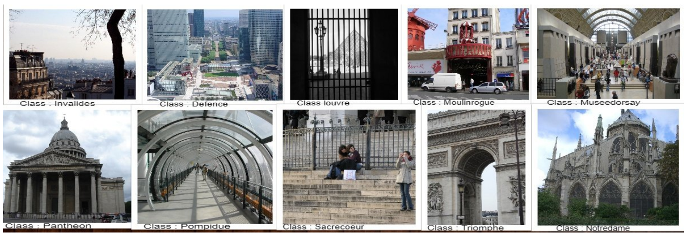

Figure 3: Paris Dataset

4. PREPROCESSING

All the images in the dataset and the query image are resized to 224*224 coloured images. Feature extraction techniques like HOG, SIFT, KAZE and SURF are applied and Normalization is performed over the extracted features. Dimensionality reduction techniques like PCA and LDA are used to reduce the dimensions and avoid 'Curse of Dimensionality'. For deciding the n_components of PCA, variance-components graphs are used. All the features are stacked together to get the complete image representation.

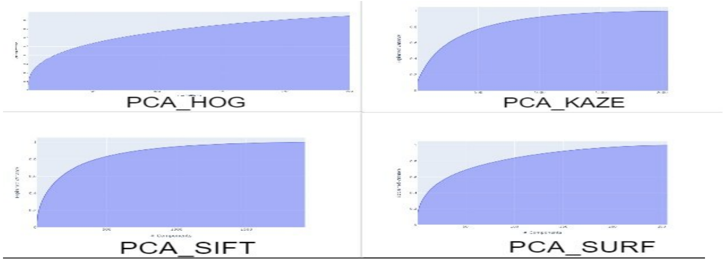

Figure 4: PCA on CBIR50 Dataset

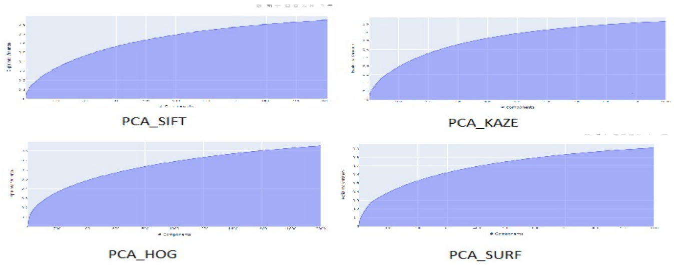

Figure 5: PCA on Oxford Dataset

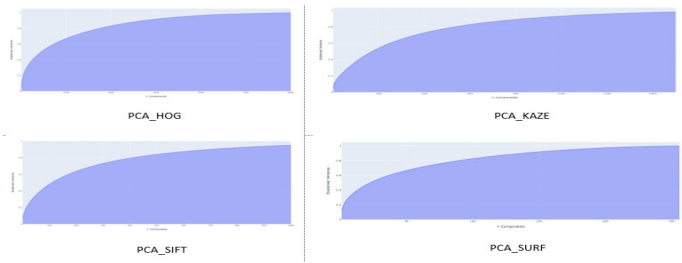

Figure 6: PCA on Paris Dataset

5. IMPLEMENTED METHOD

The implemented method follows the following algorithm:

1. All the images in the image database are reshaped to the size 224*224*3.
2. Features are extracted using feature extraction techniques like KAZE, SURF, SIFT and HOG.
3. PCA and LDA is applied to reduce the dimensions of the extracted features.
4. All these feature vectors are then combined to form a single feature vector.
5. Query image is taken as an input and all the steps from 1-4 are applied on query image as well.
6. Class to which the query image belongs is predicted using 3 models namely XGBoost, Decision Tree and SVM.
7. Once a class is predicted, cosine distances are calculated between the query image and all the images belonging to the predicted class.
8. All the distances are then sorted in an ascending order (since image with least distance will have the highest similarity) and first N images with the least distances are stored/printed. Here N is the number of matching images we want to store/print.

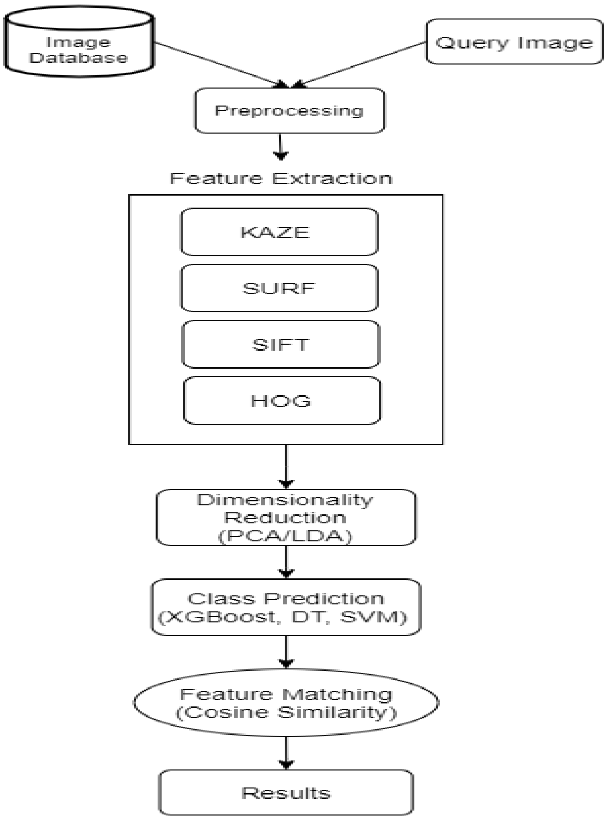
Image Database
Query Image
Preprocessing
Feature Extraction
KAZE
SURF
SIFT
HOG
Dimensionality Reduction (PCA/LDA)
Class Prediction (XGBoost, DT, SVM)
Feature Matching (Cosine Similarity)
Results

Figure 7: Implemented Method Workflow

6. BASELINE RESULTS

Results achieved after predicting classes using various models are provided in below tables:

	Precision	Recall	F1	Accuracy
SVM	0.9998	0.9987	0.9992	0.9990
XGBoost	0.9983	0.9908	0.9944	0.9940
Decision Tree	0.9998	0.9987	0.9992	0.9990

Table 1: Oxford Dataset

	Precision	Recall	F1	Accuracy
SVM	0.9753	0.9962	0.9846	0.9897
XGBoost	0.9737	0.9727	0.9731	0.9810
Decision Tree	0.9822	0.9795	0.9808	0.9842

Table 2: Paris Dataset

	Precision	Recall	F1	Accuracy
SVM	0.9995	0.9995	0.9995	0.9994
XGBoost	0.9994	0.9995	0.9994	0.9994
Decision Tree	0.9959	0.9955	0.9956	0.9959

Table 3: CBIR50 Dataset

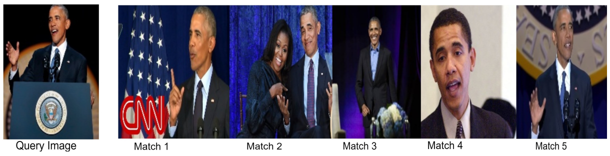

Figure 8: Query1 on CBIR50 Dataset

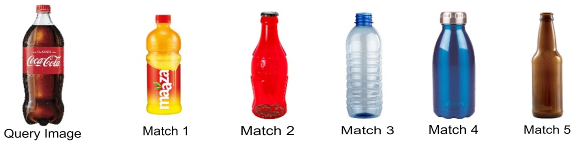

Figure 9: Query2 on CBIR50 Dataset

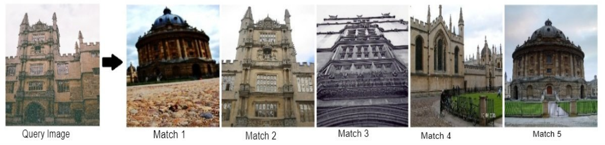

Figure 10: Query1 on Oxford Dataset

7. PROPOSED METHOD

In our previous method, a multi-feature image retrieval method was introduced by combining the features extracted using HOG, SIFT, SURF, and KAZE and using Dimensionality Reduction Techniques such as PCA and LDA. After observing our model, we came to the conclusion that retrieving

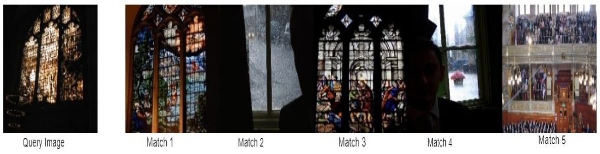

Figure 11: Query2 on Oxford Dataset

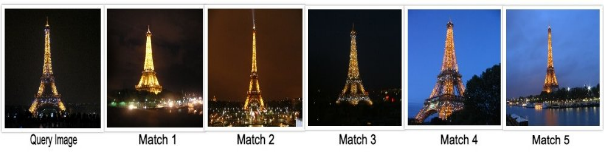

Figure 12: Query1 on Paris Dataset

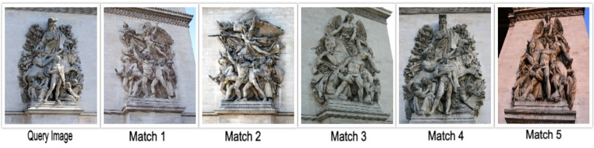

Figure 13: Query2 on Paris Dataset

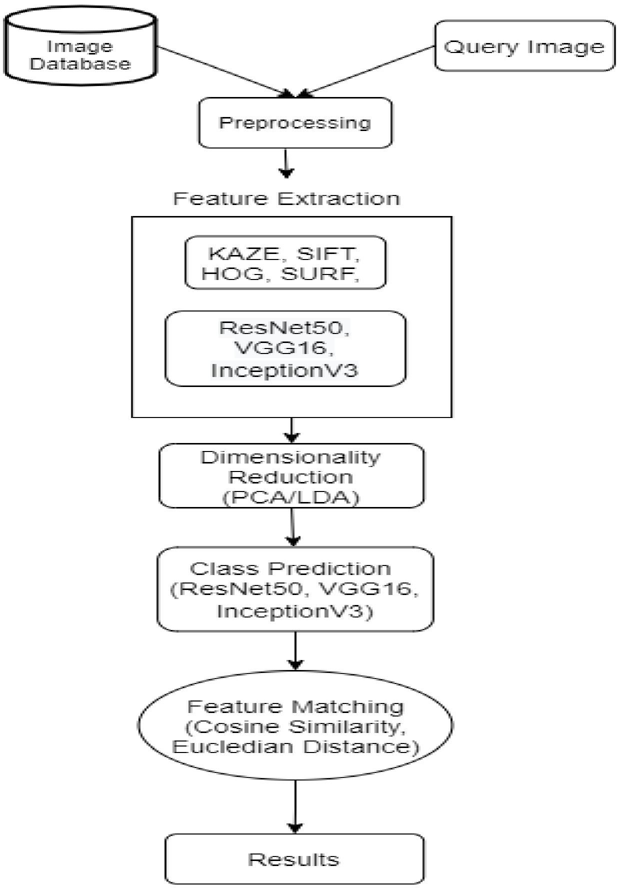
Image Database
Query Image
Preprocessing
Feature Extraction
KAZE, SIFT, HOG, SURF,
ResNet50, VGG16, InceptionV3
Dimensionality Reduction (PCA/LDA)
Class Prediction (ResNet50, VGG16, InceptionV3)
Feature Matching (Cosine Similarity, Eucledian Distance)
Results

Figure 14: Proposed Method System

images from a dataset based on feature matching is efficient only if we obtain good feature vectors from the images.

So, to achieve high precision and accuracy, we are proposing another method to retrieve images with the help of Deep Learning. In this model, each image in the image database will be represented by a Feature Vector obtained with the help of pre-trained CNN models such as VGG16, ResNet50 and InceptionV3. Ensembling will be performed by combining the Feature vectors returned by these models and the ones from techniques like HOG, SIFT, SURF and KAZE for getting much better feature vectors. After this, dimensionality reduction techniques like PCA and LDA will be applied on these feature vectors. A query image will be taken as an input and all the above techniques will be applied on it.

We will then train above Deep Learning Models (convolutional) by altering the last (Dense) layer of our Deep Neural Network which will help to predict the classes with much better accuracy. After predicting the class for query image, similarity scores between the query image and the images belonging to the predicted class will be computed using various feature matching techniques like euclidean distance and cosine similarity. Top-N images with the highest similarity or least distance will be retrieved. Here, N will be the number of images we want to retrieve.

References

[1] Swapnalini Pattanaik and D Bhalke. “Beginners to content-based image retrieval”. In: *International Journal of Science, Engineering and Technology Research* 1 (2012), pp. 40–44.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[3] Artem Babenko and Victor Lempitsky. “Aggregating deep convolutional features for image retrieval”. In: *arXiv preprint arXiv:1510.07493* (2015).

[4] Zhuozheng Wang, Kebin Jia, and Pengyu Liu. “An effective web content-based image retrieval algorithm by using SIFT feature”. In: *2009 WRI World Congress on Software Engineering*. Vol. 1. IEEE. 2009, pp. 291–295.

[5] Surbhi Jain and Joydip Dhar. “Image based search engine using deep learning”. In: *2017 Tenth International Conference on Contemporary Computing (IC3)*. IEEE. 2017, pp. 1–7.

[6] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. “One-class SVM for learning in image retrieval”. In: *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*. Vol. 1. IEEE. 2001, pp. 34–37.

[7] Roshi Choudhary, Nikita Raina, Neeshu Chaudhary, Rashmi Chauhan, and RH Goudar. “An integrated approach to content based image retrieval”. In: *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2014, pp. 2404–2410.

[8] Aman Chadha, Sushmit Mallik, and Ravdeep Johar. “Comparative study and optimization of feature-extraction techniques for content based image retrieval”. In: *arXiv preprint arXiv:1208.6335* (2012).