ACCELERATING CONTENT-BASED IMAGE RETRIEVAL: A PARALLEL COMPUTING APPROACH

REVIEW REPORT

Submitted by

Samyak Singh Chauhan(21BCE5245)
Arnav Girdhar(21BCE5250)
Jashan Jindal(21BCE5313)

Prepared For

HIGH PERFORMANCE COMPUTING (BCSE414L)

Submitted To

PROF. S.K. AYESHA

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT

"An image speaks volumes, encapsulating a wealth of information that may not always be discernible to the human eye. The exponential growth of the internet has led to an information overload, making it challenging to retrieve specific information. This is particularly true for visual data, with the internet teeming with images and other visual content such as videos and movies in various formats like JPG, PNG, BMP, and GIF. This necessitates an image search engine capable of finding related and exact images.

An Image Search Engine that employs a content-based image retrieval (CBIR) query technique requires a sample image to base its search on. CBIR aims to devise methods to index, browse, and query vast image databases using significant feature extraction and comparison methods. The CBIR system uses feature extraction to obtain the most pertinent information from the original image and represent it as a reduced set of features like texture and shape. This process involves algorithms that process the image data and store them. Given the complexity of feature extraction and data representation from image contents, various techniques and algorithms need to be employed to enhance efficiency and reduce time consumption.

However, for these algorithms to be effective in real-world applications, they need to execute in the shortest time possible to boost system performance. This can be achieved through parallel computing, which allows for the simultaneous execution of multiple parts of a program, thereby maximizing CPU utilization. This makes the program run faster and more efficiently. Therefore, implementing parallelization in image search can significantly reduce retrieval time and enhance the performance of the retrieval system, which is crucial in any search application."

INTRODUCTION

Huge collection of digital images is collected due to the improvement in the digital storage media, image capturing devices like scanners, web cameras, digital cameras and rapid development in internet. Due to these reasons there is a need to implement new programs in such a that it gives the best performance using the same resources as before and at the same time reducing the time to retrieve images from the system. Image retrieval is achieved through low level features that are extracted from the images by the extraction algorithm and then these features are represented in a form called feature vector. These feature vectors are calculated statistical values like standard deviation. Similarity is measured to rank the images by calculating the distance between the query image feature vector and feature vectors of database images. Since this algorithm performs various number of computations in order to obtain the feature vectors for the image it usually executes the computation sequentially and finally obtain the required results. This method works but in order to increase the CPU utilization of the multi-processor systems to the maximum and to reduce the computation time we can make use of multithreading. With the help of multithreading heavy processes can be divided into multiple threads and execute them at the same time. With the application of multithreading into this process instead of sequentially computing the data, each thread can take a particular function assigned to them and these threads execute simultaneously and when joined gives back the result of the computation assigned to them. Here in this method each thread can handle the calculation of various tasks like data extraction, histogram refinement, calculation of feature vectors and finally produces the results of the computation simultaneously at the same making the process execution a whole lot faster since computation is performed in parallel rather than serial execution. Thus, have achieved the goal of improving the performance of the system with the use of multithreading to enable efficient and fast execution of the program.

Applications of feature extraction from images are limitless. This data can be used in classification, recognition of images in a huge database centers where processing needs to be fast and efficient. Their significant applications in security systems as it is the basis in biometric systems. Since users usually work with a huge number of images, it is important to achieve the highest performance possible from that code. To achieve this, we make use of parallelization.

LITERATURE SURVEY

A. Color Coherent Vectors (CCV) Technique based Image Query Based Search Engine

Color's coherence is the degree to which pixels of that color are members of large similarly colored regions. They have referred to these significant regions as coherent regions and observed that they are of significant importance in characterizing images. In this technique, we compute the color coherence vector for an image by using the edge change detection and Sobel filter design and stored the matrix generated in a repository. The general concept involved in edge detection techniques and also compares various methods used for edge detection. The main reason for using edge detection is to extract the important information of the image. The computed color coherence vector matrix for this image and compared it with the color coherence vector matrix of the previous image stored in the repository. If the difference between them comes out to be greater than 1000 then the difference between those images is considered significant and hence a change is detected.

Implementation

1.) Initially the crawler crawls a portion of the Internet and store the crawled URLs in the database for the purpose of implementing the proposed research work. The source codes of these web pages, along with the images on the web page were stored in the database corresponding to their web links.

2.) The user uploads the image for which he wants to find a match. The image that has been uploaded undergoes the process of edge detection technique called the Sobel edge detection has been used to extract the content of the uploaded and web images that are being compared with the images stored in the database by using the CCV matrix algorithm. The next step involves comparing the extracted content with the help of a method called the CCV matrix technique, which checks for content similarity between the images. The images stored in the database from the crawled sites also undergo edge detection right before they are compared.

3.) After the images are matched for similarity by using the CCV matrix technique, the match results are used further. The complete web links of the images that hold a match with the uploaded image are saved in a file and each of these links is searched for in the source code of that particular website to which the image belongs. It collects all those images whose content matches with that of the uploaded image. It then ranks those images based on the percentage of content found similar to the uploaded image and displays their corresponding websites right below the image.

B. Wavelet-based Image Feature Indexing for Image Search

Wavelet transformation is a multi-level decomposition of signals. It represents a signal as a superposition of basic functions called wavelets. Wavelet-based features select sixty-four largest Haar wavelet coefficients in each of the 3 color bands and stores them in feature vector as +1 or −1 along with their position in the transformation matrix. Low-frequency coefficients tend to be more dominant than those of the high-frequency coefficients and this makes this algorithm ineffective for images with sharp color changes. Few of the wavelet coefficients in the lowest frequency band and their variances are used as a feature vector. To decrease the retrieval time, a crude selection is done based on the variances. Wavelet transform analyzes the signal at various frequency bands and is norm preserving unitary transformation. As feature vectors of [11] include information about the high-frequency bands, it will be able to group similar images better than those feature vectors which considers only the low-frequency information.

Implementation

When a query image is submitted by a user, we need to compute the feature vector as before and match it to the pre-computed feature vector in the database. This is done in 2 phases.

1.) The first phase of matching is done using the wavelet feature matching. The key vector is computed for the given query image using the previously computed sub-band histogram. Then the B+ tree is traversed down starting from the first number in the key vector. When a leaf is reached, all the images stored in the leaf are retrieved for the second stage of matching. If the leaf contains no images similar to the query image, then a partial match of the image key is carried out, i.e. tree traversal is stopped before reaching a leaf and the images of the sub-tree are retrieved for further refined matching.

2.) Having a rough and fast pass followed by refined pass increases the speed of retrieval while maintaining the quality of the retrieval. Especially in large databases, these two pass retrievals have a very affirmative effect on the retrieval speed.

3.) In the second phase, we compute the feature vector for the given query image as stated in sec II.C. We compare this query feature vector with the target images (images which were retrieved by the first pass) feature vector using chi-square distance. The images are sorted based on the distance and the first few are presented to the user.

C. Filters based Feature Analysis for CBIR using Color Histogram Method

Retrieval of the images is performed in three phases for the analysis of histogram features based on three filter methods, median, median with edge extraction and Laplacian filters. Before feature extraction some preprocessing steps are performed on image to enhance up to certain level such that the required information in image becomes more apparent for extraction of features. When the input RGB image is acquired then in the first step it is converted into grayscale for the extraction of features. Grayscale requires fewer computations as compared to color image. The grayscale image is converted to histogram equalized image. Histogram equalization is a technique to convert image's intensity levels into equal levels such that to enhance contrast and image has more details. For further enhancement of image median and Laplacian filters are used. Image retrieval for median, median with edge extraction and Laplacian will be performed in three phases using histogram method. In this study the proposed a CBIR algorithm to find the solution of the problems in the CBIR. This algorithm is based on the color histogram refinement method using median and Laplacian filters as preprocessing steps to reduce the noise and provides enhanced sharpened images with more detail information. Color histogram is divided into bins. The number of regions is determined in bins. The statistical moments mean and standard deviation are calculated in each bin by using the areas of regions to get feature vector which is used for image retrieval.

Implementation

1.) In this phase histogram equalized image is filtered with median filter to further enhance the image. For features extraction color histogram refinement method is applied on filtered image. The filtered image is quantized into histograms of 32 bins using. In each quantized bin the number of connected regions is determined. Then mean and standard deviation are calculated in each bin using the area of regions. By using L as 32 then total 64 features are extracted and a feature vector FV is generated for the image. All the images are stored in database with feature vectors. The feature vector of the user query is generated in the same way. This query image feature vector is compared with database and distance is computed between them using. Thus, we get the retrieval result for median filter using histogram technique.

2.) In this phase median filter is used with edge extraction method. As during median filtration some amount edge information is lost. To restore edge information, canny edge detection method is used after median filtration. The features for database image and query image are extracted in the same way as in phase-1 using color histogram refinement method. The distance is computed between query image and database image using to retrieve relevant images.

3.) The Laplacian filter is applied on histogram equalized image in this phase. The Laplacian filter gives a sharpened and enhanced image using real valued image. The features are extracted in same way as using color histogram refinement method. The distance is computed between images using chi-square distance.

D. Compact Composite Descriptor (CCD) Based Image Search

Compact Composite Descriptors(CCD) applies various descriptor algorithms including CEDD, FCTH, BTDH, SCD for searching exact images, since they are global image features capturing both, color and texture characteristics, at the same time they are very useful in a very compact representation which is suitable for large image databases. This system will take an input image as a query image by browsing the image database. Once the image is selected, the applying algorithm is selected by which the system should retrieve similar images.

Implementation

1.) When the user gives a query image, the block-based low-level feature from an image is exacted basically in terms of intensity and texture contrast and then clustering of this feature space is done to form meaningful patterns. Various algorithms are used for feature space clustering like CEDD, BCTH, CLD, FCTH etc. These algorithms automatically define the number of clusters.

2.) Color Edge Directivity Descriptor is a composite image descriptor that stores the color and texture information of an image in its histogram. This descriptor is capable of retrieving accurate images even if the image has undergone much deformation such as noise, transformation, smoothing, and various illumination changes. Fuzzy Color and Texture Histogram is also a composite descriptor that is similar to the CEDD descriptor that stores image color and shape and texture information in its histogram. But unlike CEDD, it captures the texture information through the Haar Transform. So usually the results from both the descriptors are somewhere similar. Edge represents an important content of an image. These descriptor stores the variation of frequency and brightness of an image into its histogram hence called as Edge Histogram Descriptor. Brightness and Texture Directionality descriptor extracts the brightness, texture characteristics, and spatial distribution into a compact ID vector and stores it into its histogram. The most important characteristic of this descriptor is that its size adapts according to the storage capabilities of the application that is using it. Thus, it is suitable for large image databases.

3.) CCD will retrieve both similar and exact image. This CCD system makes use of various descriptor algorithms for searching exact images, since they are global image features capturing both, color and texture characteristics, at the same time. This system will take an input image as a query image by browsing the image database folder. Once the image is selected, the applying algorithm is selected by which the system should retrieve similar images. The result for all the images stored in the image database folder which is in terms of deviation i.e., the difference of dissimilarity between query image and images being used to check for similarity. To match the exact images as their deviation is 0 and hence can be used for matching the exact and similar images since they are resistant to many deformation and illumination changes. This system will retrieve the exact images from a given image database folder employing compact composite descriptors, will also try to retrieve similar images based on their visual features and modified versions of the query image.

CONCLUSION

The problem with the conventional image search engine is that they search images through their metadata which is not the images, but the data attached to those images known as image annotation. So, when one searches for any desired image, it also gets a lot of nonsense images. There are various conventional image search engines that are metadata-based like Google Image, Picsearch, Yahoo Image Search, Altavista Image Search, Pixy, Web shots, and Getty Images. The solution to this problem is Reverse Image Search which will search images, not on the metadata, but will search images based on their visual qualities like color, text, and shape. When this image search engine is given a query image it will not search only the related image, but will find similar or exact images, this is the concept of Image Search. Various algorithms and techniques available for searching the images based on query images.

REFERENCES

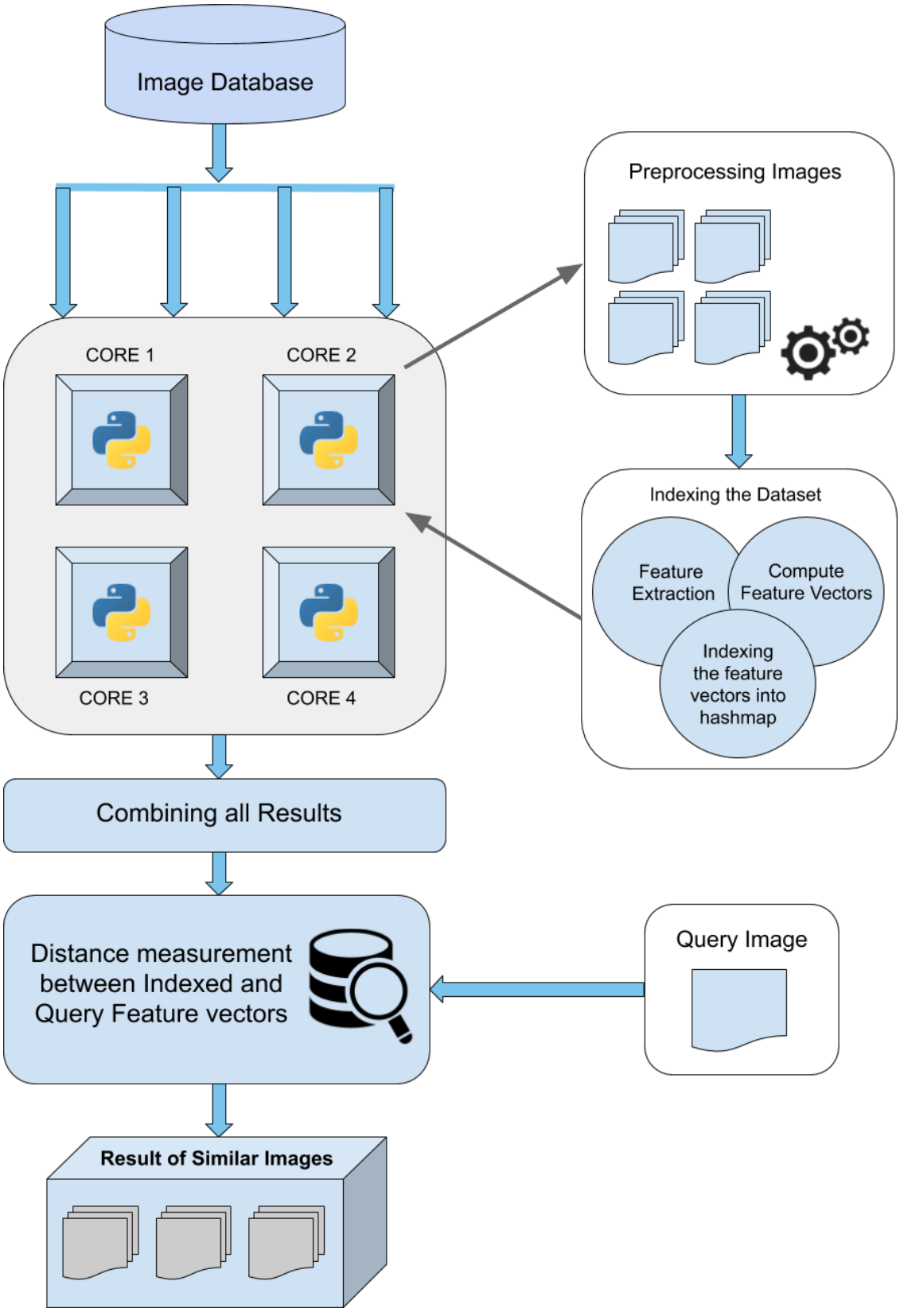
[1] G. Pass, R. Zabih, and J. Miller, "Comparing images using colour coherent vectors," Proceedings of Inforsning Science & IT Education Conference (InSITE) 2009.

[2] A. Lakshmi and S. Rakshit, "New wavelet features for image indexing and retrieval," 2010 IEEE 2nd International Advance Computing Conference (IACC), Patiala, 2010, pp. 145-150, doi: 10.1109/IADCC.2010.5423022.

[3] D. V. Ragatha and D. Yadav, "Image Query Based Search Engine Using Image Content Retrieval," 2012 UKSim 14th International Conference on Computer Modelling and Simulation, Cambridge, 2012, pp. 283-286, doi: 10.1109/UKSim.2012.48.

[4] Meshram, Kimaya & Agarkar, Ajay. (2015). Content Based Image Retrieval Systems using SIFT: A Survey. International Journal of Electronics and Communication Engineering. 2. 18-25. 10.14445/23488549/IJECE-V2I10P105.

PROPOSED ARCHITECTURE

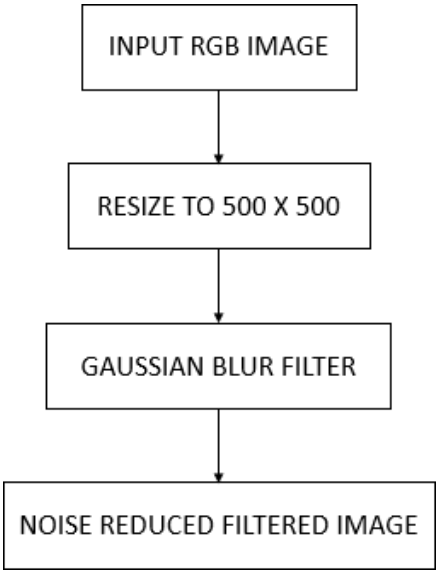


PROPOSED METHODOLOGY

The proposed architecture consists of primarily three modules which are preprocessing, indexing image, query searching. Firstly, the whole image database paths are parsed and stored in a list data structure. With the help of multithreading library, each core of the system can be assigned certain number of images thus dividing the total workload and processing the image data in parallel. Each core processes the images, indexes the features vectors it had extracted from the image and as each core completes indexing, the vectors are stored in a combined hash table with the key being the name of the image and the value being the respective feature vector of the image. The process is applied to the query image as well and it's feature vector is queried across the hash table values. The metric used to compare the query image vector with the database image vector will be a chi-square distance measurement and the distance metric results obtained with be sorted, and top 20 results will be returned.

PRE-PROCESSING:

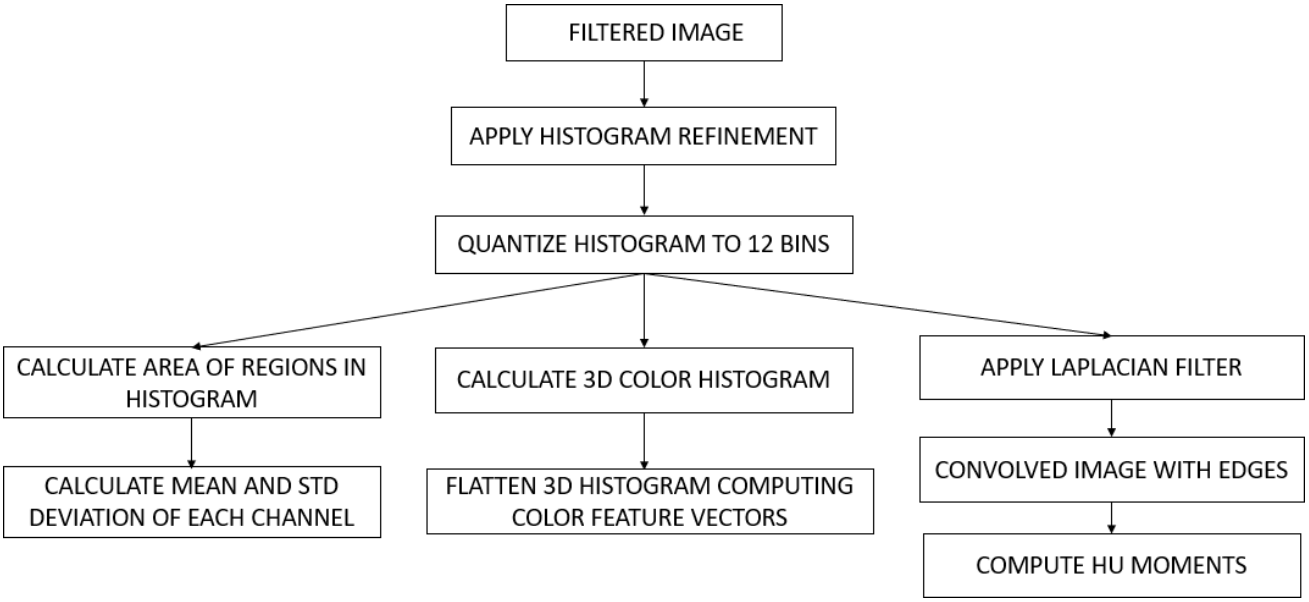
The first step in preprocessing will be process of noise reduction. Mostly images consist of some noise and unwanted information. They should be removed from images before processing for retrieval by using filters. Different filters method can be used for removal noise. A gaussian blur is applied on images for enhancement in the preprocessing step. The input image will be an RGB image of dimensions more than 2000 x 2000 which shall be reduced to 500 x 500 for ease of computation and to limit memory usage.



FEATURE EXTRACTION:

For feature extraction the color histogram refinement technique is used. Color histogram is quantized into 256 bins. Each bin is divided into connected regions of pixels using 4- neighborhood rule. The number of regions in each bin is determined. Then the area of each region is calculated. Two color moments are used to calculate features, the first-order moment called mean and second order moment called standard deviation. The mean represents the brightness of image and standard deviation represents the contrast. The dark image has low mean and bright image has high mean. The low contrast image has low and high contrast has high standard deviations. The means and standard deviations are calculated in each bin using the areas of regions.

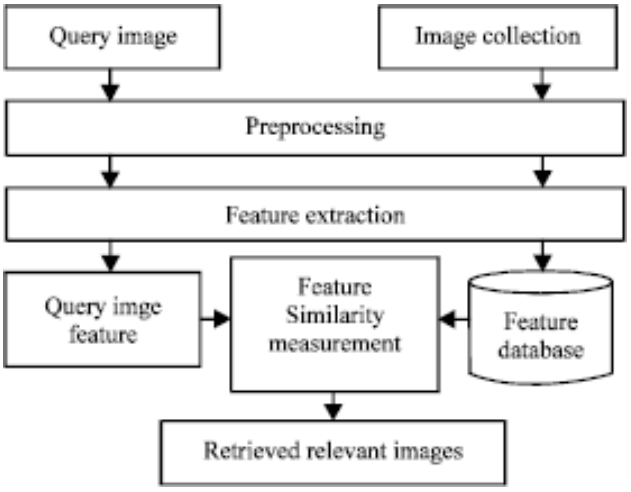
Color is the most prominent and important feature of image because it is the dominant part of human visual perception. It is used to retrieve images in CBIR. For this purpose, various color methods have been used. In these methods color histogram is popular one and mostly used method. Color histogram has the frequency of occurrence of each color in an image. Color histogram is divided into bins of color and each pixel having a specific color belongs to a color bin of that color. It has the characteristics that it represents the global information of the image. Here we use a 3D histogram of bin size of 12 for each channel having a shape of 12x12x12 which corresponds to 1728 features on flattening. This acts as a color distribution descriptor.



Also, as the color images consist of three components therefore the computational cost of feature extraction will be high. To reduce computation cost the color images are converted into grayscale. Now a convolution filter laplacian filter is applied which does edge detection, to obtain a filtered image with edges of objects in image. Hu Moments are normally extracted from the silhouette or outline of an object in an image. By describing the silhouette or outline of an object, we are able to extract a shape feature vector to represent the shape of the object.

SIMILARLITY MEASUREMENT:

As the low-level features are extracted from the images by the algorithm and then these features are represented in a form called feature vector. These feature vectors of all images are stored in a hash table. Once the database of the images with feature vectors is created, then the user can give an image as a query to retrieve the relevant images from the database. The feature vector of the query image is computed by using the same algorithm that is used for feature vectors of images in database. For the measurement of the similarity between the query image and the database images, the difference between query image feature vector and database image feature vectors is calculated. For this purpose, the Chi-Square distance metric is used to calculate the difference between the query image feature vector and database feature vectors for the similarity.



REQUIREMENT SPECIFICATION

Software Requirements

- Python Environment
- Jupyter Notebook – Anaconda Environment
- Python Threading library
- OpenCV image processing library
- Numpy data manipulation package
- Matplotlib plotting library
- Seaborn graphical plotting library

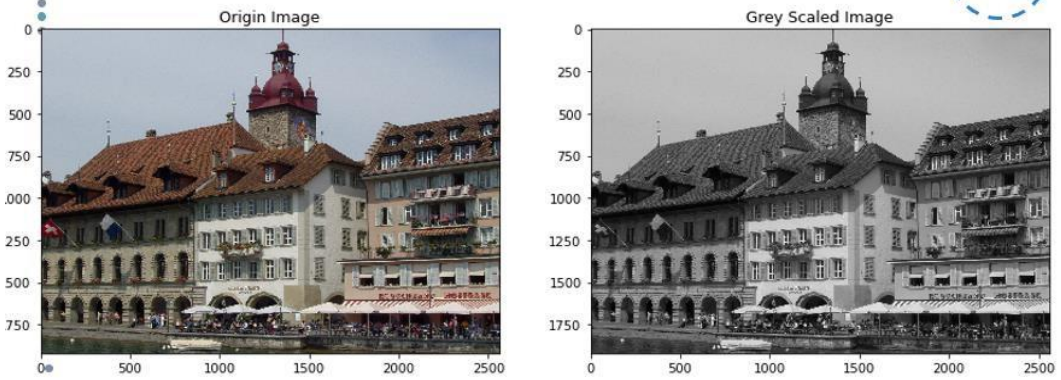
Hardware Requirements

- x86 -compatible chipset Intel or AMD
- 6 GB RAM
- 20GB of free disk space
- Windows or Linux or Mac

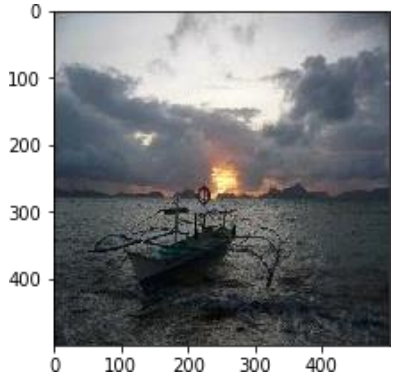
RESULT OBSERVATION

RESIZING

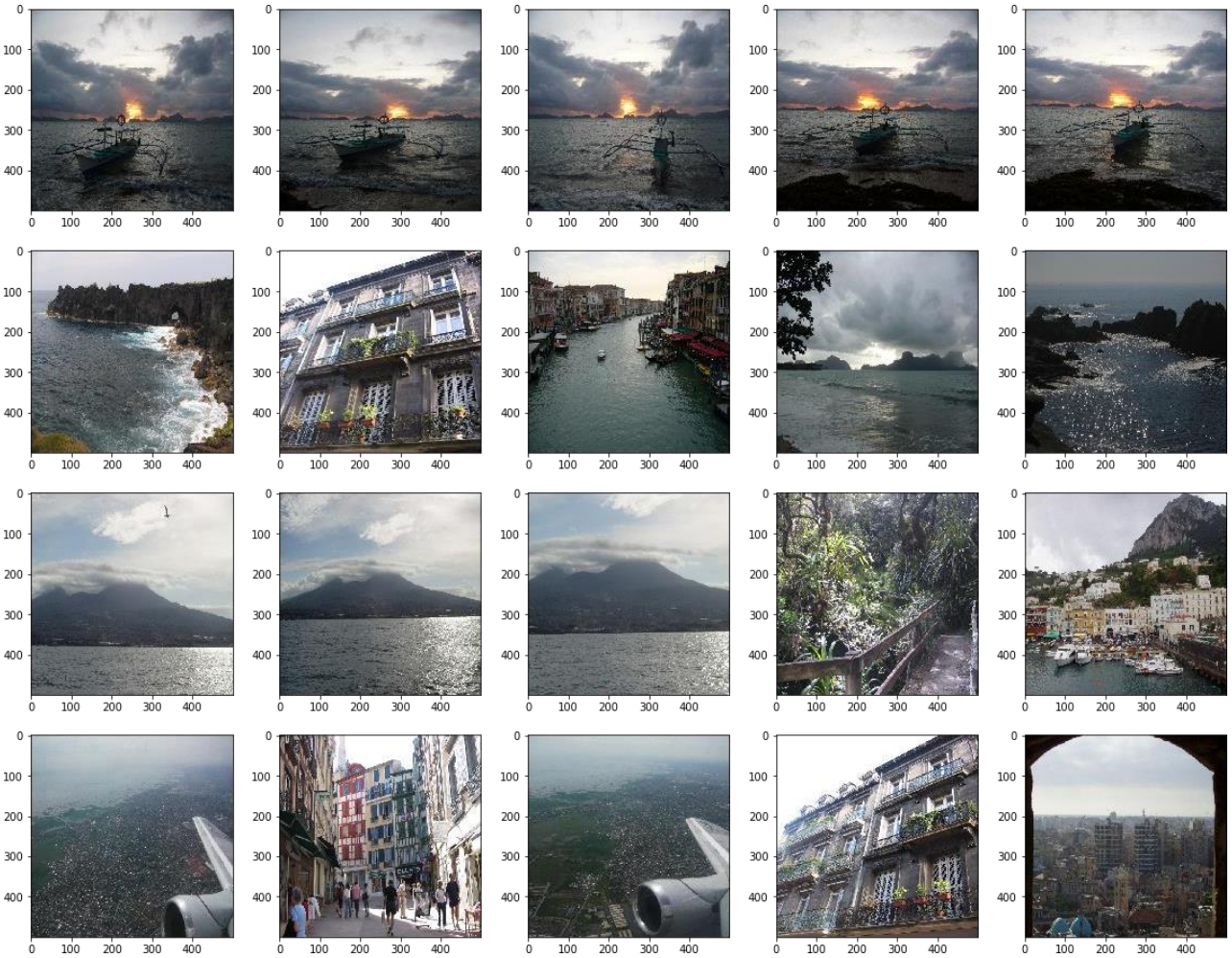
GREY SCALING



QUERY IMAGE:

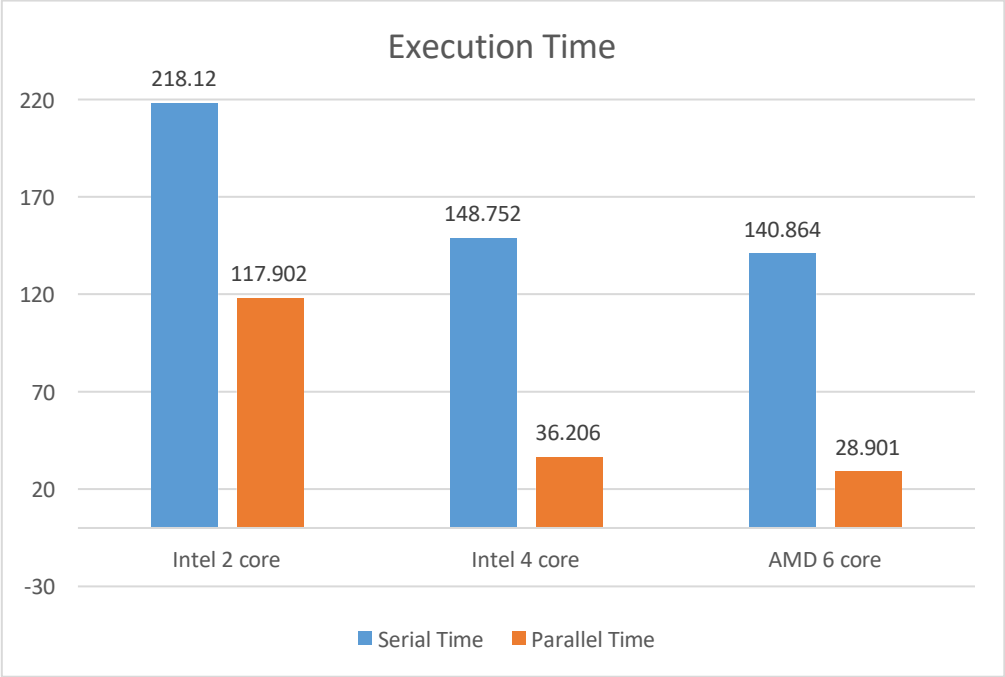


SEARCH RESULTS:

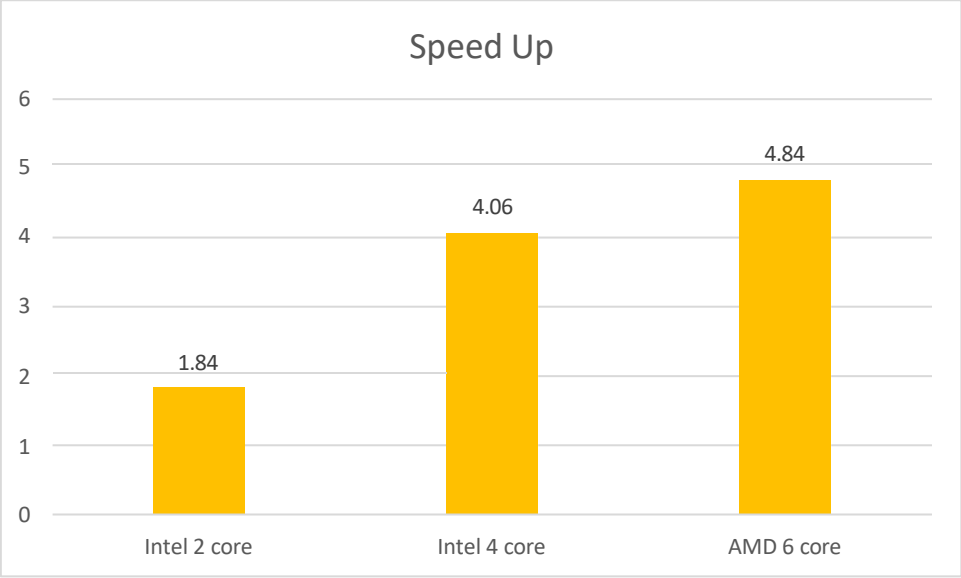


PERFORMANCE EVALUATION

SERIAL VS PARALLEL IMAGE SEARCH EXECUTION TIMES



SPEED UP ACHIEVED



CONCLUSIONS AND FUTURE WORK

By the end of this project, we will have implemented an image retrieval algorithm using parallel computing which is an application of feature extraction that executes within minimal time. By implementing multithread processing on extracting the image features the number of works completed per unit time is increased. Makes the whole process execution a lot faster. Since threads communicate more easily, they can process the input image and distribute the work amongst them and they can combine their results efficiently after obtaining results from various processes. Processes are executed in parallel thus no process need to wait until the previous process has completed its execution. Increases CPU utilization of a multi-processor system. The use of multithreading greatly increases the overall performance and throughput of the system. Faster performance of process is achieved meaning that the execution time of the process will be reduced significantly.

The user can retrieve similar images of his query image from vast collection of images with ease and low latency. This algorithm can be used to enhance an image's contrast, detail by changing the distribution of the histogram. This algorithm can be modified to implement the application of different filters on the image in real time applications since threads used in the program speed up the process till the maximum requirement.