AN APPROACH FOR FINDING OPTIMAL K FOR COLOR IMAGE SEGMENTATION USING K-MEANS ALGORITHM

A
Dissertation Report
submitted in partial fulfilment of the
requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWORK

by

Priyanka Kirsali(R102213006)

Under the guidance of

Mr. Anil Kumar

Assistant Professor

Department of Computer Science & Engineering
Centre for Information Technology
University of Petroleum & Energy Studies
Bidholi, Via Prem Nagar, Dehradun, UK
May – 2015

UPES COLLEGE OF ENGINEERING STUDIES

The innovation driven E-School

CANDIDATE’S DECLARATION

I hereby certify that the project work entitled **“An Approach for finding optimal k for color image segmentation using K-means”** in partial fulfillment of the requirements for the award of the Degree of MASTER OF TECHNOLOGY In ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWORK and submitted to the Department of Computer Science & Engineering at Center for Information Technology, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my work carried out during a period from **January**, **2015** to **April**, **2015** under the supervision of **Mr. Anil Kumar**, **Assistant professor and the Department of Computer Science & Engineering at Center for Information Technology, University of Petroleum & Energy Studies, Dehradun** .

The matter presented in this project has not been submitted by me for the award of any other degree of this or any other University.

Priyanka Kirsali
R102213006

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: _____________2015

Mr.AnilKumar
Project Guide

Mr. Amit Aggarwal
Program Head – Mtech(AI&ANN)
Center for Information Technology
University of Petroleum & Energy Studies
Dehradun – 248 001 (Uttarakhand)

ACKNOWLEDGEMENT

I wish to express our deep gratitude to our guide **Mr. Anil kumar**, for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

I sincerely thank to our respected Program Head of the Department, **Dr. Amit Agarwal**, for his great support in doing our project in **Area** at **CIT**.

I am also grateful to **Dr. Manish Prateek, Associate Dean** and **Kamal Bansal**, **Dean CoES**, **UPES** for giving us the necessary facilities to carry out our project work successfully.

I would like to thank all our **friends** for their help and constructive criticism during our project work. Finally we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

Name	**Priyanka Kirsali**
Roll No.	**R102213006**

ABSTRACT

“An Approach for finding optimal k for color image segmentation using K-means” is a project that will provide a solution for traditional k-means clustering algorithm for choosing appropriate number of clusters for color image segmentation.

We can retrieve meaningful information from images and can find region of interest with help of Clustering techniques. K-means is an important Clustering technique that works well for colour images. An important parameter for K-means is number of clusters or classes (k), which divides the image into k segments. Problem with traditional K-means is that classes (k) are fixed so the results obtained by K-means segmentation are very poor. Therefore, we need an approach with can give us better results by choosing number of clusters (k) optimally.

In our thesis work proposed algorithm will be used for colour image segmentation by choosing optimal number of classes (k) for an image. The algorithm works iteratively by finding of distance measures and then defining a validity index for clusters selection. Here initialization of cluster is selected with statistic measures, so it gives better results as compared to traditional K-means. With help of intra cluster and inter cluster distance it yield a validity index with help of which number of classes are defined.

Contents

List of Figures

1. Chapter1

2. Chapter3

3. Chapter 4

4. Chapter 6

5. Chapter 7

6. Chapter 8

List of Tables

1. Chapter5

2. Chapter8

CHAPTER 1

INTRODUCTION

1.1 Motivation

Image segmentation is important step in digital image processing as it is helpful in extracting valuable information from an image. It is a technique of selecting only the interested region from an image and discarding the other parts. Image segmentation can be done on Gray images as well as Colored images. Image Segmentation works by partitioning an image into separate regions that contains similar range of pixel values. For obtaining meaningful and useful for image segmentation results, the regions should be strongly similar to depicted objects or region of interest.

Image segmentation is can be used in our daily lives, and it seems everywhere if we want to analyse what inside the image. For example, if we have to separate black balls from red or white balls, we can make use of segmentation technique to separate these balls or objects and further each object can be analysed individually. Before serving an image for pattern recognition, image feature extraction or image compression image segmentation can be applied for image pre-processing.

Colour images segmentation can be done by many clustering methods one of which is K-means clustering. This method is very efficient in dividing image and analyses of objects in an image, but suffers with some drawbacks.

We got motivation from various research works that have been carried out in order to increase quality of results given by traditional K-means. Our goal is to research and find what we can do to improve efficiency of the algorithm. This thesis provides a validity index so that results more accurate and are formed with optimal number of clusters K.

Simple image segmentation can be depicted with figure1.1 in which an apple is image is segmented to find out the defected regions in the image.

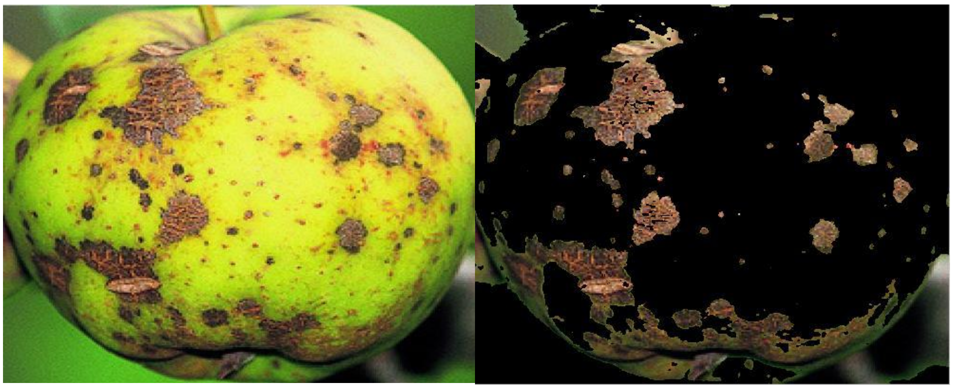


Figure 1.1: Simple color image segmentation

1.2 Thesis Outline

This report is organized as follows:

Chapter3, various Image Segmentation techniques are discussed, brief overview of K-means algorithm is provided and our problem statement is found out.

Chapter4 it provide us with the past work done in the area and different methods used to implement it.

Chapter5 gives background knowledge about the problem domain and methodology. It tells us about existing system and proposed system

Chapter6 shows the system design and flow of the proposed method.

Chapter7 provide us with implementation details which data bases are used and algorithm for the proposed method.

Chapter8 describes experiments and result of the proposed thesis work.

Chapter9 concludes the thesis of our work done in the area.

Chapter10 lastly it provides further work that can be done in the area.

CHAPTER 2
PRELIMINARIES

- **Image**: image can be termed as graph which is constructed with help of small units called pixels.
- **RGB Image**: consist of red, green and blue components or channels.
- **Clusters**: it is a collection of similar objects group together.
- **Intra-cluster distance**: it is defined as the average squared distance between the data points and centroid of a cluster i.
- **Inter-cluster distance**: it is the measure that gives the minimum difference between the cluster centroids. It is defined as the distance between different clusters.
- **Validity index:** the index which defines the validity of a cluster.
- **Image segmentation**: process of segmenting images into different regions for obtaining region of interest.

CHAPTER 3

BASICS OF IMAGE PROCESSING

For implementation of the proposed thesis for color image segmentation we should be aware of important and basic concepts of Digital Image Processing. In this section we will go through various techniques of image processing that we will use later in the project. Image processing is conducted on image to improve image quality and gaining meaningful information about the image.

3.1 Image Processing

A very initial and important step of image processing is processing of image, which involves increasing the contrast and quality of an image by modifying pixels values using different methods.

3.2 Image Segmentation

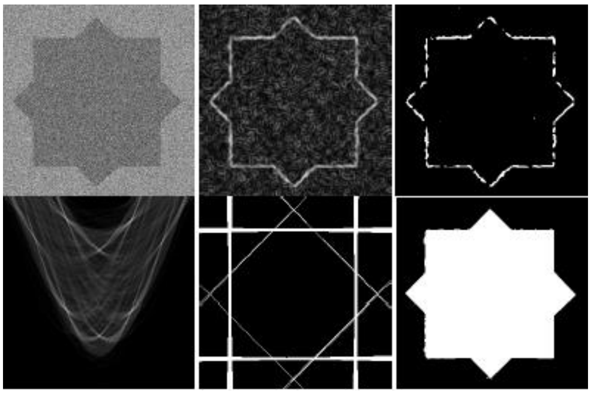
Segmentation is used for partitioning an image into separate regions, where each region is grouping of similar kinds of objects based on pixel values or intensity. Image segmentation is very helpful in finding the area of interest in an image by excluding unnecessary parts in an image. Segmentation can be done with various methods such as K-Mean clustering or Fuzzy-C mean clustering.

3.3 Image Segmentation Techniques

There are various Image segmentation techniques present that can be used for segmenting the image into different no of meaningful regions. Most of the techniques of image segmentation proceed by converting an image into binary format which is not an effective method as much of the data is lost when image is converted into binary, therefore colored segmentation techniques are more prevalent now a days. Broadly image segmentation can be grouped into two classes based on image properties namely discontinuities based and similarities based. Some of the image segmentation algorithms are described below:

3.3.1 Edge Detection Image Segmentation:

This is one of the Discontinuity based method which works by detecting sudden increase in the intensity or pixel values. These methods come in category of Edge or Boundary based methods. Edge detection techniques generally works for gray levels images by detecting edges of the image and segment it into region of interest. Edge detection can be performed in various manners such as gray histogram or gradient based technique. Edge linking and boundary detection can be done globally or locally.



:

Figure 3.1: Example of edge base segmentation

3.3.2 Threshold based Image Segmentation: this is the simplest method for image segmentation that works by setting a threshold value T for segmenting an image. Thresholding algorithm should choose a proper threshold limit T for dividing image pixels into several classes and separate foreground objects from background. Let us say that there is gray level f(p,q) of point (p,q) and m(p,q) denotes some local properties such as mean then threshold T is defined as [10]

$$T = T[p, q, f(p,q), p(p,q)] \text{ ....3.1}$$

With threshold T a threshold image g(p,q) is given as:

$$g(p,q) = \begin{cases} 1 \; if \; f(x,y) > T \\ 0 \; if \; f(x,y) \leq T. \end{cases} \text{ ....3.2}$$

Threshold can be divided into three categories namely:

- Local threshold
- Global threshold
- Adaptive threshold

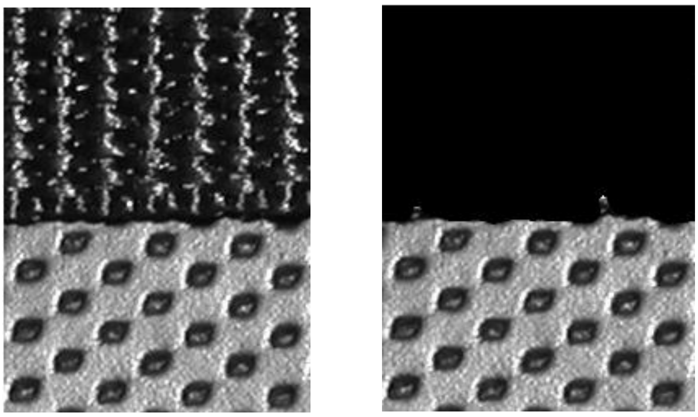


Figure 3.2: Example of threshold based segmentation

3.3.3 Region-Based Segmentation: - this method is continuity based method which works on similarities among the pixels. In this technique different sub-regions are formed from original image based on some predefined rules. Region will grow with the help of define rules and form sub images of the given one. Let us say that R determines the whole region of an image and then segmentation can be viewed as a process that partition R regions into n sub regions, $R_1, R_2, R_3 \dots \dots .. R_n$, such that

- $\coprod_{i=1}^{n} R_i = R$
- R_i is a connected region, i=1,2….,n.
- $R_i \cap R_j = \emptyset$ for all i and j, i≠j.
- P(R_i)=TRUE for i and j, i=1,2….,n.
- P($R_i \cup R_j$)=FALSE for i≠j.

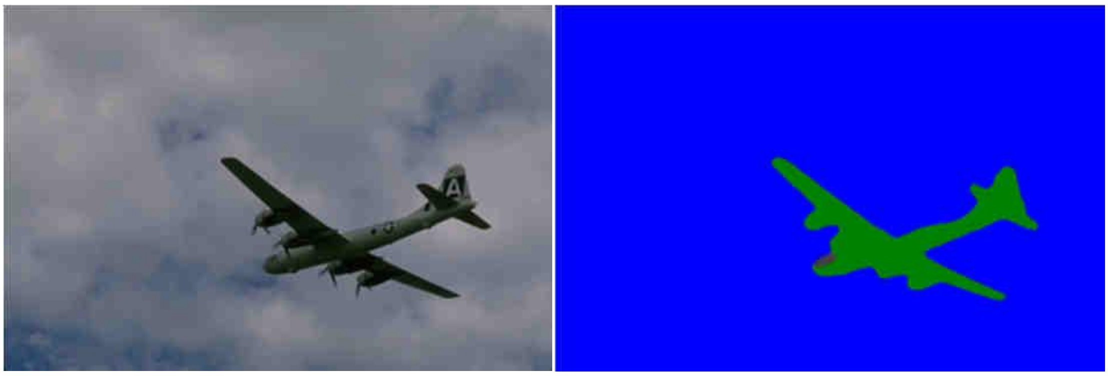


Figure 3.3: Example of region based segmentation

3.3.4 Watershed Image Segmentation: - this segmentation method first converts an image into gradient for further processing. It uses several morphological tools such as erosion, dilation for segmenting the image. Watershed is sometimes called morphological watershed. Watershed view an image into three dimensions, certain terms like watershed lines, catchment basins and regional minimum are used in watershed algorithm.

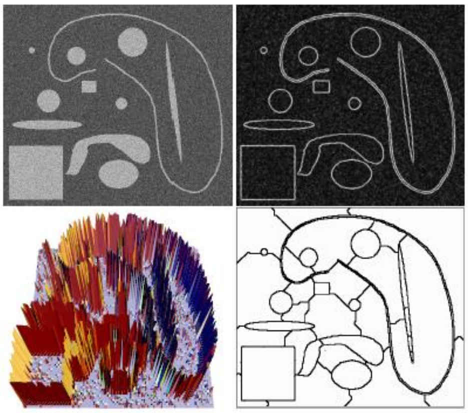


Figure 3.4: Example of watershed segmentation

3.3.5 Clustering Based Image Segmentation: clustering based is an unsupervised learning algorithm in which number of clusters (K) is defined so that image pixels can be classified in different clusters based on similarities measures. Clusters are formed based on pixels similarities that follows different properties such as size, texture etc. Clustering algorithm is also termed as grouping of objects with similarities. Clustering algorithms rely on a distance matrix between data points and cluster centroid, and is one of the essential features of clustering.Clustering techniques generally used for color image segmentation as they can work on RGB and Gray scale values of an image so there is no prior need of converting an image into binary. In later section we will explore K-means clustering technique for segmentation which will be modified in order to find more accurate and optimal results. Clustering can be categorized into various types as follows:

- hierarchical methods,
- partitioning techniques,
- density or mode seeking techniques,
- clumping techniques

3.4 K-Means Clustering Algorithm

K-means Algorithm comes under the category of clustering based image segmentation which is also called as partitioning or classification method. In k-means clustering, given a set of n data points in d-dimensional space R^d and an integer k and the problem is to determine a set of k points in R^d, called centers, so as to minimize the mean squared distance from each data point to its nearest center [5]. It partition or classifies data into different number of classes called as clusters (k). Image pixels are clustered based on their distance from centroids μ_i. K-means algorithm is an effective image segmentation technique for color images and can display segmented image in colored manner. Points around the centroid are obtained when the objective function given below in equation (1) is minimized:

$E = \sum_{i=1}^{k} \sum_{X \in C_i} (x_i - \mu_i)^2$ ………………………………. …………3.3 [2]

Where X is the data point and μ_i is the centre for clusterC_i. E is the sum of squared error of all the squared differences. For calculating new centroids and assigning data points to cluster below equations are used.

$C^{(i)} = \arg min_j \left\|(x^i - \mu_j)^2\right\|$………………………………3.4 [2]

$\mu_i = \frac{\sum_{i=1}^{m} 1\{c_{(i)}=j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c_i=j\}}$ ……………………………………………3.5 [2]

Where *k* is the number of clusters, *i* iterates over all the intensity values, j iterates over all the centroids (for each cluster) and μ_i are the centroid intensities.

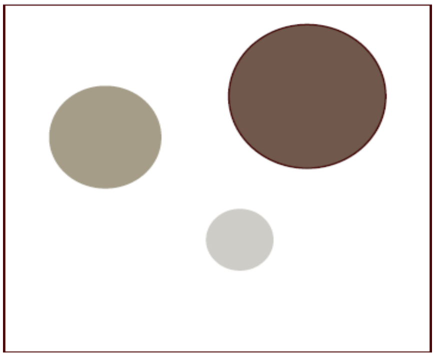


Figure 3.5: A well compact cluster

K-means process can be illustrated with help of flow chart shown in figure 3-7:-

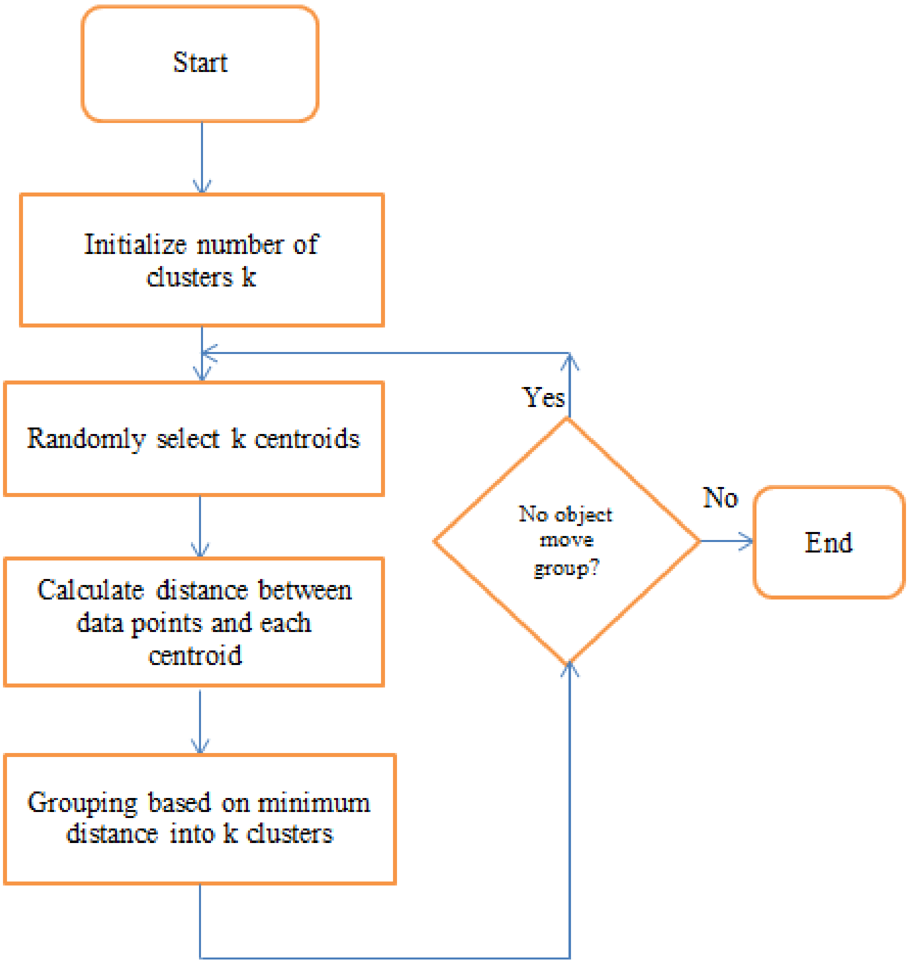

Start
Initialize number of clusters k
Randomly select k centroids
Calculate distance between data points and each centroid
Grouping based on minimum distance into k clusters
No object move group?
Yes
No
End

Figure 3.6: Flow chart for K-means algorithm

3.5 Problem Statement

There are some limitations that are faced in k-means which result in poor segmentation. The problem faced in K- means is that we have to define number of clusters as input and in traditional K-means we cannot change the clusters values as per our need. With fixed initialization of number of clusters chances are there that algorithm might give wrong segmentation results. And if results are not correct than poor segmentation will be generated. For example let us say that an image required to be clustered in K=3 but initialized cluster value is K=2, so results obtained are not correct. Therefore it is necessary to select optimal values for cluster K, so that valid results are obtained.

3.6 Aim and Objective

Aim of the project is to implement a dynamic K-means algorithm that will iteratively choose appropriate number of clusters for segmentation. Firstly we will implement the algorithm and after that create a user interface for end user for testing. Some objectives we will follow to achieve our aim are described below:

1) Study the basic concept of Digital image processing and Traditional K-mean clustering.

2) Choose appropriate method for defining number of clusters.

3) Implement the dynamic K-means algorithm.

4) Designing a user interface for testing algorithm on different image sets.

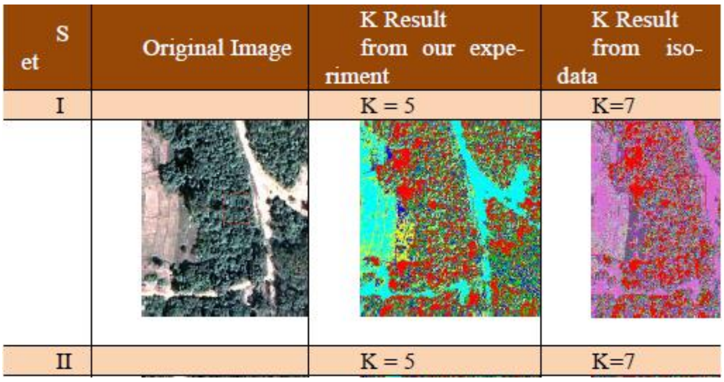
CHAPTER 4

LITERATURE REIVIEW

4.1 Related work

In this section a brief survey of the prior work done in the related field is conducted. Several research works are studied to get knowledge of how previous algorithm is implemented in order to find appropriate number of clusters. Some of the related work is described below:-

Kitti Koonsanit [1], in this paper a method has been developed which defines the initialization number of clusters/classes in satellite image clustering application using a K-means algorithm that is based on the co-occurrence matrix technique which is matrix containing number of pixel pair repetition in an image . The proposed method was tested using data from unknown number of clusters with multispectral satellite image in Thailand. The proposed algorithm worked by creating co-occurrence matrix of the grey scale image and then finding number of clusters K by applying local maximum.



S et	Original Image	K Result from our expe-riment	K Result from iso-data
I		K = 5	K=7
II		K = 5	K=7

Figure 4.1: Result of above mentioned paper

Sundararajan S [2], this paper determines an efficient numbers clusters by using Dynamic K-means and Firefly algorithm. Clusters are increased in an iterative manner if condition of distances is fulfilled and new centroids are found with help of firefly algorithm.

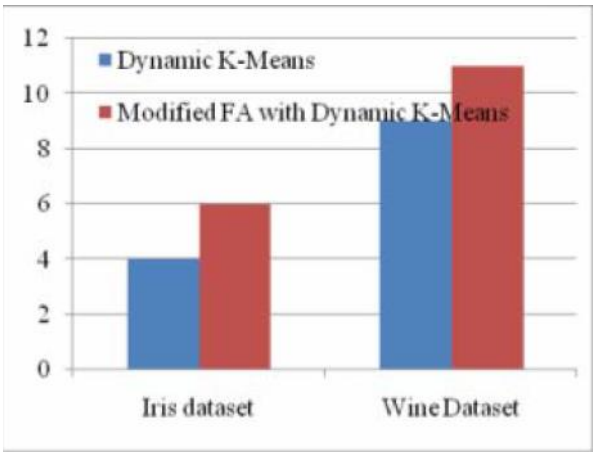


Figure 4.2: Result of above mentioned paper

Ahamed Shafeeq *et al* [3], proposed an improved K-means for data clustering for the unknown set of data. In this paper the problem of initializing of centroids initially is modified with selecting optimal number of clusters for segmentation.

Shiv Ram Dubey [4], in this paper defected fruits are analysed and segmented with help of K-means algorithm. A brief overview of how K-means algorithm works is given and later the fruits part which are defective are segmented from the image. Image is first converted into l*a*b colour space then image is fed into K-means clustering function which gives segmented image of the original one.

D T Pham [6], this paper reviews various techniques such as statistical measures that can be used for finding number of clusters K and factors by which section of clusters is affected. It also proposes a method for selecting the appropriate number of clusters.

G komarasamy [7], in this paper k-means algorithm uses modified hill-climbing search in order to reach the global optimal solution of the objective function. These Hill-climbing algorithms are iterative in nature which makes modifications by increasing the value of their objective function at each and every iteration. The experimental result of the paper shows that proposed algorithm Modified Hill-climbing aided K-Means Algorithm (MHKMA) is much more efficent than the existing algorithm KMBA and ordinary k-means.

Siddheswar Ray [8], this paper deals with determination of number of clusters for image segmentation. The basic steps involve producing all the segmented images for 2 clusters up to *Kmax* clusters, where *Kmax* represents an upper limit on the number of clusters. Then our validity measure is calculated to determine which is the best clustering by finding the minimum value for our measure. The validity measure is tested for synthetic images for which the number of clusters in known, and is also implemented for natural images. First one cluster containing all the pixels in the image is generated. Then an iterative process is applied, unless the number of clusters is equal to *Kmax*, the cluster having maximum variance is split into two. Once the cluster is split, k-means procedure is used to obtain the clustering for this new number of clusters. Once all the clusters have been formed, the validity measure can be calculated for each of them to determine what the optimal value of *K* is.

Malika Charrad [11], the R package NbClust has been developed for finding number of clusters. With help of 30 indices it find out the number of clusters and choose best number by using dominance rules. In addition, it provides a function to perform k-means and hierarchical clustering with different distance measures and aggregation methods. Any combination of validation indices and clustering methods can be requested in a single function call. This enables the user to simultaneously evaluate several clustering schemes while varying the number of clusters, to help determining the most appropriate number of clusters for the data set of interest.

Pushpa .R [12], in this paper two approaches are used for modifying k-means clustering technique. Image segmentation is done on basis of two methods edge preserving smoothing filter and anisotropic. In the paper the adaptive rate is chosen on based on certain criteria for adaptive k-means clustering.

Rupali B. Nirgude [13], in this paper image segmentation is conducted using several techniques of color image segmenation. Firstly hill climbing technique is used for finding peaks of the given image with help of the obtained histogram of the image, and then the number of classes selected which are the peak points of the calculated histogram. Then another technique that is used is K-means clustering. Third algorithm used is sequential probability ratio (SPRT). In this the neighboring pixels validity is checked with help of SPRT test .This test is used to recognize similarities characteristics by attributes as intensity, edge , pixels etc. in beginning steps it takes two regions for conducting similarities and after that it moves forward.

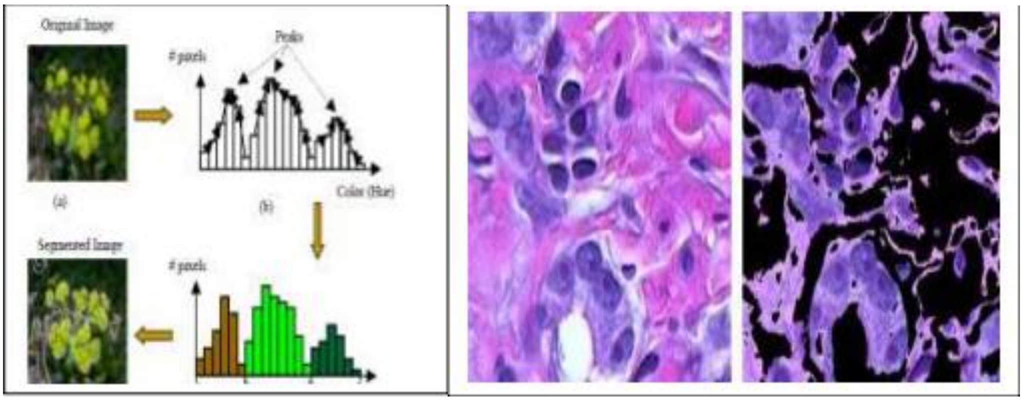


Figure 4.3: Result of above mentioned paper

Result obtained by hill climbing algorithm and k means

Amiya Halder [14], this paper uses combination of fuzzy c-mean clustering and genetic algorithm for segmenting an image. Fuzzy c mean is used for generation of population (n) and then this population is fed into genetic algorithm which gives the segmented results.

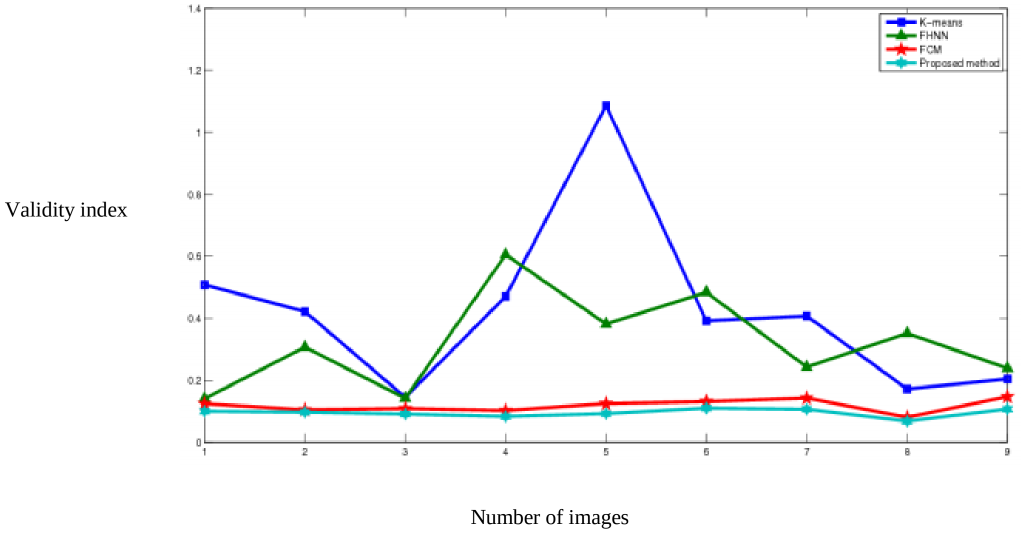


Figure 4.4: Result of above mentioned paper

Priyanka Kirsali_ [15], in this paper fuzzy c-means algorithm is used for finding the region of eye which has exudates and haemorrhages. In this approach first optic disc is segmented using largest area selection and then exudates are separated with help of fuzzy c-means clustering. The algorithm was tested on two freely available data sets databases DIARETDB0 and MESSIDOR.

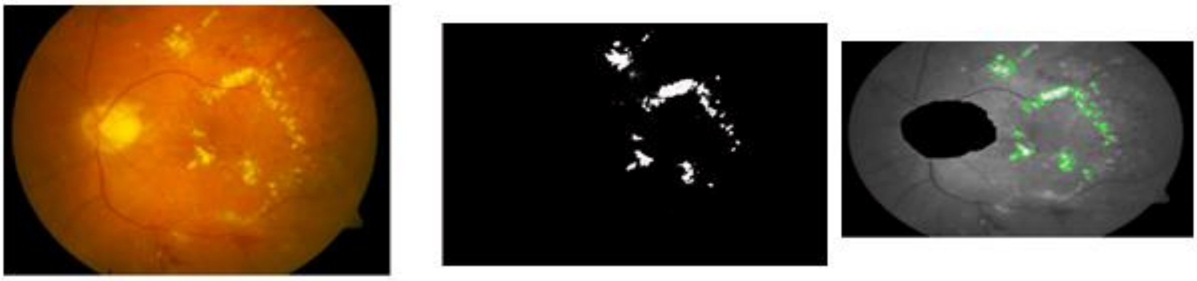


Figure 4.5: Results from above mentioned paper

Giri babu kande [16], in this paper approaches have been described to extract main features of retinal images such as optic disc using active contour technique, and detection of exudates is performed with help of Fuzzy-C mean clustering. Comparison of optic disc localization is done with methods such as GVF snake and Hough transform.

CHAPTER 5

METHODOLOGY

In this section we will have brief overview of existing system and proposed methodology for solving the problem statement and meet our objectives.

5.1 Existing System

K-means (KM) clustering is a heuristic algorithm that can minimize sum of squares of the distance from all samples emerging in clustering domain to clustering centres to seek for the minimum *k* clustering on the basis of objective function [9]. Traditional K-mean algorithm has a drawback that number of cluster initialized is fixed and it divide image into fixed number of clusters. So there is a need to propose an algorithm which can choose correct or appropriate number of clusters for image segmentation. Below table 5.1 depicts the pseudo code for traditional K-means algorithm.

Table 5.1: Pseudo code for traditional K-means algorithm.

```
Input:- image/data, and number of clusters K
Output: - data clustered into K sets.

Steps
Step1: Input an image or set of data to be clustered.
Step2: Initialize the number of clusters K.
Step3: Randomly select K centroids from the data set.
Step4: Repeat
Step5: Calculate the distance of all data points from each centroid using equation (1).
Step6: Assign the data points to the nearest cluster  $C_i$ using equation (2)
Step7: Again calculate the centroids, for each cluster by equation (3).
Step8: If no change in centroids, go to step9 else step4.
Step9: Stop.
```

5.2 Methods for selecting Number of Clusters

Different ways can be used for selection of number of classes which can be classified into following categories:

- **Selection based on variance**: variance approach is used to find out number of clusters by dividing cluster having high variance into separate parts. Cluster which has maximum variance is split into two parts.
- **Within-class and between-class measures:** another way of selecting number of clusters is by finding distance measures within the clusters and between the clusters and after that defining a validity value for selecting the number of clusters.
- **Using multiple clustering:** another possible method for cluster selection is by using multiple clustering algorithms together in order to get valid results.
- **Random data sampling:** with help of random sampling of data we can find various numbers of clusters and then we can check the validity of the cluster based on result of clustering technique. Each cluster generated with help of random data is tested and results are recorded, and cluster giving most accurate segmentation result is selected.

5.3 Proposed system

The proposed algorithm defines an index value which checks for validity of cluster k. Algorithm works by selecting number of clusters on basis of validity index. Here we will make use of within class and between class measure and find a valid index by which number of clusters is chosen. First the algorithm work iteratively for maximum number of clusters by finding intra and inter cluster distance at each iteration, after that at every iteration a validity measure is calculated and then the minimum value index is selected as number of clusters. This algorithm is more flexible and user friendly as user can get appropriate cluster values for clustering. Cluster distances are calculated with help of following equations. Basic steps of the algorithm are shown in table 5.2.

Inter cluster distance:-

$$inter = \min(\|x_i - x_j\|^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. 5.1 \text{ [2]}$$

Where i=1, 2…K -1 and j= i+1…..K

Intra cluster distance:-

$$intra = \frac{1}{N}\sum_{I=1}^{K}\sum_{x \in c_i}\|x - \mu_i\|^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. 5.2 \text{ [2]}$$

Where N is the no of pixels in an image.

After calculating the distance measures for k_{min} to k_{max} validity indexes are find out. We will use four methods for finding the validity indexes which are similar to previously find out indexes such as Dunn's index or silhouette index.

Validity index1:

Let the first validity index denoted by v_{i1}, all the indexes are find out with help of cluster internal and external measures that are found out by using above equations. For every index intra distance and inter distance are the main components, different mathematical formulas are tested for finding which index is giving appropriate results. First index is given by equation 5.3:

$$v_{i1} = \begin{cases} 1 - \left(\frac{intra}{inter}\right), if\ intra < inter \\ 0\ \ , if\ intra = inter \\ \left(\frac{intra}{inter}\right) - 1\, , if\ intra > inter \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.5.3$$

After calculating the validity index one we choose appropriate number of clusters as index of minimum value of validity index v_{i1}.

Validity index2:

Let v_{i2} be our second validity index which is a simple ratio of inter to intra cluster distance, and is taken from the idea of Dunn's index, which is given by the formula shown in equation 5.4:

$$v_{i2} = \frac{(\text{inter})}{(\text{intra})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..5.4$$

After calculating the index maximum value of index is chosen and output is the index of maximum value.

Validity index3:

Our third validity index is based on condition which is tested on intra and inter cluster. It is based on the fact that new intra cluster should be smaller than new intra cluster distance as we want to minimize the difference and new inter cluster should be greater than old inter cluster distance, the formula for validity index v_{i3} is given below:

v_{i3}= index

if(intraR(1,i)<intraR(1,i+1) && inter(1,i)>=inter(1,i+1)), condition is satisfied.

Validity index4:

The last validity which is calculated is similar to silhouette index which shows that value closer to one is the most suitable value for number of clusters. This index is given by equation 5.6:

$$v_{i4} = \frac{(inter - intra)}{\max(inter, intra)} \ldots\ldots\ldots\ldots\ldots\ldots 5.6$$

After calculating validity index four we will select the value which is closest to one and then selecting index of that value.

Proposed algorithm is explained with help of the Pseudo code by table 5-2:

Table 5.2: Pseudo code for the proposed algorithm

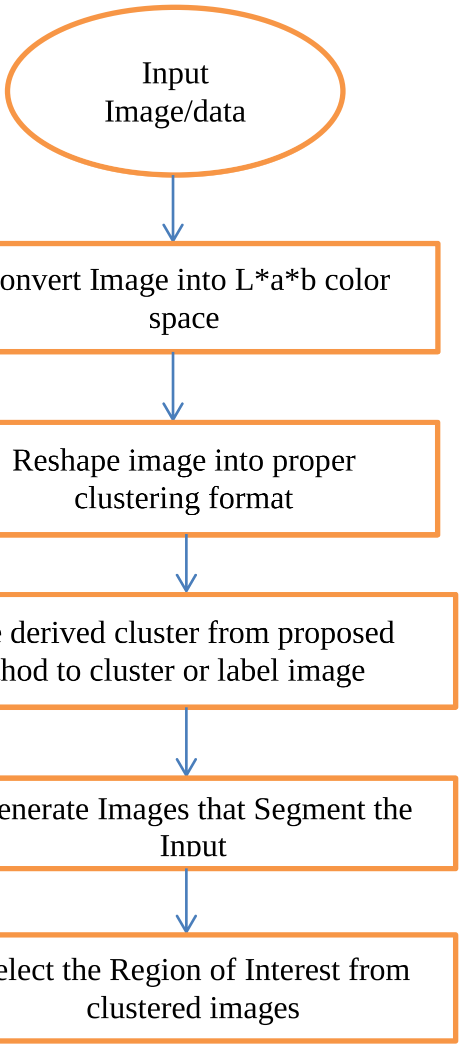
```
Input:- image/data, and number of clusters K
Output: - data clustered into K sets.

Steps
Step1: Input an image or set of data to be clustered.
Step2: Initialize the number of clusters K to Kmax.
Step3: Randomly select K centroids from the data set.
Step4: Repeat
Step5: Calculate the distance of all data points from each centroid using equation (3.3).
Step6: Assign the data points to the nearest cluster $C_i$ using equation (3.4)
Step7: Again calculate the centroids, for each cluster by equation (3.5)
Step8: If no change in centroids, go to step9 else step4.
Step9: Find the intra distance using equation (5.1)
Step10:Find the inter cluster distance using equation (5.2)
Step11:end
Step12:find validity indexes
 Step13: Display(K) from validity indexes
Step14: Stop
```

CHAPTER 6

SYSTEM DESIGN

6.1 Approach

Design of the proposed method is divided into different modules. These steps are depicted in below flow chart and explained in latter section:-



Input Image/data
Convert Image into L*a*b color space
Reshape image into proper clustering format
Use derived cluster from proposed method to cluster or label image
Generate Images that Segment the Input
Select the Region of Interest from clustered images

Figure 6.1: Flow Chart for Image segmentation with proposed algorithm

6.2 Input Image

In first step input image is given which is in RGB color space. An example of clustering using Dynamic K-means is shown with help of an input image.

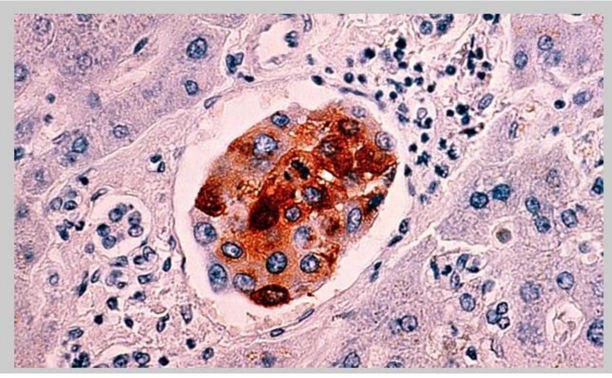


Figure 6.2: Input Image

6.3 Conversion to L*a*b color space

Transform Image from RGB to L*a*b* colour Space. We have used L*a*b* color space because it consists of a luminosity layer in 'L*' channel and two chromaticity layer in 'a*' and 'b*’ channels. Using L*a*b* colour space is computationally efficient because all of the colour information is present in the 'a*' and 'b*' layers only

6.4 Image Reshape

Reshape the image into proper format so that clustering can be done in efficient manner. An m*n matrix is converted into m*1 double matrix which list the entire pixel as feature or attribute of an image.

6.5 Apply Dynamic K-means

In this step apply the proposed dynamic K-means clustering for image segmentation. An optimal value of K is selected with help of this algorithm and at last we get a set of K clusters containing image parts. It gives an output clustered image with every pixel having a label to the cluster it belongs to that is, every pixel of the image will be labelled with its cluster index

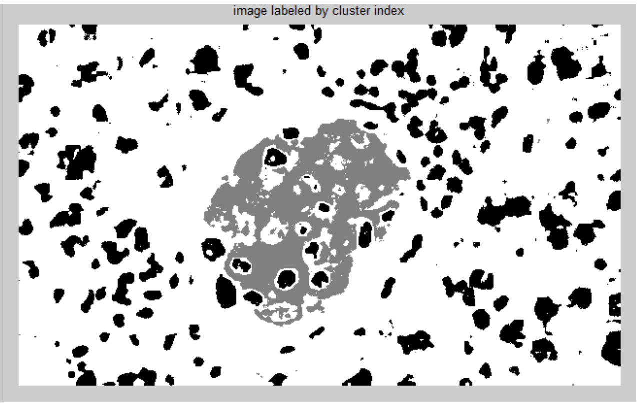


Figure 6.3: Image labeled with cluster index

6.6 Generate Segmented Images

Generate Images that Segment the Input Image by Colour. We have to separate the pixels in image by colour using pixel labels, which will result different images based on the number of clusters.

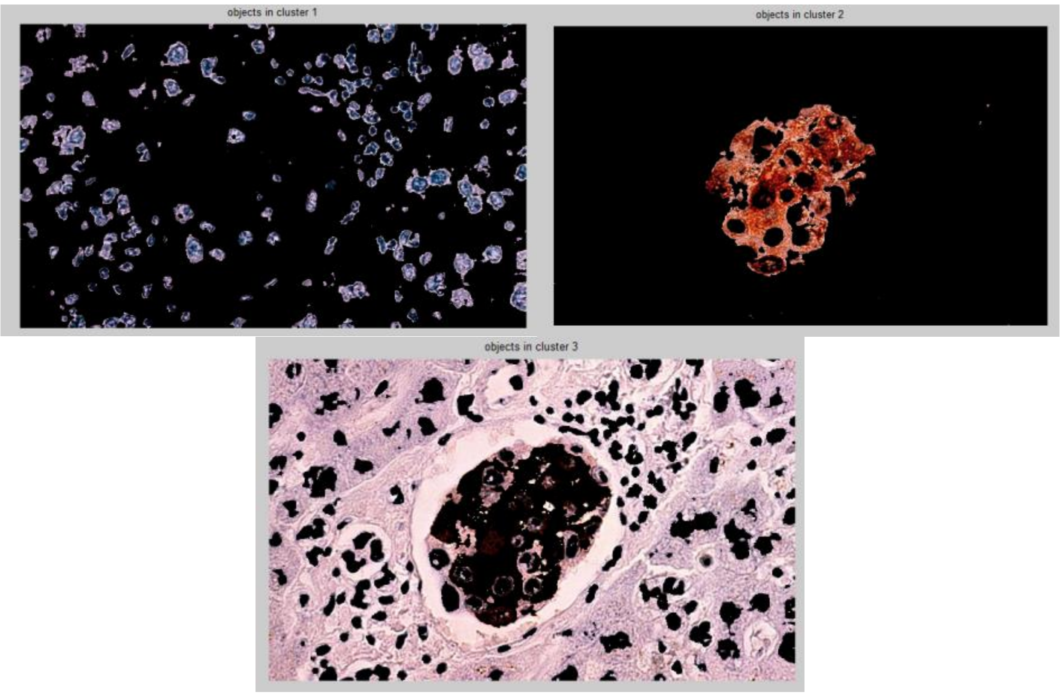


Figure 6.4: Segmented cluster images with K=3

CHAPTER 7

IMPLEMENTATION

7.1 Algorithm

Input: D : image to be clustered converted into double

k : 2 to k_{max}

Output: C={c_1, c_2,….,c_k} (set of cluster centroids)

labels={l(e) | e = 1,2,….,n} (set of cluster labels E)

intra and inter distances

```
Input image (I) that need to be clustered
D ←double(I)
[row1 col1] ←size(D)
h←1,g←1
for each k: 2 to $k_{max}$
    [label ,C]←MyKMeans(D,k)
    Initialize Intra_distance=0
    for l←1 to k
        for i←1 to row1
            if (lablel,1) is equal to 1
                intra_distance← intradistance+ abs $(\mathrm{data}(\mathrm{i},1)-\mathrm{C}(\mathrm{l},1))$^2
                intraR(1,h)←intra_distance
                h← $\mathrm{h}+1$
            end
        end
    end
    p←1
    for l←1 to k-1
        for m←l+1 to k
            inter_distance(p,1) ←abs$(\mathrm{c}(\mathrm{l},1)-\mathrm{c}(\mathrm{m},1))$^2
            p=p++
        end
    end
    inter_distance_min← min(intra_distance)
    inter(1,g)←inter_distance_min
    g←g+1
end
s=1
for i←k-1
    if intraR(1,i)<inter(1,i)
        vi1←1-(intraR(1,i)./inter(1,i))
    elseif intra(1,i)> inter(1,i)
        vi1←(intraR(1,i)./ inter(1,i))1
    else
        vi←0
    end
end
[val numb_clust1] ←min(si1(1,:))
for i←k-1
    vi2=(inter(1,i)./intraR(1,i))
end
[val1 numb_clust2] ←max(vi2(1,:))
for i←k-1
    if intraR(1,i)<intraR(1,i+1) and inter (1,i)>=inter(1,i+1)
        vi3=(i)
    end
end
d=1
for i←k-1
    vi3(1,d)=(inter(1,i)– intraR(1,i))./max(inter(1,i),intraR(1,i));
    d←d+1
end
t=1
for i←k-1
    tmp2(1,2) ←abs(vi4(1,i)-1)
    [val2 numb_clust3] ←min(tmp2)
```

```
t ← t+1
end
```

MyKMeans (D, k)

Input: D : dataset to be clustered

K : number of clusters to be formed

m : maximum number of iterations

Output: $C=\{c_1, c_2,\ldots,c_k\}$ (set of cluster centroids)

$L=\{l(e) \mid e = 1,2,\ldots,n\}$ (set of cluster labels E)

```
for each ci ∈ C do
    $c_i \leftarrow e_j \in E$ (random selection)
end
for each $e_i \in E$ do
    $l(e_i) \leftarrow$ argminDistance($e_i$,$c_j$) $j \in \{1,\ldots,k\}$
end
changed ← false
iter = 0
repeat
    for each $c_i \in C$ do
        UpdateCluster($c_i$)
    end
    for each $e_i \in E$ do
        minDist ← argminDistance($e_i$, $c_j$) $j \in \{1,\ldots,k\}$
        if minDist $\neq l(e_i)$ then
            $l(e_i) \leftarrow$ minDist
            changed ← true
        end
    end
    iter ++
until changed =true and iter $\leq$ m
```

7.2 Tools and Techniques

In this project we have use Matlab tool kit to implement our project. Matlab consist of Image processing tool box that can be very helpful. We have use this tool box only for simple functions like image read etc. MATLAB is a high-performance statistical language for technical computing. It combines computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation [17]. The name MATLAB stands for matrix laboratory. MATLAB was originally written to provide easy access to matrix software developed by the LINPACK and EISPACK projects, which together represent the state-of-the-art in software for matrix computation.

MATLAB provides various toolboxes related to different fields such as image processing, neural net etc. it is a very helpful statistical tool kit.

7.3 Image processing toolbox

Image Processing Toolbox™ provides various standards built in algorithms, function and many more features for image processing, image analysis and design. We can perform various operations such as image analysis, image segmentation, and image enhancement with help of this toolbox.

7.4 Application of image processing toolbox

Matlab image processing toolbox provides various applications such as:

- Image improvement
- Image segmentation
- Histogram
- Removing noisy data from image
- Reading image

7.5 Function of image processing toolbox used in the project

In this section, we will describe some function of Image Processing Toolbox that used in our projects. [18]

- **rgb2Lab:** convert image from rgb color space to lab.
- **imread:** for reading an image from specific location into matlab as a matrix.
- **max:** to find maximum value from number of elements.
- **min:** to find minimum value from number of elements.
- **find:** for finding index value of an element in an array.
- **imshow:** for displaying an image on screen.
- **reshape:** reshape an image into given number of rows and columns.
- **randperm:** used for generating random numbers with help of different permutation.

- **Size:** this function gives the number of rows and columns of the given input matrix.

7.6 User Interface

A graphical user interface (GUI) is a graphical display containing controls, called components that enable a user to perform interactive tasks. GUI components can include menus, toolbars, push buttons, radio buttons, list boxes, and sliders. GUIs created using MATLAB® tools can also perform any type of computation, read and write data files, communicate with other GUIs, and display data as tables or as plots. [19]

In this project, the implementation of user interface aims for the user to use the application easily and effectively because users tend to use applications which are flexible, accurate and easy to use. We aim to implement the user interface that able to accept the input from user, process the input according to the algorithm, display several processing image, and display the clustered images.

7.6.1 Explanation of User Interface

The figures below shows selected screen shots which are home screen and result screen:

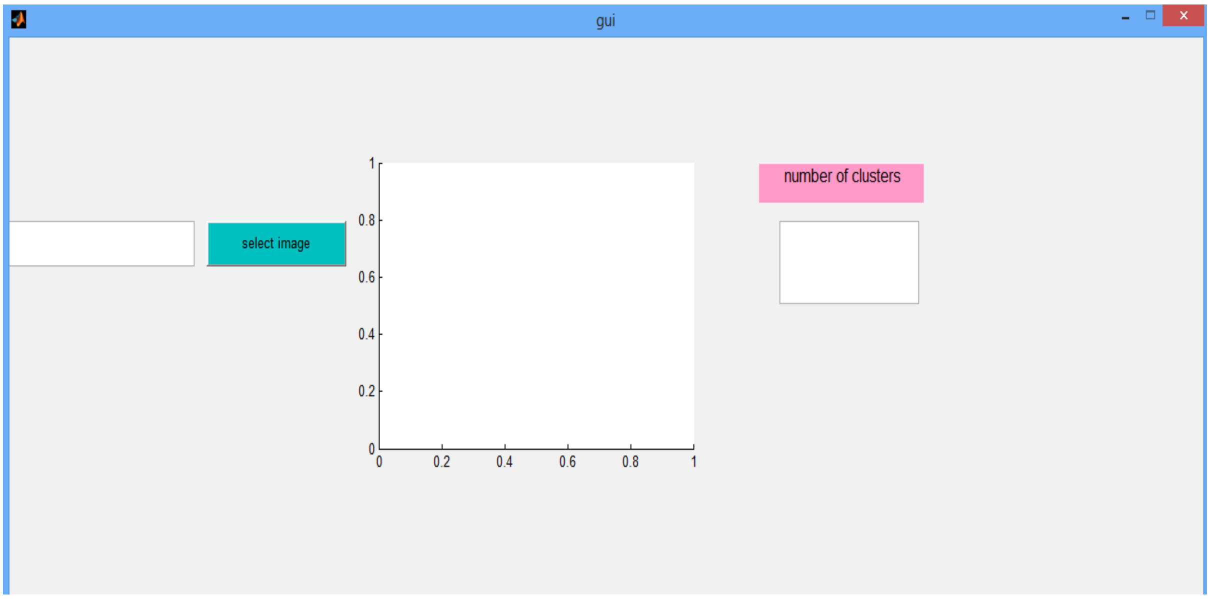
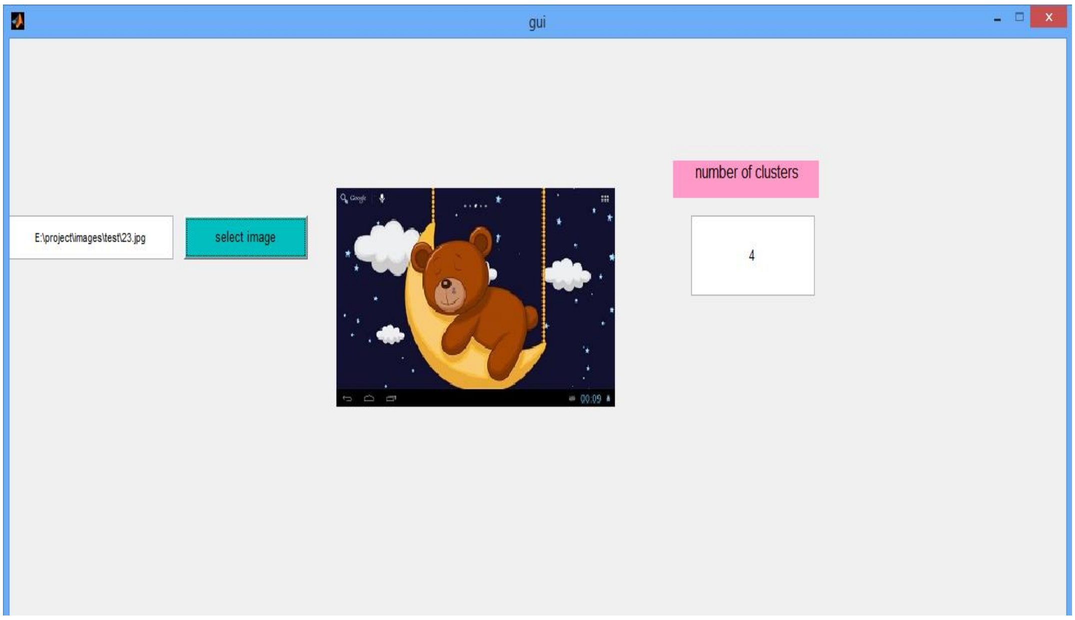


Figure 7.1: Home Screen for showing number of clusters

The home screen composed of

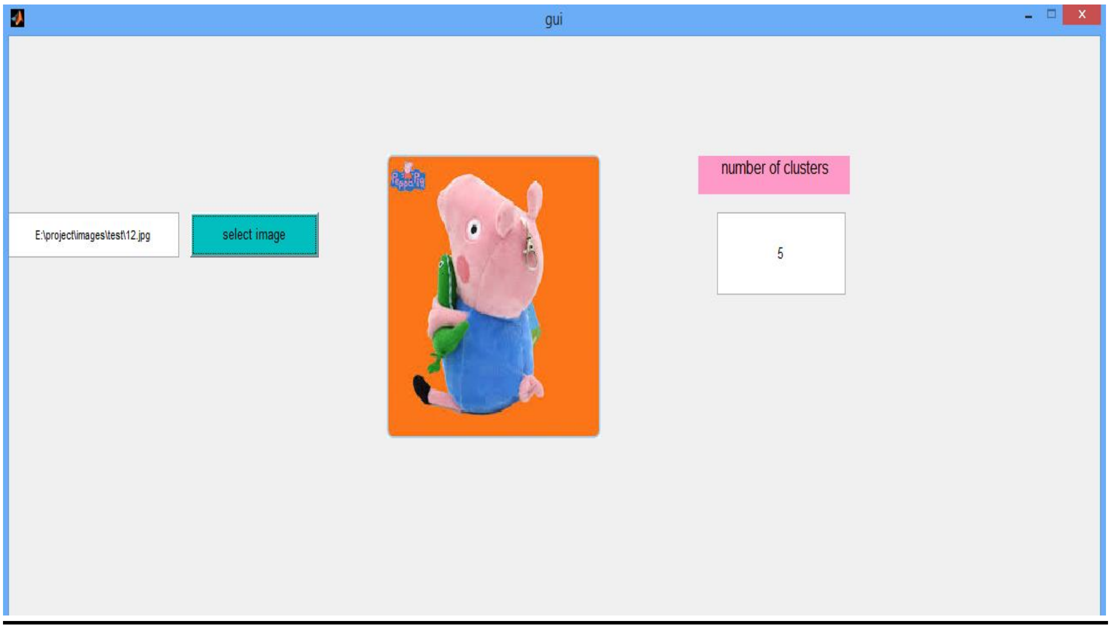
- Axes1 to show the processing images.
- Processing button for browsing image and process the image.
- It also consist of a text box for displaying number of clusters on screen



E:\project\images\test\23.jpg

4

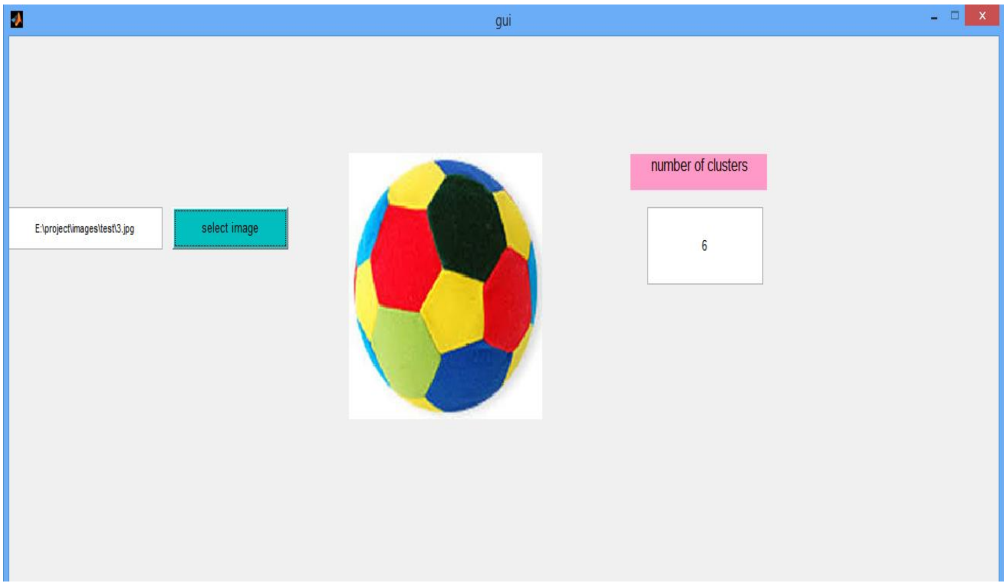
Figure 7.2: Result Screen 1



E:\project\images\test\12.jpg

5

Figure 7.3: Result Screen 2



E:\project\images\test\3.jpg

number of clusters

6

Figure 7.4: Result Screen 3

The result screen would show the images during processing from:

1. Original image inputted by user.

2. Number of cluster.

3. Segmented image on different axes.

CHAPTER 8

EXPERIMENTS AND RESULTS

In this chapter, several experiments are conducted on proposed algorithm and results are obtained. We have implemented an algorithm which proposed four different validity measuring index for finding optimal number of clusters for a colored image from which number of cluster selected by majority rules. We have conducted the experiment for k ranging from 2 to 18 (kmax). We have first applied an iterative process on the image converted into lab colored space and perform clustering from min number of cluster to the maximum number. After that we have find the distance measures both internal and external, and on basis of these distances we have define four validity indexes and index occurring many time is selected on basis of majority rule.

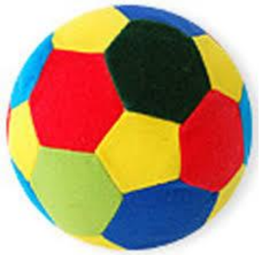
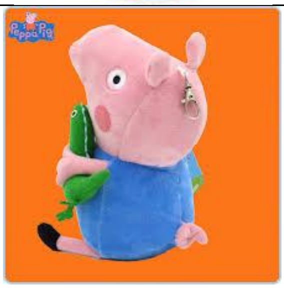
8.1 Image data base for segmentation

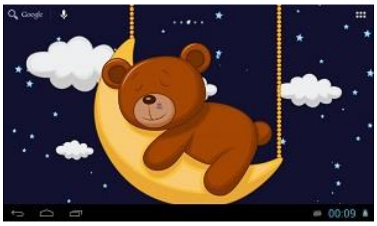
For performing our proposed method we have downloaded several object datasets from Google search engine. These objects are images of birds or different objects. All the images used are colored ones and experiments are conducted on these images.

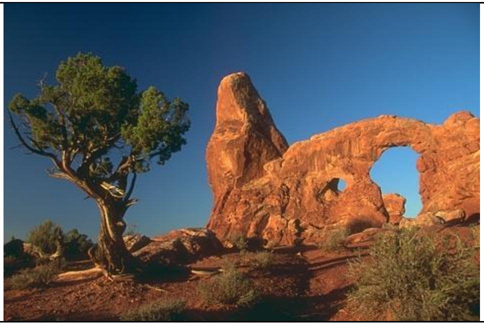
8.2 Finding K for different set of Images

Below table 8-3 depicts the entire test that has been conducted on the images. Number of cluster required is shown with help of integer values.

Table 8.1: Examples of Evaluation Results

Serial no	Image name	Image	Number of clusters
1	Test_image1		6
2	Test_image2		5

3	Test_image3		4
4	Test_image4		6
5	Test_image5		16
6	Test_image6		7
7	Test_image7		8
8	Test_image8		3

9	Test_image9		6
10	Test_image10		4

We have got the number of cluster shown in table 8.1, now we can perform segmentation of the images with help of K-means algorithm. Now we have idea of how many clusters we want for our image segmentation.

Figure 8.1 will show us segmentation on one of the tested image with the calculated number of classes.

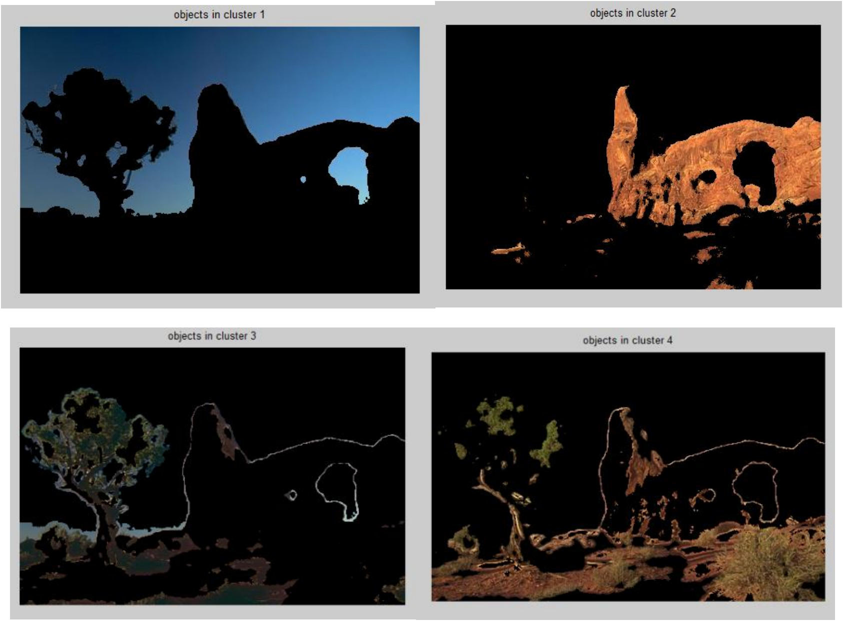
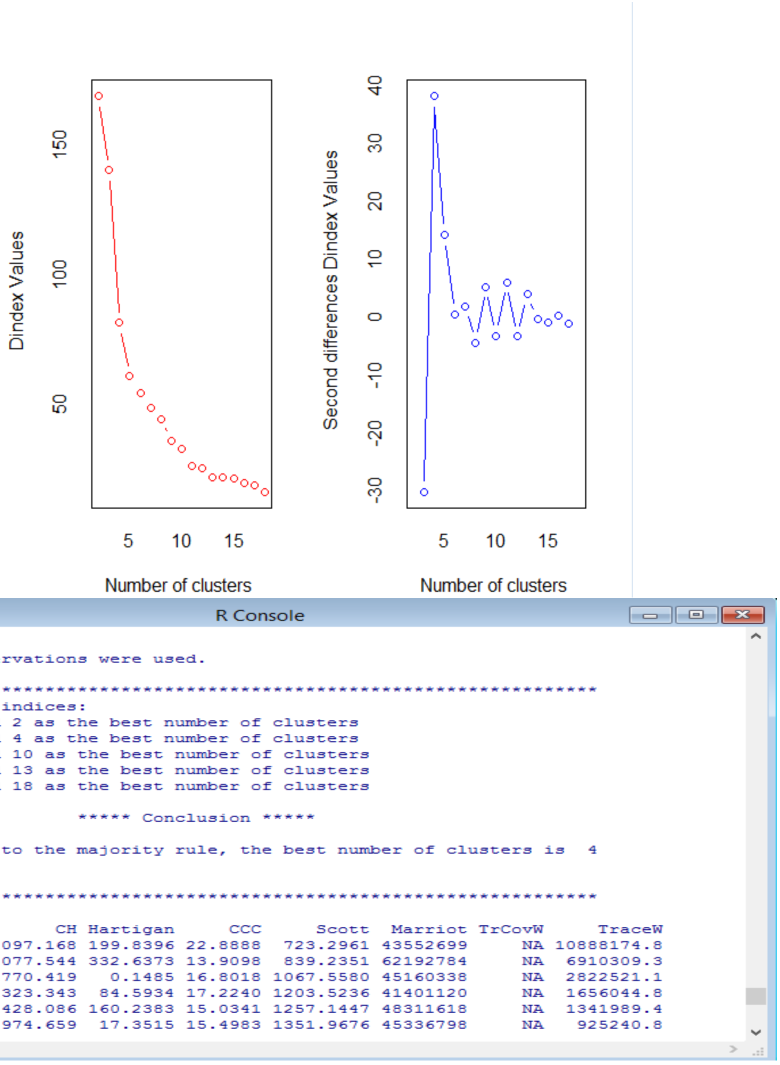


Figure 8.1: Segmentation result with k=4

8.3 Comparison

The results obtained from our thesis work were compared with the NbClust package for finding in R. Similar to Matlab, R is a data analysis tools for performing various calculations and obtaining desired results. R is open source software and can easily be downloaded from web. R provides a library package NbClust which is used for giving number of clusters for different uses. It uses 30 indices as validity indexes for calculating number of clusters. With its majority rules it defines number of classes for a given data set.

We will convert our image into data matrix and then give it into R NbClust package and compare its results with our proposed method, and then we can know that how much correctness we have obtained from our proposed work.



```
All 255 observations were used.

*******************************************************************
* Among all indices:
* 1 proposed 2 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 1 proposed 13 as the best number of clusters
* 1 proposed 18 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4


*******************************************************************
$All.index
       KL       CH Hartigan     CCC     Scott  Marriot TrCovW     TraceW
2  6.5617 1097.168 199.8396 22.8888  723.2961 43552699     NA 10888174.8
3  0.8750 1077.544 332.6373 13.9098  839.2351 62192784     NA  6910309.3
4  0.5431 1770.419   0.1485 16.8018 1067.5580 45160338     NA  2822521.1
5  4.7361 1323.343  84.5934 17.2240 1203.5236 41401120     NA  1656044.8
6  0.4107 1428.086 160.2383 15.0341 1257.1447 48311618     NA  1341989.4
7  0.9401 1974.659  17.3515 15.4983 1351.9676 45336798     NA   925240.8
```

Figure 8.2: Result obtained by R for test_image10

Table 8.2: Comparison Table

Image name	Kl index	CH index	Hartigan index	Scott index	R optimal result	Proposed Thesis Result
Test_image1	12	13	3	3	3	6
Test_image2	2	13	11	12	2	5
Test_image3	3	12	3	3	3	4
Test_image4	7	15	3	3	3	6
Test_image5	2	7	4	3	2	16
Test_image6	10	17	3	3	3	7
Test_image7	7	18	15	4	15	8
Test_image8	12	17	4	5	2	3
Test_image9	7	9	3	3	3	6
Test_image10	10	18	4	2	4	4

From the above comparison result we know that R result did not show much variation and its result are similar as it is selected by majority indexes. Our result showed variation for different images and resulted in dispersed number of cluster. Our accuracy matched was about 95% percent for finding number of cluster after we have tested it on various images.

CHAPTER 9

CONCLUSION

In our thesis an approach for finding optimal k for color image segmentation using K-means we have implemented various validity indexes for finding optimal K for K-mean and then use that k for image segmentation. Cluster is found with help of various distance measure such as inter and intra clusters. Further with the found number of cluster image is segmented into different segments. We were able to achieve 95% accurate result with help of the proposed method and in some cases it gave better results than the R NbClust package.

CHAPTER 10

Future work and Improvement

Following the Major project, we improved the application performance by applying various techniques. Then we tested the result which is quite effective. We implement user friendly interface so that it can be easily used by user. Although, the results from our application are quite satisfied, there are some improvements that can be made in the future for this project such as:

- The application should certainly detect the number of clusters. Because the application sometimes may detect wrong number of clusters.
- The application should also consider initialization of centroid points.

11. REFERENCES

[1] Kitti Koonsanit, Chuleerat Jaruskulchai, and Apisit Eiumnoh, "Determination of the Initialization Number of Clusters in K-means Clustering Application Using Co-Occurrence Statistics Techniques for Multispectral Satellite Imagery" *,International Journal of Information and Electronics Engineering,* Vol. 2, No. 5, September 2012

[2] Sundararajan and Karthikeyan ," An Efficient hybrid approach for data clustering using dynamic k-means algorithm and firefly algorithm", *ARPN Journal of Engineering and Applied Sciences,VOL.* 9, NO. 8, AUGUST 2014.

[3] Ahamed Shafeeq B.M. and Hareesha K.S. 2012," Dynamic Clustering of Data with Modified K-Means Algorithm", *International Conference on Information and Computer Networks* 2012.

[4] Shiv Ram Dubey, Pushka Dixit," Infected Fruit Part Detection using K-Means Clustering Segmentation Technique", *International Journal of Artificial Intelligence and Interactive Multimedia,* Vol. 2, Nº 2.

[5] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE,Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE, "An Efficient k-Means Clustering Algorithm : Analysis and Implementation", *Ieee Transactions On Pattern Analysis And Machine Intelligence*, VOL. 24, NO. 7, JULY 2002.

[6] D T Pham, S S Dimov, and C D Nguyen, "Selection of K in K-means clustering", *Proc. IMechE* Vol. 219 Part C: J. Mechanical Engineering Science.

[7] G Komarasamy,Amitabh Wahi, "A New Algorithm For Selection Of Better K Value Using Modified Hill Climbing In K-Means Algorithm", *Journal of Theoretical and Applied Information Technology* , 30th September 2013. Vol. 55 No.3.

[8] Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation".

[9] S. Prakash kumar and K. S. Ramaswami, "Efficient Cluster Validation with K-Family Clusters on Quality Assessment", *European Journal of Scientific Research,* 2011, pp.25-36.

[10] Rafael C. Gonzalez, *Digital Image processing* [second edition].

[11] Malika Charrad,Nadia Ghazzali,Azam Niknaf,NbClust: “An R Package for Determining theRelevant Number of Clusters in a Data Set”, *Journal of Statistical Software* October 2014, Volume 61, Issue 6.

[12] Pushpa .R. Suri, Mahak, “Image Segmentation With Modified K-MeansClustering Method”, *International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-1, Issue-2, June 2012

[13] Rupali B. Nirgude1, Shweta Jain,”Color Image Segmentation with Different Image Segmentation Techniques” *,International Journal of Engineering Research and General Science* Volume 2, Issue 4, June-July, 2014.

[14] Amiya Halder, “Dynamic Image Segmentation using Fuzzy C-Means based Genetic Algorithm”, *International Journal of Computer Applications* (0975 – 8887) Volume 28– No.6, August 2011.

[15] Priyanka Kirsali_, K. Sambasivarao “An Automatic Detection System for the Detection of Optic Disc and Pathologies in Retinal Images”, *IEEE International Symposium on Signal Processing and Information Technology* (ISSPIT 2014).

[16] Giri babu kande , P.Venkata, T Satya,”Features extraction in Digital Fundus Image”, *Journal of medical and Biological engineering*, 29(3).

[17] MATLAB language of Technical computing, www.mathworks.in/products/matlab/

[18] Function Refrences. Image Processing Toolbox, R2012a Documentation, Mathworks,RetrievedFromhttp://www.mathworks.com/help/toolbox/images/ref/f3-23960.html

[19] GraphicalUserInterfacefrom:http://www.mathworks.com/help/matlab/creating_guis/what-is-a-gui.html.