# Enhanced K-Means Clustering Algorithm for Color Image Segmentation

*Thesis submitted in partial fulfillment of the requirements for the award of degree of*

**Master of Technology**
in
**Computer Science and Application**

*Submitted By*
**Shreyansh Ojha**
**(Roll No. 601103024)**

Under the supervision of
**Dr. Swarnajyoti Patra**
Assistant Professor, SMCA

SCHOOL OF MATHEMATICS AND COMPUTER APPLICATIONS
THAPAR UNIVERSITY
PATIALA – 147004

**July 2013**

# Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*Enhanced K-Means Clustering Algorithm For Color Image Segmentation* ", in partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Applications submitted in School of Mathematics and Computer Applications, Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of **Dr. Swarnajyoti Patra** and refers other researcher's work which are duly listed in the reference section.

The matter presented in this thesis has not been submitted for award of any other degree of this or any other University.

(Shreyansh Ojha)

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

**(Dr. Swarnajyoti Patra)**
Assistant Professor
SMCA

**Countersigned by:**

**(Dr. Rajesh Kumar)**
Head
School of Mathematics and Computer Applications
Thapar University
Patiala

**(Dr. S. K. Mohapatra)**
Dean (Academic Affairs)
Thapar University
Patiala

# Acknowledgement

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life.

This work would not have been possible without the encouragement and able guidance of my supervisor **Dr. Swarnajyoti Patra**. I thank my supervisor for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Rajesh Kumar**, Associate Professor and Head, School of Mathematics and Computer Applications, for motivation and inspiration that triggered me for the thesis work.

I will be failing in my duty if I don't express my gratitude to **Dr. S. K. Mohapatra,** Senior Professor and Dean of Academic Affairs the University, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of School of Mathematics and Computer Applications Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable.

Last but not least, I would like to thank my parents for their wonderful love and encouragement, without their blessings none of this would have been possible. I would also like to thank my brother, since he insisted that I should do so. I would also like to thank my close friends for their constant support.

# Abstract

K-Means Clustering Algorithm is one of the most popular and unsupervised learning algorithm , this research work details the implementation of new adaptive technique for color image segmentation that is the enhanced version of K-Means Clustering Algorithm. The standard K-Means algorithm produces accurate segmentation results only when applied to images defined by homogenous regions with respect to texture and color, and the K supplied to the standard algorithm is user defined which can lead to the empty clusters in the segmentation. In addition, the initialization of the K-Means algorithm is problematic and usually the initial cluster centers are randomly picked, hence there's a need to enhance the K-Means clustering algorithm by first predefining the optimal number of clusters and merging the similar looking cluster until it reaches to the optimal number of cluster.

The main contribution of this work is the enhancement of the K-Means Clustering algorithm that includes the primary features that describe the color smoothness and the quality of the clusters in the process of pixel assignment. The resulting color segmentation scheme has been applied to a large number of natural images and the experimental data indicates the robustness of the new developed segmentation algorithm.

# Table of Contents

# List of Figures

# List of Tables

# Abbreviation

| | |
|---|---|
| **DBI** | Davies Bouldin Index |
| **DI** | Dunn Index |
| **PI** | Pixel Intensity |
| **CCTV** | Closed Circuit Television |
| **CPU** | Centre Processing Unit |
| **ID** | Identification |
| **MV** | Machine Vision |
| **MRI** | Magnetic Resonance Imagin |
| **EM** | Expectation Maximization |

# Chapter1: Introduction

This chapter includes basic introduction of image, image processing, and image segmentation, application of image segmentation and image segmentation algorithms. It also consists a brief introduction of the clustering and the classification of the clustering.

## 1.1 Image

A digital image is a two-dimensional array of small square regions known as pixels. In the case of a monochrome or gray-scale image, the intensity of each pixel is represented by a numeric value. It is also a representation of a two-dimensional image as a finite set of digital values, called picture elements or pixels. Digital images can be created by a variety of input devices and techniques: Digital cameras, Scanners, Coordinate measuring machines. There are various types of digital images. Binary image, Grayscale image and Color image.

### 1.1.1 Binary Image

A binary image is a digital image that has only two possible values, 0 and 1, for each pixel. Binary images are also called bi-level or two-level. A binary image is usually stored in memory as a bitmap, a packed array of bits. Binary images often arise in digital image processing as masks or as the result of certain operations such as segmentation and thresholding.

### 1.1.2 Grayscale Image

Gray-scale images typically contain values in the range from 0 to 255, with 0 representing black, 255 representing white and values in between representing shades of gray.

### 1.1.3 Color Image.

A color image can be represented by a two-dimensional array of Red, Green and Blue triples. Typically, each number in the triple also ranges from 0 to 255, where 0 indicates

that none of that primary color is present in that pixel and 255 indicates a maximum amount of that primary color.

### 1.2 Image Processing

In imaging science, image processing is any form of signal processing for which the input is an image, such as a photograph or video frame the output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. Image processing usually refers to digital image processing but optical and analog image processing also are possible. This article is about general techniques that apply to all of them. The acquisition of images (producing the input image in the first place) is referred to as imaging. Image processing is referred to processing of a 2D picture by a computer. An image defined in the "real world" is considered to be a function of two real variables, for example, a(x,y) with a as the amplitude (e.g. brightness) of the image at the real coordinate position (x,y). Modern digital technology has made it possible to manipulate multi-dimensional signals with systems that range from simple digital circuits to advanced parallel computers. The goal of this manipulation can be divided into three categories:

An image may be considered to contain sub-images sometimes referred to as regions-of-interest, or simply regions. This concept reflects the fact that images frequently contain collections of objects each of which can be the basis for a region. In a sophisticated image processing system it should be possible to apply specific image processing operations to selected regions. Thus one part of an image (region) might be processed to suppress motion blur while another part might be processed to improve color rendition. Sequence of image processing:

The most requirements for image processing of images is that the images be available in digitized form, that is, arrays of finite length binary words. For digitization, the given Image is sampled on a discrete grid and each sample or pixel is quantized using a finite number of bits. The digitized image is processed by a computer. To display a digital image, it is first converted into analog signal, which is scanned onto a display.

Closely related to image processing are computer graphics and computer vision. In computer graphics, images are manually made from physical models of objects, environments, and lighting, instead of being acquired (via imaging devices such as cameras) from natural scenes, as in most animated movies. Computer vision, on the other hand, is often considered high-level image processing out of which a machine/computer/software intends to decipher the physical contents of an image or a sequence of images (e.g., videos or 3D full-body magnetic resonance scans).

In modern sciences and technologies, images also gain much broader scopes due to the ever growing importance of scientific visualization (of often large-scale complex scientific/experimental data). Examples include microarray data in genetic research, or real-time multi-asset portfolio trading in finance.

Before going to processing an image, it is converted into a digital form. Digitization includes sampling of image and quantization of sampled values. After converting the image into bit information, processing is performed. This processing technique may be, Image enhancement, Image reconstruction, and Image compression, Image Segmentation.

### 1.2.1 Image Enhancement or Image Sharpening

It refers to accentuation, or sharpening, of image features such as boundaries, or contrast to make a graphic display more useful for display & analysis. This process does not increase the inherent information content in data. It includes gray level & contrast manipulation, noise reduction, edge crispening and sharpening, filtering, interpolation and magnification, pseudo coloring, and so on.

### 1.2.2 Image Restoration

It is concerned with filtering the observed image to minimize the effect of degradations. Effectiveness of image restoration depends on the extent and accuracy of the knowledge of degradation process as well as on filter design. Image restoration differs from image enhancement in that the latter is concerned with more extraction or accentuation of image features.

### 1.2.3 Image Compression

It is concerned with minimizing the no of bits required to represent an image. Application of compression are in broadcast TV, remote sensing via satellite, military communication via aircraft, radar, teleconferencing, facsimile transmission, for educational & business documents , medical images that arise in computer tomography, magnetic resonance imaging and digital radiology, motion , pictures ,satellite images, weather maps, geological surveys and so on.

### 1.2.4 Image Segmentation

Segmentation is the process of partitioning an image into non-intersecting regions such that each region is homogeneous and the union of no two adjacent regions is homogeneous for some applications, such as image recognition or compression we cannot process the whole image directly for the reason that it is efficient and unpractical. Therefore, several image segmentation algorithms were proposed is to classify or cluster an image into several parts or can say regions. There are lots of image segmentation algorithms exist and be extensively applied in science and daily life. According to their segmentation method, we can approximately categorize them into region-based segmentation, data clustering and edge base segmentation. Image segmentation is useful in many applications. It can identify the regions of interest in scene or annotate the data . Region-based segmentation includes the seeded and unseeded region growing algorithms. The goal of segmentation is typically to locate certain objects of interest which may be depicted in the image, segmentation could therefore be seen as a computer vision problem. Several general-purpose algorithms and techniques have been developed for image segmentation. To be useful, these techniques must typically be combined with a domain's specific knowledge in order to effectively solve the domain's segmentation problems.

#### 1.2.4.1 Threshold Technique

It makes decisions based on local pixel information and is effective when the intensity levels of the objects fall squarely outside the range of levels in the background. Because

spatial information is ignored, however, blurred region boundaries can create havoc. Region boundaries and edges are closely related because there is a sharp adjustment in intensity at the region boundaries. Edge detection techniques have therefore been used as the base of another segmentation technique. The edges identified by edge detection are often disconnected. Closed region boundaries are needed to segment an object from an image. Discontinuities are bridged if the distance between the two edges is within some predetermined threshold [1].

#### 1.2.4.2 Edge Based Methods

This methods center around contour detection: their weakness in connecting together broken contour lines make them, too, prone to failure in the presence of blurring.[2]

#### 1.2.4.3 Region-Based Methods

A Region-based method usually proceeds as follows: the image is portioned into connected regions by grouping neighboring pixels of similar intensity levels. Adjacent regions are then merged under some criterion involving perhaps homogeneity or sharpness of region boundaries. Over stringent criteria create fragmentation, lenient ones overlook blurred boundaries and over merge. [3]

#### 1.2.4.4 Connectivity-preserving relaxation-based

This segmentation method, usually referred to as the active contour model, starts with some initial boundary shape represented in the form of spline curves and literatively modifies it by applying various shrink/expansion operations according to some energy function. Although the energy-minimizing model is not new, coupling it with the maintenance of an elastic contour model gives it an interesting new twist, as usual with such methods, getting trapped into a local minimum is a risk against which one must guard.

#### 1.2.4.5 Graph Partioning Methods

Graphs can effectively be used for image segmentation. Usually a pixel or a group of pixels are vertices and edges define the dissimilarity among the neighbourhood pixels.

Some popular algorithms of this category are random walker, minimum mean cut, minimum spanning tree-based algorithm, normalized cut, etc. The "normalized cuts" method was first proposed by Shi and Malik in 1997. In this method, the image being segmented is modelled as a weighted, undirected graph. Each pixel is a node in the graph, and an edge is formed between every pair of pixels. The weight of an edge is a measure of the similarity between the pixels. The image is partitioned into disjoint sets by removing the edges connecting the segments. The optimal partitioning of the graph is the one that minimizes the weights of the edges that were removed.

#### 1.2.4.6 Mathematical Morphology

Mathematical morphology examines the geometrical structure of an image by probing it with small patterns, called 'structuring elements', of varying size and shape. This procedure results in nonlinear image operators which are well-suited to exploring geometrical and topological structures.

Mathematical Morphology is a tool for extracting image components that are useful for representation and description. Morphology can provide boundaries of objects, their skeletons, and their convex hulls. It is also useful for many pre- and post-processing techniques, especially in edge thinning and pruning.

Most morphological operations are based on simple expanding and shrinking operations. The primary application of morphology occurs in binary images, though it is also used on grey level images. It can also be useful on range images. (A range image is one where grey levels represent the distance from the sensor to the objects in the scene rather than the intensity of light reflected from them).

Mathematical Morphology has several advantages over other techniques especially when applied to image processing like preserves edge information, works by using shape-based processing, can be designed to be idempotent, computationally efficient. Morphology has been used in a wide range of applications. A few of the possible applications are image enhancement, image restoration, Edge detection, texture analysis, noise reduction. [4]

### 1.2.4.7 Split and Merge Method

Split-and-merge segmentation is based on a quadtree partition of an image. It is sometimes called quadtree segmentation.

This method starts at the root of the tree that represents the whole image. If it is found non-uniform (not homogeneous), then it is split into four son-squares (the splitting process), and so on so forth. Conversely, if four son-squares are homogeneous, they can be merged as several connected components (the merging process). The node in the tree is a segmented node. This process continues recursively until no further splits or merges are possible. [5]

### 1.2.4.8 Clustering Methods

Clustering is the grouping of objects or set of objects in such a way that objects in the same group which can be called as cluster are similar to each other than to those in other clusters. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are coherent internally, but clearly dissimilar to the objects belonging to other clusters.

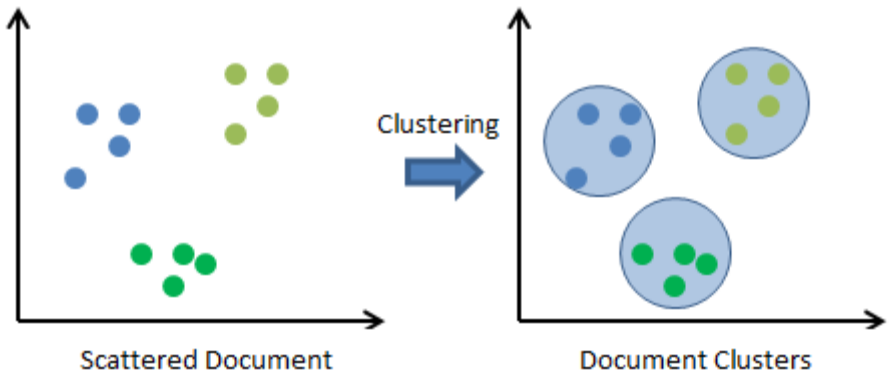


***Figure 1.1*** *Clustering*

Segmentation of an image entails the division or separation of the image into regions of similar attribute. The most basic attribute for segmentation of an image is its luminance amplitude for a monochrome image and color components for a color image. Clustering is one of the methods used for segmentation. In this case we easily identify the 3 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering, here we are going to deal with is distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Clustering algorithms may be classified as listed below

- Flat clustering (Creates a set of clusters without any explicit structure that would relate clusters to each other; It's also called exclusive clustering)
- Hierarchical clustering (Creates a hierarchy of clusters)
- Hard clustering (Assigns each document/object as a member of exactly one cluster)
- Soft clustering (Distribute the document/object over all clusters)

Algorithms

1. Agglomerative (Hierarchical clustering)
2. K-Means (Flat clustering, Hard clustering)
3. EM Algorithm (Flat clustering, Soft clustering)

Hierarchical Agglomerative Clustering (HAC) and K-Means algorithm have been applied to image clustering in a straightforward way. Typically it usages normalized, weighted vectors and cosine similarity. Here, we have illustrated the k-means algorithm using a set of points in n-dimensional vector space for text clustering. [6]

Expectation Maximization (EM) is one of the most common algorithms used for density estimation of data points in an unsupervised setting. The algorithm relies on finding the

maximum likelihood estimates of parameters when the data model depends on certain latent variables. In EM, alternating steps of Expectation (E) and Maximization (M) are performed iteratively till the results converge. The E step computes an expectation of the likelihood by including the latent variables as if they were observed, and maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the last E step. The parameters found on the M step are then used to begin another E step, and the process is repeated until convergence. [7]

The K-means algorithm is an iterative technique that is use to segment the image into k clusters. The basic algorithm [8]:

1. Pick K cluster centers, either randomly or based on some heuristic
2. Assign each pixel in the image to the cluster that minimizes the distance between the pixel and the cluster center
3. Re-compute the cluster centers by averaging all of the pixels in the cluster
4. Repeat steps 2 and 3 until convergence is attained (e.g. no pixels change clusters)

### 1.2.5 Application of Image Segmentation

Some of the practical applications of image segmentation are:

#### 1.2.5.1 Content-based image retrieval

Content-based image retrieval also known as query by image content and content-based visual information retrieval is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases Content based image retrieval is opposed to concept based approaches. Content-based" means that the search will analyze the actual contents of the image rather than the metadata such as keywords, tags, and/or descriptions associated with the image. The term 'content' in this context might refer to colors, shapes, textures, or any other information that can be derived from the image itself. Content-Based Image Retrieval is desirable because most web based image search engines rely purely on metadata and this produces a lot of garbage in the results.

#### 1.2.5.2 Machine vision

Machine vision is the technology and methods used to provide imaging-based automatic inspection and analysis for such applications as automatic inspection, process control, and robot guidance in industry. The scope of MV is broad. Imaging the primary uses for machine vision are automatic inspection and industrial robot guidance. Common MV applications include quality assurance, sorting, material handling, robot guidance, and optical gauging.

#### 1.2.5.3 Medical Imaging

Medical imaging is the technique and process used to create images of the human body (or parts and function thereof) for clinical purposes medical procedures seeking to reveal, diagnose, or examine disease or medical science including the study of normal anatomy and physiology. Although imaging of removed organs and tissues can be performed for medical reasons, such procedures are not usually referred to as medical imaging, but rather are a part of pathology. As a discipline and in its widest sense, it is part of biological imaging and incorporates radiology (in the wider sense), nuclear medicine, investigative radiological sciences, endoscopy, (medical) thermography, medical photography, and microscopy (e.g. for human pathological investigations). A majority of diagnostic imaging centers are located in California, followed by Texas and Florida. Measurement and recording techniques which are not primarily designed to produce images, such as electroencephalography, magnetoen and cephalography, electrocardiography, and others, but which produce data susceptible to be represented as maps (i.e., containing positional information), can be seen as forms of medical imaging. Up until 2010, 5 billion medical imaging studies had been conducted worldwide. Radiation exposure from medical imaging in 2006 made up about 50% of total ionizing radiation exposure in the United States. In the clinical context, "invisible light" medical imaging is generally equated to radiology or "clinical imaging" and the medical practitioner responsible for interpreting (and sometimes acquiring) the image is a radiologist. "Visible light" medical imaging involves digital video or still pictures that can be seen without special equipment. Dermatology and wound care are two modalities

that utilize visible light imagery. Diagnostic radiography designates the technical aspects of medical imaging and in particular the acquisition of medical images. The *radiographer* or *radiologic technologist* is usually responsible for acquiring medical images of diagnostic quality, although some radiological interventions are performed by radiologists. While radiology is an evaluation of anatomy, nuclear medicine provides functional assessment.

As a field of scientific investigation, medical imaging constitutes a sub-discipline of biomedical engineering, medical physics or medicine depending on the context: Research and development in the area of instrumentation, image acquisition e.g. radiography, modeling and quantification are usually the preserve of biomedical engineering, medical physics, and computer science; Research into the application and interpretation of medical images is usually the preserve of radiology and the medical sub-discipline relevant to medical condition or area of medical science (neuroscience, cardiology, psychiatry, psychology, etc.) under investigation. Many of the techniques developed for medical imaging also have scientific and industrial applications.

Medical imaging is often perceived to designate the set of techniques that noninvasively produce images of the internal aspect of the body. In this restricted sense, medical imaging can be seen as the solution of mathematical inverse problems. This means that cause (the properties of living tissue) is inferred from effect (the observed signal). In the case of ultra sonography the probe consists of ultrasonic pressure waves and echoes inside the tissue show the internal structure. In the case of projection radiography, the probe is X-ray radiation which is absorbed at different rates in different tissue types such as bone, muscle and fat.

The term noninvasive is a term based on the fact that following medical imaging modalities do not penetrate the skin physically. But on the electromagnetic and radiation level, they are quite invasive. From the high energy photons in X-Ray Computed Tomography, to the Tesla coils of an MRI device, these modalities alter the physical and chemical environment of the body in order to obtain data.

#### 1.2.5.4 Object detection

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

#### 1.2.5.5 Face Detection

Face detection is a computer technology that determines the locations and sizes of human faces in arbitrary (digital) images. It detects facial features and ignores anything else, such as buildings, trees and bodies. Face detection can be regarded as a specific case of object-class detection. In object-class detection, the task is to find the locations and sizes of all objects in an image that belong to a given class. Examples include upper torsos, pedestrians, and cars. Face detection can be regarded as a more general case of face localization. In face localization, the task is to find the locations and sizes of a known number of faces (usually one). In face detection, one does not have this additional information. Early face-detection algorithms focused on the detection of frontal human faces, whereas newer algorithms attempt to solve the more general and difficult problem of multi-view face detection. That is, the detection of faces that are either rotated along the axis from the face to the observer (in-plane rotation), or rotated along the vertical or left-right axis (out-of-plane rotation), or both. The newer algorithms take into account variations in the image or video by factors such as face appearance, lighting, and pose.

#### 1.2.5.6 Face Recognition

A facial recognition system is a computer application for automatically identifying or verifying a person from a digital image or a video frame from a video source. One of the ways to do this is by comparing selected facial features from the image and a facial database. It is typically used in security systems and can be compared to other biometrics such as fingerprint or eye iris recognition systems. Some facial recognition algorithms

identify facial features by extracting landmarks, or features, from an image of the subject's face. For example, an algorithm may analyze the relative position, size, and/or shape of the eyes, nose, cheekbones, and jaw. These features are then used to search for other images with matching features. Other algorithms normalize a gallery of face images and then compress the face data, only saving the data in the image that is useful for face detection. A probe image is then compared with the face data. One of the earliest successful systems is based on template matching techniques applied to a set of salient facial features, providing a sort of compressed face representation. Recognition algorithms can be divided into two main approaches, geometric, which looks at distinguishing features, or photometric, which is a statistical approach that distills an image into values and compares the values with templates to eliminate variances.

#### 1.2.5.7 Fingerprint recognition

Fingerprint recognition or fingerprint authentication refers to the automated method of verifying a match between two human fingerprints. Fingerprints are one of many forms of biometrics used to identify individuals and verify their identity. This article touches on two major classes of algorithms and four sensor designs optical, ultrasonic, passive capacitance, and active capacitance.

#### 1.2.5.8 Iris recognition

Iris recognition is an automated method of biometric identification that uses mathematical pattern-recognition techniques on video images of the irides of an individual's eyes, whose complex random patterns are unique and can be seen from some distance. Not to be confused with another, less prevalent, ocular-based technology, retina scanning, iris recognition uses camera technology with subtle infrared illumination to acquire images of the detail-rich, intricate structures of the iris. Digital templates encoded from these patterns by mathematical and statistical algorithms allow the identification of an individual or someone pretending to be that individual. Databases of enrolled templates are searched by matcher engines at speeds measured in the millions of templates per second per (single-core) CPU, and with infinitesimally small False Match rates. Many millions of persons in several countries around the world have been enrolled

in iris recognition systems, for convenience purposes such as passport-free automated border-crossings, and some national ID systems based on this technology are being deployed. A key advantage of iris recognition, besides its speed of matching and its extreme resistance to False Matches is the stability of the iris as an internal, protected, yet externally visible organ of the eye.

#### 1.2.5.9 Video surveillance

Closed-circuit television (CCTV) is the use of video cameras to transmit a signal to a specific place, on a limited set of monitors. It differs from broadcast television in that the signal is not openly transmitted, though it may employ point to point (P2P), point to multipoint, or mesh wireless links. Though almost all video cameras fit this definition, the term is most often applied to those used for surveillance in areas that may need monitoring such as banks, casinos, airports, military installations, and convenience stores. Video telephony is seldom called "CCTV" but the use of video in distance education, where it is an important tool, is often so called.

## 1.3 Thesis Outline

This thesis is organized into five chapters. Chapter 1 describes about the basic concepts of Image, Image processing, Image Segmentation, application of image segmentation, a brief introduction to methods of the image segmentation and clustering. Chapter 2 describes information related to the history of K-means Clustering algorithm, literature survey which has been done during this thesis. Chapter 3 describes the motivation behind the thesis, discusses the problem statement and its objectives. Chapter 4 explains all the results obtain from the algorithm that has been developed for color image segmentation using enhanced K-means clustering algorithm. Chapter 5 summarizes the conclusions drawn from the work done along with the directions regarding the future work.

# Chapter 2: Literature survey

Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. Image segmentation is a fundamental task in computer vision. Although many methods are proposed, it is still difficult to accurately segment an arbitrary image by any method alone. In recent years, more and more attention has been paid to combining segmentation algorithms and information from multiple feature spaces (e.g. colour, texture, and pattern) in order to improve segmentation results. This chapter aims to review previous work on K-means Clustering Algorithm and determination of number of clusters in K-Means Clustering algorithm.

## 2.1 K-Means Clustering Algorithm

K-Means was first used by James MacQueen in 1967.[8] This standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, and then 1965, E.W.Forgy published essentially the same method that is why it is sometimes referred as Lloyd-Forgy too[9]. A more efficient version then was proposed and published by Hartigan and Wong in 1975/1979 in Fortran [10-11]. The k-means clustering algorithm is known to be efficient in clustering large data sets. is one of the simplest and the best known unsupervised learning algorithms that solve the well-known clustering problem. The K-Means algorithm aims to partition a set of objects, based on their attributes/features, into k clusters, where k is a predefined or user-defined constant. The main idea is to define k centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related (in terms of similarity function; similarity can be measured by using different methods such as cosine similarity, Euclidean distance, Extended Jaccard) to all objects in that cluster. In this case, distance is the squared or absolute difference between a pixel and a cluster center. The difference is typically based on pixel color, intensity, texture, and location, or a weighted combination of these factors. *K* can be selected manually, randomly, or by a heuristic. This algorithm is guaranteed to

converge, but it may not return the optimal solution. The quality of the solution depends on the initial set of clusters and the value of *K*. K-Means can be thought of as an algorithm relying on hard assignment of information to a given set of partitions. At every pass of the algorithm, each data value is assigned to the nearest partition based upon some similarity parameter such as Euclidean distance of intensity. The partitions are then recalculated based on these hard assignments. With each successive pass, a data value can switch partitions, thus altering the values of the partitions at every pass. K-Means algorithms typically converge to a solution very quickly as opposed to other clustering algorithms.[12]

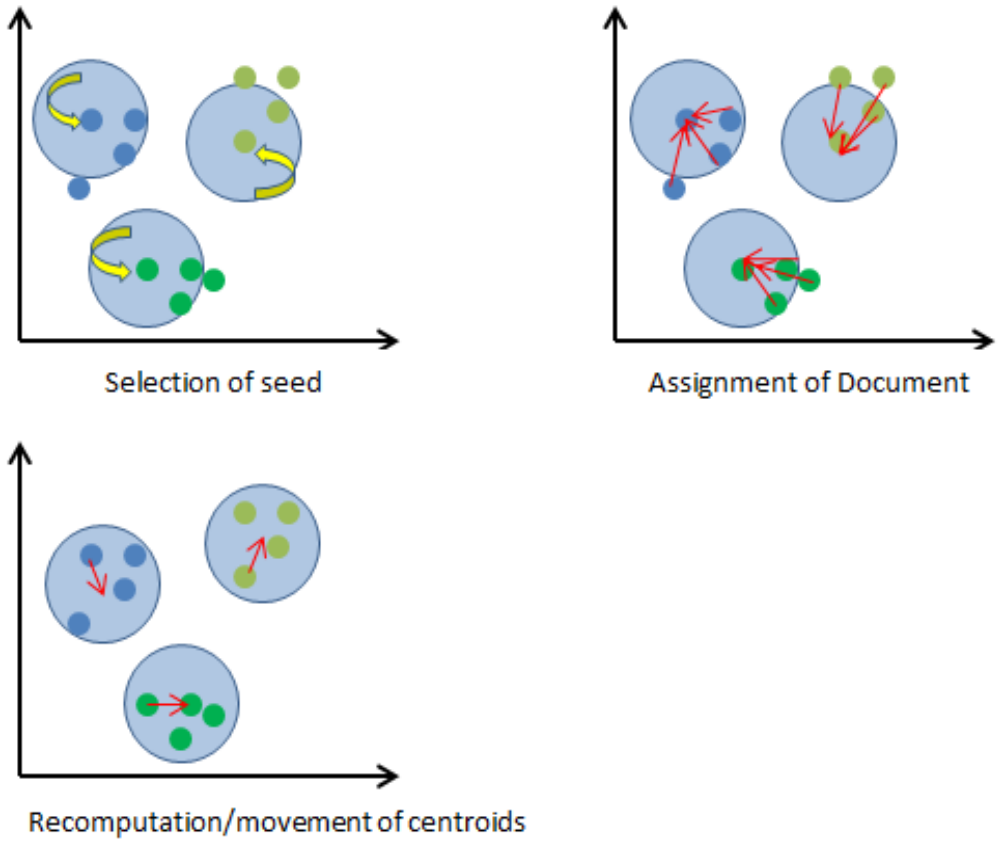


***Figure 2.1*** *Clustering Of data set in two dimensional feature using K-means clustering Algorithm.*

The K-Means algorithm a straight forward and widely-used clustering algorithm. Given a set of objects (Image Pixel), the goal of clustering or segmentation is to divide these objects into groups or "clusters" such that objects within a group tend to be more similar to one another as compared to objects belonging to different groups. In other words, clustering algorithms place similar points in the same cluster while placing dissimilar points in different clusters. Note that, in contrast to *supervised* tasks such as regression or classification where there is a notion of a target value or class label, the objects that form the inputs to a clustering procedure do not come with an associated target. Therefore clustering is often referred to as unsupervised learning. Because there is no need for labeled data, unsupervised algorithms are suitable for many applications where labeled data is difficult to obtain [13]. Unsupervised tasks such as clustering are also often used to explore and characterize the dataset before running a supervised learning task. Since clustering makes no use of class labels, some notion of similarity must be defined based on the attributes of the objects. The definition of similarity and the method in which points are clustered differ based on the clustering algorithm being applied. Thus, different clustering algorithms are suited to different types of data sets and different purposes. The "best" clustering algorithm to use therefore depends on the application. It is not uncommon to try several different algorithms and choose depending on which is most useful. This clustering algorithm is convergent and its aim is to optimize the partitioning decisions based on a user-defined initial set of clusters. The applications of clustering algorithms to the segmentation of complex color textured images are restricted by two problems. The first problem is generated by the starting condition that is the initialization of the initial cluster centers while the second is generated by the fact that no spatial or regional cohesion is applied during the space partitioning process.

The K-Means method aims to minimize the sum of square distanced between all points and the cluster centre. There are few steps described by Tou and Gonzalez [14].

1. Choose K initial cluster centers, $z_1(1)$, $z_2(1)$,……, $z_k(1)$.
2. At the k-th iterative step, distribute the samples {x} among the K clusters using the relation

$$x \epsilon C_j(k) \;\; if \;\; \left\| x - z_j(k) \right\| < \left\| x - z_i(k) \right\|$$

For all $i = 1,2, \ldots \ldots, k;\ \ i \neq j$; where $C_j(k)$ denote the set of samples whose cluster centre is $z_j(k)$

3. Compute the new cluster centers $z_j(k+1)$, $j = 1,2, \ldots \ldots, k$ such that the sum of the squared distance from all points in $C_j(k)$ to the new cluster centre is minimized. The measure which minimizes this is simply the same mean of $C_j(k)$. Therefore the new cluster centre is given by

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \epsilon C_j(k)} \quad j = 1,2, \ldots \ldots, k$$

Where $N_j$ is the number of samples in $C_j(k)$.

4. if $z_j(k+1) = z_j(k) for\ j = 1,2, \ldots \ldots, K$ then the algorithm has converged and the procedure is terminated

Otherwise go to step 2

It is obvious in this description that the final clustering will depend on the initial cluster centers chosen and on the value of K. the latter is one of the most concern since this requires some prior knowledge of the number of clusters present in the data, which in practice, is highly unlikely.

## 2.2 Methods for finding the optimal no. of clusters and cluster validity

In this section , we present the research topics of finding the optimal number of clusters for color image segmentation. While clustering and segmentation algorithms are unsupervised learning processes, users are usually required to set some parameters for these algorithms. These parameters vary from one algorithm to another, but most clustering/segmentation algorithms require a parameter that either directly or indirectly specifies the number of clusters/segments. This parameter is typically either k , the number of clusters/segments to return, or some other parameter that indirectly controls the number of clusters to return, such as an error threshold. Setting these parameters requires either detailed pre-existing knowledge of the data, or time-consuming trial and error. The latter case still requires that the user has sufficient domain knowledge to know what a good clustering “looks” like. However, if the data set is very large or is multi-dimensional, human verification could become difficult. To find a reasonable number of

clusters, many existing methods must be run repeatedly with different parameters, and are impractical for real-world data sets that are often quite large. We desire an algorithm that can efficiently determine a reasonable number of clusters/segments to return from any hierarchical clustering/segmentation algorithm. We have to identify the correct number of clusters to return from a hierarchical clustering/segmentation algorithm. The definition of a "cluster" is not well-defined, and measuring cluster quality is subjective. Thus, there are many clustering algorithms with unique evaluation functions and correspondingly unique notions of what a good cluster "looks" like. In cluster analysis, a fundamental problem is to determine the best estimate of the number of clusters, which has a deterministic effect on the clustering results. However, a limitation in current applications is that no convincingly acceptable solution to the best-number-of-clusters problem is available due to high complexity of real data sets. Choosing an appropriate clustering method is another critical step in clustering. A large number of clustering methods are available for cluster analysis. However a fundamental problem in applying most of the existing clustering approaches is that the number of clusters needs to be pre-specified before the clustering is conducted. The clustering results may heavily depend on the number of clusters specified. It is necessary to provide educated guidance for determining the number of clusters in order to achieve appropriate clustering results. At the current stage of research, none of the existing methods of choosing the optimal estimate of the number of clusters is completely satisfactory. The gap method was recently proposed by Tibshirani, *et al*. [15]. The main idea of the gap method is to compare the within cluster dispersions in the observed data to the expected within cluster dispersions assuming that the data came from an appropriate null reference distribution. Simulation results reported by Tibshirani, *et al.* indicated that the gap method is a potentially powerful approach in estimating the number of clusters for a data set. However, recent studies have shown that there are situation where the gap method may perform poorly. For example, when the data contain clusters which consist of objects from well separated exponential populations. The correct choice of $k$ is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing $k$ without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered

its own cluster (i.e., when *k* equals the number of data points, *n*). Intuitively then, the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster. If an appropriate value of *k* is not apparent from prior knowledge of the properties of the data set, it must be chosen somehow. There are several methods of finding the optimal number of clusters. [16]

### 2.2.1 Davies Bouldin Index

The Davies–Bouldin index (DBI) was introduced by David L. Davies and Donald W. Bouldin in 1979[17] is a metric for evaluating clustering algorithms. This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset. This has a drawback that a good value reported by this method does not imply the best information retrieval.

Let $R_{ij}$ be a measure of how good the clustering scheme is and let $M_{ij}$ be the separation between the $i^{th}$ and the $j^{th}$ cluster, which ideally has to be as large as possible, and $S_i$ the within cluster scatter for cluster I, which should be as low as possible. Hence the Davies Bouldin Index is defined as the ration of S and M. such that these properties are conserved:

1. $R_{i,j} \geq 0$.
2. $R_{i,j} = R_{j,i}$.
3. if $S_j \geq S_k$ and $M_{i,j} = M_{i,k}$ then $R_i, j > R_{i,k}$.
4. and if $S_j = S_k$ and $M_{i,j} \leq M_{i,k}$ then $R_{i,j} > R_{i,k}$.

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

This is the symmetry condition. Due to such formulation, the lower the value is the better the separation of the clusters and the tightness inside the clusters.

$$D_i \equiv \max_{j:i \neq j} R_{i,j}$$

If N is the number of clusters:

$$\mathrm{DB} \equiv \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathrm{D_i}$$

DB is calling the Davies Bouldin Index. This is dependent both on the data as well as the algorithm. $D_i$ chooses the worst case scenario, and this value is equal to $R_{i,j}$ for the most similar cluster to cluster *i*. there could be many variations to this formulation, like choosing the average of the cluster similarity, weighted average and so on.

### 2.2.2 Dunn's Index

The Dunn index (DI) was introduced by J.C. Dunn in 1974[18] is a metric for evaluating clustering algorithms. Its an internal evaluation scheme where the result is based on the clustered data itself. In this indices the aim is to identify the sets of clusters which are compact with a little bit variance between the members of the cluster, and well separated where the means of different clusters are sufficiently far apart as compared to the within cluster variance. For a given assignment of clusters, a higher Dunn index indicates better clustering. One of the drawbacks of using this, is the computational cost as the number of clusters and dimensionally of the data increase.

Being defined in this way, the DI depends on m, the number of clusters in the set . if the number of clusters is not known apriori, the m for which the DI is the highest can be chosen as the number of clusters. There is also some flexibility when it comes to the definition of d(x,y) where any of the well known metric can be used, like Manhattan distance or Euclidean distance based on the geometry of the clustering problem. This formulation has a peculiar problem, in that if one of the clusters is badly behaved, where the others are tightly packed, since the denominator contains a max term instead of an average term, the DI for that set of clusters will be uncharacteristically low. This is thus some sort of a worst case indicator, and has to be used keep that in mind. There are ready implementations of DI in some vector based programming languages.

$$DI_m = \min_{1 \le i \le m} \left\{ \min_{1 \le j \le m, j \neq i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le k \le m} \Delta_k} \right\} \right\}$$

### 2.2.3 Validity Measure

There have been many criteria developed for determining cluster validity [17-23]. All of which have a common goal to find the clustering which results in compact clusters which are well separated. The DBI [17], is a function of the ratio of the sum of within cluster scatter to between cluster separations. The objective is to minimize this measure as we want to minimize the within cluster scatter and maximize the between cluster separation. Bezdek and Pal[19] have given a generalization of Dunn's index[18]. Considering the five different measures of distance function between clusters and three different measures of cluster diameter. They obtained fifteen different values of the generalized DI.

Since K-Means method aims to minimize the sum of squared distances from all points to their cluster centers, this should result in compact clusters. We can therefore use the distances of the points from their cluster centre to determine whether the clusters are compact. For this purpose we use the intra cluster distance measure which is simply the distance between a point and its cluster centre and we take the average of all of these distances defined as

$$\text{intra} = \frac{1}{N}\sum_{i=1}^{k}\sum_{x\epsilon\, C_i}\|x - Z_i\|^2$$

where N is the number of pixels in the image, K is the number of cluster $C_i$. we obviously want to minimize the measure. We can also measure the inter cluster distance or the distance between clusters, which we want to be as big as possible. We calculate this as the distance between cluster centers, and take the minimum of this value, defined as

$$\text{inter} = \min\left(\left\|Z_i - Z_j\right\|^2\right), i = 1,2,\ldots..K-1, j = i+1,\ldots..,K$$

we take only the minimum of this value as we want the smallest of this distance to be maximized, and the other larger values will automatically be bigger than this value. We want both of these measures to help us determine if we have a good clustering .

$$\text{validity} = \frac{\text{intra}}{\text{inter}}$$

We want to minimize the intra cluster distance and this measure is in the numerator, we consequently want to minimize the validity measure, we also want to maximize the inter cluster distance measure and since this in the denominator we again want to minimize the validity measure. Therefore, the clustering which gives a minimum value for the validity measure will tell us what the ideal value of K is in the K-Means procedure.

**Description of Validity Measure**

There are numbers of color spaces exist in which the segmentation of images can be performed. The method and the result are all based on the three features of image that is red, green blue color space but the method can be easily implemented in any color space, we are concerned with the segmented image for 2 up to $K_{max}$ clusters, where $K_{max}$ is an upper limit on the number of clusters, and then we need to calculate the validity measure which can determine the best clustering and the optimal value of K, we do this by first forming one cluster which contains all the pixels of the image, and then an iterative process starts which last till the number of cluster reaches to $K_{max.,}$ those clusters which are having the maximum variance are split into two clusters. Once the cluster is split then we apply K-means clustering algorithm to obtain the clustering for the new number of clusters. Once we have all the clusters formed then the validity measure is calculated for each of them to determine what the optimal value of K is.

As K-Means algorithm aims is to minimize the average intra cluster distance. It is most likely that the cluster having maximum variance will be separated by the K-Means procedure when the number of clusters is increased. Hence, when we need the number of clusters to be increased , we split the clusters whose having the maximum variance so the K-Means clustering algorithms is given good starting cluster centers. [24]

Calculation of the variance of the three components for cluster $C_i$ as

$$\sigma_{ij}^2 = \frac{1}{N_i} \sum_{x \epsilon C_i} \left(x - Z_{ij}\right)^2 , i = 1,2, \ldots . , K \; and \; j = 1,2,3$$

Where $N_i$ is the number of pixels in cluster $C_i$ and x is the vector representing each pixel's red, green and blue components as $x_1$, $x_2$ and $x_3$ respectively. This will give us three variance values, but ultimately we need just one value, which we will use to compare the variance of each cluster, hence we can take the average variance of the three components by adding them up and then dividing the sum by 3 . this will give us the following variance values.

$$\sigma_i^2 = \frac{1}{3}\sum_{j=1}^{3} \sigma_{ij}^2 \qquad i = 1,2, \dots \dots, K.$$

After splitting a cluster, we consider all three of the red, green and blue components. Given the cluster $C_i$ whose cluster centre is $z_i$ , we can split this cluster into two new cluster such that the new cluster centers can be defined as $z_i^{`}$ and $z_i^{``}$. The two new cluster centers are then calculated as

$$Z_i^{`} = (Z_{i1} - a_1,\ Z_{i2} - a_2,\ Z_{i3} - a_3)$$

$$Z_i^{``} = (Z_{i1} + a_1,\ Z_{i2} + a_2,\ Z_{i3} + a_3)$$

Where $a_1$, $a_2$, and $a_3$ are constants. The values of these constants are determined by taking into account the minimum and maximum values for each color component occurring in the cluster. The constants, $a_j$ will be the values which are half of the smaller of $(z_{ij} - \min_j)$ and $(z_{ij} - \max_j)$ where $\min_j$ is the minimum value for the *j-th* color component and $\max_j$ is the maximum value for *j-th* component. This will make two new cluster centers being well separated, but also still well within the original cluster.

# Chapter 3: Problem Statement

This chapter includes the problem statement and objectives of the thesis.

## 3.1 Problem Statement

The K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters; it is well suited to generating globular clusters. The K-Means method is numerical, unsupervised, non deterministic and iterative. The main disadvantage of the k-means algorithm is that the number of clusters, K must be supplied as a parameter. The learning algorithms require apriori specification of the number of cluster centers. If there are two highly overlapping data then K-Means will not be able to resolve that there are two clusters because of the use of Exclusive Assignment. The learning algorithm provides the local optima of the squared error function. Randomly choosing of the cluster center cannot lead us to the fruitful result. Difficult in comparing the quality of clusters which results in noisy image, hence we need to improve the quality of the clusters; in K-Means clustering algorithm different initial partitions can result in different final clusters. For the large number of clusters the K-Means clustering algorithm can make several empty clusters, if no points have been allocated to a cluster in the assignment step which can result in noisy image, so we need a criteria through which we can remove the problem of empty clusters creation. The problem statement of this thesis is to remove the empty clusters and consider each pixel as the data points and then apply K-Means Clustering algorithm to generate the large number of clusters and then apply some technique to merge the similar looking clusters and remove the empty clusters.

## 3.2 Thesis Objectives

Our Thesis objective is to enhance the existing K-Means Clustering Algorithm for Color Image Segmentation. we need to find the method which through which we can find the optimal number of cluster for K-Means clustering then validate the validity of the clusters, after that we need to find the criteria through which we can merge the similar

cluster which can be similar in terms of the variance or the distance measure. We have to assume each pixel of the input image as a data point, hence there will be n data points or we can say n patterns $x_1, x_2, x_3, \ldots, x_n$ generated from the input image which can have n number of pixels. After that we need to apply K-Means clustering algorithm to the whole set of patterns so that we can generate the large number of clusters, where each cluster contains the similar type of data points, then we need to find the method to select the two clusters from these large number of clusters for merging . The merging will be based on some similarity between each of the existing cluster pairs, to find the similarity between the existing cluster we need to consider the variance of each individual clusters and also the distance between the two cluster centers. We need to develop an algorithm which can merge the similar looking clusters based on similarity criteria.

# Chapter 4: Proposed Work

This chapter contains the details of the work that has been done to meet all the thesis objectives.

## 4.1 Enhanced K-Means clustering Algorithm for Color Image Segmentation.

As we know that in K-means algorithm the number of clusters, K must be supplied as a parameter hence validity measure which is based on the intra-cluster distance measures which allows the number of clusters to be determined automatically. The basic procedure involves producing all the segmented images for 2 clusters up to $K_{max}$ clusters, where $K_{max}$ 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. Validity measure is simply the ratio of intra cluster distance measure to the inter cluster distance, intra cluster distance is defined as the distance between a point and its cluster centre within a cluster whereas inter cluster distance is the distance between the two clusters. While performing the image segmentation for a given input image we first regroup the pixels together to form a set of coherent image regions. After taking color image as input we first calculate the total number of pixels in the image. For the similarity of the pixels, we can measure it on the basis of different feature like intensity, color, texture , local entropy , etc. but here in the proposed algorithm we have taken the three color feature for the similarity of the pixels. Individual features or the combinations of them can be used to represent an image pixel. Hence we can associate a feature vector x to each of the pixel of input image. Clustering is then performed on the set of feature vectors so as to group them. Finally clustering result is mapped back to the original spatial domain to obtain the segmented image.

We start our work by first assuming each pixel of the input image as a data point or we can say patterns, suppose an input image has n pixel hence n patterns or data points $x_1, x_2, \ldots\ldots, x_n,$ can be generated from the input image. First, apply the K-Means Clustering

algorithm to the whole data set to generate the large number of clusters. Each cluster hence contain the similar type of data points, then we apply the validity measure which is already discussed in literature survey to find the optimal number of clusters and the validity of the clusters, then from these large number of clusters which has been found by K-Means clustering algorithms, two clusters are selected for merging by using the similarity criterion function. This criterion function measures the similarity between each existing cluster pairs, the similarity is considered by using the variance of each individual cluster, the variance of compound cluster which is the merged cluster and the square distance between the two cluster centers. Then the cluster merging process continues until we get the pre-defined number of cluster which is the optimal number of cluster which we found by using validity measure. In contrast to Bhattacharyya distance [25] or Jeffries-Matusita distance [26] based criterion functions (Mak and Barnard, 1996; Richards and Jia. 2006), the proposed simple criterion function does not need to compute the inverse of the covariance matrix.

Let $C_1, C_2, \ldots\ldots\ldots, C_k$ be the intial k clusters obtained by applying K-Means clustering algorithm . $\mu_i$ and $\sigma_i^2$ are the mean and variance of cluster $C_i$. the similarity criterion function $S(C_i, C_j)$ measures the similarity between cluster pair $C_i$ and $C_j$. It is defined as follows:

$$S(C_i, C_j) = \frac{\left|\left(\sigma_i^2 + \sigma_j^2\right) - \sigma_{ij}^2\right| \times d_{ij}^2}{N} \qquad (1)$$

Where $\sigma_{ij}^2$ is the variance of the merged cluster $C_i$ and $C_j$, $d_{ij}^2 = \left\|\mu_i - \mu_j\right\|^2$ and the N mentioned above is $N = arg\max_{ij}\{d_{ij}^2\}$ is used as a normalization factor. If the patterns are modeled by m features, the variance $\sigma_{ij}^2$ of the cluster $C_{ij}$ is computed as:

$$\sigma_{ij}^2 = \sqrt{\sum_{l=1}^{m}\left(\sigma_{ij},_l^2\right)^2} \qquad (2)$$

Where $\sigma_{ij},_l^2$ is the variance of $l^{th}$ feature of the patterns which belong to the cluster $C_{ij}$ and is computer as Ross (1987):

$$\sigma_{ij,l}^2 = \frac{n_i \times (\sigma_{i,l}^2 + d_{i,l}^2) + n_j \times (\sigma_{j,l}^2 + d_{j,l}^2)}{n_i + n_j} \quad (3)$$

Where

$$d_{i,l}^2 = \left\| \mu_{i,l} - \mu_{ij,l} \right\|^2 \quad (4)$$

$$d_{j,l}^2 = \left\| \mu_{j,l} - \mu_{ij,l} \right\|^2 \quad (5)$$

$$\mu_{ij,l} = \frac{n_i \times \mu_{i,l} + n_j \times \mu_{j,l}}{n_i + n_j} \quad (6)$$

And $n_i$ and $n_j$ denote the number of points included in $C_i$ and $C_j$, respectively.

The similarity criterion function defined in Eq.(1) measures the similarity between cluster pair $C_i$ and $C_j$ by considering three terms:

I. The variances ($\sigma_i^2$ *and* $\sigma_j^2$) of each individual cluster $C_i$ and $C_j$,

II. The variance ($\sigma_{ij}^2$) which is obtained by merging the cluster pair $C_i$ and $C_j$ , and

III. The Euclidian distance $d_{ij}^2$ between the cluster centers $\mu_i$ and $\mu_j$.

From Eq.(1) we can say that, if the difference between the sum of the variance $\sigma_i^2 + \sigma_j^2$ Of two individual clusters and the variance ($\sigma_{ij}^2$) of the merged clusters is small and the distance between two individual cluster centers is small, then the data point included in the cluster pair are very similar. So the cluster pair associated with the smallest value of *S* (.) is the best candidate for merging.

Before applying the above discussed similarity criterion function on two selected clusters, first we need to find the empty clusters, we can remove all empty clusters easily from these group of clusters by just exchanging the first found empty cluster to the last cluster and then remove the last cluster. For example if cluster number 8 is the group of 30 clusters, then we can exchange the cluster number 8 to the last cluster that is cluster number 30 and then we can see that now cluster number 30 is at the position of cluster number 8 and cluster number 8 is at the position of cluster number 30, now remove this cluster number 30, now we are left with total 29 clusters, we can again find if any empty cluster left or not. Then we can move ahead to find the similar clusters which are very

similar in terms of minimum distance between them or the individual cluster variance which is already discussed above.

## 4.2 Proposed Algorithm

This section explains how this algorithm actually works. We are providing a pseudo code which explains the working of this algorithm. We have taken an input image which can be any color image which has already been used for many color image segmentations in past years. We are consider with the three color feature of the color image that is red green and blue, validity measure has been used here for finding the optimal number of cluster and also it tells us about the quality of the cluster.

Algorithm : Proposed Method

Input : (i) $x_1, x_2, \ldots\ldots, x_n$ are the n data points (or input patterns) are generated from the input image , which has n number of pixels. Each pixel corresponds to a data point..

(ii) r is the pre-defined number of clusters which we found using validity measure.

Output: A segmented Color Image.

Step 1. Apply the K-Means Clustering Algorithm to the whole input data points And the selecting a a value of K is large enough ($k > r$). Let $C_1, C_2, \ldots, C_k$ be the cluster obtained by applying the K-Means Clustering Algorithm.

Step 2. Measure the similarity between each cluster pairs that is $C_i$ and $C_j$ , $\forall(i, j)$ and $i \neq j$ by using Eq. (1).

Step 3. Find the cluster pair $C_p$ and $C_q$ such that $\arg \min_{\forall(i,j), i \neq j} \{S(C_i, C_j)\}$

Step 4. Merge the cluster $C_p$ and $C_q$ and the compute the new mean and variance Of the merged cluster $C_{pq}$ by using Eq. (6) and Eq. (3). respectively.

Step 5 k = k – 1.

Step 6 If $k \neq r$ then goto step 2.

Step 7 the data points in the r clusters are mapped back to the special domain to Segment the image into different regions.

Step 8 Stop

## 4.3 Experimental Results

The algorithm has been applied on a number of images. Experiments were carried out on a large number of different kinds of images. The present investigation was done by considering three feature of image red, green and blue value of a pixel. However, please not that in general we can take any number of features.

Input Image

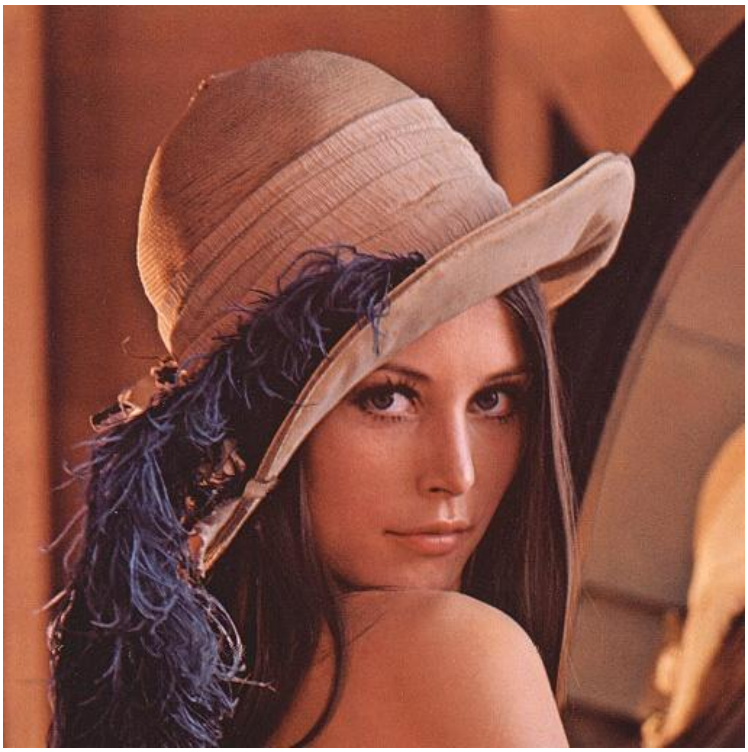

***Figure 4.1*** *Lena (Original Image)*

Above is the original image, now we will first apply the normal K-Means Clustering algorithm by taking the large number of clusters. Suppose here the number of cluster k = 30. Hence after applying K-Means clustering algorithm the new image will be segmented into 30 different cluster, not that in these 30 clusters there can be one or any number of empty clusters. As in K-Means, if there is no pixel assign to any one the cluster then it becomes the empty cluster, it doesn't remove the empty clusters which can result in noisy image .

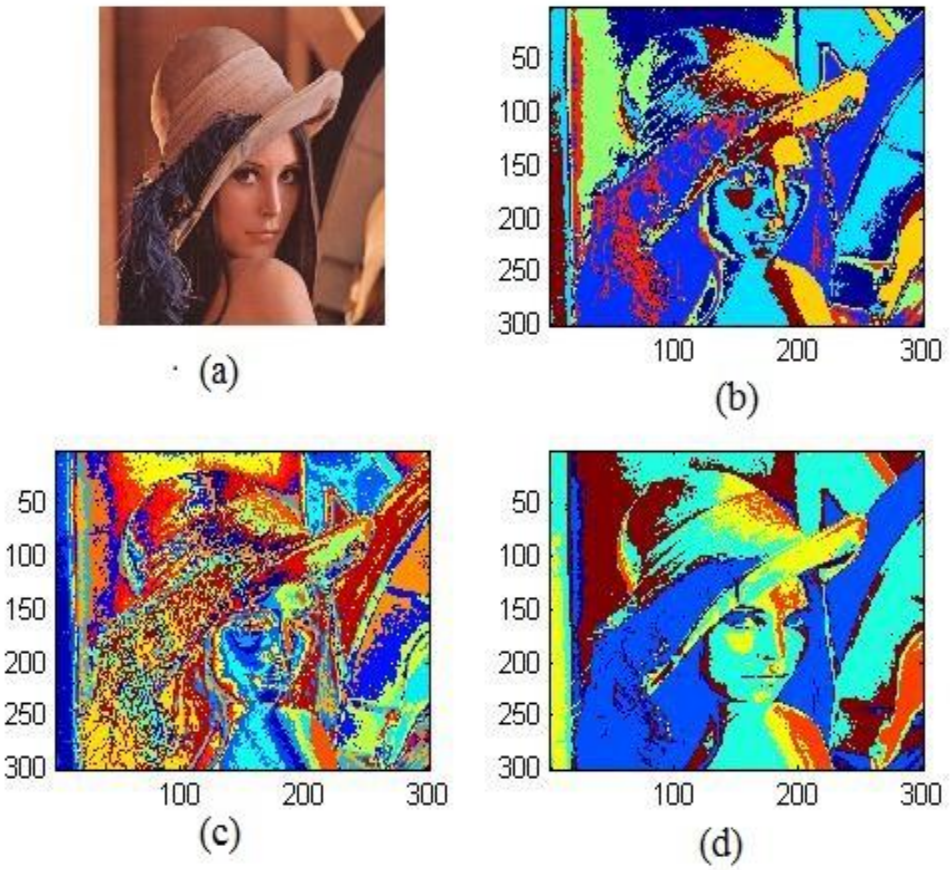


**Figure 4.2** *(a) Original Lena Image (b) Segmented Image Using K-Means ( K =7 ) (c) Segmented Image Using K-Means (K = 30) (d) Segmented Image Using Enhanced K-Means( K = 7).*

Here in this First image is original image , Second Image is a segmented image by using K-Means Clustering Algorithm and K supplied to this algorithm is found by using Validity Measure, third image is segmented using K-Means clustering algorithm where K supplied to it is 30 ( a large number of clusters) and the last image if segmented by using the Enhanced K-Means Clustering Algorithm which was our proposed work , now we will test results for few more images and then will compare our proposed method that is the Enhanced K-means clustering algorithm to Original K-Means clustering algorithm. The comparison will be done by using the validity measure, we can use the validity measure to compare both these algorithms, and minimum the value of validity measure proves that the image is better segmented.

### 4.4 Result Analysis

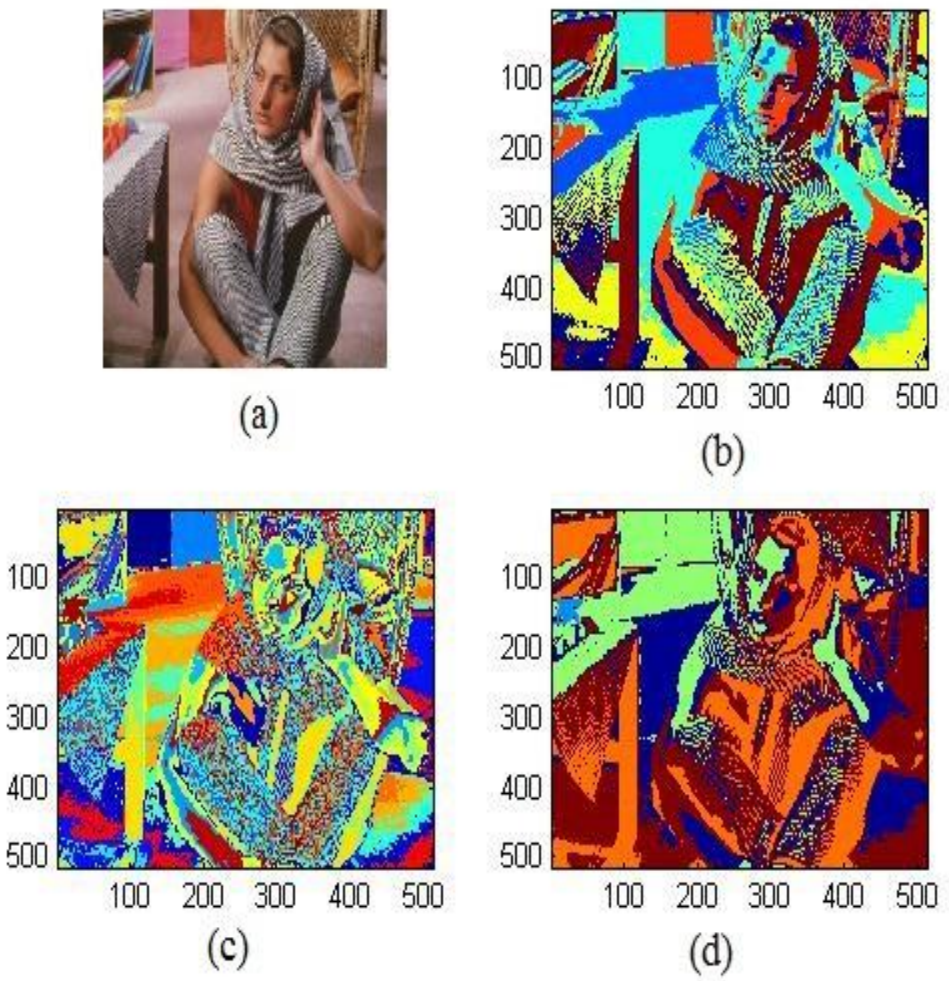


**Figure 4.3** *(a) Original Barbara Image (b) Segmented Image Using K-Means ( K =6) (c) Segmented Image Using K-Means (K = 30) (d) Segmented Image Using Enhanced K-Means( K = 6).*

Here the optimal number of cluster found using validity measure is 6, hence the second image is clustered using K-Means where K = 6, third image is normally segmented using K-Means clustering algorithm where K = 30, and fourth image is segmented using Enhanced K-Means algorithm where K = 6 found using validity measure.

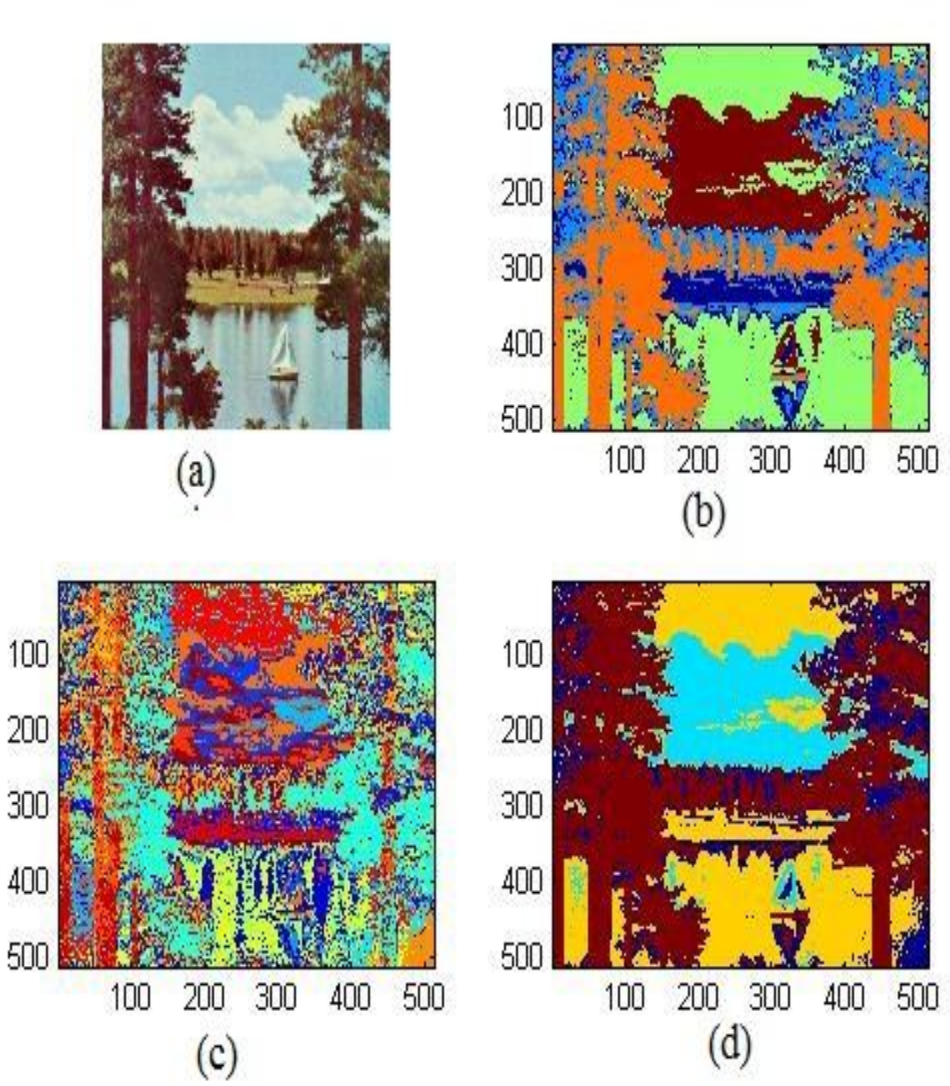


**Figure 4.4** *(a) Original Sailboat image (b) Segmented Image Using K-Means ( K =5) (c) Segmented Image Using K-Means (K = 30) (d) Segmented Image Using Enhanced K-Means( K = 5).*

Here the optimal number of cluster found using validity measure is 5, hence the second image is clustered using K-Means where K = 5, third image is normally segmented using K-Means clustering algorithm where K = 30, and fourth image is segmented using Enhanced K-Means algorithm where K = 5 found using validity measure.

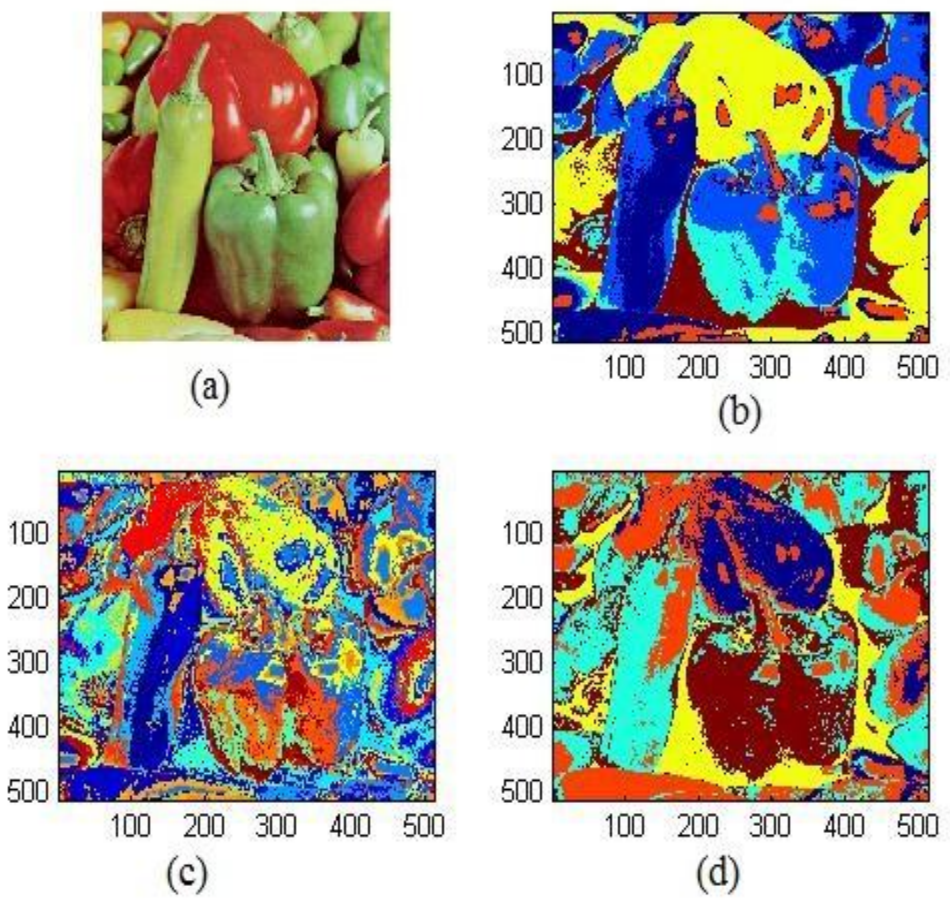


**Figure 4.5** *(a) Original Peppers image (b) Segmented Image Using K-Means ( K =6 ) (c) Segmented Image Using K-Means (K = 30) (d) Segmented Image Using Enhanced K-Means( K = 6).*

Here the optimal number of cluster found using validity measure is 6, hence the second image is clustered using K-Means where K = 6, third image is normally segmented using K-Means clustering algorithm where K = 30, and fourth image is segmented using Enhanced K-Means algorithm where K = 6 found using validity measure.

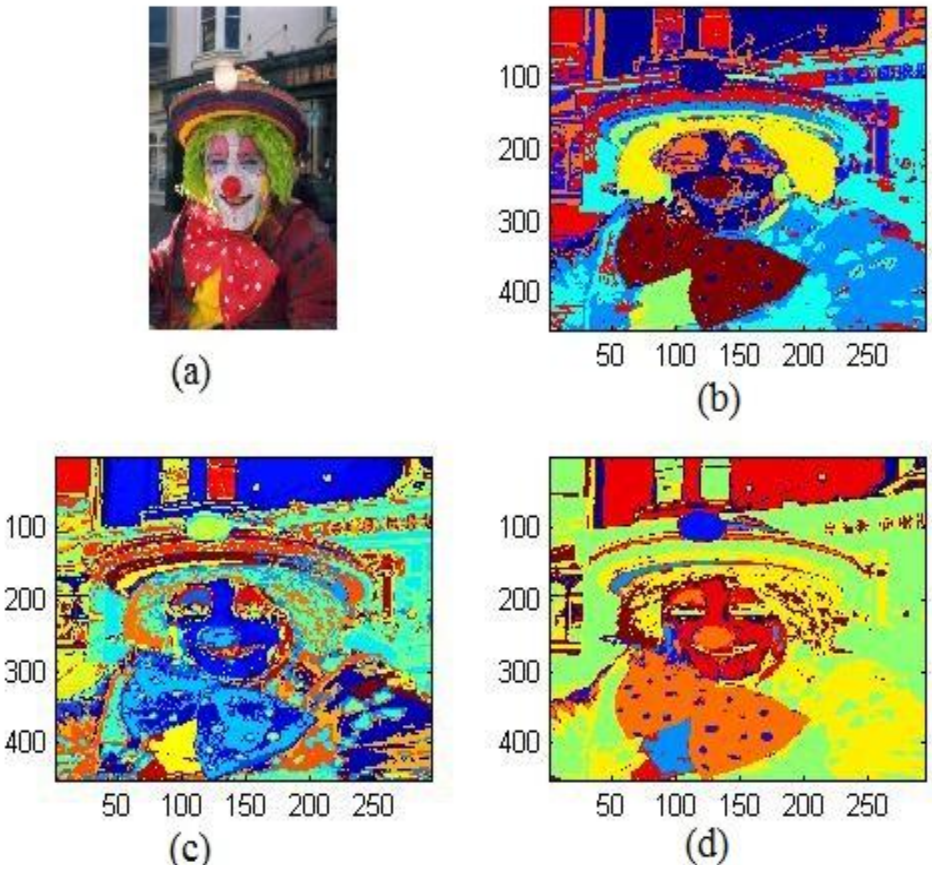


**Figure 4.6** *(a) Original Joker Image (b) Segmented Image Using K-Means ( K =9 ) (c) Segmented Image Using K-Means (K = 30) (d) Segmented Image Using Enhanced K-Means( K = 9).*

Here the optimal number of cluster found using validity measure is 9, hence the second image is clustered using K-Means where K = 9, third image is normally segmented using K-Means clustering algorithm where K = 30, and fourth image is segmented using Enhanced K-Means algorithm where K = 9 found using validity measure

**Table 4.1** Comparison between K-Means Clustering Algorithm and Proposed Algorithm using Validity Measure

| **COVER IMAGE NAME** | **VALIDITY MEASURE OF K-MEANS CLUSTERING ALGORITHM** | **VALIDITY MEASURE OF PROPOSED ALGORITHM** |
|---|---|---|
| Lena | 0.1975 | 0.0293 |
| Barbara | 0.1867 | 0.0478 |
| Sailboat | 0.1916 | 0.2174 |
| Peppers | 0.2183 | 0.0970 |
| Joker | 0.2917 | 0.2536 |

The comparison above shown is done in terms of validity measure, the validity measure of each of the algorithm that is the K-Means Clustering Algorithm and Proposed Algorithm has been calculated. The less validity measure shows that the proposed algorithm is working better then the K-Means clustering Algorithm, validity measure shows that the cluster is optimal and cluster validity.

# Chapter 5: Conclusion and Future Scope

## 5.1 Conclusion

Research has proven that the proposed algorithm is better than the conventional K-Means Clustering Algorithm for color image segmentation, the validity measure of nearly all the images has been better than the conventional K-Means clustering algorithm, the conventional K-means algorithm uses user defined number of cluster which use to cause noisy image, but in the proposed algorithm, it uses the method for determining the number of optimal cluster. It also removes the problem of empty clusters problem from conventional K-Means clustering algorithm where there was issue that if no pixel is assigned to a cluster then that cluster remains empty.

This algorithm has been implemented in Matlab. The algorithm is working properly for all types of color images, Experimental results obtained from this algorithm are satisfactory, giving better clusters and better segmented image then the conventional K-Means clustering algorithm. Formula generated for merging the similar looking clusters gives better segmented images. Hence, this algorithm produces better segmented image then the conventional K-Means clustering algorithm for color image segmentation.

## 5.2 Future Scope

Following are the future direction in which our proposed work can be carried out:

- The proposed algorithm can be compare with other clustering algorithms so that the efficiency of the algorithm can be measured
- To work with the non-globular clusters.
- To reduce the problem of outliers.

# REFERENCES


1. Histogram-Based Optimal Multiple Thresholding Technique , Prof. Hisham Al-Rawi, Dr. Jane J. Stephan ,College of Information Technology / University Of Bahrain –Bahrain ,National Computer Center-Baghdad- IRAQ.
2. N. Senthilkumaran and R.Rajesh, "Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches" International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
3. Tranos Zuva, Oludayo O. Olugbara, Sunday O. Ojo and Seleman M. Ngwira, "Image Segmentation, Available Techniques, Developments and Open Issues" Canadian Journal on Image Processing and Computer Vision Vol. 2, No. 3, March 2011.
4. J. Serra. Image Analysis and Mathematical Morphology. Academic Press, 1982.
5. S.L. Horowitz and T. Pavlidis, Picture Segmentation by a Directed Split and Merge Procedure, Proc. ICPR, 1974, Denmark, pp.424-433.
6. I. D. Guedalia, M. London, and M. Werman, "An on-line agglomerative clustering method for nonstationary data", *Neural Comput.*, vol. 11, pp.521 - 540 1999
7. A. P. Dempster , N. M. Laird and D.B. Rubin "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statiscal Soc.*, vol. 39, no. 1, pp.1 -38 1977
8. J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, Proc. Fifth Berkeley Symp. Math. Statistics and Probability, vol. 1, pp. 281-296, 1967.
9. S. P. Lloyd, "Least squares quantization in PCM," unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meet., Atlantic City, NJ, Sept. 1957. Also, IEEE Trans. Inform. Theory (Special Issue on Quantization), vol. IT-28, ppp. 129-137, Mar. 1982.
10. J.A. Hartigan (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

11. Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". *Journal of the Royal Statistical Society, Series C* **28** (1): 100–108. JSTOR 2346830
12. Deng Y. and Manjunath B.S. (2001). Unsupervised segmentation of color-texture regions in images and video, IEEE Trans. Pattern Analysis Machine Intell, 23(8): 800-810.
13. Khan S. and Ahmad A., (2004). Cluster center initialization algorithm for K-Means clustering, Pattern Recognition Letters, 25(11): 1293-1302.
14. J.T. Tou and R.C. Gonzalez, Pattern Recognition Principles, Massachusetts: Addison – Wesley, 1974.
15. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of data clusters via the gap statistic. Journal of the Royal Statistical Society B, 63:411{423, 2001.
16. N.R. Pal and S.K Pal, A review on image segmentation techniques, Pattern Recognition Vol. 25, PP. 1277-1294, 1993.
17. D.L. Davies and D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Machine Intell. Vol. 1, PP, 224-227, 1979.
18. J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern., vol.3,PP. 32-57,1973.
19. J.C. Bezdek and N.R. Pal, Some new indexes of cluster validity, IEEE Trans. Syst. Man. Cybern., vol. 28, PP. 301-315, 1998.
20. G.W. Milligan, Clustering Validation: Results and implications for applied analyses, In P.Aravie, L.J Hubert and G. De Soete (EDS), clustering and Classification, Singapore: World Scientific, pp. 341-375, 1996.
21. G.W. Milligan and M.C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika, Vol. 50, 159-179, 1985.
22. M.C. Cooper and G.W. Milligan, The effect of measurement error on determining the number of clusters in cluster analysis, In W. Gayl and M. Schader (Eds), Data, Expert Knowledge and Decisions, Berlin: Springer-Verlag, pp. 319-328, 1988

23. N.R. Pal and J.C. Bexdek, On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Systems, Vol. 3, pp. 370-379, 1995.
24. Siddheswar Ray and Rose H. “Turi, Determination of Number of Clusters in K-Means clustering and Application in color image segmentation”, industrial conference practice and Research Techniques.
25. Bhattacharyya, A. (1943). "On a measure of divergence between two statistical populations defined by their probability distributions". *Bulletin of the Calcutta Mathematical Society* **35**: 99–109. MR 0010358
26. H. Jeffreys *Theory of Probability,* 1948 :Oxford Univ. Press