IMAGE CLUSTERING: COMPARISION OF TWO COLOR SEGMENTATION TECHNIQUES

PICHAIYAN SUBRAMANIAM KAVITHA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

IMAGE CLUSTERING: COMPARISION OF TWO COLOR SEGMENTATION TECHNIQUES

PICHAIYAN SUBRAMANIAM KAVITHA

This report is submitted in partial fulfillment of the requirements for the Master of Computer Science (Software Engineering and Intelligence)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITY TEKNIKAL MALAYSIA MELAKA
2010

BORANG PENGESAHAN STATUS THESIS*

JUDUL: **IMAGE CLUSTERING: COMPARISION OF TWO COLOR SEGMENTATION TECHNIQUES**

SESI PENGAJIAN: *2010/2011*

Saya :*PICHAIYAN SUBRAMANIAM KAVITHA*

(HURUF BESAR)

Mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan peojek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

_____ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

_✓_ TIDAK TERHAD

(TANDA TANGAN PENULIS)
Alamat Tetap: G17,Pangsapuri Bukit Beruang Bestari
Taman Bukit Beruang Bestari
75450 Bukit Beruang Melaka
Tarikh: 13.12.2010

(TANDA TANGAN PENYELIA)
PROF. DR. NANNA SURYANA HERMAN
Nama Penyelia
Tarikh: 13.12.2010

CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana ** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

DECLARATION

I hereby declare that this project report entitled "**IMAGE CLUSTERING: COMPARISON OF TWO COLOR SEGMENTATION TECHNIQUES**" is the result of my own research except as cited in the references. The report has not been accepted for any degree and is not concurrently submitted in candidature of any other degrees.

Signature : .....

Name : PICHAIYAN SUBRAMANIAM KAVITHA

Date : 13.12.2010

APPROVAL

I hereby declare that I have read through this project report and in my opinion this project report is sufficient in terms of scope and quality for the award of the degree of Master of Computer Science (Software Engineering and Intelligence).

Signature : .....

Supervisor : PROF. DR. NANNA SURYANA HERMAN

Date : 18 - 12 - 2010

DEDICATION

To
my beloved husband : Dr.C.Senthilpari,
my children: S.Balavikas and S.Haresh vardanan

Thank you for being my inspiration and the full support given.

Pichaiyan Subramaniam Kavitha

ACKNOWLEGEMENT

The Immense pleasure and joy one derives on completion of this project is beyond the void of pinnacle description. It is the drag of the concerned person to validate his respects and to acknowledge the advice, valuable guidance and powerful assistance that are received from all stamps for such a hectic process.

First and foremost, I would like to say my sincere thank to my supervisor YBhg Professor Dr Nanna Suryana Bin Herman of University Teknikal Malaysia (UTeM), for all the advice, support and assistance throughout accomplishing my research. His recommendations and suggestions were valuable for the successful completion of my thesis.

In this context, I am greatly indebted to number of people who have encouraged me, motivated me and graciously contributed their knowledge, time and energy.

Words fail me to express my appreciation to my husband C.Senthilpari whose dedication, love and persistent confidence in me, has taken the load off my shoulder. I owe him for being unselfishly let his intelligence, passions, and ambitions collide with mine.

ABSTRACT

The clustering research is regarding the area of data mining and implementation of the clustering algorithms. The image clustering is major part of data mining where study about how to binds the similar data together in a cluster and show the meaningful data. There are many algorithm for analysing clustering each having its own method to do clustering. This clustering technique increasingly common and has yield many insights into segmentation factors, would effect image functioning and performance. The enormous researches going on extract image with background subtraction. We focus on the outlier detection and background subtraction on image. This project proposed a two color segmentation techniques such as K-means and Fuzzy C-means clustering algorithm that are accurately segment the desired images, which have the same color as the pre-selected pixels with background subtraction. In the software development testing we examine image based clustering, as we can used clustering by distance base, by pixel (red, green, blue) value etc., The problem is solved by region based method which is based on connect component and background detection techniques. The appropriate Java codes are developed for solve this task. The developed patterns are applied in the field of real-time analysis. Finally, the algorithm found, which would solve the image segmentation problem.

ABSTRAK

Penyelidikan pengelompokan adalah mengenai bidang Perlombongan data dan pelaksanaan algoritma pengelompokan. Pengelompokan imej merupakan bahagian utama dari perlombongan data di mana kajian tentang cara mengikat data yang sama bersama dalam sebuah kelompok dan menunjukkan data yang penuh makna. Terdapat banyak algoritma untuk menganalisis pengelompokan di mana setiap kaedah mempunyai kaedah tersendiri untuk membuat pengelompokan. Teknik pengelompokan menjadi semakin biasa dan menghasilkan banyak wawasan dalam faktor segmentasi, dan akan memberi kesan yang menjejaskan fungsi dan prestasi imej. Penyelidikan yang besar berlaku pada ekstrak gambar dengan pengurangan latar belakang.Fokus kami adalah pada pengesanan luaran dan pengurangan latar belakang pada imej. Projek ini mencadangkan dua teknik segmentasi warna seperti algoritma pengelompokan K-means dan Fuzzy C-means yang mengsegmen imej yang diinginkandengan tepat, yang mempunyai warna yang sama seperti piksel pra-pilih dengan latar belakang pengurangan. Dalam ujian pembangunan perisian, kami meneliti pengelompokan berasaskan imej, seperti mana kita boleh menggunakan pengelompokan berasaskan jarak, piksel (merah,hijau,biru) dan lain-lain nilai. Masalah ini diselesaikan dengan kaedah berasaskan wilayah yang didasarkan pada teknik komponen yang bersambung dan pengesanan latar belakang. Kod Java yang sesuai dibangunkan untuk menyelesaikan tugas ini. Pola yang dibangunkan digunakan dalam bidang analisa masa-nyata. Akhirnya, satu algoritma ditemui, yang akan menyelesaikan masalah segmentasi imej.

TABLE OF CONTENTS

CHAPTER	SUBJECT		PAGE
	DECLARATION		**iv**
	APPROVAL		**v**
	DEDICATION		**vi**
	ACNOWLEDGEMENT		**vii**
	ABSTRACT		**viii**
	ABSTRAK		**ix**
	TABLE OF CONTENTS		**x**
	LIST OF APPENDICES		**xv**
	LIST OF FIGURES		**xvi**
	LIST OF ABBREVATIONS		**xviii**
CHAPTER			
1	**INTRODUCTION**		**1**
	1.0	Introduction and Background of Study	1
	1.1	Problem statement	8
	1.2	Research Questions	12
	1.3	Objectives	12
	1.4	Hypothesis and Assumptions	13
	1.5	General Methodology	14

1.6 Scope of Study 15

1.7 Significance of Study 15

1.8 Summary 16

CHAPTER

2 LITERATURE REVIEW 17

2.0 Introduction 17

2.1 Cluster Analysis 18

2.2 Image Segmentation 21

2.3 Uses and Requirement of Clustering 22

2.4 K-Means Clustering 26

2.4.1 Advantages of using K-Means 28

2.4.2 K-Means application 29

2.5 Fuzzy C-means clustering 29

2.5.1 Advantages of using Fuzzy 32

2.5.2 Fuzzy C-Means application 32

2.6 The Designing platform 33

2.6.1 Strength of java Swing 35

2.7 Texture Features 36

2.8 Color Features 37

2.9 Outlier Detection 40

2.10 Background subtraction 41

2.11 Summary 43

CHAPTER

3 METHODOLOGY **44**

3.0 Introduction 44

3.1 K-means Clustering Algorithm 45

3.2 Fuzzy C-means Clustering algorithm 47

3.2.1 Fuzzy Factor 48

3.2.2 Ideal Number of Clusters 'c' 49

3.2.3 Significance of Membership Function in Cluster Analysis 49

3.3 Determination of Java technology for K-means and Fuzzy C-means 50

3.3.1 K-means & Fuzzy C-means testing in Java 50

3.4 Summary 52

CHAPTER

4 MODEL DEVELOPMENT AND RESULT **53**

4.0 Introduction 53

4.1 Implementation of Methods 54

4.1.1 K-Means Clustering Algorithm 55

4.1.2 Fuzzy C-Means Clustering Algorithm 56

4.2 Implementation of algorithm in Java workbench 58

4.2.1 Classes Involved 58

4.2.2 Kmeansflag.java 58

4.2.3 Fuzzyflag.java 59

	4.2.4 Get Image .java	60
4.3	Classes and Image Packages	60
	4.3.1 Class Diagrams of the System	62
4.4	Solution of connected component and background problem	63
	4.4.1 Work of Connect component	63
	4.4.2 Work of Background Problem	63
4.5	Eclipse platform	64
	4.5.1 Eclipse Workbench	67
4.6	Setting Input image and Output parameter	71
	4.6.1 Input image (a),(b),(c)(d)- K_means and Fuzzy C-means	72
4.7	Output image of K-means and Fuzzy C-means	72
	4.7.1 Compare of K-means clustering and Fuzzy C-means clustering image	74
4.8	Testing	75
	4.8.1 White Box Testing	75
	4.8.2 Unit Testing	76
	4.8.3 Black Box Testing and Boundary Value Testing	77
	4.8.4 Equivalence Partitioning Testing	77
4.9	Software coding for K-means	78
4.10	Software coding for Fuzzy C-means	97
4.11	Summary	117

CHAPTER

5 **CONCLUSION AND FUTURE RESEARCH RECOMMENDATION** **118**

5.0 Introduction 118

5.1 Summary of Research 118

5.2 Recommendations for Future Research 120

REFERENCES **121**

APPENDIX A **130**

APPENDIX B **131**

APPENDIX C **132**

APPENDIX D **133**

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Location of K-means and Fuzzy C-means output of .jpg and, .txt file	130
B	Flowchart for image segmentation	131
C	Flowchart for the algorithm	132
D	Low level feature	133

LIST OF FIGURES

DIAGRAM	TITLE	PAGE
1.1	General view of K-means	6
1.2	Example of the Outlier Detection	10
1.3	Example of the different background	11
1.4	General Methodology	14
3.1	Overview of the system	51
4.1	Class Diagrams of the System	62
4.2	Plane and segmented image	64
4.3	The Eclipse Rich Client Platform (RCP) is a subset of the Eclipse Platform	65
4.4	Eclipse Platform architecture	66
4.5	Workbench showing Java perspective	69
4.6	Display of java class file	70
4.7	Coding for our algorithm	70
4.8	Execution of our program	71
4.9	Input image of K-means and Fuzzy C-means	72
4.10	K-means and Fuzzy C-means clustered image	73

4.11	Comparison of background detection between pixels based K-means and Fuzzy C-means Model in RGB color and our proposed approach	74
4.12	White box testing	75
4.13	Black box testing	77
A	Location of output file	130
B	Flow chart for image segmentation	131
C	Flow chart for algorithm	132

LIST OF ABBREVIATIONS

K-means clustering

The K-means method which is most popular and widely used in various fields due to the simple, easily implemented, has the ability to cluster large amounts of data, capable of dealing with outlier data

FCM Fuzzy C-means

Fuzzy *C*-Means is one of the algorithms which is clustering based on optimizing an objective function and being sensitive to initial conditions

RGB Red(R), Green (G) and Blue (B)

CBIR Content Based Image Retrieval

JPG image file extension (.jpg)

CHAPTER 1

INTRODUCTION

1.0 Introduction and Background of Study

The set of data objects for analysis for unlike in classification, where class label of an object is known. This quit common in large database, because assigning class labels to large number of objects can be very costly process. The process of grouping the data into the class or cluster is called as clustering. So that, the objects within the cluster have high similarities in comparison to one another but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values, which is describing the object. Clustering has its root in many areas, including data mining, statistics, biology, and image processing etc.

The image segmentation is preliminary implementation of clustering. According to Pappas T. (1992), the image segmentation may be defined as the process of dividing an image into disjoint homogeneous regions. These homogeneous regions usually contains similar objects of interest or part of them. The extent of homogeneity of the segmented regions can be measured using some image property (e. g. pixel intensity). On the other hand, clustering can be expressed as

optimal partitioning of a given set of n data points into '**c**' subgroups. The data point belongs to the same groups that are similar to each other as possible. Whereas data points from two different groups share the maximum difference. Image segmentation can be treated as a clustering problem where a feature is describing each pixel corresponds to a pattern and each image region (i.e. a segment) corresponds to a cluster. Therefore many clustering algorithms have widely been used to solve the segmentation problem.

Image segmentation is a recent relevant research area in computer vision and hundreds of segmentation algorithms have been proposed in the past 30 years. Many segmentation methods are based on two basic properties of the pixels in relation to their local neighborhoods that are discontinuity and similarity. Method based on pixel discontinuity are called boundary-based methods and methods based on pixel similarity are called region-based methods.

According to Pavlidis, T., Liow, Y (1990), it is well experienced that such segmentation techniques, which based on boundary or region information alone. Hence, in the last few years, there has been a tendency towards algorithms which take advantage of the complementary nature of such information. Haralick, R., Shapiro, L (1985) proposed that reviewing the different works on region-based segmentation (surveys on image segmentation). It is interesting to note the evolution of region-based segmentation methods, which were initially focused on grey-level images, gradually incorporated color, and more recently texture. In view of Drimbarean, A., Whelan, P.(2001) the color and texture are fundamental features in defining visual perception and experiments have demonstrated that the inclusion of color can be increase the texture segmentation/classification results without

significantly complicating the feature extraction algorithms. Nevertheless, most of the literature deals with segmentation based on either color or texture, and there is a limited number of systems which consider both properties together. Nowadays the following areas are widely used in Image base Clustering.

(i) Security (nonsense object detection) .

(ii) Traffic conjunction detection.

(iii) Distracting meaningful object from image.

(iv) Spatial information by set-lights

As a branch of statistics, cluster analysis has been extensively studied for many years, focusing mainly on distance-based cluster analysis. Cluster analysis tools on k- means, k- methods, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS (Statistical Package for the Social Sciences) and SAS (Special Air Service). In machine learning, the clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labelled training examples. According to Jain A.K. and Dubes R.C. (1988), clustering is a form of learning by observation, rather then learning by examples. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases. Khan S. and Ahmad A., (2004) described that the image active themes of research are focuses on the scalability of clustering methods, which illustrate the effectiveness of methods for clustering complex shapes, types of data, high-dimensional clustering techniques, methods for clustering mixed

numerical and categorical data in large database. Clustering is a challenging field of research in which its potential applications pose their own special requirements.

According to Drimbarean A. and Whelan P.F. (2001), Zöller H.T. and Buhmann J.M. (2002), Many clustering algorithms work well on small data sets containing fewer than several hundred data objects. However, a large database may contain millions of objects. Clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.

Many clustering algorithms have been exist in the literature. The clustering algorithm is difficult to provide a crisp categorization of clustering methods because these categories may overlap, so that a method may be features from several categories. Clustering algorithms may be classified as listed in below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

This project is used only exclusive algorithms and overlapping clustering algorithms among above said algorithms. In the Exclusive clustering, data are grouped in an exclusive way. So that, the certain datum belongs to a definite cluster then it could not be included in another cluster. K-Means clustering as an example.

As per reference of J.L. Marroquin, F. Girosi, (1993), K. Atsushi, N. Masayuki, (1998), A. Murli, L. D'Amore, V.D. Simone, (1999) and S. Ray, R.H. Turi (1999),

the K- meaning algorithms using the exclusive clustering technique. This type of method used in real time application such as traffic light signal.

In the Overlapping Clustering type, use fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value. Fuzzy C-Means clustering as an example.

This project implemented two clustering algorithm on the image that are listed in below.

- K-means clustering
- Fuzzy C-means clustering

J. MacQueen (1967) and then by J. A. Hartigan and M. A. Wong around 1975 found that K-Means clustering method which is most popular and widely used in various fields due to the simple, easily implemented, has the ability to cluster large amounts of data, capable of dealing with outlier data.The linear time complexity O (nKT) is suitable structure formula where n is the number of documents, K is the number of clusters, and T is the number of iteration. K-Means algorithm serves to classify an object that has a similarity (grouping process called clustering) on the basis of K clusters, where K is the positive integers. K-means is a partitioning method clustering that separates data into different groups. K-Means able to minimize the average distance of each data to its cluster With partitioning is iterative.

The characteristics of K-Means algorithm is very sensitive in determining the initial cluster center for K-Means clustering center generating initial random. At the

beginning of the generation of random focal point is approaching the final solution the cluster centers, The K-Means has a high possibility to find the appropriate cluster centers. Conversely, if the initial center point is far from the final solution the cluster centers, then the likely cause of inappropriate clustering.

Initial step process of K-Means algorithm that determines the center of each cluster is almost a kind of later called the centroid. Centroid bias determined at random,which does the calculation the distance between the centroid of each cluster of existing and then classified each cluster based on the nearest distance from each object to the centroid. Then count back centroid, do this repeatedly until the centroid position does not move anymore. The procedure step of the K-means illustrated in Fig 1.1 which is describing the basic steps of algorithms.

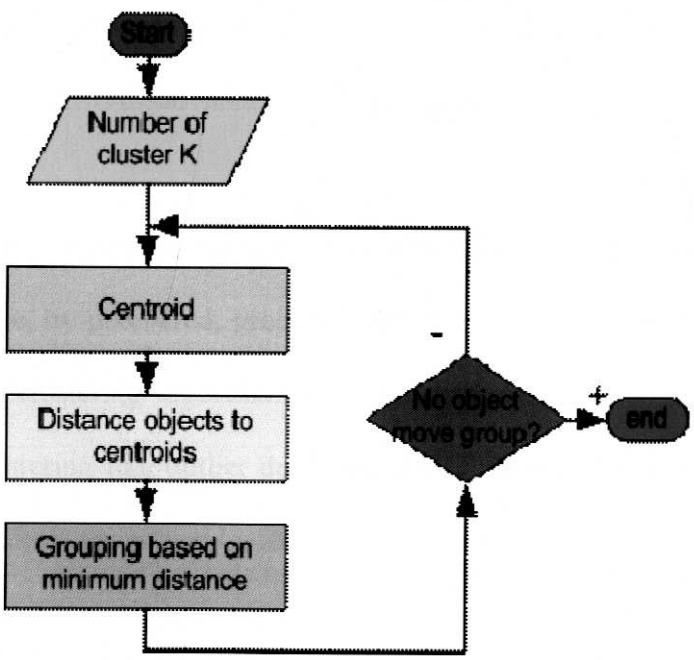

Fig 1.1 General view of K-means

The fuzzy C-Means algorithm is very similar to the K-means algorithm. Dunn in 1973 and Bezdek in 1981 found that the fuzzy C-Means developed by the