IDENTIFYING THE COUNT OF CLUSTERS IN UNLACED DATASETS

PREETHI.S

(12PCS012)

A Project Report submitted to

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641043

In Partial Fulfillment of the Requirements for the Master's Degree in Computer Science

March, 2014

TABLE OF CONTENTS

CHAPTER No	TITLE	PAGE No

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVE OF THE PROJECT

The Project entitled "**Identifying the count of clusters in unlaced datasets"** has been designed and developed by using **Microsoft Visual Studio Dot Net 2008** Orcas Version and **Microsoft Access** as Back End Tool. The main objective of the project is to determine the **Number of clusters** in unlabeled data sets during cluster analysis.

Cluster Analysis:

A general question in the data mining community is how to organize observed data into meaningful structures or taxonomies. Considering Clustering analysis, it aims at grouping objects of a similar kind into their respective categories.

Pre-Clustering Tendency Assessment

The selection of the number of clusters is an important and challenging issue in cluster analysis. A number of attempts have been made to estimate c in a given data set. Most methods are post clustering measures of cluster validity, i.e., they attempt to choose the best partition from a set of alternative partitions. In contrast, tendency assessment attempts to estimate c before clustering occurs. Our focus is on pre clustering tendency assessment, but for completeness, we briefly summarize some existing approaches to the post clustering cluster validity problem, before describing visual methods for cluster tendency assessment.

To overcome post clustering tendency assessment, Dark Block Extraction (DBE) is introduced for automatically estimating the number of clusters in unlabeled data sets, which is based on the existing algorithm for Visual Assessment of Cluster Tendency (VAT) of a data set, using several common image and signal processing techniques.

1.2 LITERATURE SURVEY

Table 1.1 Literature Survey

S.NO	TITLE	CONCEPT USED	IMPLEMENTATION OF THE CONCEPT
1.	Visual Assessment of clustering Tendency for Rectangular Dissimilarity Matrices	❖ Algorithm coVAT: Visual Assessment of Co-Cluster Tendency in Rectangular Dissimilarity Data ❖ Scalable coVAT for VL Rectangular Dissimilarity Data	After Segmentation and classification completes, the revised VAT is going to be used to identify large collections of data with proxy controller support.
2.	VAT: A Tool for Visual Assessment of (Cluster) Tendency	**"VAT Ordering and displaying algorithm"** -new approach for visually assessing cluster tendency using ordered dissimilarity images.	Considering minimal spanning tree of a weighted graph, Dundas chart is proposed to overcome computationally expensive for large data sets.
3.	bigVAT:Visual assessment of cluster tendency for large data sets	**big VAT ALGORITHM -** • bigVAT combines the quasi-ordering technique used by reVAT with an image display of the set of profile graphs displaying the clustering tendency information with a VAT-	Proxy Preview Imaging Area is initialized to chose any random images which allowed for segmentation and Classification.

		like image	

S.NO	TITLE	CONCEPT USED	IMPLEMENTATION OF THE CONCEPT
4.	Scalable visual assessment of cluster tendency for large data sets	**"New Scalable VAT Algorithm is implemented"** The visual assessment of cluster tendency (VAT) tool has been successful in determining potential cluster structure of various data sets, but it can be computationally expensive for large data sets.	Back Propagation of Artificial Neural Network is implemented to overcome this problem which automatically determines the potential cluster structure of various data sets.
5.	Automatically Determining the Number of Clusters in Unlabeled Data Sets	❖ VAT ❖ DAB ❖ DBE is more reliable than CCE. ❖ It overcomes the confusing problem in CCE of where to cut	Determines the number of clusters automatically using post-cluster clustering tendency

		the Histogram	

CHAPTER 2

SYSTEM ANALYSIS

2.1 EXISTING SYSTEM

The basic aim of the system analysis is to get the clear understanding of the needs, what exactly is the need from the software and what are the constraints on the solutions. Analysis leads to the actual specification.

A general question in the data mining community is how to organize observed data into meaningful structures (or taxonomies). As an exploratory data analysis tool, cluster analysis aims at grouping objects of a similar kind into their respective categories. There have been a large number of clustering algorithms reported in the literature.

In general, clustering of unlabeled data poses three major problems:

- Assessing cluster tendency, i.e., how many clusters to seek? or what is the value of number of clusters?
- Partitioning the data into number of clusters meaningful groups, and
- Validating the number of clusters discovered.

The first problem is determining the number of clusters ‘c’ prior to clustering. Many clustering algorithms require the number of clusters ‘c’ as an input parameter, so the quality of the resulting clusters is largely dependent on the estimation of ‘c’. For some applications, users

can determine the number of clusters with domain knowledge. However, in many situations, the value of ‘c ’is unknown and needs to be estimated from the data themselves.

A new method called Dark Block Extraction (DBE) which uses a pre-clustering method, i.e., it does not require the data to be clustered, nor does it find the clusters in the data.

Steps followed for Dark Block Extraction Algorithm :

- Generating a VAT image of an input dissimilarity matrix,
- Performing image segmentation on the VAT image to obtain a binary image, followed by directional morphological filtering
- Apply a distance transform to the filtered binary image and projecting the pixel values onto the main diagonal axis of the image to form a projection signal, and
- Smooth the projection signal, computing its first-order derivative, and then detecting major peaks and valleys in the resulting signal to decide the number of clusters.

2.1.1 DRAWBACKS OF THE EXISTING SYSTEM

Table 2.1 Drawbacks of the Existing System

reVAT	It becomes hard to mentally integrate the information in a set of ‘c’ profile graphs when viewed sequentially. Clusters in the data are not compact and well separated, the ‘c’ profile graphs is pretty confusing.
bigVAT	❖ Solves the Large data problem suffered by VAT ❖ Solves the interpretation problem solved by reVAT.
Cluster Counter Extraction (CCE) Algorithm	❖ CCE is not applicable for Iris and Face Extraction process in real time examples. ❖ Complexity in where to cut the histogram

Dark Block Extraction Algorithm(DBE) (CURRENT BASE PAPER)	❖ DBE is more reliable than CCE. ❖ It overcomes the confusing problem in CCE of where to cut the Histogram Drawbacks : Slightly overestimated or under estimated value of 'c', it provides the initial estimation of the cluster number.

2.2 PROPOSED SYSTEM

A new method called Trusted Pre Cluster Count (TPCC) is introduced for automatically estimating the number of clusters in unlabeled data sets, which is based on an existing algorithm for Visual Assessment of Cluster Tendency (VAT) of a data set, using several common image and signal processing techniques such as reVAT, bigVAT, Dark Block Extraction Algorithm. Its basic steps include:

Steps followed for Trusted Pre Cluster Count Algorithm:

- ❖ Generating a VAT image.
- ❖ Performing image segmentation on the VAT image.
- ❖ Divide the image into matrix pixels as row 'r' X column 'c'.
- ❖ Create a proxy controller to store all matrix pixels values.
- ❖ Calculate each pixel value and compare each pixel relates with neighbor pixel.
- ❖ Group the similar result's pixels.
- ❖ Calculate the number of groups separated which is equal to number of clusters.
- ❖ The resulted cluster count will be the perfect pre cluster count value where it produces the VAT image into a super quality image.

2.2.1 ADVANTAGES OF PROPOSED SYSTEM

- TPCC is an advanced method of detecting the number of clusters in a pre defined manner in order to give more accuracy to the segmented image.
- TPCC is a pre-clustering method, i.e., it does not require the data to be clustered, nor does it find clusters in the data.
- By using TPCC method, the segmented image is well clearly classified into pixel transformations by maintaining the entire pixel data in a proxy structure ,i.e., in an array format. So the calculations process is very little to find the density of the image in order to produce accuracy to the image.

2.3 FEASIBILITY STUDY

Feasibility study is a test of the system proposal according to the workability, impact of the organization, ability to meet user's needs and effective use of resources. The feasibility study must satisfy the following factors:

- User demonstrable needs
- Problem worth solving
- Method of solving problem.

2.3.1 ECONOMICAL FEASIBILITY

Economic feasibility is the most frequently used method of evaluating the effectiveness of a candidate system. The procedure is to determine the savings and benefits from the candidate system and compare the costs. If the benefits outweigh the costs then it is decided to go ahead with the project. Otherwise, further justification or alterations in the proposed system should be made to have a chance of being approved. It is an on-going effort that improves the accuracy at each phase of the system life cycle.

In the economic feasibility study, the following points are found:

- The automated system will be costly.
- Maintenance also involves some investment in terms of money.

Once the computerized system is installed, it can cater to the needs of the customer and the business without manual work, which is more cost-effective for the management. Therefore, the automated system is economically feasible.

2.3.2 OPERATIONAL FEASIBILITY

People are inherently resistant to change the computers have been known to facilitate change. It is common knowledge that computer installations have a lot to do with the turnover transfer retaining and changes to employee job status. Therefore it is understandable that the introduction of the candidate system requires special effort to educate and train the staff on a new way of conducting business. But since ultimately the introduction of a new system will only reduce the staff's workload, staff's may have no objection to install a computerized system and of course will be eager to extend their co-operation.

The main solutions are :

- Measuring the worth of the system being developed compared to the existing system.
- The system avoids all the possible dissatisfaction.
- The reduction of cost affects the performance of the system.
- The system is uniformly accepted by all type of users.

2.3.3 TECHNICAL FEASIBILITY

Technical feasibility centers on the existing system It involves financial considerations to accommodate technical enhancements. If the budget is a serious constraint, then the project will be judged not feasible. So the user having realized the advantages, benefits and economic feasibility of the new system is ready to afford the extra expense that may arise for the satisfaction of all the hardware and software requirements.

CHAPTER 3

SYSTEM SPECIFICATION

3.1 HARDWARE REQUIREMENTS

Processors will continue to get faster, smaller and cheaper, where as memory will continue to get faster, larger and cheaper.

Processor **:** Pentium IV

Processor Speed **:** 1.7 GHz

Memory (RAM) **:** 256 MB

Hard Disk **:** 10 GB

Monitor **:** Dell Color Monitor

Keyboard **:** 104 keys Intex Keyboard

Mouse **:** Intex Optical Mouse

3.2 SOFTWARE REQUIREMENTS

When an application project is considered the three basic software requirements are the platform in which the project is developed, the front-end tool that provides the interaction with the users and the back-end tool that stores the data.

Operating System **:** Windows XP

Software Tools **:** Microsoft Visual Basic .Net 2008

Application Server **:** Internet Information Server

Database : Microsoft Office Access 2007

CHAPTER 4

SOFTWARE DESCRIPTION

4.1 FRONT END

VB.NET

Visual Basic .NET is a major component of Microsoft Visual Studio .NET suite..Net is a Framework in which Windows applications may be developed and run. .NET must go back in time and follow the development of Windows and the advent of Windows programming.

The .NET version of Visual Basic is a new improved version with more features and additions. After these new additions, VB qualifies to become a full object-oriented language such as C++.

VB.NET is the following version of VB 6.0. Microsoft .NET is a new programming and operating framework introduced by Microsoft. All .NET supported languages access a common .NET library to develop applications and share common tools to execute applications. Programming with Visual Basic using .NET is called VB.NET. VB.NET, the following version of VB 6.0 is an improved, stable, and full Object Oriented language.

VB 6.0 wasn't a true object-oriented language because there was no support for inheritance, overloading, and interfaces. VB.NET supports inheritance, overloading, and interfaces.

Multithreading and Exception handling was two major weeks' areas of VB 6.0. In VB.NET, the user can develop multithreaded applications as the user can do in C++ and C# and it also supports structured exception handling.

FEATURES

- Object Oriented Programming language.
- Support of inheritance, overloading, interfaces, shared members and constructors.
- Supports all CLS features such as accessing and working with .NET classes, interaction with other .NET languages, Meta data support, common data types, and delegates.
- Multithreading support.
- Structured exception handling.

ASP.NET

ASP.NET is the latest version of Microsoft's Active Server Pages technology (ASP). ASP+ is the other name for ASP.NET. ASP+ is just an early name used by Microsoft when they developed ASP.Net.

ASP.NET provides a unified Web development model that includes the services necessary for developers to build enterprise-class Web applications. ASP.NET has been designed to work seamlessly with WYSIWYG (What You See Is What You Get) HTML editors and other programming tools, including Microsoft VisualStudio.NET. Not only does it make Web development easier, but it also provides all the benefits that these tools have to offer, including a GUI that developers can user to drop server controls onto a Web page and fully integrated debugging support.

FEATURES

- Programmable controls
- Event-driven programming
- XML-based components

ADO.NET

ADO.NET is as set of classes that expose data access services to the .NET programmer. ADO.NET provides a rich set of components for creating distributed, data sharing applications. It is an integral part of the .NET Framework, providing access to relational data, XML, and application data. ADO.NET supports a variety of development needs, including the creation of front-end database clients and middle-tier business objects used by applications, tools, languages, or Internet browsers.

ADO.NET Architecture

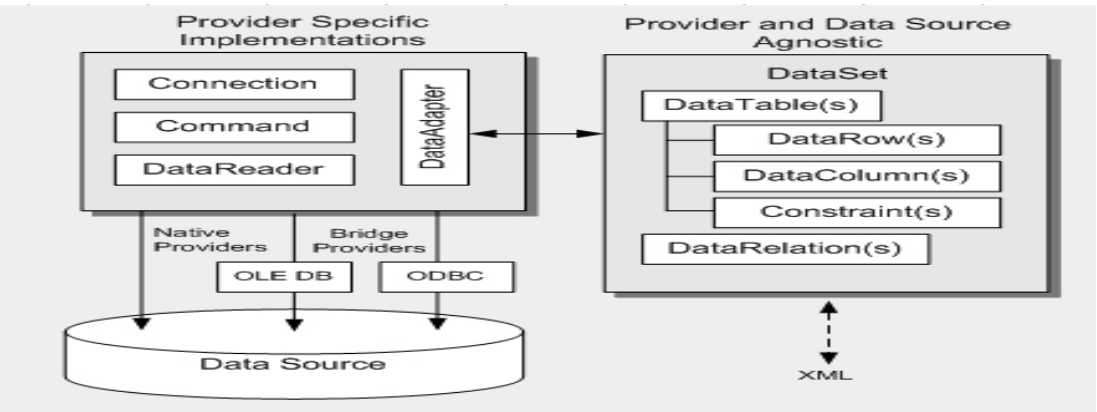

Fig 4.1 ADO.NET Architecture

ADO.NET provides consistent access to data sources such as Microsoft SQL Server, as well as data sources exposed through OLE DB and XML. Data-sharing consumer applications can use ADO.NET to connect to these data sources and retrieve, manipulate, and update data.

ADO.NET cleanly factors data access from data manipulation into discrete components that can be used separately or in tandem. ADO.NET includes .NET Framework data provides for connecting to a database, executing commands, and retrieving results. Those results are either processed directly, or placed in an ADO.NET **Dataset** object in order to be exposed to the user in an ad-hoc manner, combined with data from multiple sources, or remoted between tiers. The ADO.NET **Dataset** object can also be used independently of a .NET Framework data provider to manage data local to the application or sourced from XML.

4.3 BACK-END

SQL SERVER

Microsoft SQL Server extends the performance, reliability, quality, and ease-of-use of Microsoft SQL Server version. Microsoft SQL Server includes several new features that make it an excellent database platform for large-scale online transactional processing (OLTP), data warehousing, and e-commerce applications.

The OLAP Services feature available in SQL Server version is now called SQL Server Analysis Services. The term OLAP Services has been replaced with the term Analysis Services. Analysis Services also includes a new data mining component.

The Repository component available in SQL Server version is now called Microsoft SQL Server Meta Data Services. References to the component now use the term Meta Data Services. The term repository is used only in reference to the repository engine within Meta Data Services.

SQL Server Replication Services are used by SQL Server to replicate and synchronize database objects, either in entirety or a subset of the objects present, across replication agents, which might be other database servers across the network, or database caches on the client side. Replication follows a publisher/subscriber model, i.e., the changes are sent out by one database server ("publisher") and are received by others ("subscribers").

CHAPTER 5

PROJECT DESCRIPTION

5.1 PROBLEM DEFINITION

In general, clustering of unlabeled data poses three major problems (where C is the number of clusters):

1 How many cluster to seek? Or what is the value of ‘c’?

2 Partitioning the data into ‘c’ meaningful groups.

3 Validating the C Clusters discovered.

Here we are addressing the first problem i.e., determining the number of clusters ‘c’ prior to clustering.

5.2 OVERVIEW OF THE PROJECT

The Project entitled “I**dentifying the Count of Clusters in Unlaced Datasets”** has been designed and developed by using **Microsoft Visual Studio Dot Net 2008** Orcas Version and **SQL Server** as Back End Tool.

5.3 MODULES DESCRIPTION

5.3.1 Authentication and Authorization Form

The authentication is the major part for any kind of software. Generally authentication is used for security purpose to protect from intruders. Here, two walls majorly acting for security named as authentication wall and authorization wall. Authentication wall filters the users by providing username and password. After successful entering into the authentication wall, the authorization wall will check the entered user having administrator rights or normal rights. Based upon the rights, the access permission will be allowed.

5.3.2 Image Master

Image Master acts as the gateway for preview imaging. The main advantage of using this module is it allows the scanned machine print document and scanned hand written document to stored in the centralized database with the sufficient details related to that scanned tiff images like image details, image taken time and date and so on.

5.3.3 Preview Imaging

The preview imaging will be very helpful to the user in order to preview a bunch of scanned images having a mingled collection of machine print scanned images and hand written scanned images stored in the centralized database already. It is mainly used to select the required image soon to allow it for segmentation and classification process.

5.3.4 Segmentation

Once the scanned images allowed for segmentation, two panels will be available for easy segmentation of the scanned images which is present in the form of left hand side and the right hand side . In the left hand side of the panel, the selected image from preview imaging will be displayed. The functionality embedded in the left hand side panel is to allow the user by segmenting the image in the form of placing mouse cursor point as scoring point. Once the user left click the mouse, the pointed area is calculated as x1 and y1 points and when the user release the mouse click, the released area is calculated as x2 and y2 points. By using the calculation ratio, it initially analyzes the bounded region of the selected portion which is present inside the (x1,y1) and (x2,y2) segments. Now the segmented image will be displayed in the right hand side panel and in the bottom of the right side panel, the visual zoom scaling will be placed for more clear visualization of the segmented image to the user. In the segmented area, there is a button placed called comparative study, in which if the user clicks that button, the scanned segmented image in the right hand side panel will be allowed to compare with Existing reVAT and bigVAT methodology and the proposed Dark Block Extraction (DBE).

5.3.5 Proxy Controller

Proxy controller is one of the main array tools which are allowed to store any particular data for time being transmission.

5.3.6 Implementing reVAT Methodology

The (unordered) reVAT image I(R) shows the same basic relationships between pairs of data as the ordered VAT image I(R). Each image suggests that these data possess five potential clusters (based on the dark blocks along the diagonals). Although potential clusters are not in the same order along the diagonal, still able to get a sense of the relative sizes of the clusters and the fuzzy relationships between the clusters.

reVAT Algorithm steps include :

- Choose segmented image from right hand side panel from the segmentation form.
- Analyze the total number of pixel values and based on that pixel value, create a proxy controller which stores each pixel value of a reVAT processing image
- Assign black to pixels where intensity level is low and assign white to pixels where intensity level is high.
- Create a for loop and start the reVAT calculation from 0^{th} position of initial pixel proxy value to n^{th} position of proxy value.
- Considering reVAT algorithm process, the getpixel( ) function is used to trace pixel value of 0^{th} pixel proxy position and the pixel value will be allowed for dividend into ‘R’, ‘G’, ‘B’ values and the values will be formulated and the formulated value will be displayed in the corresponding list view position and the calculated pixel is replaced in the same position by using the function called setpixel( ).
- Once the entire process completed, the output will be displayed in the Screen.

5.3.7 Implementing bigVAT Methodology

bigVAT combines the quasi-ordering technique used by reVAT with an image display of the set of profile graphs displaying the clustering tendency information with a VAT-like image. Several numerical examples are given to illustrate and support the new technique.

bigVAT Algorithm steps include :

- Choose segmented image from right hand side panel from the segmentation form.
- Analyze the total number of pixel values and based on that pixel value, create a proxy controller which stores each pixel value of a bigVAT processing image.
- Total Proxy Value Count acts as the threshold value =á‘ from Selected Segmented image ‘M’.
- Considering bigVAT algorithm process, the getpixel( ) function is used to trace pixel value of 0^{th} pixel proxy position and the pixel value will be allowed for dividend into ‘R’, ‘G’, ‘B’ values. Perform a distance transform to obtain a gray scale image and scale the pixel values.
- Project the pixel values of the image on to the main diagonal axis to form a projection signal. Smooth the signal to obtain the filtered signal by an average filter and the values will be formulated and the formulated value will be displayed in the corresponding list view position and the calculated pixel is replaced in the same position by using the function called setpixel( ).
- Once the entire process completed, the output will be displayed in the screen.

5.3.8 Implementing Dark Block Extraction Methodology

One of the major problems in cluster analysis is the determination of the number of clusters in unlabeled data, which is a basic input for most clustering algorithms. In this module, we investigate a new method called **Dark Block Extraction (DBE)** for automatically estimating the number of clusters in unlabeled data sets, which is based on an existing algorithm for Visual Assessment of Cluster Tendency (VAT) of a data set, using several common image and signal processing techniques.

Dark Block Extraction Algorithm steps include :

- generating a VAT image of an input dissimilarity matrix,
- performing image segmentation on the VAT image to obtain a binary image, followed by directional morphological filtering,
- applying a distance transform to the filtered binary image and projecting the pixel values onto the main diagonal axis of the image to form a projection signal
- smoothing the projection signal, computing its first-order derivative, and then detecting major peaks and valleys in the resulting signal to decide the number of clusters.

5.3.9 Implementing Proposed Trusted Pre Cluster Count

- TPCC is an advanced method of detecting the number of clusters in a pre defined manner in order to give more accuracy to the segmented image.
- TPCC is a pre-clustering method, i.e., it does not require the data to be clustered, nor does it find clusters in the data.
- By using TPCC method, the segmented image is well clearly classified into pixel transformations by maintaining the entire pixel data in a proxy structure ,i.e., in an array format. So the calculations process is very little to find the density of the image in order to produce accuracy to the image.

TPCC Algorithm steps include:

- Generating a VAT image.
- Performing image segmentation on the VAT image.
- Divide the image into matrix pixels as row ‘r’ X column ‘c’.
- Create a proxy controller to store all matrix pixels values.
- Calculate each pixel value and compare each pixel relate with neighbor pixel.
- Group the similar result’s pixels .

- Calculate the number of groups separated which is equal to number of clusters.
- The resulted cluster count will be the perfect pre cluster count value where it produces the VAT image into a super quality image.

5.3.10 Accuracy Comparison of reVAT-bigVAT-DBE

reVAT, bigVAT and DBE algorithms are implemented separately and are merged together in order to make accurate comparison among all these three algorithms.

5.3.11 Accuracy Comparison of bigVAT and DBE

The dissimilarity matrices of bigVAT and Dark Block Extraction algorithms are integrated and their evaluated pixel values are compared with the neighborhood pixel values. Based on their similarities cluster count is generated and their accuracy is compared.

5.3.12 Accuracy Comparison of DBE and TPCC

The dissimilarity matrices of Dark Block Extraction and Trusted Pre Cluster Count algorithms are integrated and their evaluated pixel values are compared with the neighborhood pixel values. Based on their similarities cluster count is generated and their accuracy is compared.

5.3 DATA FLOW DIAGRAM

Context Level Diagram

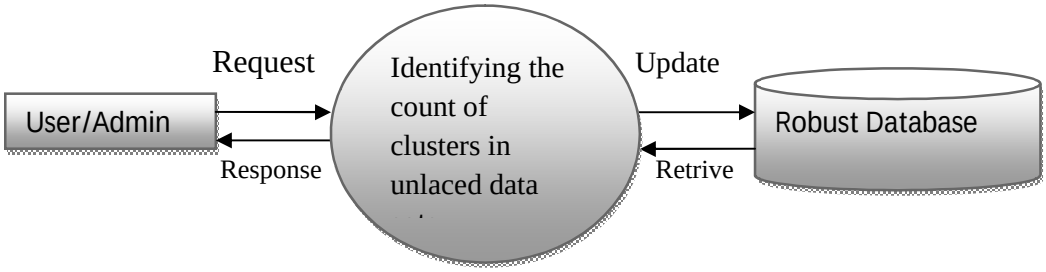

Fig 5.1 Context Level Diagram

LEVEL 0 DFD

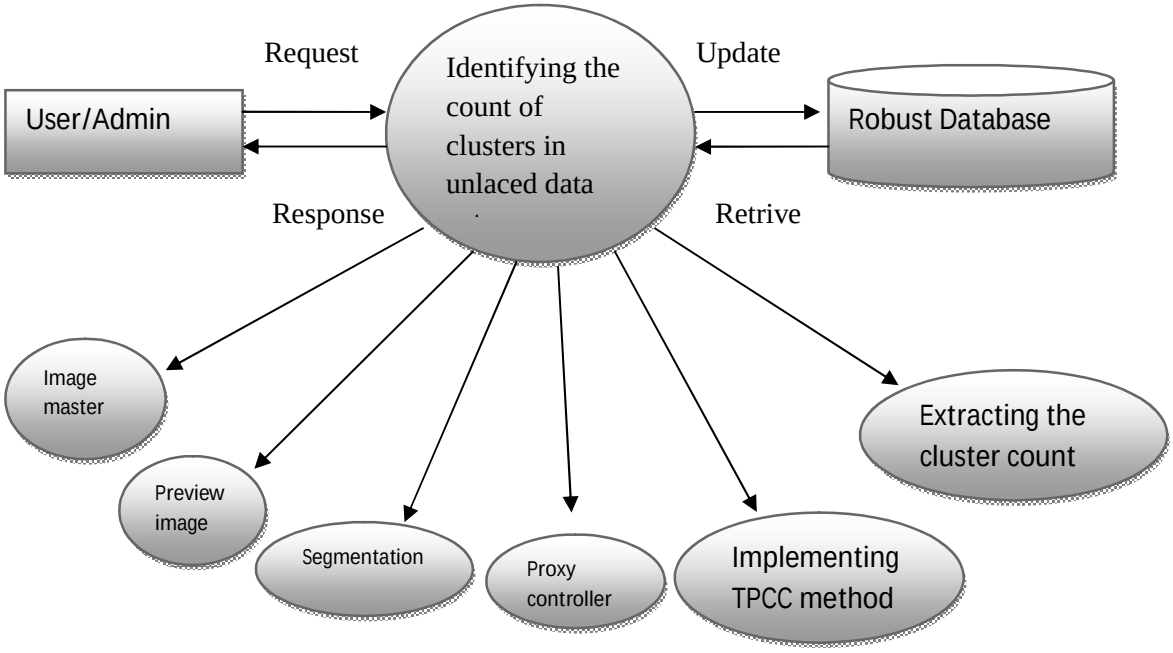

Fig 5.2 Level 0 DFD

LEVEL 1 DFD

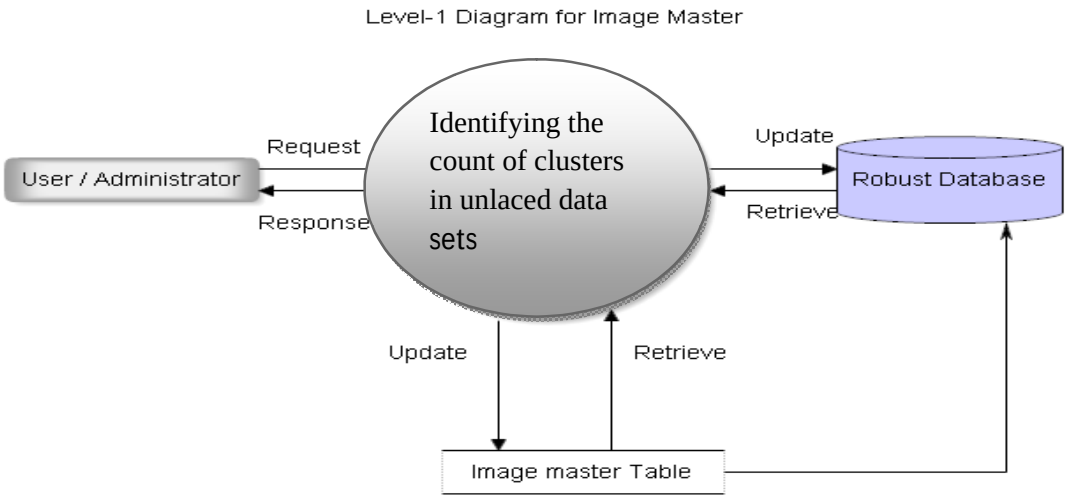

Fig 5.3 Level 1 DFD for Image Master

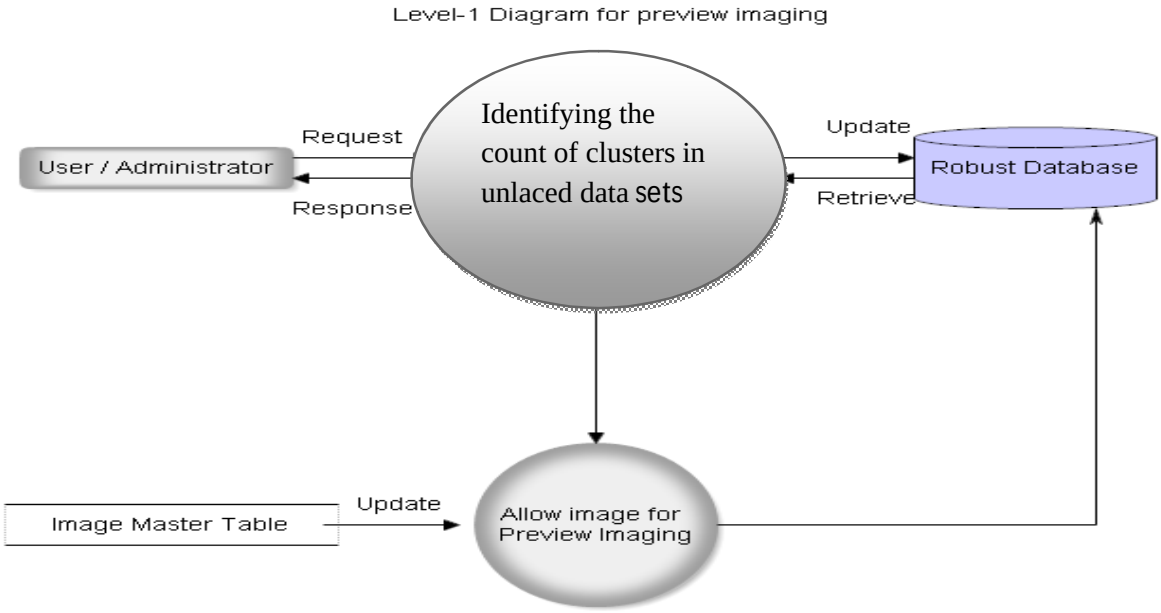

Fig 5.4 Level 1 DFD of Preview Imaging

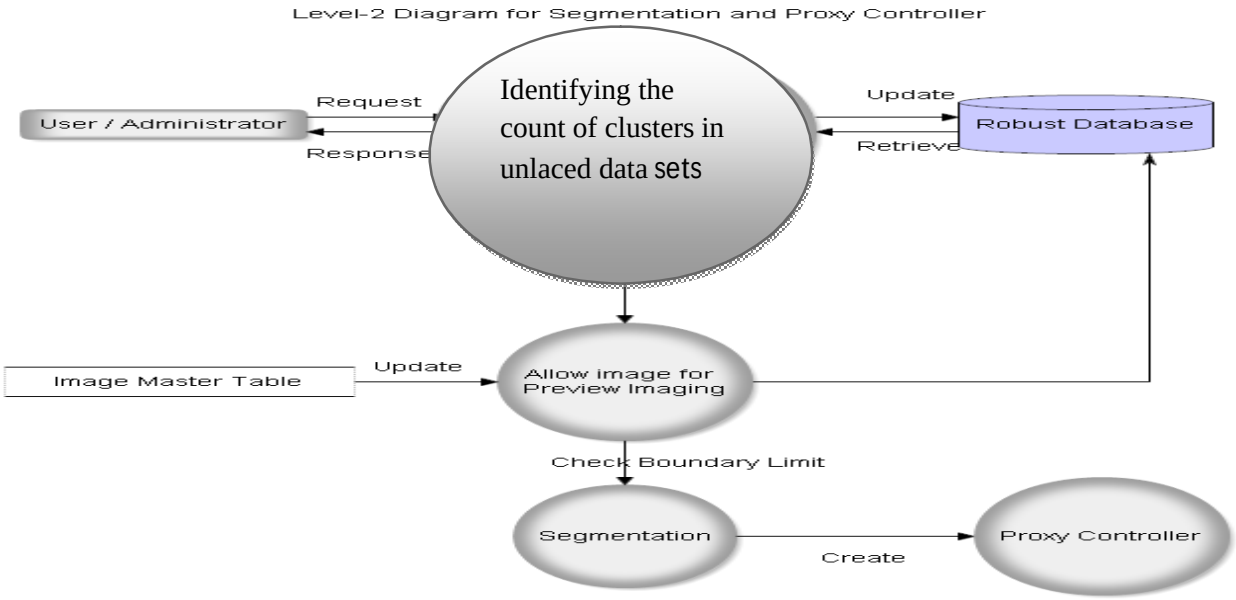

Fig 5.5 Level 2 DFD of Segmentation and Proxy Controller

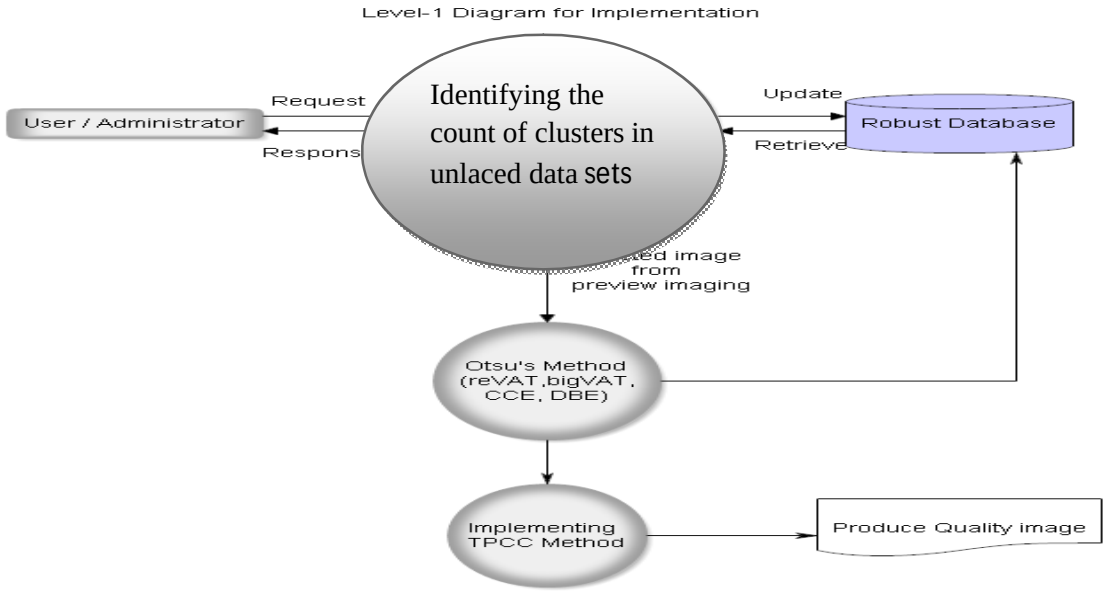

Fig: 5.6 Level 1 DFD of Implementation

5.5 DATABASE DESIGNS

5.5.1 Data Constraints

All business in the world runs on business data being gathered stored and analyzed. Business managers determine a set of rules that must be applied to the data being stored to ensure its integrity.

Types of Data Constraints

There are two types of data constraints that can be applied to data being inserted into a database table .One type of constraint is called an I/O constraint. The other type of constraint is called a business rule constraint.

- **I/O Constraints**

The input /output data constraint is further divided into two distinctly different constraints.

The Primary Key Constraint

Here the data constraint attached to a column ensures:

- That the data entered in the table column is unique across the entire column.
- That none of the cells belonging to the table column are left empty.

The Foreign Key Constraint

Foreign constraint establishes a relationship between records across a master and a detail table. The relationship ensures.

- Records cannot be inserted in a detail table if corresponding records in the master table does not exist.
- Records of the master table cannot be deleted if corresponding records in the detail table exist.

Business Rule Constraints

The Database allows the application of business rules to table columns. Business managers determine business rules.

The Database allows programmers to define constraints at:

- Column Level
- Table Level

Column Level Constraints

If data constraints are defined along with the column definition where creating or altering a table structure, they are column level constraints.

Table Level Constraints

If data constraints are defined after defining all the table columns when creating or altering a table structure, it is a table level constraint.

TABLE: 5.1 IMG_MAS **Primary Key :** Code

FIELD NAME	DATA TYPE	SIZE	DESCRIPTION
Code	Numeric	10	Image Code
Img_name	Varchar	50	Image Name
Descrip	Varchar	250	Image Description
Img_path	Varchar	250	Image Path

5.6.2 NORMALIZATION

In relational database design, the process of organizing data to minimize redundancy is called normalization. Normalization usually involves dividing a, database into two or more tables and defining relationships between tables.

The objective is to isolate data so that additions, deletions and modifications of a field can be made in just one table and then propagated through the rest of the database via defined relationships.

There are three normal forms, each with increasing levels of normalization:

First Normal Form (1NF) : Every cell in the table must have only one value (i.e.,) it should not have multiple values.

Second Normal Form (2NF): All non-key attributes must be fully functional dependent on the primary key and not just the part of the key.

Third Normal Form (3NF) : The database must be in second normal form and non-prime attribute should be transitively dependent on the primary key.

Database is generally normalized up to 3NF, as every cell in the table has only one value i.e. it does not have multiple values. All non-key attributes are fully dependent on the primary key and not just the part of the key and non-prime attribute is transitively dependent on the primary key.

5.6 INPUT DESIGN

Input design is the process of converting the user-oriented description of the computer based business information into program-oriented specification. The goal of designing input data is to make the automation as easy and free from errors as possible.

Considering Input design, the following module will take place

- ✓ **Image Master :**

  Image Master acts as the gateway for preview imaging. The main advantage of using this module is it allows the scanned machine print document and scanned hand

written document to stored in the centralized database with the sufficient details related to that scanned tiff images like image details, image taken time and date and so on.

- ✓ **Preview Imaging :**

The preview imaging will be very helpful to the user in order to preview a bunch of scanned images having a mingled collection of machine print scanned images and hand written scanned images stored in the centralized database already. It is mainly used to select the required image soon to allow it for segmentation and classification process.

- ✓ **Segmentation :**

Once the scanned images allowed for segmentation, two panels will be available for easy segmentation of the scanned images which is present in the form of left hand side and the right hand side . In the left hand side of the panel, the selected image from preview imaging will be displayed. The functionality embedded in the left hand side panel is to allow the user by segmenting the image in the form of placing mouse cursor point as scoring point. Once the user left click the mouse, the pointed area is calculated as x1 and y1 points and when the user release the mouse click, the released area is calculated as x2 and y2 points. By using the calculation ratio, it initially analyzes the bounded region of the selected portion which is present inside the (x1,y1) and (x2,y2) segments. Now the segmented image will be displayed in the right hand side panel and in the bottom of the right side panel, the visual zoom scaling will be placed for more clear visualization of the segmented image to the user. In the segmented area, there is a button placed called comparative study, in which if the user clicks that button, the scanned segmented image in the right hand side panel will be allowed to compare with Existing reVAT and bigVAT methodology .

5.7 **OUTPUT DESIGN**

Output Design is the most important and direct source of information to the user. The output design is an ongoing activity during study phase. The objectives of the output design define the contents and format of all documents and reports in an attractive and useful format.

Considering Output Design, the following output will be extracted from the Module 6, Module 7, Module 8.

- The pixel values will be listed out in the corresponding reVAT list view based on Module-6 technique
- The pixel values will be listed out in the corresponding bigVAT list view based on Module-7 technique
- The pixel values will be listed out in the corresponding Dark Block Extraction Algorithm list view based on Module-8 technique
- The accuracy comparison and cluster count comparison.

CHAPTER 6

SYSTEM TESTING

The philosophy behind testing is to find errors. The common view of testing is to bring the program without errors. Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and code generation. Once the source code has been generated, software must be tested to uncover as many errors as possible before delivery to the customer. In order to find the highest possible number of errors, tests must be conducted systematically and test cases must be designed using disciplined techniques.

6.1 Validation Testing

Validation testing provides the final assurance that software meets all functional, behavioral and performance requirements. Validation testing can be defined in many ways, but a simple definition is that validations succeed when the software functions in a manner that is expected by the user. The software once validated must be combined with other system element. System testing verifies that all elements combine properly and that overall system function and performance is achieved. After the integration of the modules, the validation test was carried out over by the system. It was found that all the modules work well together and meet the overall system function and performance. According to this testing marketing is error free.

Example : Considering Login Form, validating username and password.

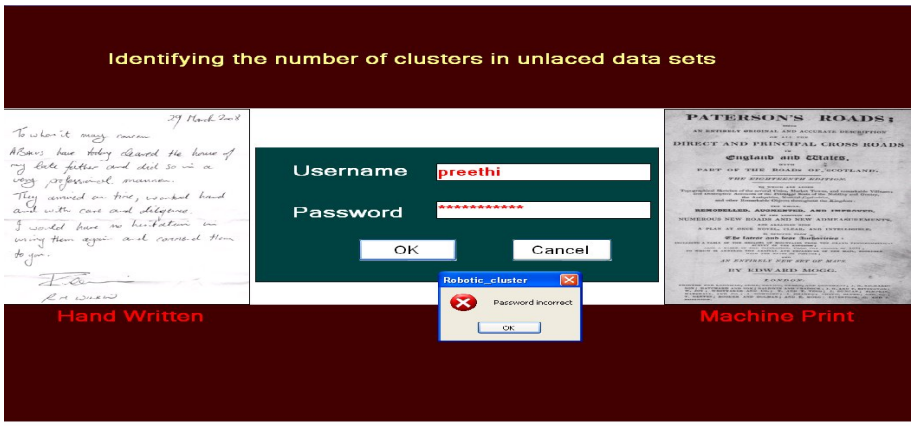

Fig 6.1 Validation Testing

6.2 Integration Testing

Integration testing is a systematic technique for constructing the program structure while at the same time conducting test to uncover errors associated with interfacing. The objective is to take unit - tested modules and build a program structure that has been dictated by design. Careful test planning is required to determine the extent and nature of system testing to be performed and to establish criteria by which the result will be evaluated.

All the modules were integrated after the completion of unit test. While Top - Down Integration was followed, the modules are integrated by moving downward through the control hierarchy, beginning with the main module. Since the modules were unit - tested for no errors, the integration of those modules was found perfect and working fine. As a next step to integration, other modules were integrated with the former modules.

After the successful integration of the modules, the system was found to be running with no uncovered errors, and also all the modules were working as per the design of the system, without any deviation from the features of the proposed system design.

Example : Considering Integration Testing, selected image from the preview is moving to Segmentation form.

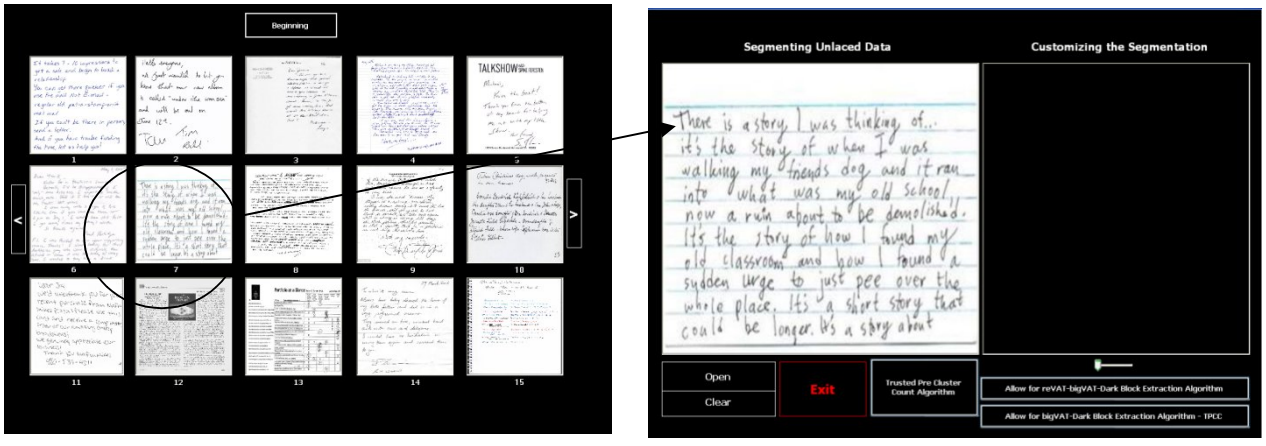

Fig 6.2 Integration Testing

6.3 Acceptance Testing

Acceptance testing involves planning and execution of functional tests, performance tests and stress tests in order to demonstrate that the implemented system satisfies its requirements. When custom software is built for one customer, a series of acceptance tests are conducted to enable the customer to validate all requirements.

In fact acceptance cumulative errors that might degrade the system over time will incorporate test cases developed during integration testing. Additional testing cases are added to achieve the desired level functional, performance and stress testing of the entire system.

6.4 Unit testing

Unit testing focuses verification effort on the smallest unit of the software. Using the detailed design description as design a guide, important control path are tested to uncover errors within the boundary of the module. This testing was carried out during programming stage itself. After testing each every field in the modules, the modulus of the project is tested separately. Unit testing focuses verification efforts on the smallest unit of software design and field. This is known as field - testing. According to unit testing lab is error free.

6.5 TEST CASES

The test that occurs as part of unit testing is given below.

- **Interface**

Tested to ensure the information properly flows in and out of the program unit under test.

- **Local Data Structures**

The temporarily stored data in this module have been checked for integrity. It was seen that no lose of data or misinterpretation of data was taking place in this module.

- **Boundary Conditions**

The data to this module have fixed length and are known to have a particular range of values. The input data with corresponding lower bound and upper bound values and also the

values in between the range, and was found that the module operates well with the boundary conditions.

- **Independent Paths**

The module was tested for independent paths to bound values and also the values in between the range, operates well with the boundary conditions.

- **Error Handling Paths**

The module was tested for error handling conditions. The module was given wrong input and was checked for error paths. It was found that the module was able to produce appropriate error messages for all the wrong inputs given to the module.

CHAPTER 7

SYSTEM IMPLEMENTATION

System Implementation is the stage of the project when the theoretical design is tuned into working system. If the implementation system stage is not carefully controlled and planned, it can cause chaos. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the users a confidence that the system will work and be effective.

The implementation stage in a project involves:

- Careful Planning investigation of the current system, checking constraints and the implementation.
- Training the staffs in the newly developed system.

A software application in general is implemented after navigating the complete life cycle method of a project. Various life cycle processes such as requirement analysis, design phase, verification, testing and finally followed by the implementation phase results in a successful project management. The software application which is basically a Windows based application has been successfully implemented after passing various life cycle processes mentioned above.

As the software is to be implemented in a high standard industrial sector, various factors such as application environment, user management, security, reliability and finally performance are taken as key factors throughout the design phase. These factors are analyzed step by step and the positive as well as negative outcomes are noted down before the final implementation.

Security and authentication is maintained in both user level as well as the management level. The data is stored in Access 2000 as RDBMS, which is highly reliable and simpler to use, the user level security is managed with the help of password options and sessions, which finally ensures that all the transactions are made securely.

The application's validations are made, taken into account of the entry levels available in various modules. Possible restrictions like number formatting, date formatting and confirmations for both save and update options ensures the correct data to be fed into the database. Thus all the

aspects are charted out and the complete project study is practically implemented successfully for the end users.

7.1 SYSTEM FLOW DIAGRAM

An overall representation of the system can be represented by using system flow diagram. In a system flow diagram the source and the destination are depicted by a rectangle. The arrow in a system flow diagram represents the flow of data from one source to the other.

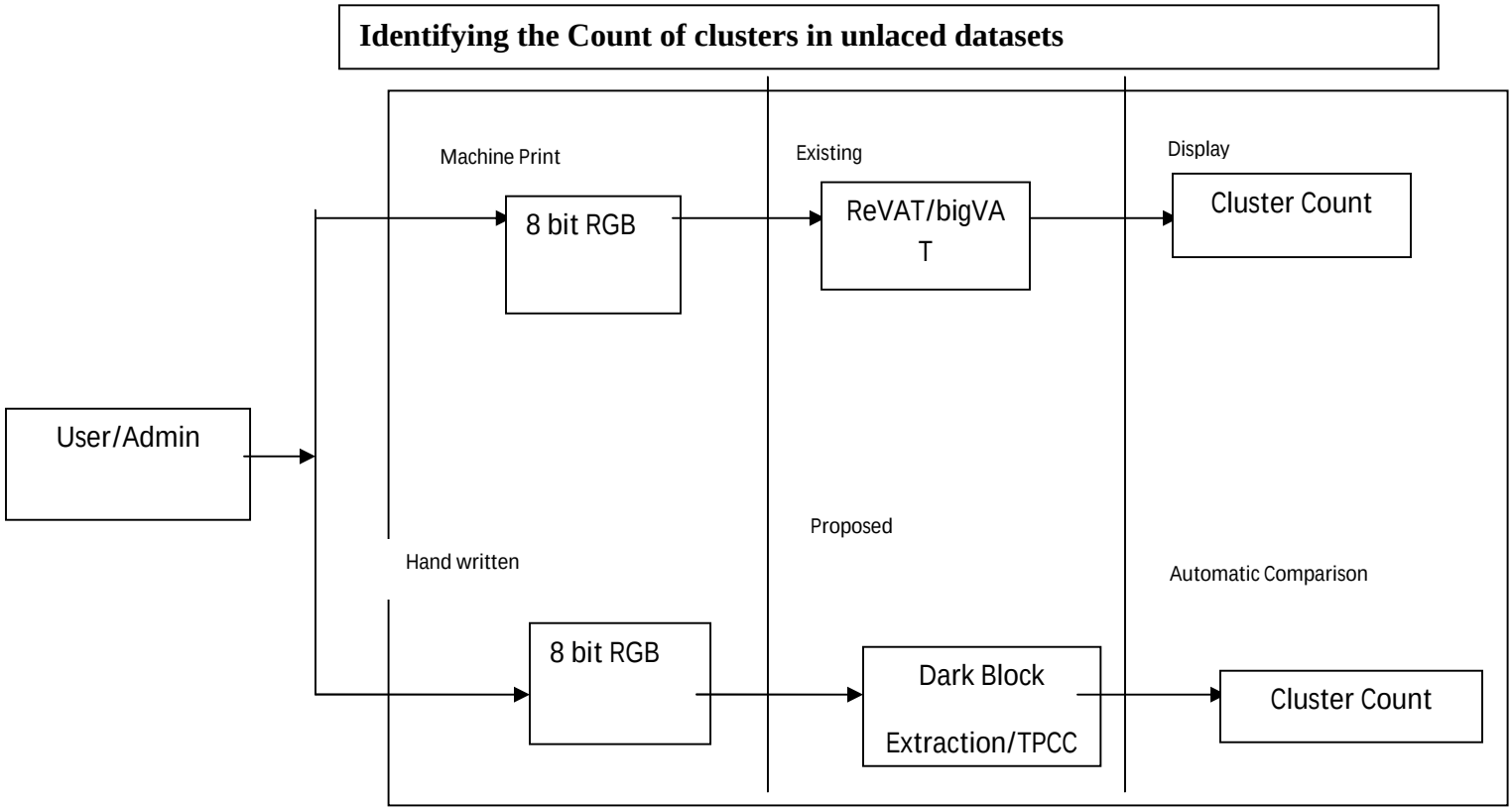

Fig 7.1 System Flow Diagram

7.2 SYSTEM ARCHITECTURE:

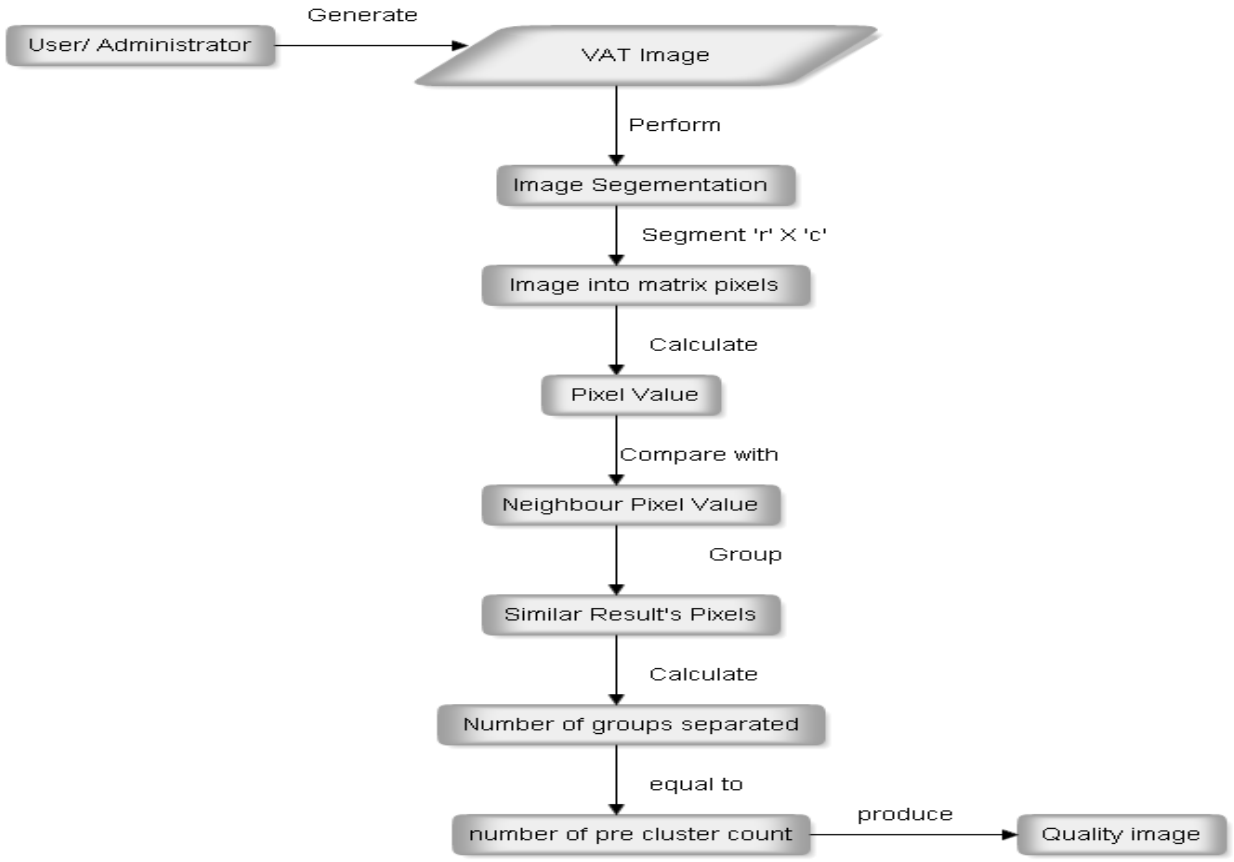

Fig: 7.2 System Architecture

CHAPTER 8

CONCLUSION AND FUTURE ENHANCEMENTS

8.1 CONCLUSION

Experiments confirm what many users of clustering believe: that most methods prefer "larger" rather than "smaller" clusters. Thus the cluster number (corresponding to the detection of major peaks) extracted by TPCC appears to be increasingly reliable. This is understandable since if increases, the segmented binary image will be less noisy, which is naturally helpful to subsequent processing. As long as the filter sizes are set to be less than the minimum meaningful cluster size, the larger they are, the more reliable the estimation of the cluster number should be. TPCC will probably reach its useful limit when the RDI formed by any reordering of D is not from a well structured dissimilarity matrix. In our experiments, we used the simple euclidean distance to compute pairwise dissimilarities when the input data are feature vectors. The euclidean distance may not be suitable for high dimensional or complex data. It is that TPCC does not eliminate the need for cluster validity, but it simply improves the probability of success. The initialization of the **Trusted Pre Cluster Count algorithm** for object data clustering is highly useful.

7.2 FUTURE ENHANCEMENT

TPCC is more reliable than DBE . The coding has been done cautiously so that any developer can follow the programs easily with the knowledge of the convention followed hence it is easy to be maintained. It should not be too hard to find an approximate center sample for each meaningful cluster from any well structured RDI. Extrapolating an approximate centroid for each cluster will not only speed up the termination of the clustering algorithm but also reduce the need for multiple runs with randomized initializations. Inferring the approximate sizes of each cluster. Although DBE is not in and of itself a clustering method, it may provide some useful information on object labels, especially for objects around the peak in the projection signals. If such label information could be used, only the remaining boundary objects need to be clustered, thus reducing the amount of data to be clustered.

CHAPTER 9

APPENDICES

APPENDIX 1

SCREEN SHOTS

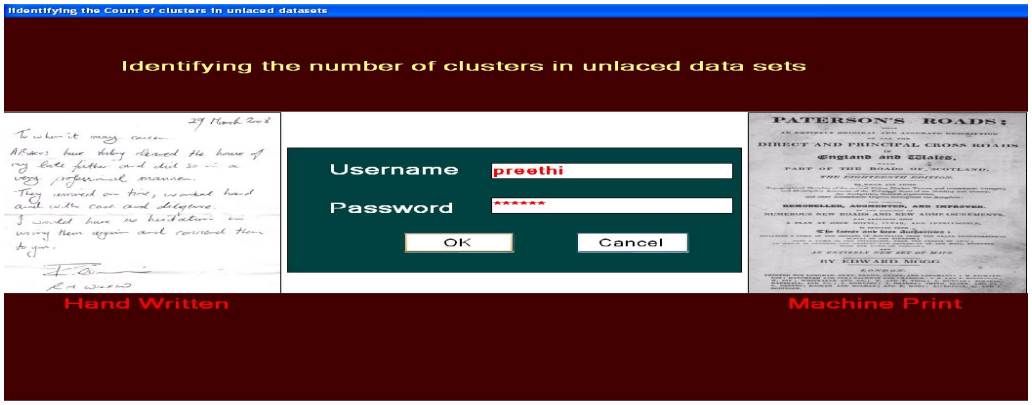

Fig: A 2.1 LOGIN FORM

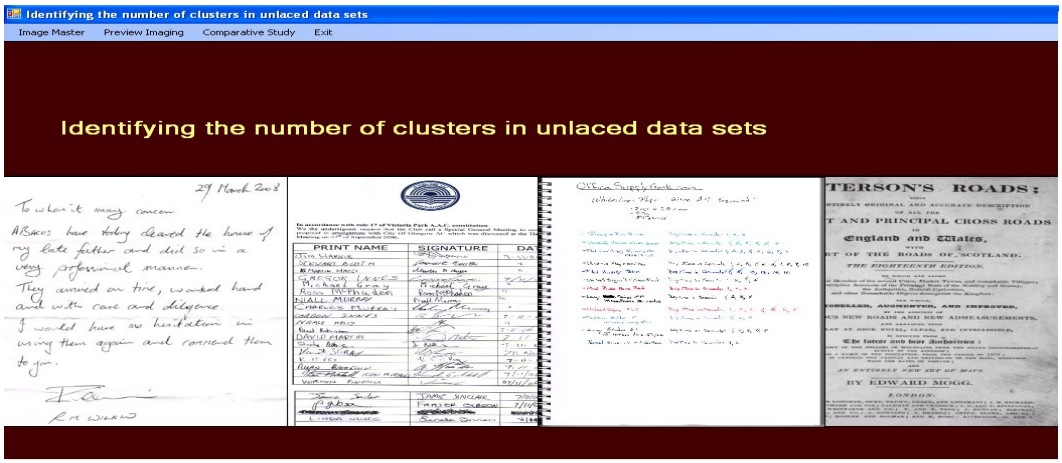

Fig: A 2.2 MENU FORM

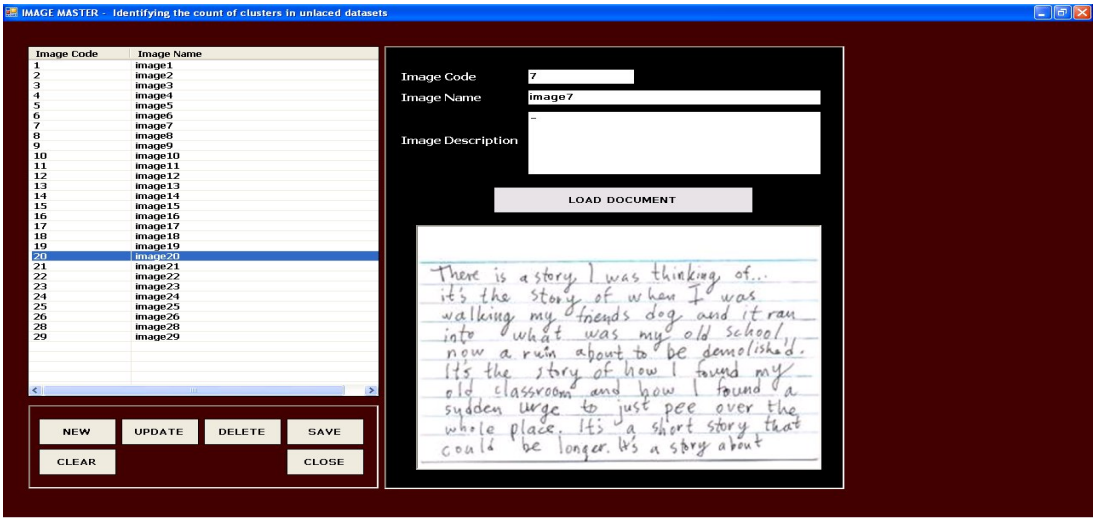

Fig: A 2.3 IMAGE MASTER

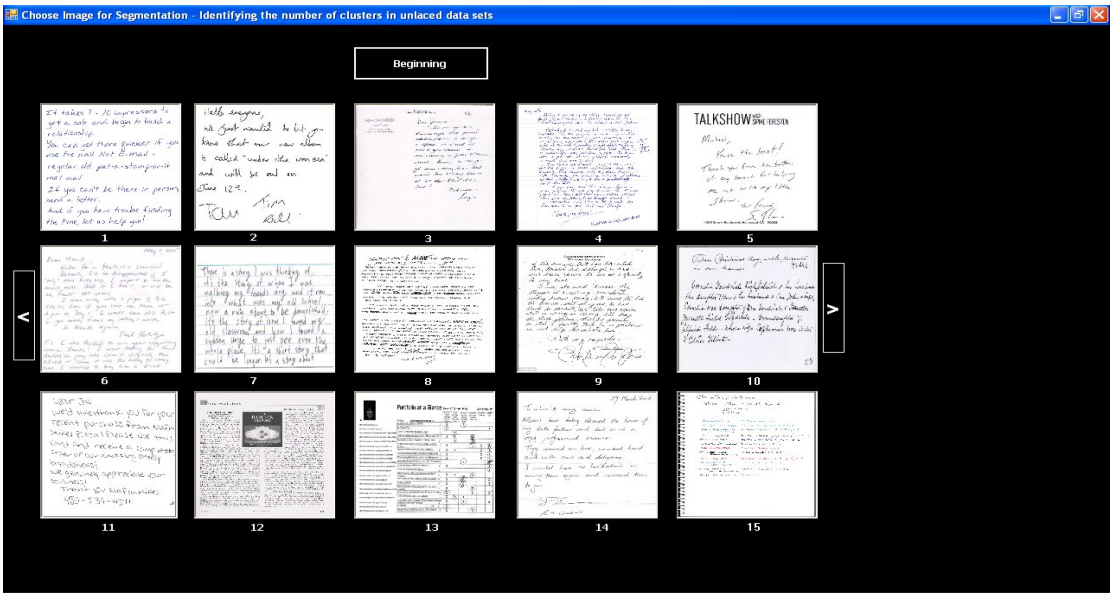

Fig: A 2.4 PREVIEW IMAGING

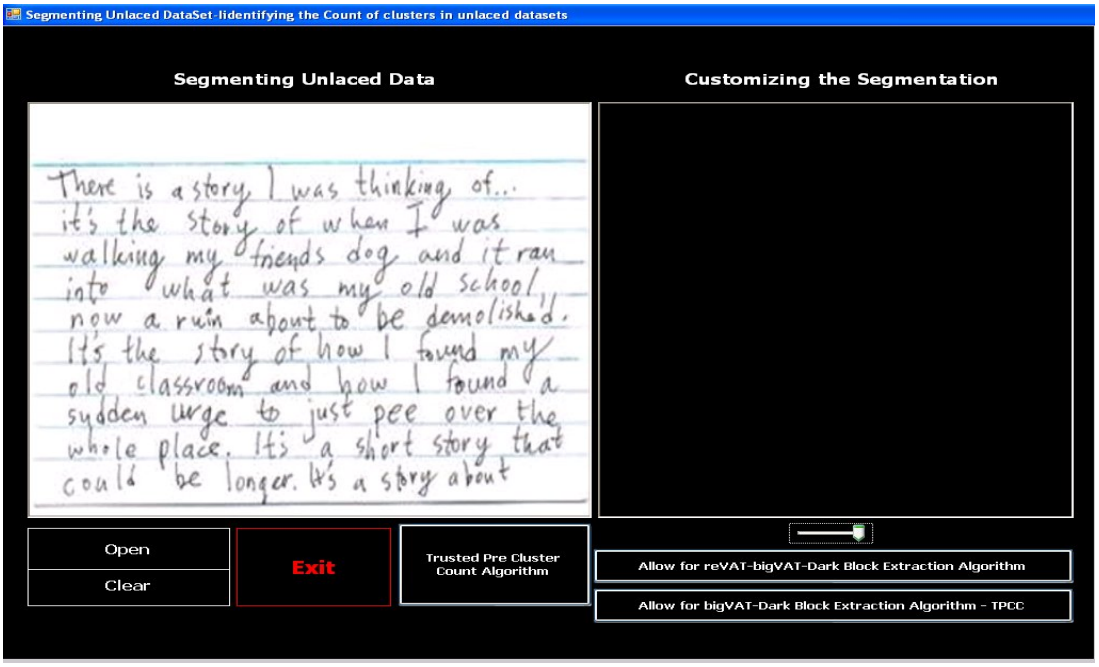

Fig A 2.5 SEGMENTING UNLACED DATA

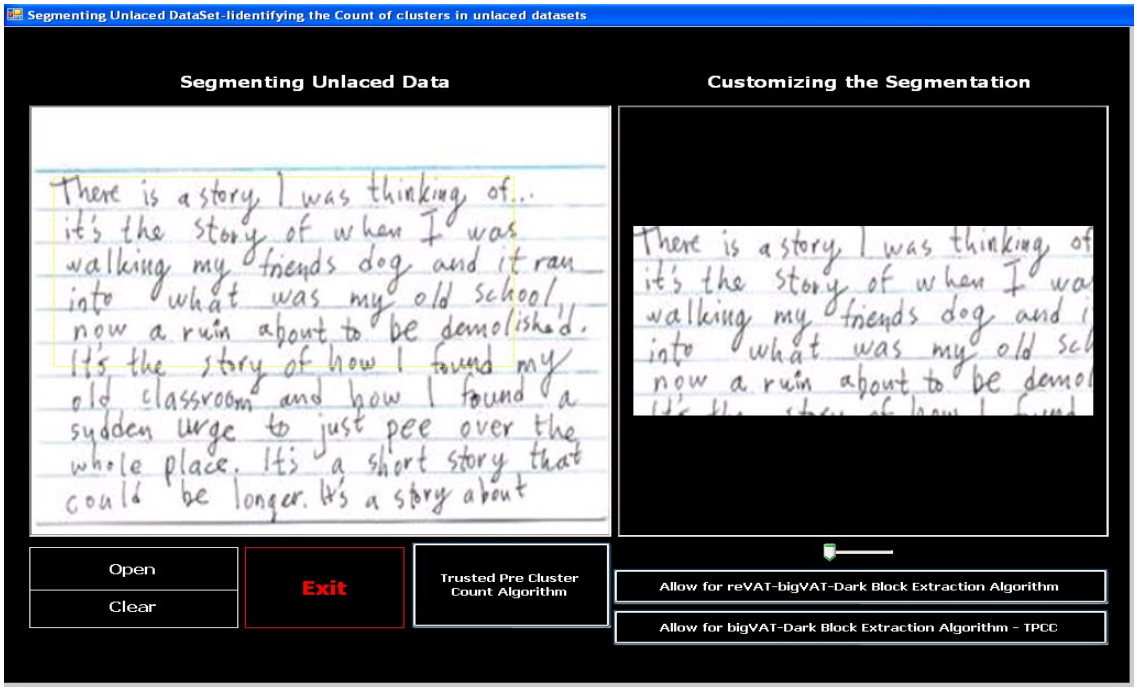

Fig A 2.6 AFTER SEGMENTING

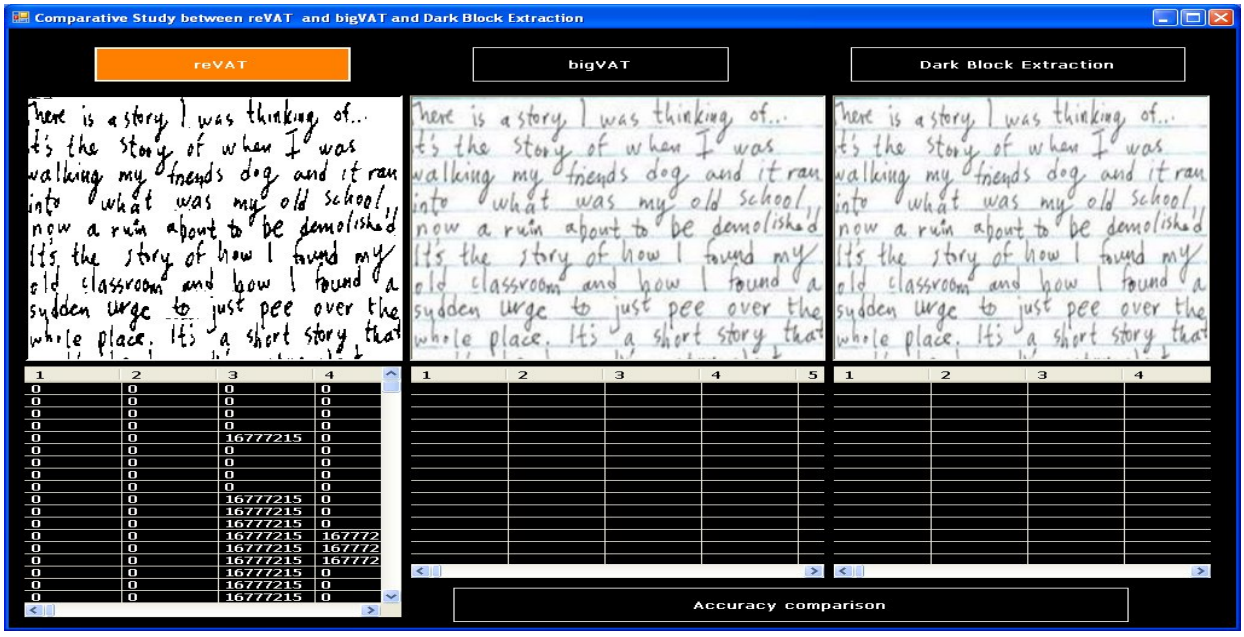

Fig A 2.7 reVAT Output

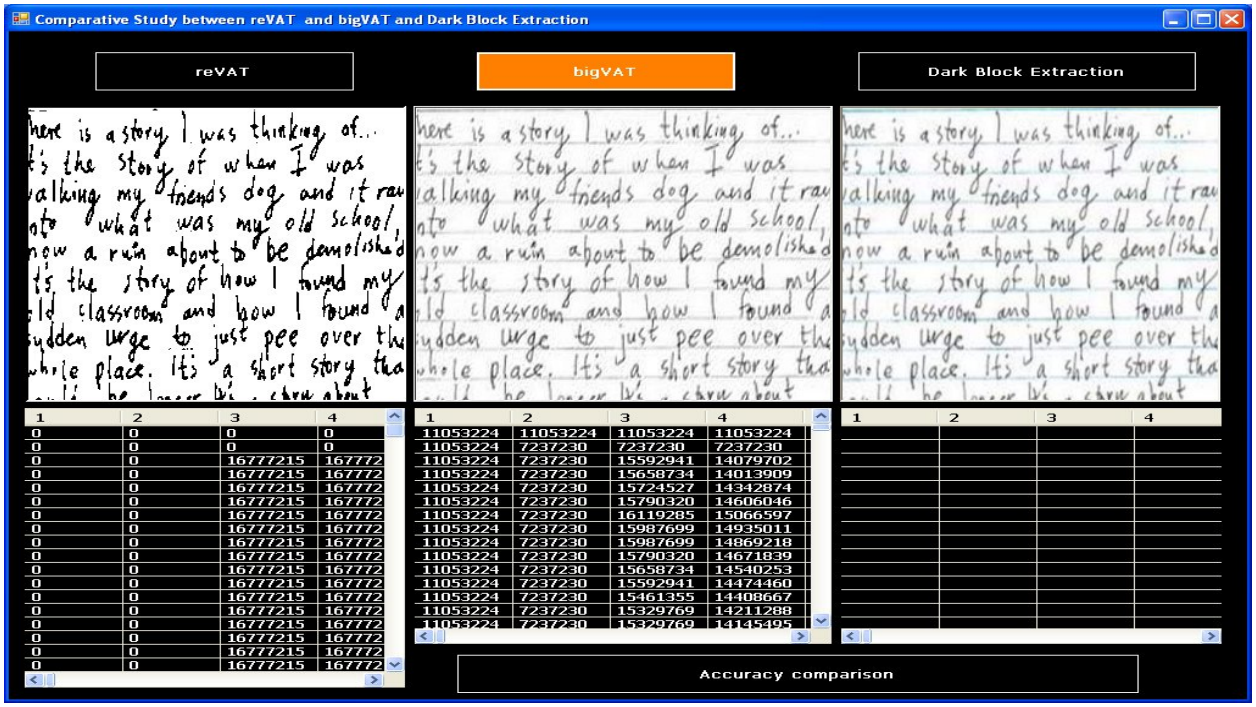

Fig A 2.8 bigVAT Output

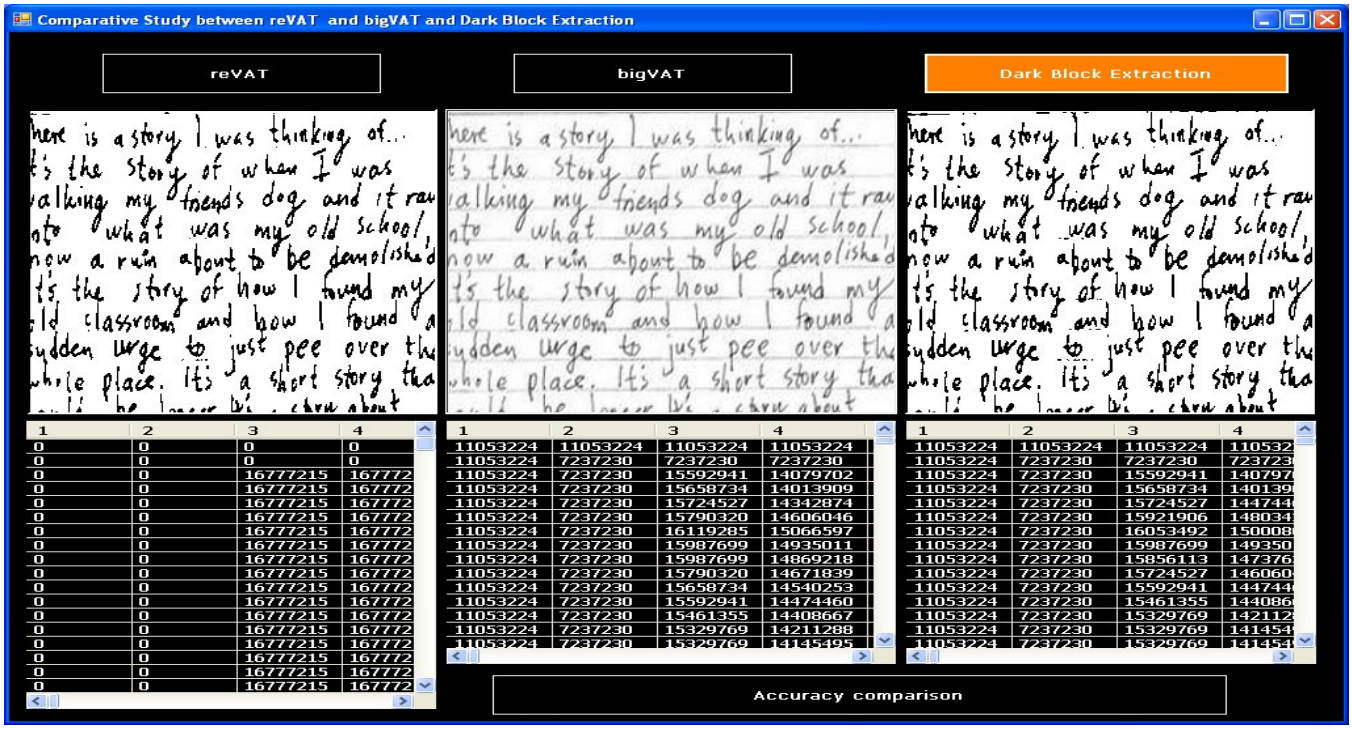

Fig A 2.9 Dark Block Extraction Output

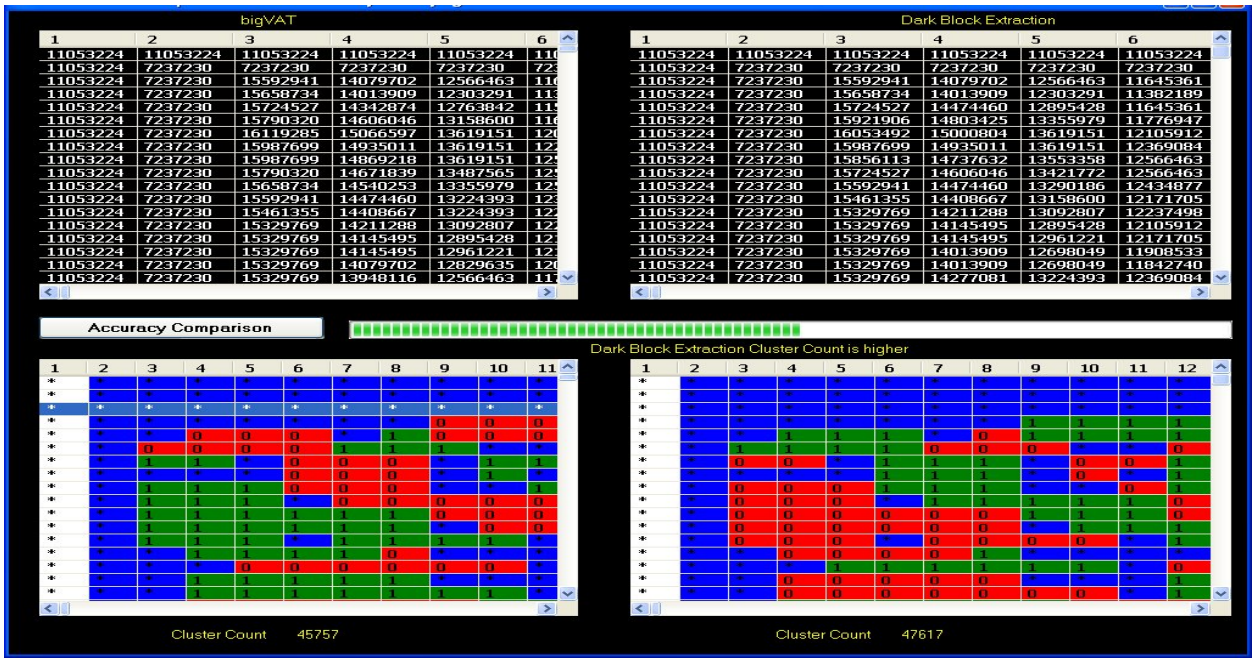

Fig A 2.10 Cluster Count Comparison

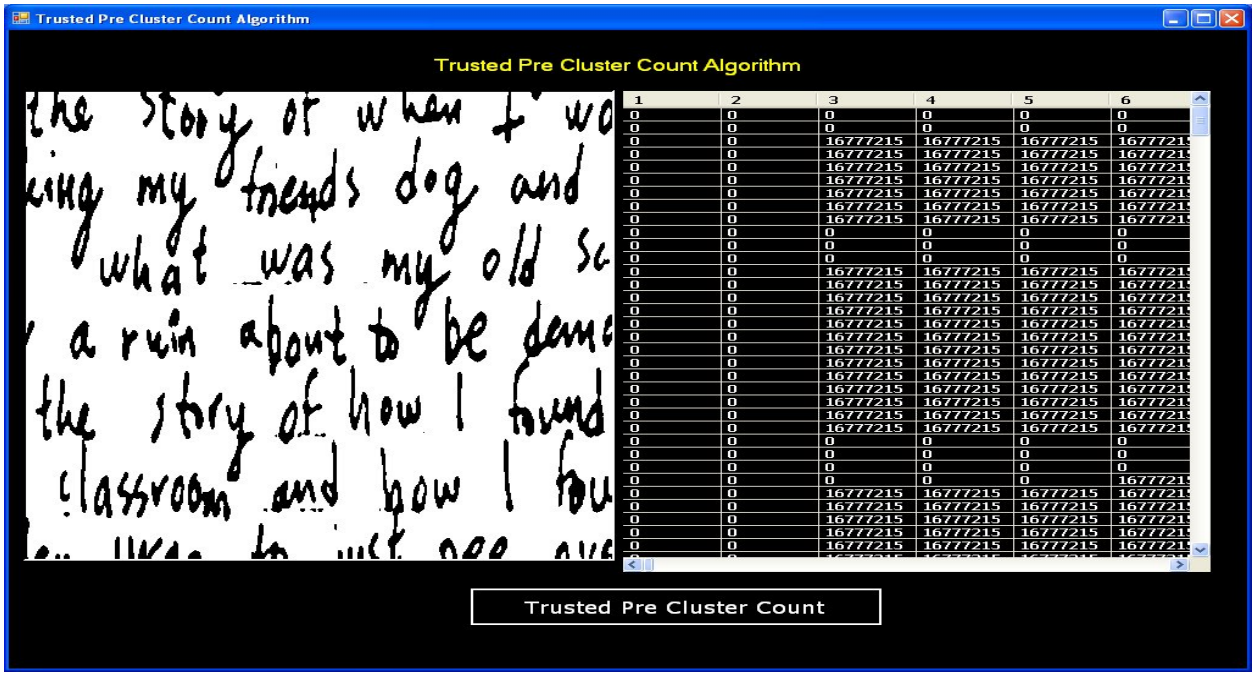

Fig A 2.11 Implementing Trusted Pre Cluster Count

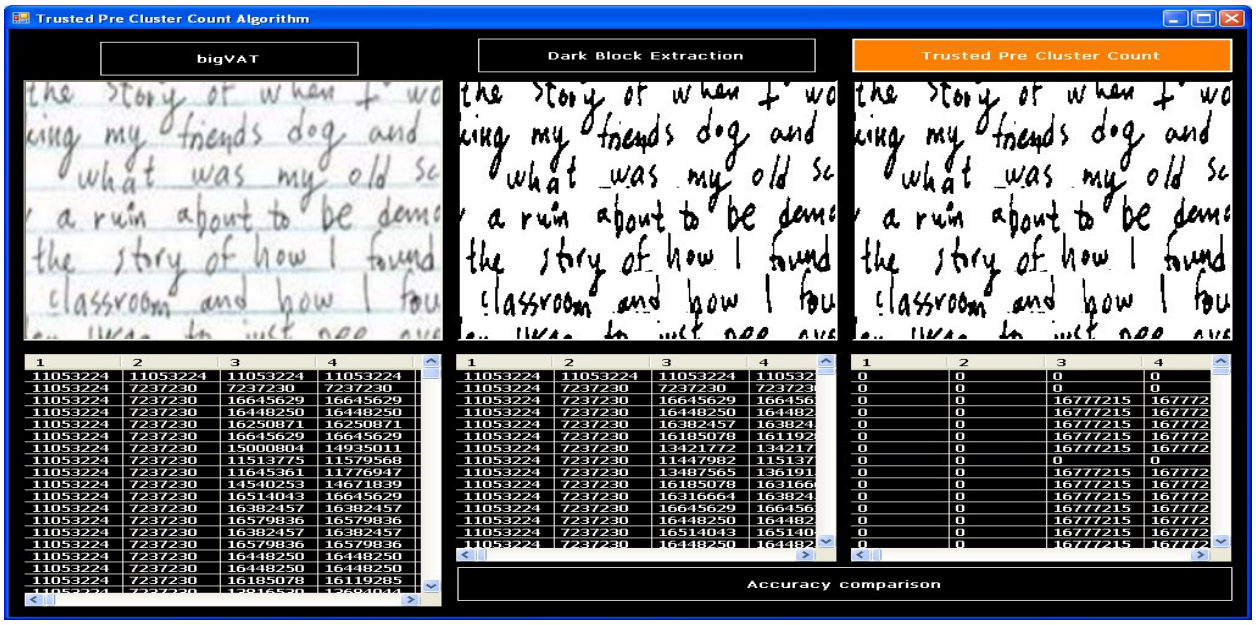

Fig A 2.12 bigVAT , Dark Block Extraction and

Trusted Pre Cluster Count Comparison

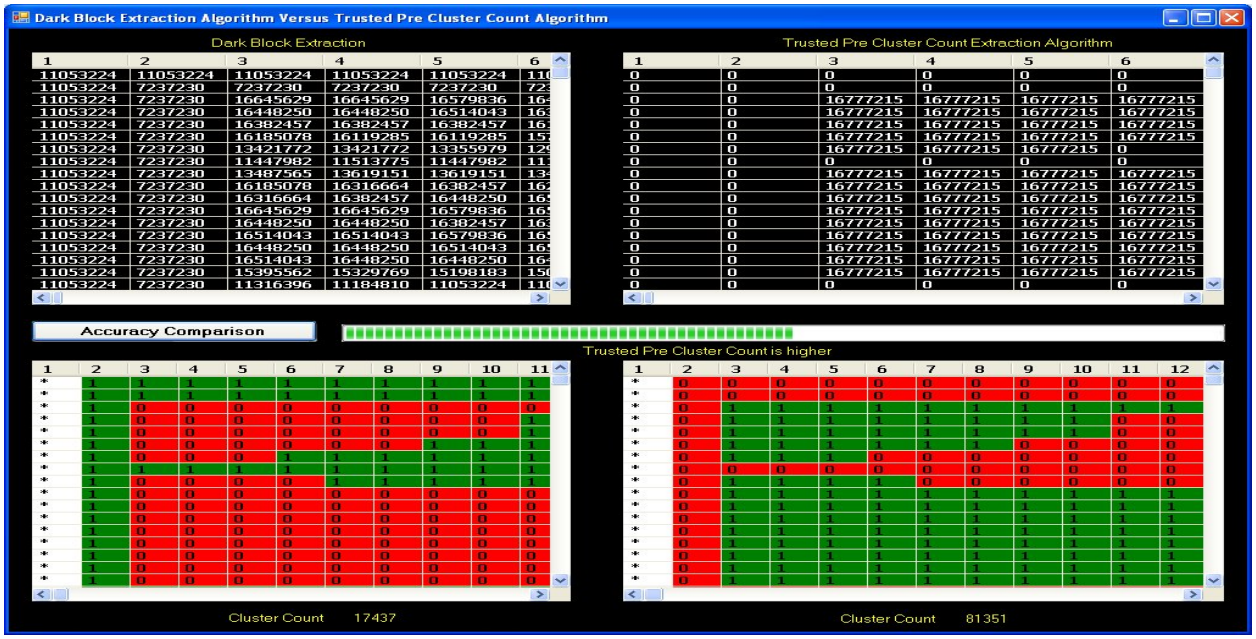

Fig A 2.13 DBE and Trusted Pre Cluster Count Comparison

CHAPTER 9

BIBLIOGRAPHY

1. Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, and James Bezdek, (2012) "Automatically Determining the Number of Clusters in Unlabeled Data Sets," IEEE TRANS. KNOWLEDGE AND DATA ENGINEERING, VOL. 21,pp.335-350

2. Bezdek,J.C. Hathaway,R.J. and Huband,J. (2011) "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices," IEEE Trans. Fuzzy Systems, vol. 15, no. 5, pp. 890-903.

3. Bezdek,J.C, and Hathaway,R. (2012) "VAT: A Tool for Visual Assessment of (Cluster) Tendency," Proc. Int'l Joint Conf. Neural Networks (IJCNN '02), pp. 2225-2230.

4. Huband,J. Bezdek ,J.C. and Hathaway,R. (2005) "bigVAT: Visual Assessment of Cluster Tendency for Large Data Sets," Pattern Recognition, vol. 38, no. 11, pp. 1875- 1886.

5. Hathaway,R.. Bezdek ,J.C. and Huband,J. (2006) 'Scalable VisualAssessment of Cluster Tendency,'Pattern Recognition, vol. 39,pp. 1315-1324.