"DATA REDUNDANCY REMOVAL IN OBJECT DATABASES USING CLUSTERING TECHNIQUES"

A

Dissertation

submitted in partial fulfillment of the

requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

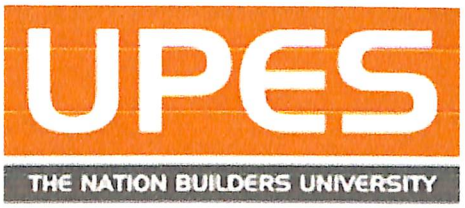
ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWROK

by

Name	Roll No.
Rittika Raichaudhuri	R102214010

under the guidance of

Mr. Anil Kumar



Department of Computer Science & Engineering

Centre for Information Technology

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, UK

April – 2016

The innovation driven
E-School

CANDIDATE'S DECLARATION

I hereby certify that the project work entitled **" Data Redundancy Removal In Image Databases Using Clustering Techniques"** in partial fulfilment of the requirements for the award of the Degree of MASTER OF TECHNOLOGY in ARTIFICIAL INTELLIGENCE AND ARTIFICIAL NEURAL NETWORK and submitted to the Department of Computer Science & Engineering at Center for Information Technology, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my/ our work carried out during a period from **January, 2016** to **April, 2016** under the supervision of **Anil Kumar, Assistant Professor, CIT, COES, UPES**.

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

Rittika Raichaudhuri

Rittika Raichaudhuri
Roll No. 102214010

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

Date: April 2016

12/04/16.

Anil Kumar
Thesis Guide

Center Of Information Technology
University Of Petroleum & Energy Studies
Dehradun – 248 001 (Uttarakhand)

12/04/16

Dr. Amit Agarwal
Program Head – M.Tech (AI and ANN)
Center for Information Technology
University of Petroleum & Energy Studies
Dehradun – 248 001 (Uttarakhand)

ACKNOWLEDGEMENT

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

I am heartily thankful to my course coordinator, **Vishal Kaushik**, for the precise evaluation of the milestone activities during the project timeline and the qualitative and timely feedback towards the improvement of the project.

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

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

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

Rittika Raichaudhuri

R102214010

ABSTRACT

"Data Redundancy Removal In Image Databases Using Clustering Techniques," is a project which is going to deal with removing redundancy from an image data base thereby providing a clean dataset which could be used in further applications.

Duplication of image templates in the object database creates redundancy, which will affect the training process and in turn the classifier performance in machine learning tasks. Hence the removal of redundant data is of prior importance. Image templates which appears analogous may consist of variant information content. In similar, visually diverse images may comprise of analogous data. Hence it is practically impossible to remove the repetitive pattern/images manually. Here, this work targets at the removal of such redundant data existing in large object database using clustering techniques. Clustering is a task of grouping set of objects, in such a way that the objects belonging to the same group exhibits similar characteristics comparing with the objects belonging to another cluster (group). This project aims at the removal of redundancy in object database for the generation of an optimized and efficient database in the field of machine learning and data mining.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Supervised learning is a machine learning process of inferring a function from a labeled set of training data. A training dataset consists of all those data which are used for learning that is to fit the parameter of the classifier. Therefore, it can be inferred that the effectiveness of the supervised learning algorithm depends greatly on the precision of the training dataset. However, training datasets can be populated with unwanted, redundant and noisy data making the dataset unfit for accurate learning.

The main aim of this project is to detect and remove redundant and unwanted data from the image database and generate an optimized and efficient training dataset, with maximum intra-variance information for training machine learning algorithms. Clustering techniques are used for the redundancy removal purpose. Clustering is the process of grouping data objects in such a way that the data points in the same group share the property of similarity and the property of dissimilarity with those from different groups. Clustering algorithms can be further classified on the basis of clustering models as Hierarchical clustering, Centroid-based clustering, Density Based clustering, Distribution- based clustering, Partition- based clustering

1.1 History

The very idea of machine being able to do their work on their own and similar to that of humans led to the coining of the concept of Artificial Intelligence. In order for machines to work independently they should be able to learn by themselves and function accordingly to perform a particular query. This marked the dawn of Machine learning technique which allows computers to learn from a given data without being explicitly programed to do so. It explores the study and creation of algorithm than can learn and make predictions on and as well as infer knowledge from a given set of data.

Machine learning, can further be divided into two broad categories depending on the type of learning nature of the algorithm:

- *Supervised Learning:* in this the algorithm is provided with a labeled set of training data and from this the algorithm has to learn the relationship between the dataset.
- Unsupervised Learning: in this the algorithm is provided with an unlabeled set of training data and from this the algorithm has to learn and infer the relationship between the dataset.

A supervised learning algorithm examines the training data and infers the desired knowledge from the data, which can be used for mapping new data sets (testing data). A training data is made up of a set of training examples which are a pair of an input object (typically a vector) and a desired output value (also called the supervisory signal). So it can be inferred that the training data plays an important role as the entire prediction depends on this data set. Hence it is important to provide a clean training data set for learning purpose which does not contain redundant data or inconsistent data. The nature of these data sets can either be numerical, images, hybrid or complex. Years of research has been spent to do the same.

Many techniques have been adopted to clean the data before providing it for training. DanasinghAsir et al. [1], in this research used feature selection method along with clustering technique to remove redundancy from training data set. Most researchers have adopted different feature selection and clustering techniques to remove redundancy from a dataset. Clustering methods like BIRCH, SOM, K-means algorithm etc. have been adopted to cluster the similar data points in a dataset [2,3].

Images or commonly known as pictures forms an important part of human life. A lot of data or information can either be stored in images or can be retrieved from images which can be further used for analysis purpose. Many fields like medicine, military, space research, education uses images as data. Image processing and pattern recognition techniques another sub-field of artificial intelligence are used to extract information from such images.

Image processing provides with algorithms which when applied on images produces or extracts vital information from images. Information extracted through image processing can be further used by classification algorithm, pattern recognition algorithm, clustering algorithm etc. to achieve some goals.

In this project Self-Organizing Map clustering technique will be used to group similar images in order to identify redundant as well as similar images. To achieve this, a significant amount of information will be extracted through image processing steps like feature extraction, from each image and compared to identify and detect redundant images.

1.2 Requirement Analysis

Requirement analysis, also known as requirement engineering, is amethod of defining user's expectations and needs from a particular system being developed which willthen be used by the user themselves. These requirements, must be quantifiable, relevant and detailed in nature which could be easily understood by the system designers and engineers.

Requirement analysis can be divided into the following categories (a) Functional Requirements (b) Non-Functional Requirements (c) Software Requirement and (d) Hardware Requirement

1.2.1 Functional Requirements:

Functional requirements specify what the system should do. It specifies the main characteristics that the system should have that fulfills the user's needs. Functional requirements can be defined in terms of calculations, technical details, data manipulation and processing and other specific functionality that outlines what a system is supposed to accomplish. The functional requirements of this project are as follows:

- Should be able to cluster images which are similar in nature
- Slight change in orientation or scaling of the images should not affect the clustering.
- The system should work for all three types of data sets that is it should provide good results for all the data sets.
- The clustering technique should not be alone dependent on the feature selection method.

1.2.2 Non-Functional Requirements:

Non-functional requirements are those that specifies how the system should work that is its constraints and its abilities. It provides an overall description of the property of the system as a

whole or of a particular aspect and not a specific function. The non-functional requirements of this project are:

- Works only for static images.
- Image should be an RGB image.
- Image should be digital.
- The System should be able to handle minor errors and not crash with unwanted input.
- The system should produce a result that is authentic in every sense.
- The system will assist even in remote areas.
- The system should work for all three data sets (Vehicle, Pedestrian, TSR).
- The system should be able to run on machine that can fulfill its basic requirements.

1.2.3 Software Requirement:

Table 1: Software Requirement

ITEM	APPLIED FOR
Softwareresources:	
MatLab	Development Tool

1.2.4 Hardware Requirement:

- At least 1.6 GHz Pentium Processor or Intel compatible processor.
- At least 4 GB RAM.
- 32-bit (x86 processor) or 64-bit (x64 processor)
- A video graphics card.
- At least 10 GB free hard disk space.

1.3 Main Objective

- Data redundancy removal from the object database and for including maximum variance information that is required for generating an optimized data base, which will assist the object detection and recognition tasks.

- Clustering technique is used for redundancy removal purpose. Clustering technique, groups the data with similar information, which helps the user to identify the image templates belonging to same class.
- Redundancy Removal through similarity distance between two images within the same cluster would further detect the most similar and duplicate images thereby deleting only those images from the cluster which make the dataset noisy. This step will ensure no good data is lost.

1.4 Sub Objective

- Detection and removal of redundant data from the image database using clustering techniques.
- This technique reduces the human effort of removing similar image templates for creating an optimized data base.
- It is helpful in identifying the data that are redundant in information content but appears to be visually different and vice versa.
- To avoid data inconsistency and information corruption while training machine learning algorithms

1.5 Scope of the System

The scope of this project is to be able to build such a system that will be able to extract important and relevant features from the images and identify and cluster together similar images. The project scope also involves in finding hidden information and providing a data set that consists of assorted images. This system will be able to provide results that can be easily interpreted by the concerned user. It will not only reduce the manual labor but also provide faster results. Additionally, the image clusters can also be used by other algorithms which may require clustered datasets like these.

1.6 Feasibility Study

Feasibility analysis is the process of (a) recognizing the candidate system, (b) assessing all the options and (c) then selecting the most feasible system. This is done by studying all work done in

the area under investigation or be reviewing the feasibility of any proposed ideas which may not exist practically related to the new system. It is a test of a system proposal according to its workability, impression on the organization, ability to satisfy the user needs, and effective use of resources. The objective of feasibility study is not to solve the problem but to acquire a sense of its scope.

Three key considerations involved in the feasibility analysis are: economical, technical and operational.

1.6.1 Technical Feasibility

- The users of our system need no additional training and the product requires minimum hardware requirements.
- Training datasets will be more efficient and would train the algorithm more efficiently.

1.6.2 Economic Feasibility

Once the hardware and software requirements get fulfilled, there is no need for the user of our system to spend for any additional overhead. For the user, the product will be economically feasible in the following aspects.

- No expert will be required to interpret the feature that is extracted by the system.
- Our product will reduce the time that is wasted in manual processes.
- Easily available application.

1.6.3 Operational Feasibility

The system will reduce the time consumed to detect and group similar images. It will also be able to group those images which may visually appear to be different but are actually similar in nature and vice-versa. Hence operational feasibility is assured.

2. SYSTEM ANALYSIS

Duplication of the original data in a database is known as Data Redundancy. But the meaning of the term is not limited to only data duplication, it also occurs when highly similar form of data exists in a database. Data redundancy sources a lot of complications like it makes the data set inconsistent, it may also lead to data corruption and anomalies. Due to this it becomes difficult to extract information from the dataset making the data set useless. Hence it is essential to remove redundancy from a dataset such that it can be further used for analysis. Removing redundancy through manual procedures is a tedious as well as a time consuming task.

The proposed system solves the above stated problem by providing a semi-automated system that will group similar images in an image data base and remove the duplicate images thereby returning a clean dataset which can be used by any other algorithm for further processing.

2.1 Existing System

The goal of this study was to develop a system that could group similar images from very large datasets using the BIRCH clustering algorithm [2]. BIRCH algorithm is one such algorithm that is capable of handling noisy datasets. This algorithm, first scans the dataset once and builds an initial in-memory CF tree and then proceeds by applying clustering over this CF tree. BIRCH can easily take care of large clustering problems by concentrating on densely occupied areas, and by using a condensed summary. It employs measurements that captures the natural closeness phenomena of the data.

Through this paper the author provides a solution of finding near duplicate images from very large datasets using clustering techniques [3]. To achieve this the first step that was adopted was to extract important features from each individual images using the PCA-SIFT feature extraction algorithm. The next step was to index these local descriptors or features using Locality Sensitive Hashing Technique. In the final step weighted graphical method was applied to cluster the similar images in one cluster.

This study portrays the concept of grouping similar and duplicate images in a large scale unstructured dataset which consists of one million images obtained through random searches over the internet(Google)[4]. The first stage was to represent images as sets of binary visual attributes. For this purpose, 'Bag of Visual Words,' representation was used. Then in the next stage was to mine groups of similar images using LCM (Linear Time Closed Item-set Miner) Algorithm. This method scales linearly both in time and in memory as it required only three minutes and around 150 Megabytes and detected around 80 thousand groups of duplicate images in a database of 1 million images.

This paper presented a clustering based indexing technique which grouped similar images into a single cluster using colour features as the clustering criteria [5]. This method was proposed as a solution to image retrieval problems. According to this paper one can simply extract the feature vectors from the query image and match it with the predefined clusters obtained from the above stated algorithm and extract images from only those clusters which have the maximum hit.

The author presents an innovative method of displaying only those images in an ordered fashion which match the query image in an image retrieval system [6].In this the first step was to select a collection of neighboring target images for a query image using Nearest-neighbor method (NNM). Next, it constructs a weighted undirected graph containing the query image and its neighboring target images. Finally, the normalized cut (N-cut) method was used for image clustering.

In this paper the authors have used large image data sets [7]. First they applied the k-means clustering algorithm over the feature vectors of the images to group similar images together and then density based clustering algorithm was applied to obtain better clustering results. In the next step a super hyperplane classifier support vector machine (SVM) was used which classify all the outlier left from density based clustering. Density based clustering grouped the images as per the nearest feature sets but did not group outliers This method improves the efficiency of image grouping and gives better results.

The goal of this paper was to provide an in depth study on the different data mining issues [8]. The authors attempted to identify the unique research issues that can arise while performing image mining. A decent image mining system comprehends the following functions: image

storage, image processing, feature extraction, image indexing and retrieval, patterns and knowledge discovery. This paper discusses all the bottle necks that may appear while applying datamining algorithms for grouping or clustering image files.

In this paper a comparative study on different feature extraction technique based on moments invariants was presented [9]. Also different feature extraction methods were combined together and the results were tested accordingly. The data set consisted of 2000 handwritten (Devanagari) numerals. The different moment invariant methods that were tested were Correlation Coefficient, Principal Component Axes (PCA) and Perturbed Moments.

According to the author of this paper images in a dataset could be clustered together using Self Organizing Map(SOM), Artificial Neural Network Method [10]. It performed the experiment using 250 colour (RGB) images which were first converted into Gray images. It then proceeded by finding the colour histogram and then performing the feature vector selection using two methods (a) PCA (Principal Component Analysis) and (b) LSA (Latent Semantic Analysis). The final step was to perform clustering on these feature vectors using the SOM clustering technique. According to this study report highest accuracy (which equaled to 88%) was obtained when using PCA combined with SOM and by selecting 100 feature vectors from each image as compared to LSA combined with SOM which could produce accuracy up to 74%.

As per this paper a new approach towards image retrieval system was adopted where SOM algorithm was used to organize images according to their similarities [11]. Feature vectors in the form of shape features such as roundness, rectangularity, elipticity, eccentricity, bending energy were used for each individual images. Finally clustering was performed on these feature vectors and images were organized accordingly.

2.3 Motivations

In present day scenario a lot of information is stored and transferred in the form of images and videos. Some of these image and video databases are also used for analysis and information extraction purpose which is known as image or video mining. Also a single image holds more information than any numerical data set.

- Images or video file require a lot of storage space and redundancy of these files in a database causes lot of wastage in storage space.
- Mining information from any dataset requires the data set to be clean that is the data set should be free of any redundancy and noise.
- Also unsupervised learning algorithms in machine learning require the training data sets to train the algorithm for mapping the relation between the input and target parameters. These training datasets should be clean and invariant in nature.
- Removing redundancy through manual technique may end in producing faulty results as. Also manually it would take a lot of time and effort to remove redundancy from image datasets.
- In fact, clustering techniques are also used to check the image authenticity which automatically identifies whether the query image is a fabrication or a simple copy of the original one, [13].
- Sometimes fast and efficient grouping or clustering similar images can be useful for other applications as well like image retrieval purpose, [6].

2.3 Proposed System

The system developed have the following incorporated functionalities:

- The user can load 'N' number of cropped images which either belong to the vehicle, pedestrian or TSR datasets which could be of any format.
- The images have to be a still image.
- The images will then be converted into gray scale images.
- Then the next step is feature extraction and creation of the feature vectors.
- Hu's moment invariance feature extraction or Histogram of Orientation Gradient method will be used for extracting features. The one method that suits the best will be selected for satisfying the same purpose.
- Finally clustering algorithm will be applied on these feature vectors and clustering operation will be performed.
- Clustering will be achieved using the SOM clustering Algorithm.
- The result would be folders containing similar images after clustering.

- Then computing the correlation (similarity distance measure) between the images within a particular cluster and removing the most similar images from the clusters.
- The final step would be to provide a dataset that would be clean and without any redundancy

2.4 Modules

The proposed system can be broken down into three segments: (a) Colour Space Conversion: in this segment the RGB images are converted into Gray scale images, (b) Image Processing segment: features of each image will be extracted and stored as feature vectors (c) Image Mining segment: where clustering algorithm will be applied and similar features will be clustered together as per the extracted features and (d) Redundancy Removal: on each clustered group a correlation method in the form of similarity distance measure will be applied to seek the most similar images and finally remove them and provide a clean dataset. The feature set of each image will serve as the input parameter for the clustering algorithm.

2.4.1 Colour Space Conversion:

In this step each RGB image in the data set is converted into a Gray image. This is done by extracting individual intensity value representing each colour channel (R, G and B) of a particular pixel and then by replacing it with the weighted average of these intensity values into that particular pixel.

2.4.2 Image Processing (Feature Extraction):

Image processing refers to application of mathematical operators on images to extract information from the images. In this input is an image, or a video, and the output of image processing may be either an image or a set of characteristics or parameters defining the image. Feature extraction is the process of reducing dimensionality that efficiently represents and describes interesting points of an image as a compact feature vector. Feature extraction technique can be based on color, shape or texture features.

- *Color Feature Extraction Technique:* is the process of extracting information from the colour properties of an image. It is performed by calculating the colour histogram (which

can be defined as frequency of occurrence of each colour pertaining to a particular colour range) of an image.

- *Shape Feature Extraction Technique:*in this technique information about the shape of the object in an image is extracted. Some parameters that are used to define the shape of an object arecenter of gravity/centroid, eccentricity, circularity ratios, ellipticvariance, solidity etc.
- *Texture Feature Extraction Technique:*The texture feature of an image contains information as regards to contrast, uniformity, rigidity, regularity, etc.

2.4.3 Image Mining (Clustering):

Image mining is the process of searching and discovering valued information and knowledge from an image or a series of images. This method draws its basic principles from the concepts of data mining, machine learning, statistics, pattern recognition and 'soft' computing.

Clustering belongs to the group of unsupervised learning problems, that deals with providing a structure to a unlabeled data set. The structuring of data basically means clubbing the data into groups or clusters that are similar in nature but show certain level of dissimilarity with data points present in other clusters.

Some well-known cluster models include:

- *Connectivity models:* builds models based on the distance connectivity (for example, hierarchical clustering).
- *Centroid models:* this algorithm represents each cluster by a single mean vector (for example, k-means).
- *Density based models:* defines clusters as connected dense regions in the data space (for example, DBSCAN and OPTICS).

2.4.4 Redundancy Removal (Data Correlation)

Redundancy is the process of creating duplicate files of an already existing file and storing them. This causes the data set to have unnecessary copy of the original data. Redundancy removal is the process of removing this duplicate data from the data set.

Data correlation refers to the process of detecting the relationship that exists between two data points or variables. Correlations are useful since they define a relationship that could be exploited by other applications or practices. Relationship between two data points means the dependency of the two data point over each other. It can either be a strong or a weak correlation between two data points. There are various ways of identifying the correlation with similarity distance measure being one such method.

Similarity measure is a real value that quantifies the relation or similarity between two data points. It is usually computed using a distance formula like Euclidean Distance, Chord Distance, Manhattan Distance, Mahalanobis Distance, etc.

3. SYSTEM DESIGN

The dataset related to pedestrian, vehicle and TSR would be collected. Algorithms based on image processing techniques for feature extraction would be designed. Manual feeding of the datasets, in the form of digitized RGB color photographs would be done for feature extraction. Then the extracted features will be used to perform clustering and the result would be folders containing similar images. Now all the images in a cluster are not bad data, some of the data from these clusters are worth preserving. So in order to find the images that serve as actual noise a similarity distance measure that belongs to data correlation technique is used.

3.1 Use-Case Model

Use case is used to describe a systems behavior as how it responds to a request that originates from an outside source that does not belong to that system. In other words, it describes "who" as in the user can do "What" with the system under consideration. This practice is used to capture a system's behavioral requirements by detailing any scenario driven threats that may occur due to the functional requirements.

A use-case is designed using the following:

- **Actors:**an actor could be a person, an organization or any external system that interacts with the system directly. Actors are basically the user of the system. Actors are drawn by a stick figure.
- **Associations**: Associations between actors and the use cases are depictedin the diagram through solid lines. An association exists whenever an actor is involved with any kind of interaction as described by the use case. Associations are demonstrated as lines connecting the use case and the actor with an arrow head on one end of the line. The arrow head is often used to indicate the direction of the interaction.

- **System boundary boxes**: A rectangular box enclosing the use caseis called as the system boundary box which indicates the scope of the system being built. Anything within the box represents functionality that is in scope and anything outside the box is not.

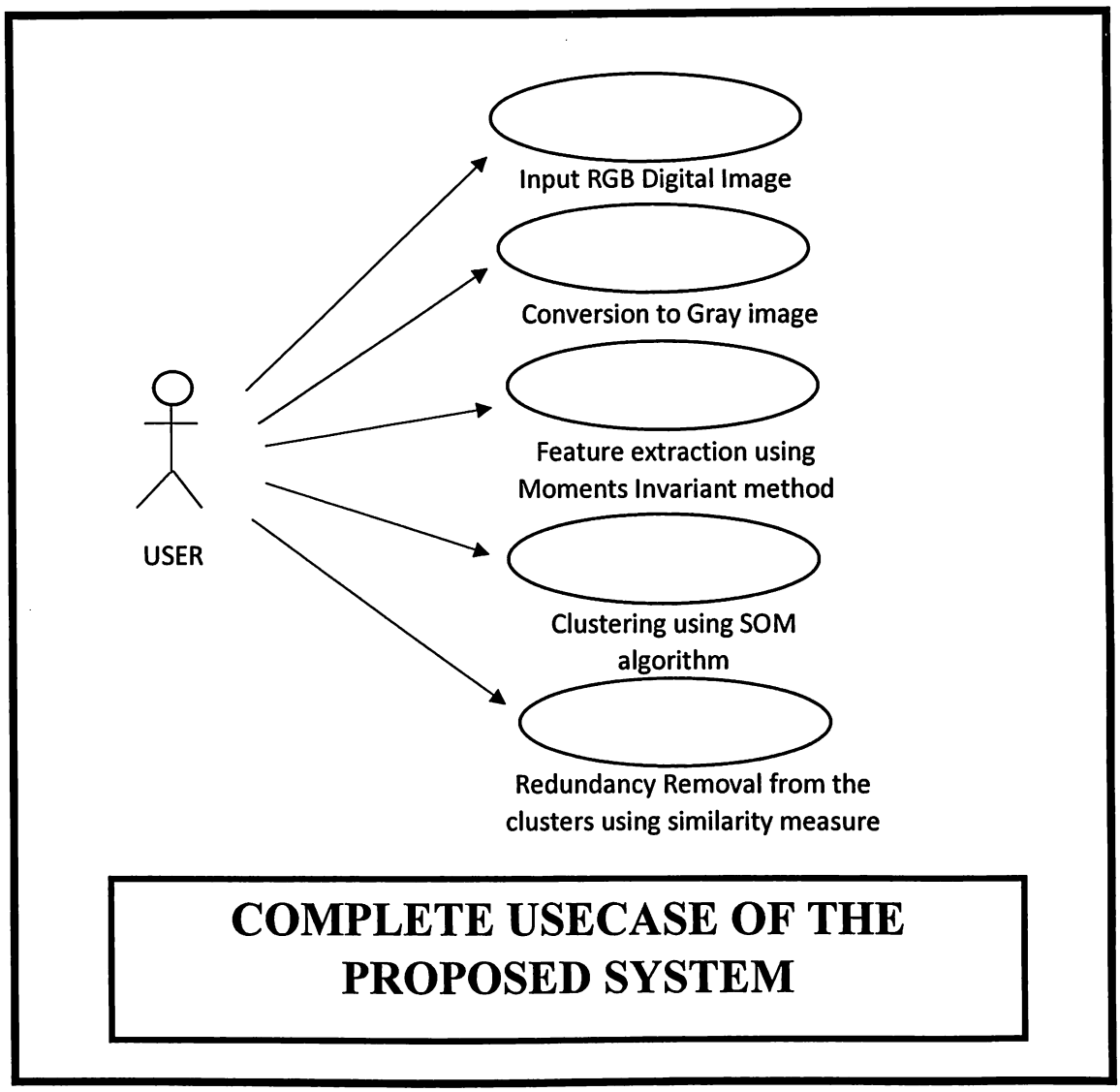


Fig. 3.1. Use-case Diagram of the Proposed System

ACTORS1: User of the system.

USECASE1: Inputs the 'N' number of RGB images.

USECASE2: Conversion to Gray image

USECASE3: Feature extraction using Moments Invariant method.

USECASE4: Clustering using SOM algorithm.

USECASE5: To remove the redundant images and produce a clean dataset.

3.2 Flow Chart

A flowchart is a kind of diagram that represents an algorithm or the workflow of the whole process, by displaying the steps as boxes of various types (as rectangle, ellipse, parallelogram each depicting a different functionality), and their order by connecting them with arrows. This illustrative representation exemplifies a solution model to a given problem. Flowcharts are used in analyzing, designing, documenting or managing a process or program in various fields.

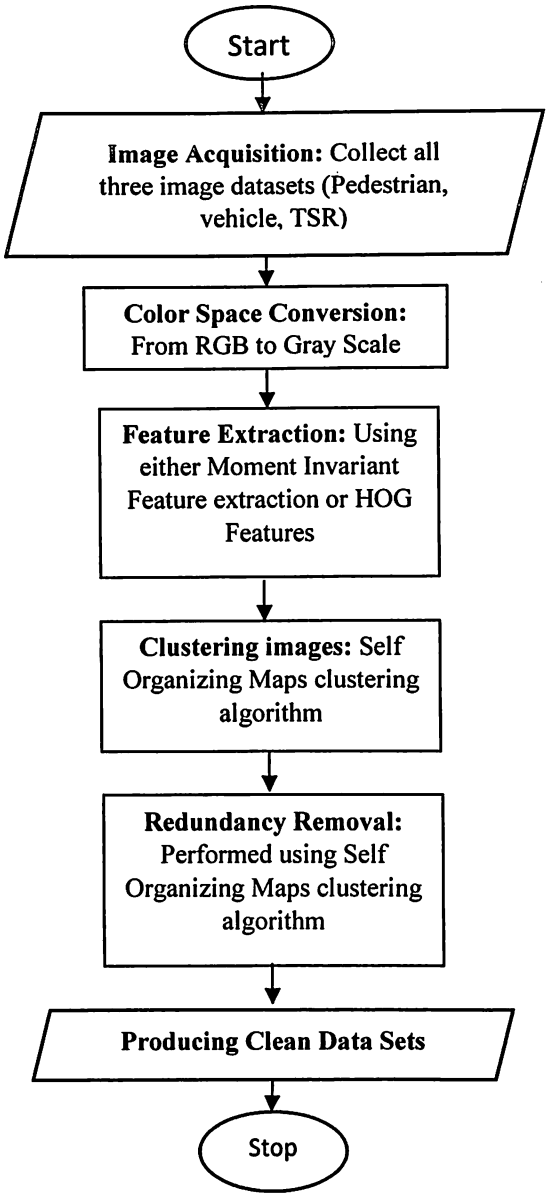


Fig. 3.2 Flowchart of the proposed System

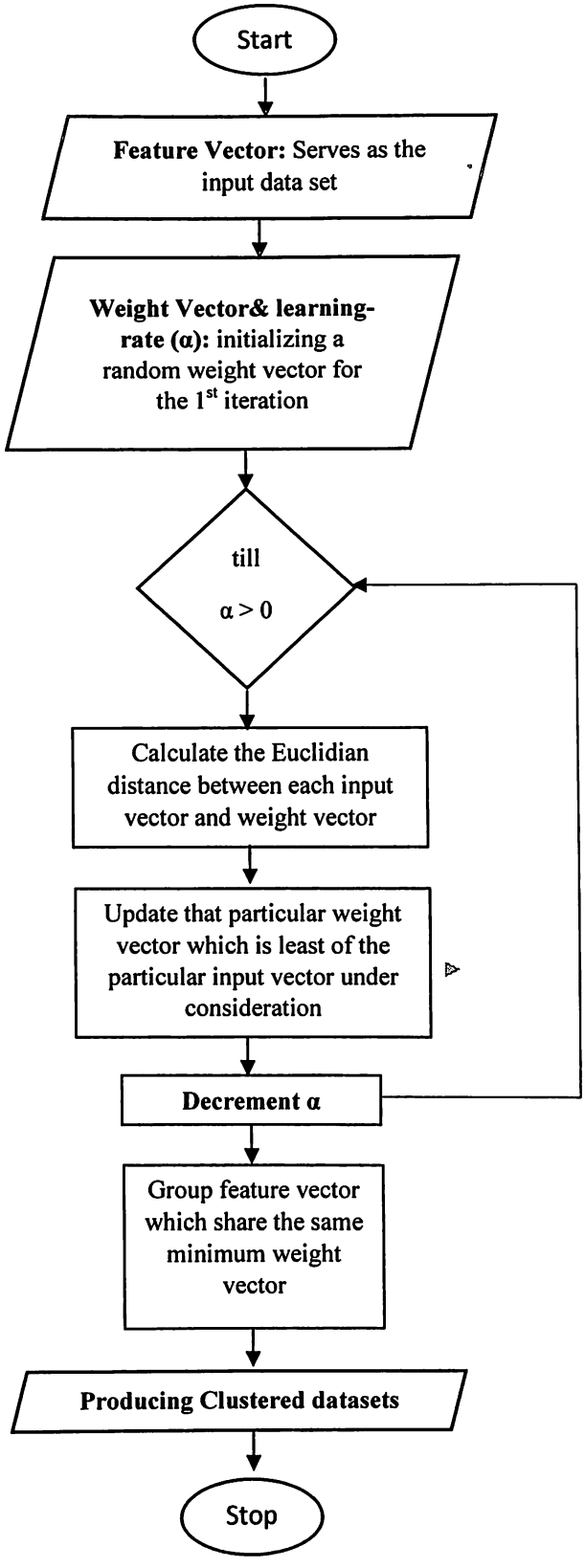


Fig. 3.3 Flowchart of the SOM Clustering Algorithm

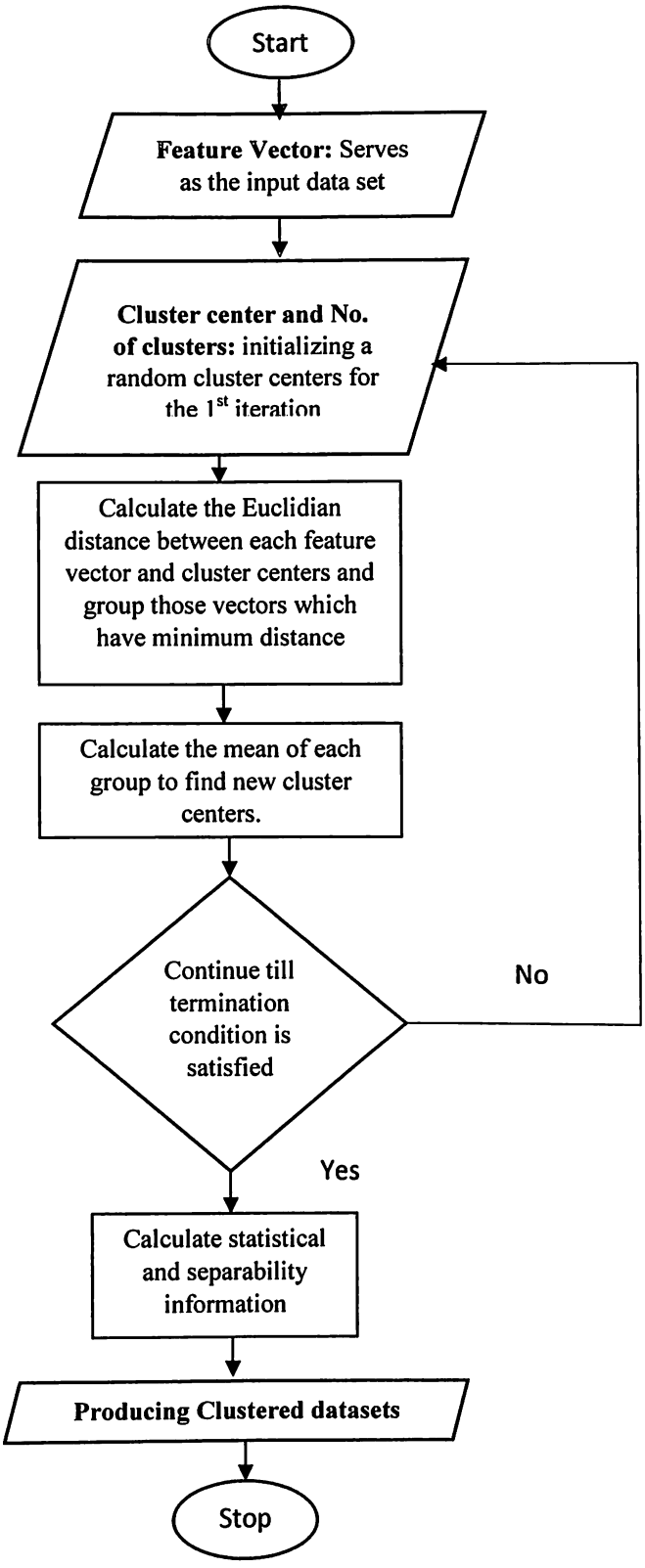


Fig. 3.4 Flowchart of the K-means Clustering Algorithm

3.3 Block Diagram

Block diagram is a diagrammatic depiction of the proposed system which highlights the principal components or functions by representing it with blocks and lines. The flow of the data is shown using lines with arrow head, and the arrow head shows the direction of the data flow which in turn portrays the relationship between the blocks. They are frequently used in engineering the hardware design, electronic design, software design, and process flow diagrams. They however provide a less detailed description of the system and only defines the overall concepts of the system. They do not provide a detailed report of the implementation.

Block diagrams use rudimentary geometric shapes like boxes and circles.

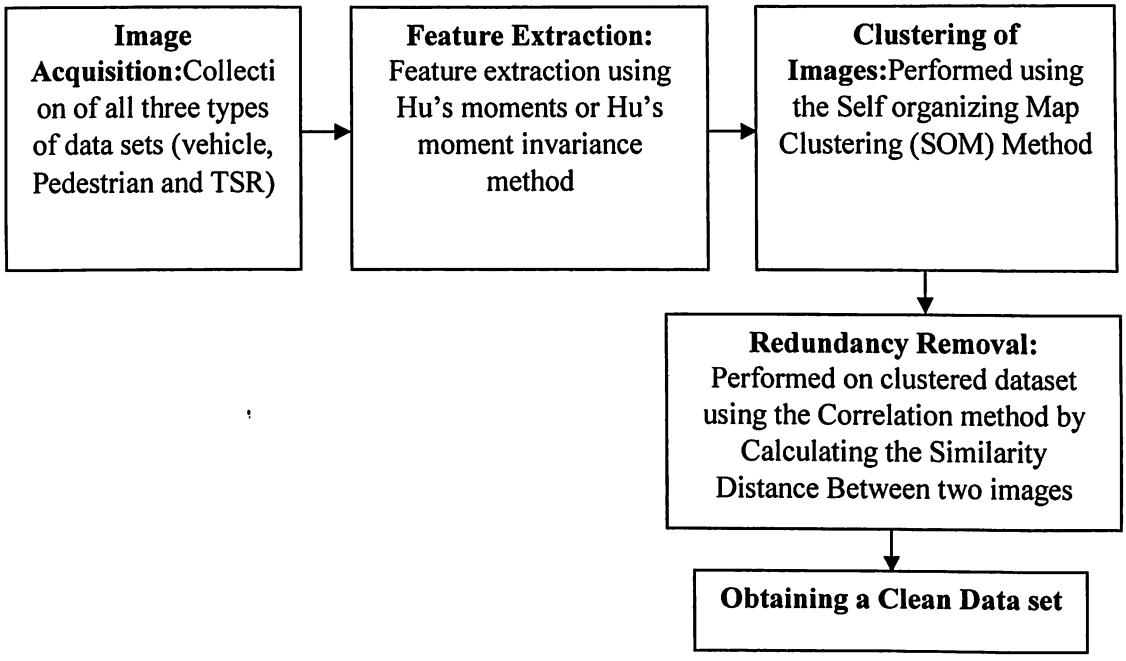


Fig. 3.5. The Basic Block Diagram of the Proposed System

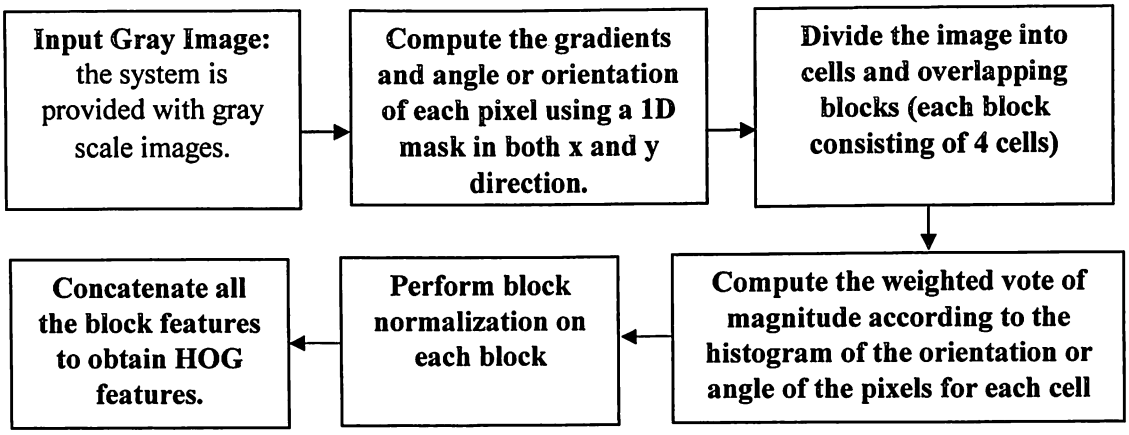


Fig. 3.6. The Basic Block Diagram of the Hog Feature Extraction Method

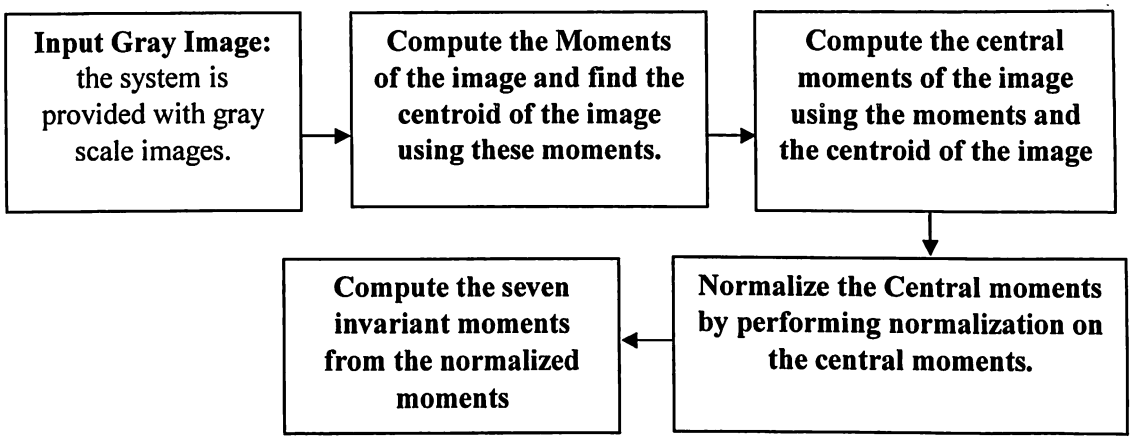


Fig. 3.7 The Basic Block Diagram of the Moment Invariant Feature Extraction Method

4. REDUNDANCY REMOVAL

4.1 Colour Space conversion

In this step if the image is in RGB color space it would be converted into Gray scale color space. In RGB color space of a digital image the color components of an object's color are all correlated with the amount of light hitting the object. This makes object discrimination difficult in RGB color space. In RGB images each pixel consist of intensities contributed by three different colour channels namely red, blue and green. Whereas, gray images consist of pixels with intensities belonging to only one channel. These images are also known as black-and-white images, are composed of shades of gray, varying from black at the weakest intensity (0 intensity value) to white at the strongest intensity (255 intensity value). Due to this application of image processing algorithm on gray scale images becomes easier as each pixel consists of only one intensity value.

The following formula was used to convert RGB image to Gray image:

$$Gray\_Value = 0.2126 * R + 0.7152 * G + 0.0722 * B$$

4.2 Feature Extraction

Feature extraction is the procedure of outlining a set of necessary features, or image characteristics that forms the core element which when represented in an efficient or meaningful manner give the required information that is important for analysis and classification purpose. Feature extraction technique can be based on color, shape or texture features. This technique can be broken down into two parts (a) feature construction and (b) feature selection.

In this project both Moment Invariant as well as Histogram of Orientation Gradient (HOG) feature extraction method was used to perform feature extraction from images and for creating the feature vectors. The feature extraction method which was more suitable for performing clustering was finally chosen.

Moment Invariant Feature ExtractionMethod:

An image moment is obtained by calculating the weighted average (also known as moment) of the image pixel intensities that outlines the properties of the object which is under consideration. Image moments highlight simple properties of the image like area or total intensity of the image, the image centroid, or information about its orientation (angle).

Computing Moments:

A (p + q)th ordered moment of an image that dependent on the scaling, rotation and translation factor of that image can be calculated using the following formula:

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x,y)\, dx\, dy \quad (1)$$

for digitized signals it is of the form

$$m_{pq} = \sum\sum x^p y^q f(x,y)\, dx\, dy \quad (2)$$

where f(x,y) represent the image and 'x', 'y' are the coordinate points on the image.

Computing Central Moments:

The moments defined in equation (1) are not invariant in nature that is when the image f(x,y) is subjected to translation, rotation or scaling the moments obtained will also be effected. So central moments are calculated which are translation invariant using the following formula:

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - x_c)^p\, (y - y_c)^q\, f(x,y)\, dx\, dy \quad (3)$$

for digitized signals it is of the form

$$\mu_{pq} = \sum\sum (x - x_c)^p\, (y - y_c)^q\, f(x,y)\, dx\, dy \quad (4)$$

where x_c and y_c corresponds to the centroid of the image f(x,y) and are defined as

$$x_c = m_{10}/m_{00} \quad \& y_c = m_{01}/m_{00}$$

The central moments μ_{pq} that is computed using the centroid of the image f(x,y) is comparable to the m_{pq} but, whose center has been shifted to the centroid of the image.

Normalizing Central Moments:

Scale invariance can be obtained by normalizing these central moments. The normalized central moments are defined as follows:

$$\square_{pq} = \mu_{pq} / (\mu_{00})^{\square} \quad (5)$$

where

$$\square = 1 + ((p + q)/2)$$

Computing Moment Invariants:

Based on normalized central moments, Hu[15] introduced seven moment invariants:

$$W_1 = \square_{20} + \square_{02} \quad (6)$$

$$W_2 = (\square_{20} - \square_{02})^2 + 4\square_{11} \quad (7)$$

$$W_3 = (\square_{30} + 3\square_{12})^2 + (3\square_{21} + \square_{03})^2 \quad (8)$$

$$W_4 = (\square_{30} - \square_{12})^2 + (\square_{21} - \square_{03})^2 \quad (9)$$

$$W_5 = (\square_{30} - 3\square_{12})(\square_{30} + \square_{12})[(\square_{30} + \square_{12})^2 - 3(\square_{21} + \square_{03})^2] + (3\square_{21} - \square_{03})(\square_{21} + \square_{03}) \quad (10)$$
$$[3(\square_{30} + \square_{12})^2 - (\square_{21} + \square_{03})^2]$$

$$W_6 = (\square_{20} - \square_{02})^2 [(\square_{30} + \square_{12})^2 - (\square_{21} + \square_{03})^2] + 4\square_{11}(\square_{30} + \square_{12})(\square_{21} + \square_{03}) \quad (11)$$

$$W_7 = (3\square_{21} - \square_{03})(\square_{30} + \square_{12})[(\square_{30} + \square_{12})^2 - 3(\square_{21} + \square_{03})^2] - (\square_{30} - 3\square_{12})(\square_{21} + \square_{03})[3(\square_{30} + \square_{12})^2 - (\square_{21} + \square_{03})^2] \quad (12)$$

The first Hu's moment W_1 which highlights moment of inertia around centroid of the image. The W_7, $W_6$is Skew Invariant in nature. W_3, W_4 is rotation invariant moment.

Histogram of Orientation Gradient:

The theoretical concept of HOG revolves around the fact that the object property such as appearance and shape within an image can be described using the distribution of the intensity gradient or the direction of the edge. This technique proceeds by counting the occurrences of gradient orientation in localized portions known as cells of an image. In this method the

weighted vote of the gradient to their respective orientation bin (orientation Histogram) is calculated for a localized area (cells) of an image.

HOG is calculated using the following steps:

Gradient Computation:

The first step is to compute the gradient value for each pixel using some gradient mask. Different types of gradient mask can be applied ranging from Sobel mask to normal 1-Dimensional gradient mask. But it is seen that 1-D gradient masks work the best for this algorithm. This mask is applied in both horizontal as well as in vertical direction.

1-D gradient mask is of the form

$$G_x = [-1, 0, 1] \quad \& G_y = [-1, 0, 1]^T$$

Consider an image F then the x and y derivatives are obtained by convolving the above mask with the image

$$F_x = I * G_x \& F_y = I * G_y \tag{13}$$

Magnitude of the Gradient is given by :

$$| G | = \sqrt{((F_x)^2 + (F_y)^2)} \tag{14}$$

Orientation of the gradient is given by:

$$\square = \arctan (F_y / F_x) \tag{15}$$

Orientation Binning:

For this step we divide the given image F into cells. Then each pixel of a cellcasts a weighted vote to a particular orientation based histogram bin based on the orientation value that is found in the previous step. The histogram channel is evenly distributed over 0 to180 degrees depending on the gradient value that is pre-calculated. This method works best when the orientation histogram channel can be distributed evenly over 9 channels. The weighted vote can be in the form of the gradient magnitude of that pixel.

Descriptor Blocks:

The changes in illumination and contrast is calculated by normalizing the gradient strengths locally. This is done by cells together into larger, spatially connected blocks. These blocks overlap each other such that each cell contributes at least four times to the final descriptor except the cells that form the border of the image. Commonly blocks are used that are made up of four cells. Hereafter the HOG descriptor is formed by concatenating the vector components of each normalized cell histograms that are obtained from allthe block regions.

Block normalization:

Block normalization can be done using any one of the three different methods. Let v be the non-normalized vector containing all histograms in a given block, ||v|| be the sum of square of all the vectors and e be some small constant. Then the normalization factor can be one of the following: (16)

$$L_2 \text{ norm} : f = v / \sqrt{(||v|| + e^2)}$$

$$L_1 \text{ norm} : f = v / (\sqrt{||v|| + e}) \quad (17)$$

$$L_1 \text{ norm} : f = \sqrt{(v / (||v|| + e))} \quad (18)$$

4.3 Clustering

Clustering can be defined as the technique that groups similar data into one cluster and dissimilar to the data that belongs to a different cluster. It is a technique that belongs to the concept of statistical data analysis. Clustering is not an algorithm in itself it is a technique that can be achieved using many algorithms.

Self-Organizing Maps Clustering technique was used for clustering purpose in this project. It is an unsupervised clustering technique and gets its roots from the family of neural network and is based on the theory of competitive learning. When equatedwith other categories of centroid-based clustering this technique is quite different as its goal is to first find a set of centroids and to then assign each object in the data set to their respective centroidswhich provides the best approximation towards that object.

The SOM learning algorithm is centered on nearest neighbor competition and weight adaptation procedure through many iteration steps to reduce the difference between input feature vector (X) and weight vector (W). This is done by calculating the distance between the input vector and the weight vector byusing the Euclidian distance and then selecting the minimum weight vector which corresponds to that input vector.

Euclidian distance which is calculated by the given formula

$$\text{Dist} = (X(t) - W(t))^2 \tag{19}$$

Where t = iteration number, X(t) = input feature vector, W(t) = weight vector

Then that particular weight vector is updated using the following formula.

$$\Delta W_k(t) = W_k(t) + \alpha(X(t) - W_k(t)) \tag{20}$$

where $0 < \alpha < 1$ is a learning rate which decreases with each iteration.

This process continues either till the learning rate reaches zero or till the algorithm converges and clusters are obtained.

4.4 Comparative study of SOM and K-means Clustering method

- SOM algorithm explores the dataset thoroughly before the final clustered data is delivered. Due to this the problem of local minima can be avoided through this algorithm
- Unlike k-means, in SOM each unit will move towards only those units which have the least distance with the same weight vector.
- Also in K-means it is a must that all the clusters should contain some data points even if it may belong to some other cluster. However, in SOM it is not necessary that all the clusters should contain data points. Some clusters can be empty as well.

5. IMPLEMENTATION

5.1 Data Set

The data set consists of commercial vehicles and cars where each dataset consists of approximately 800 images. A commercial vehicle falls in the category of vehicle which are specifically used for carrying goods or paid passengers. These are basically lorries, buses, trucks, etc. Personal car dataset consists of SUV's.

5.2 Algorithm

Algorithm provides a step by step method of solving a given problem. It helps by providing a systematic procedure of solving a problem. It helps in determining the approximate time complexity of solving that particular problem. Also one can determine the feasibility of the path adopted to solve the given problem.

5.2.1 HOG Feature Extraction

Step 1: Input in the form of gray images of size (M x N).

Step 2: Compute the gradient magnitude and the orientation of the gradient using.

$$| G | = \sqrt{((F_x)^2 + (F_y)^2)}$$

$$F_x = I * G_x \& F_y = I * G_y$$

Step 3: Divide the gradient image into cells of (m x n) size

Step 4: For each cell compute the weighted vote of each pixel for each orientation based histogram bin. Number of histogram bins is 9 and orientation ranges from 0 to 180 degrees.

Step 5: Perform Block normalization where each block consists of four cells.

$$L_2 \text{ norm} : f = v / \sqrt{(||v|| + e^2)}$$

Step 6: Concatenate the block features to get HOG feature vector.

5.2.2 Self-Organizing Maps Clustering Algorithm

Step 1: Initialize input parameters like feature vectors X (n training parameters), weight vector W, alpha learning rate α.

Step 2: Continue till learning rate reaches zero ($\alpha = 0$).

Step 3: For i = 0 to n

Step 4: Calculate the Euclidian distance between weight vector and input (feature) parameters.

$$Dist = (X(t) - W(t))^2$$

Step 5: Find the weight with minimum distance.

Step 6: Update the weight vector which is minimum.

$$\Delta W_k(t) = W_k(t) + \alpha(X(t) - W_k(t))$$

Step 7: Decrease the value of alpha α.

6. OUTPUT SCREENS

6.1 Commercial Vehicles Dataset

A commercial vehicle falls in the category of vehicle which are specifically used for carrying goods or paid passengers.

SOM Clustering with Moments Invariance Features

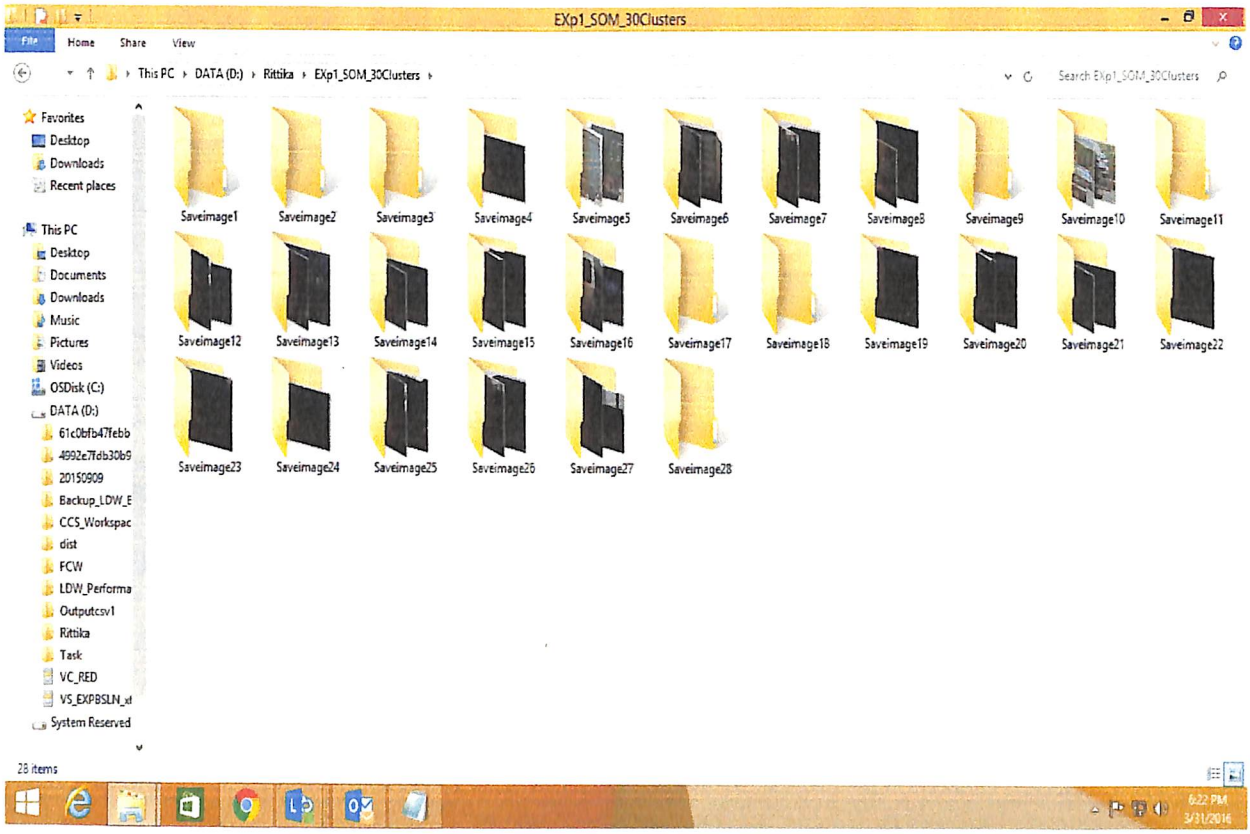


Fig. 6.1. Output Screen SOM clustering and Moments Invariance method.

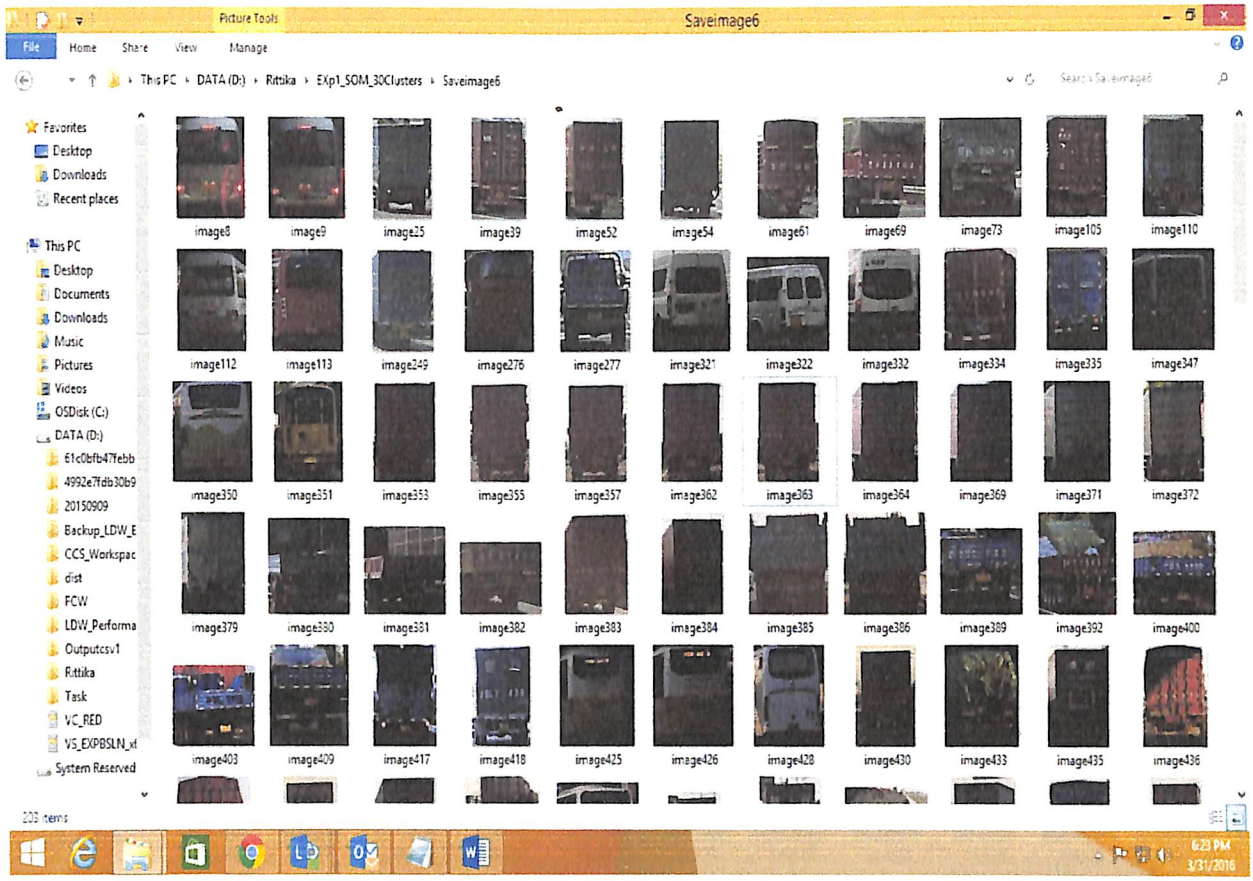


Fig. 6.2 Output Screen SOM clustering the duplicate images in the cluster number 6.

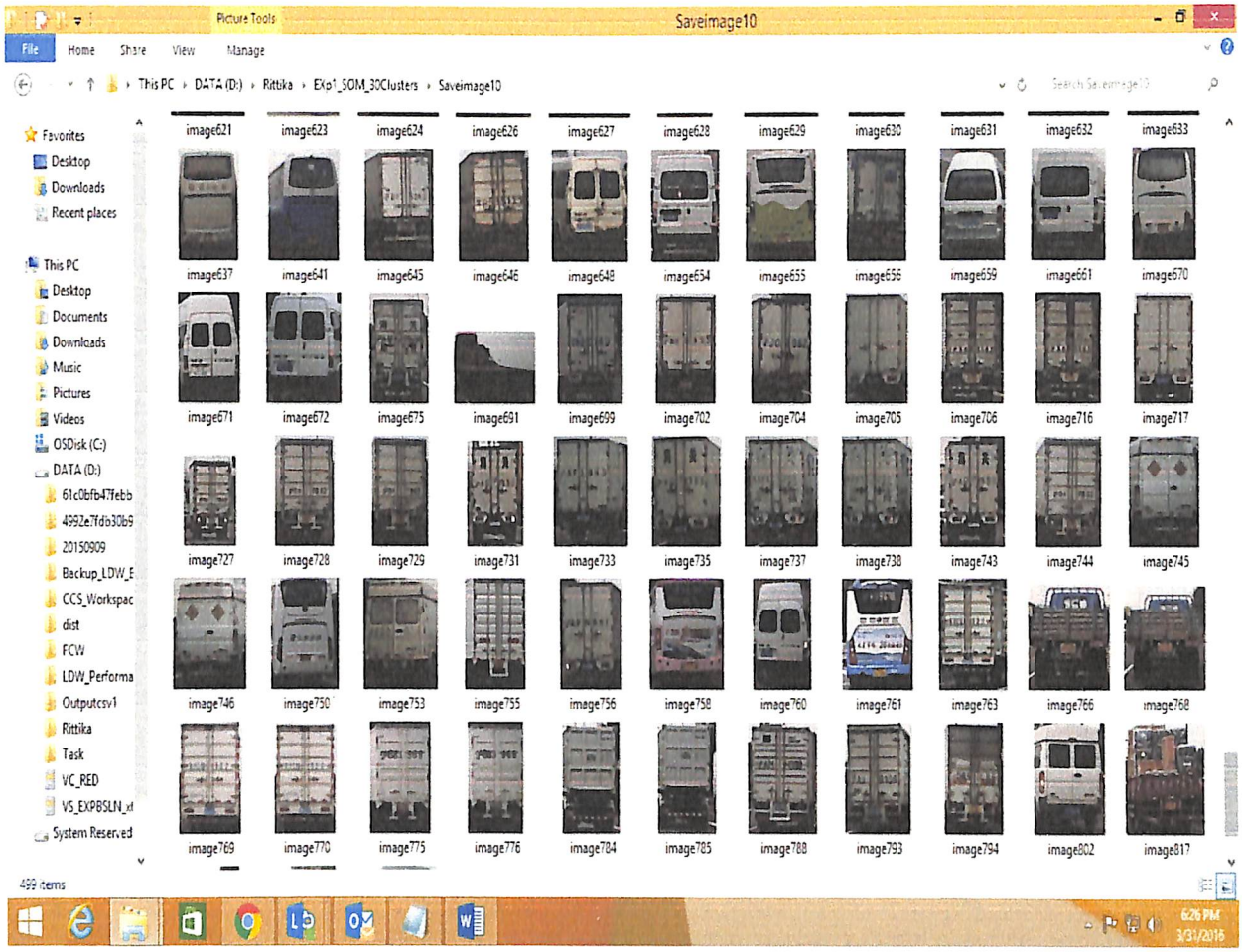


Fig. 6.3 Output Screen of clustered images using SOM clustering duplicate images in the cluster number 10.

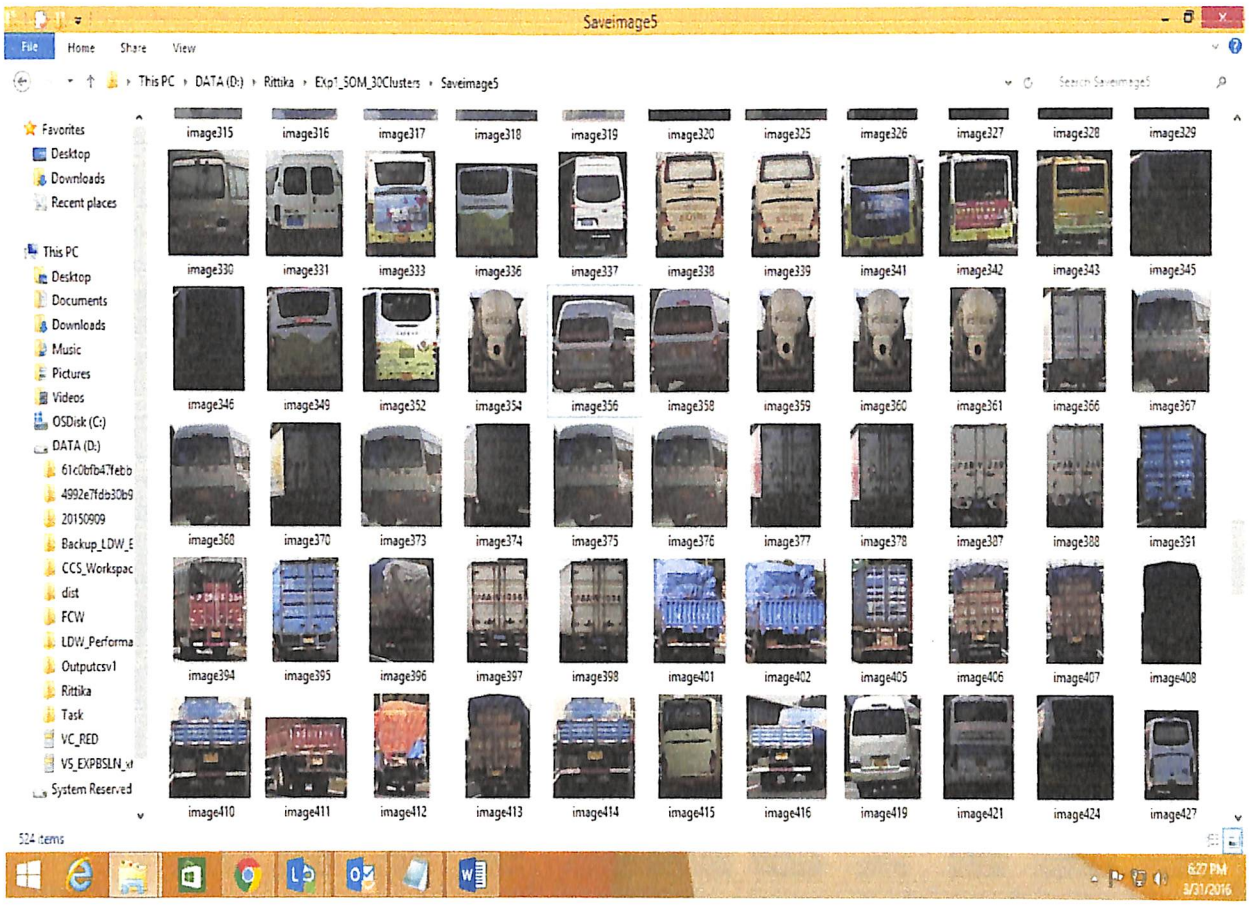


Fig. 6.4 Output Screen of clustered images using SOM clustering the duplicate images in the cluster number 20.

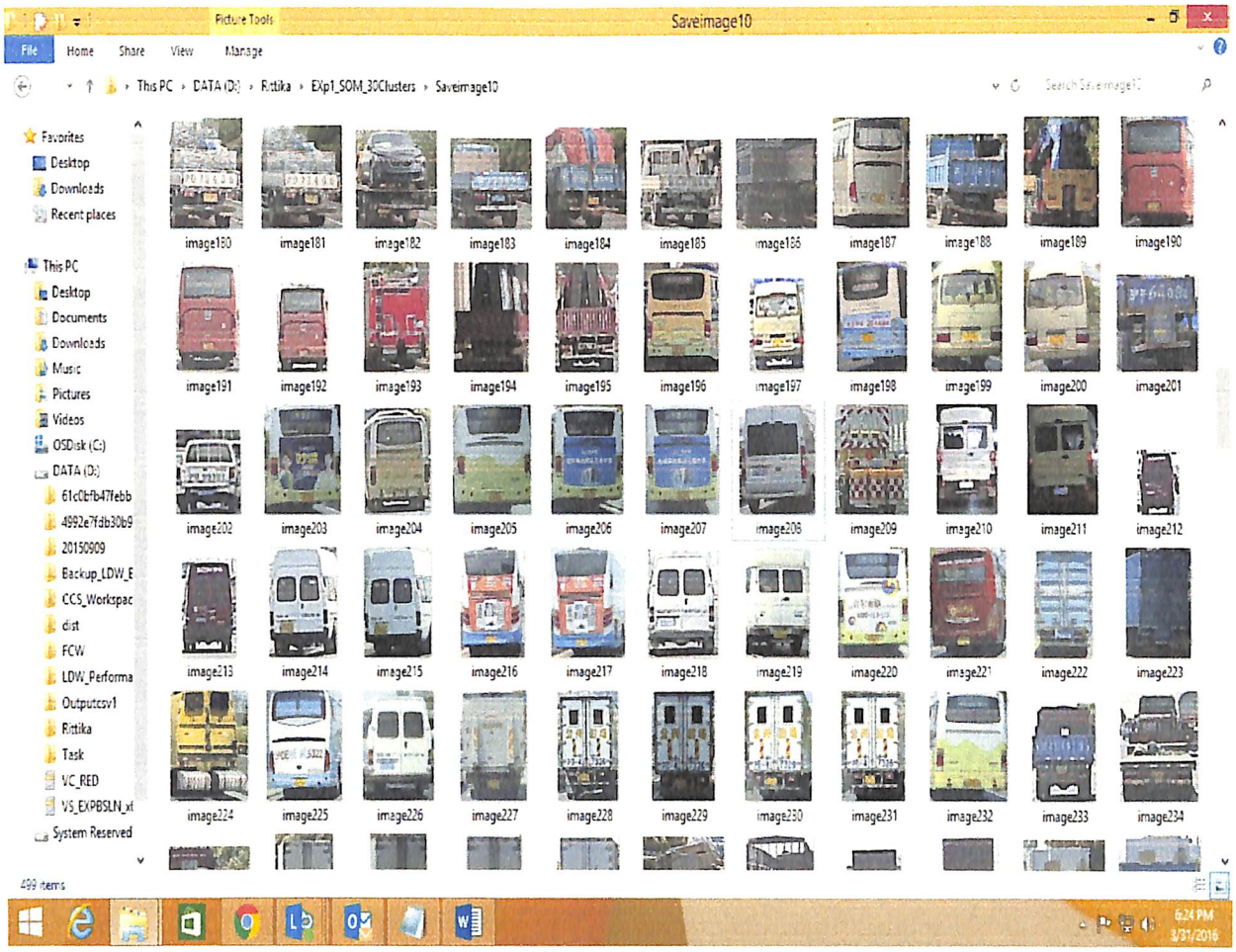


Fig. 6.5 Output Screen of clustered images using SOM clustering the duplicate images in the cluster number 27.

K-Means Clustering with Moments Invariance Features

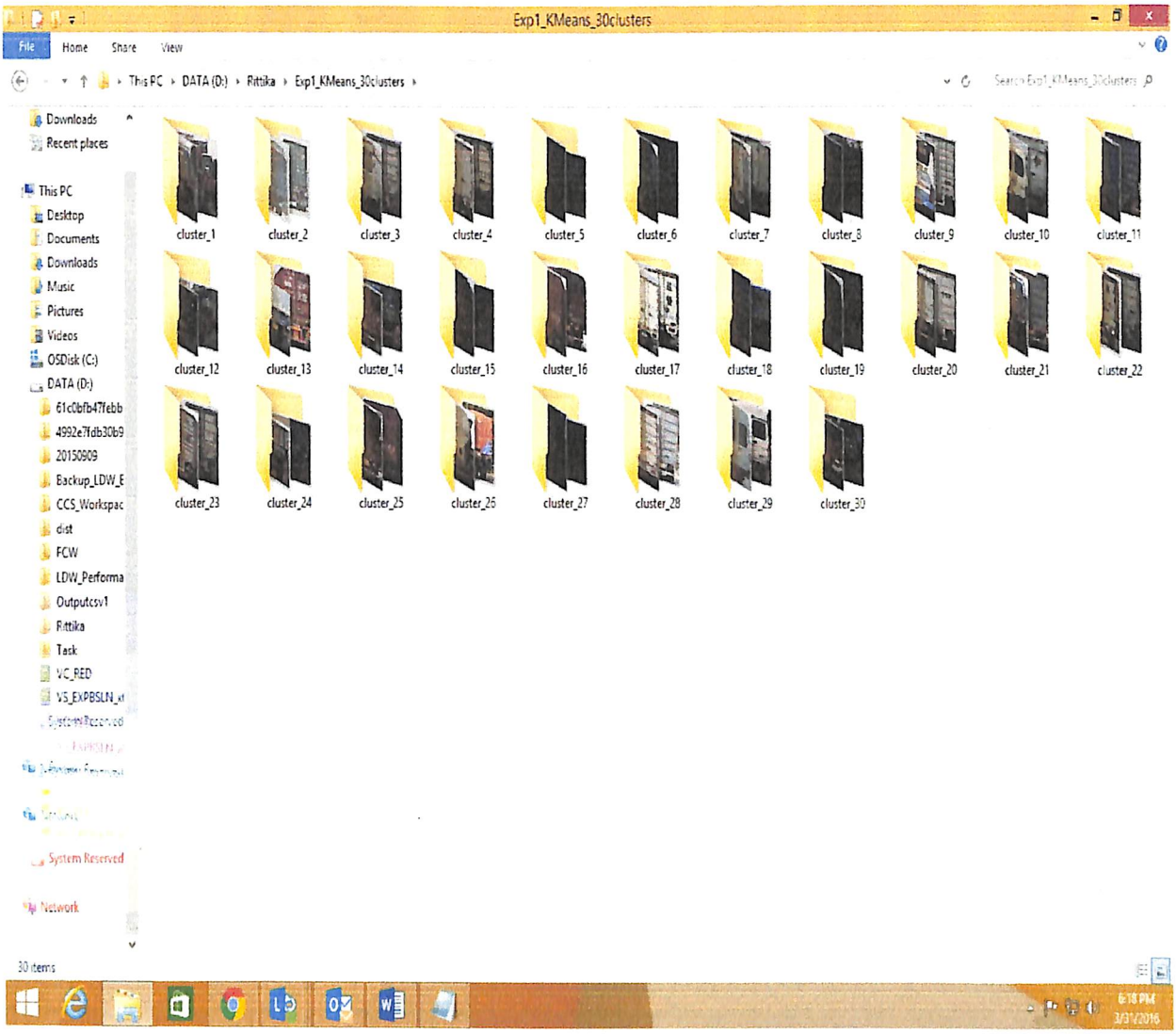


Fig. 6.6 Output Screen of clustered images using K-Means clustering and Moments Invariance feature extraction.

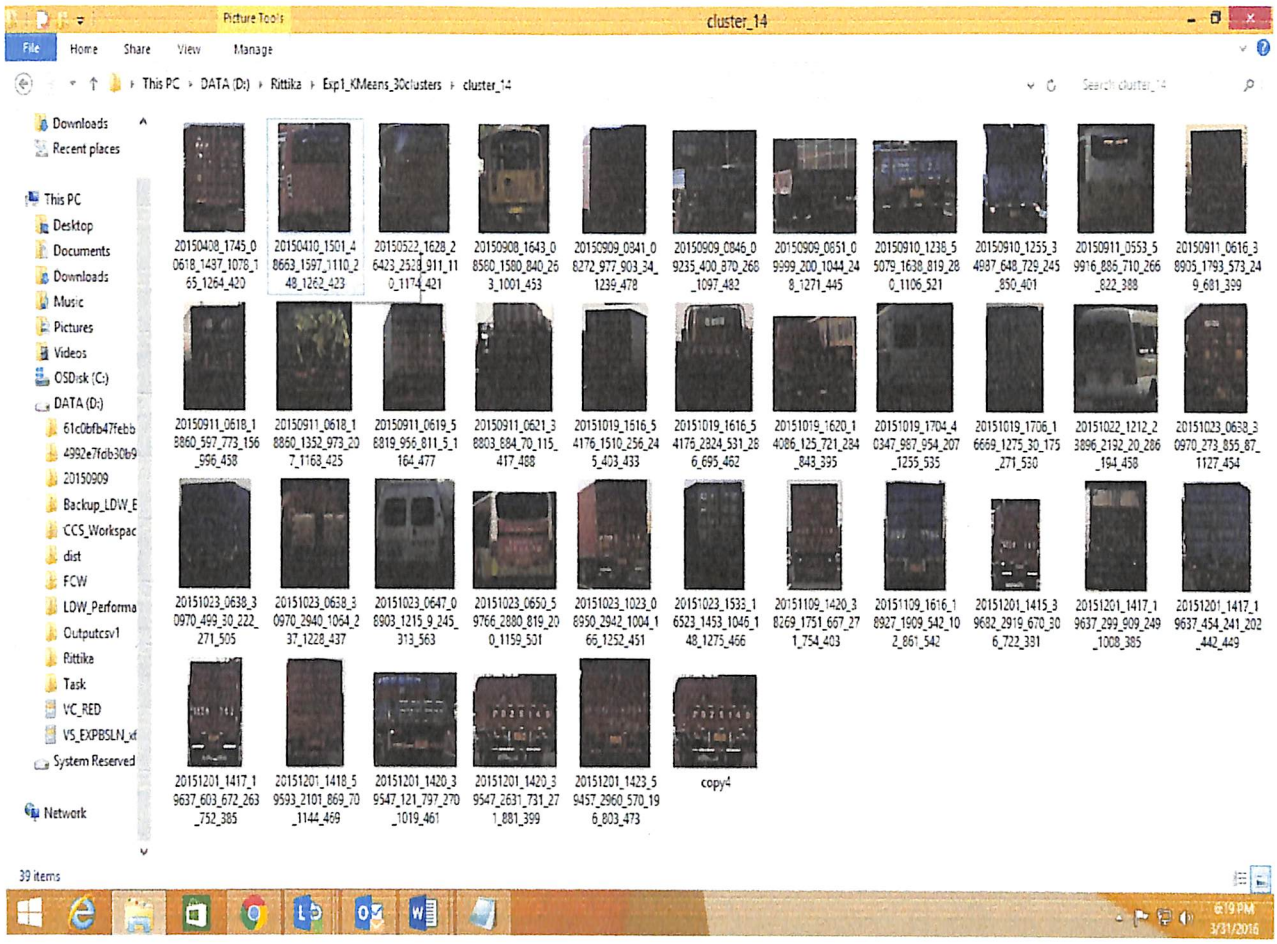


Fig. 6.7 Output Screen of clustered images using K-Means in the cluster number 14.

6.2 Personal Cars Dataset

SOM Clustering with Moments Invariance Features

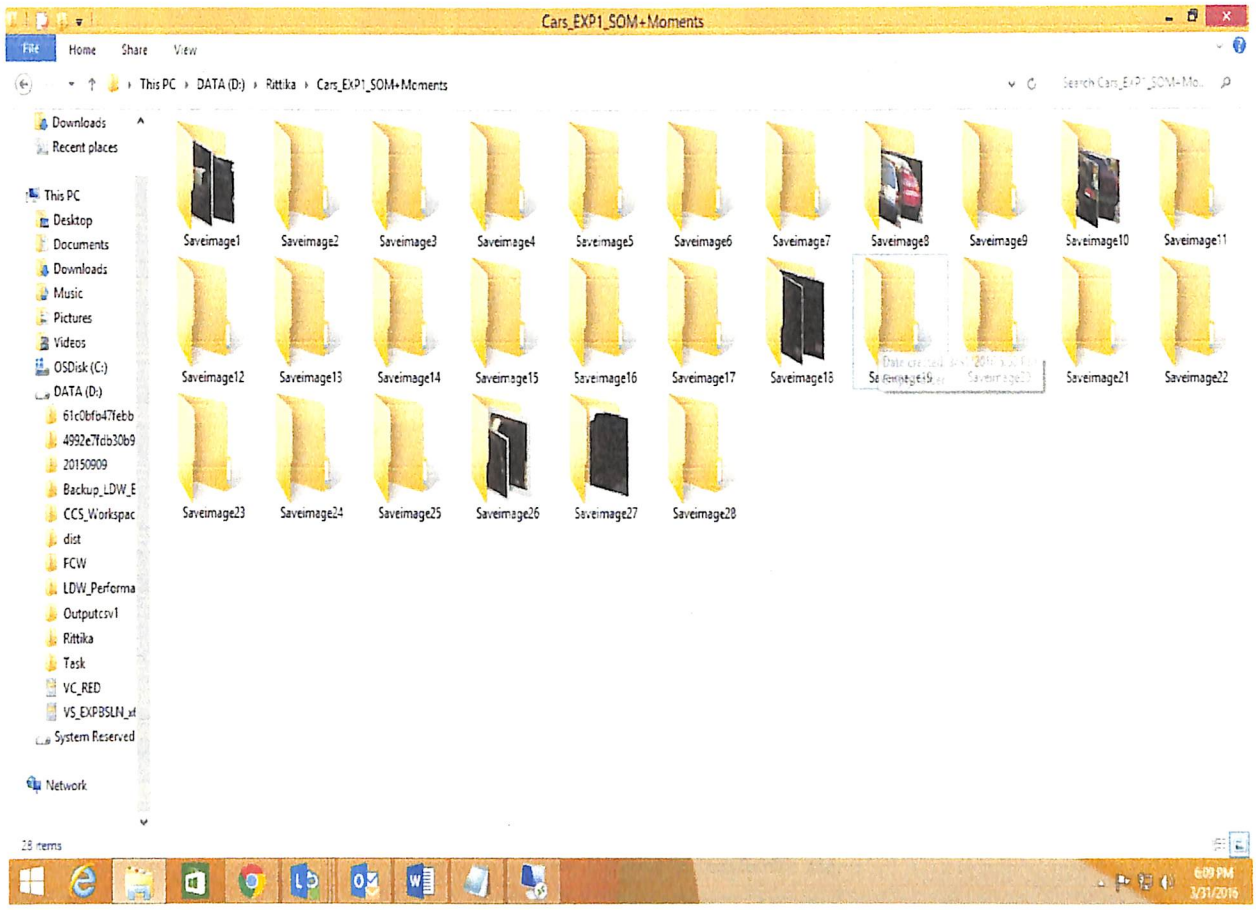


Fig. 6.8 Output Screen of clustered images using SOM clustering and Moments Invariance Feature extraction method.

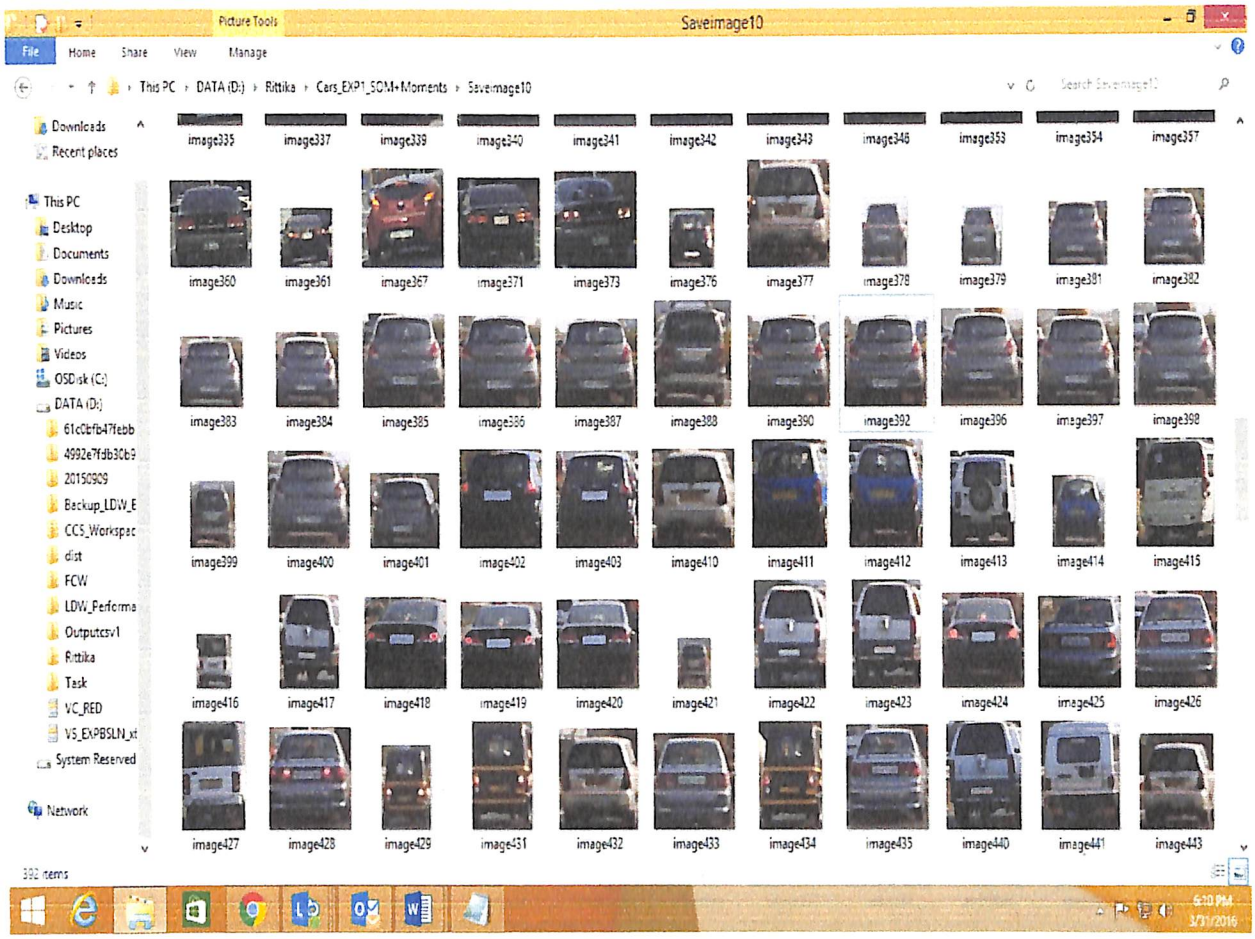


Fig. 6.9 Output Screen of clustered images using SOM clustering the duplicate images in the cluster number 10.

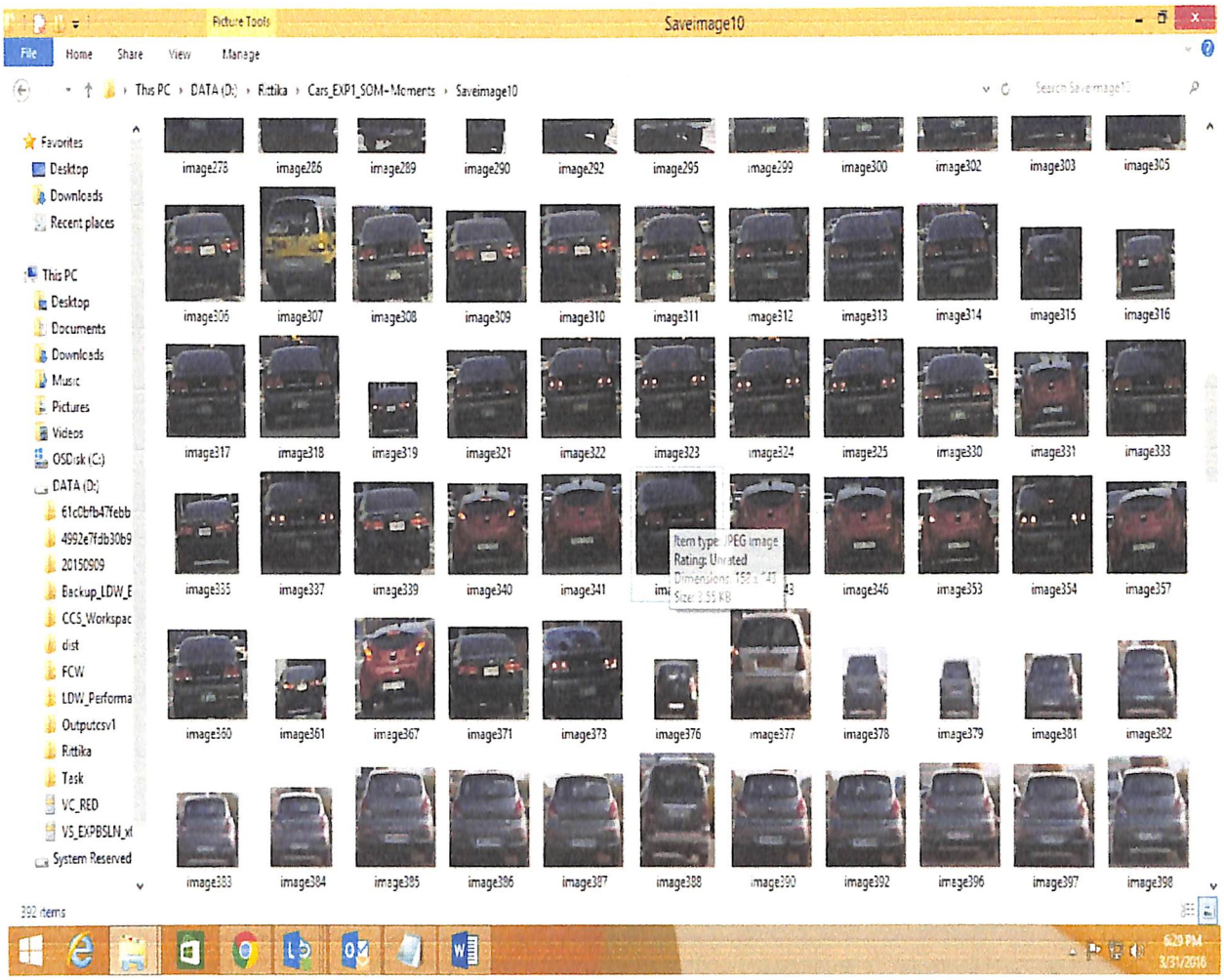


Fig. 6.10 Output Screen of clustered images using SOM clustering the duplicate images in the cluster number 26.

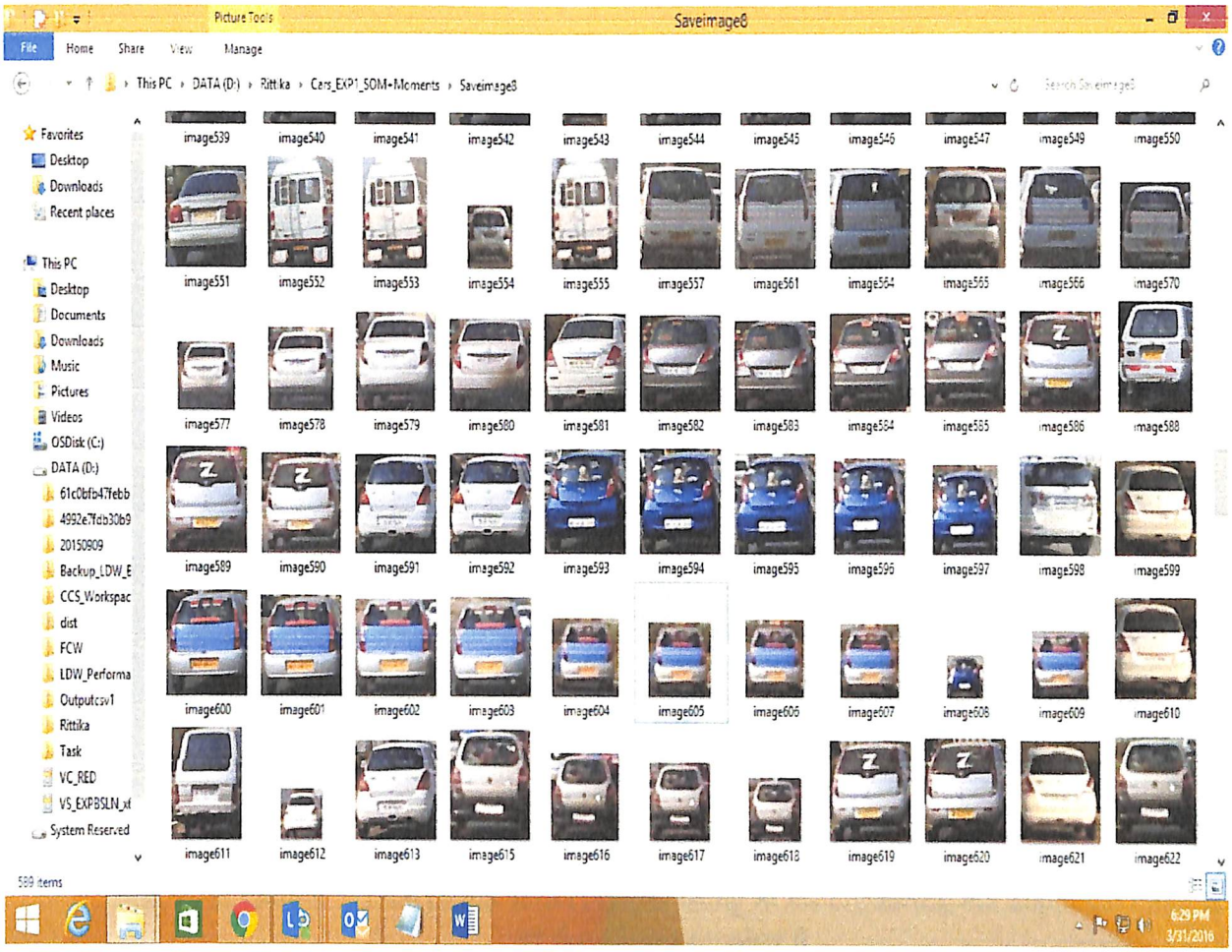


Fig. 6.11 Output Screen of clustered images using SOM clustering the duplicate images in the cluster number 18.

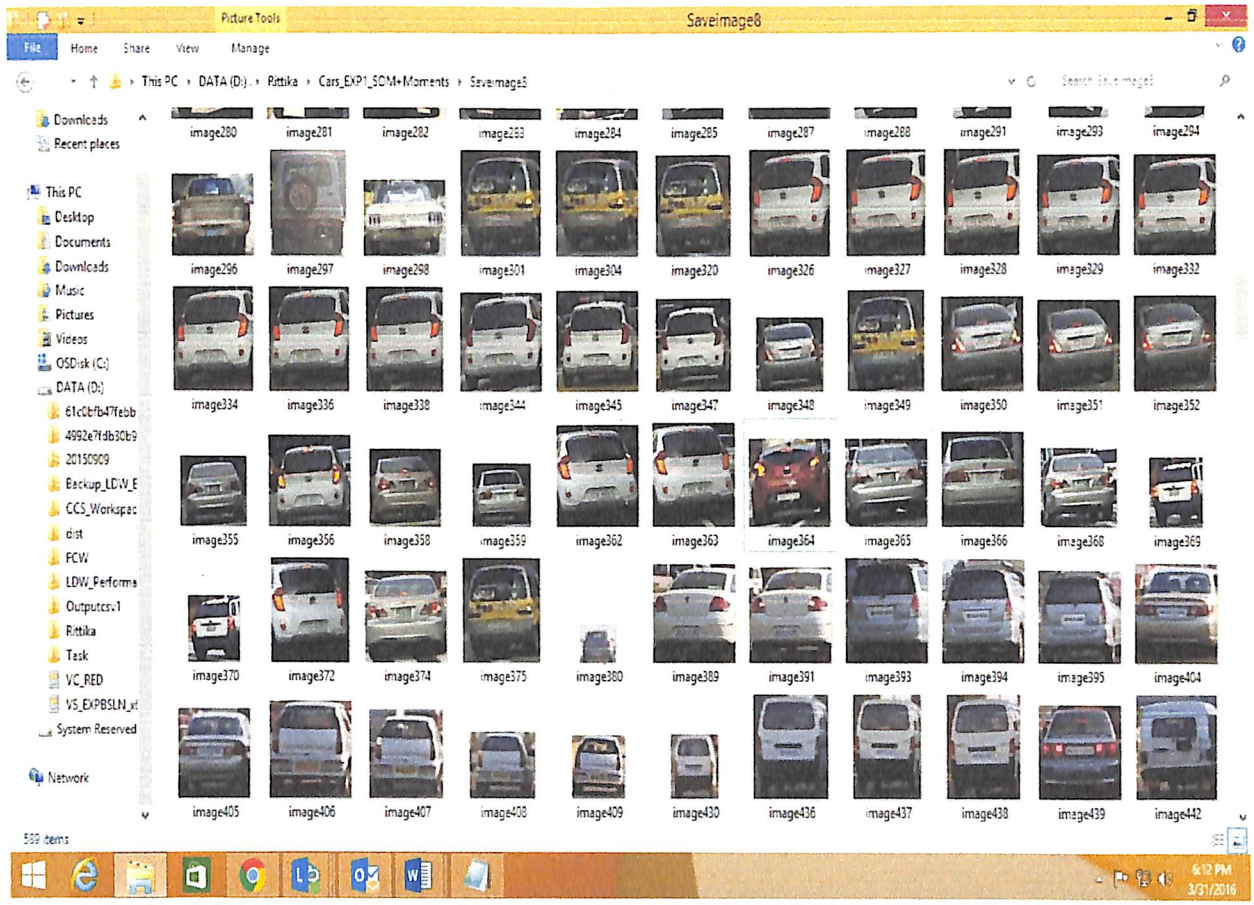


Fig. 6.12 Output Screen of clustered images using SOM clustering the duplicate images in the cluster number 8.

K-Means Clustering with Moments Invariance Features

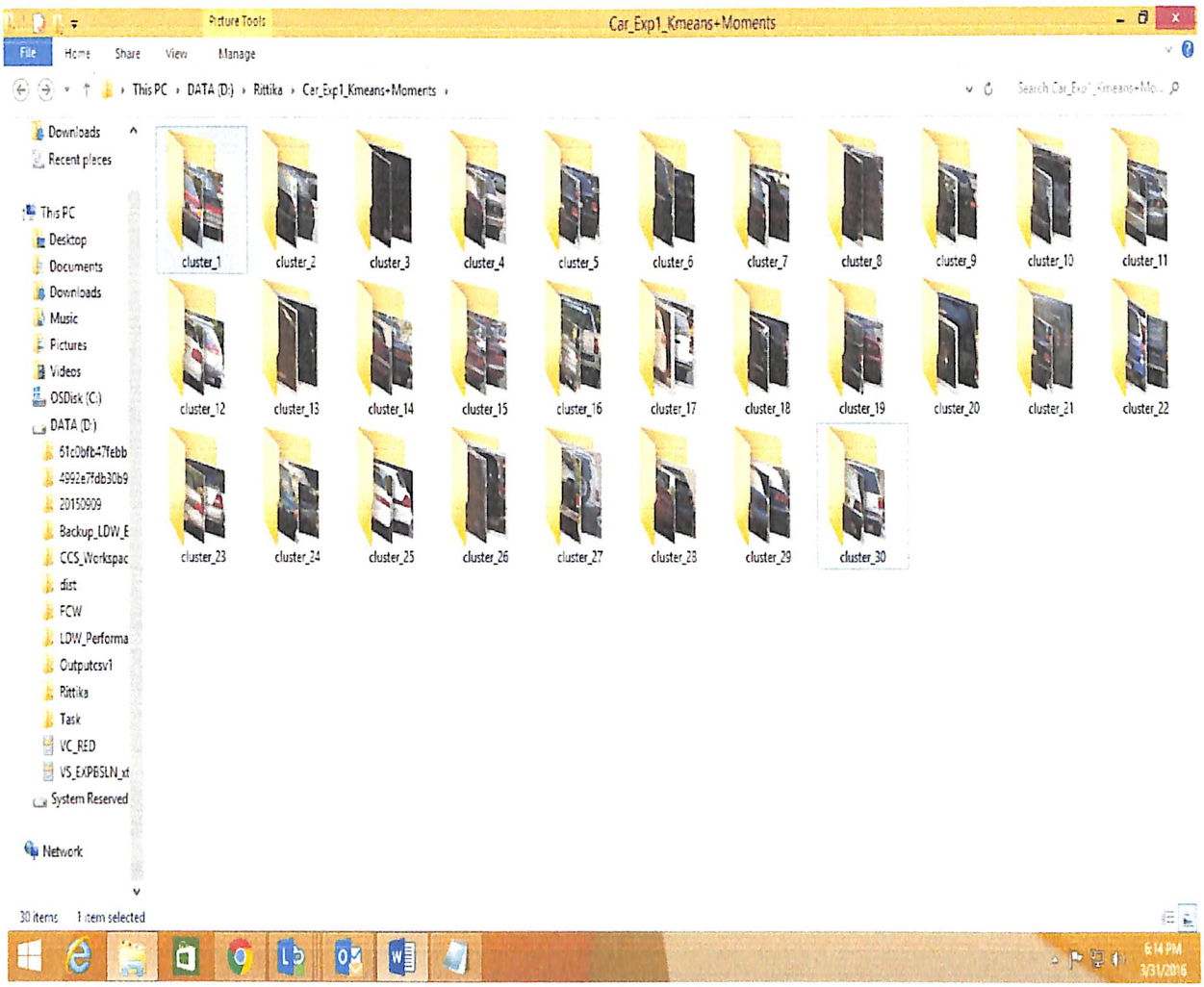


Fig. 6.13 Output Screen of clustered images using K-Means clustering and Moments Invariance feature extraction.

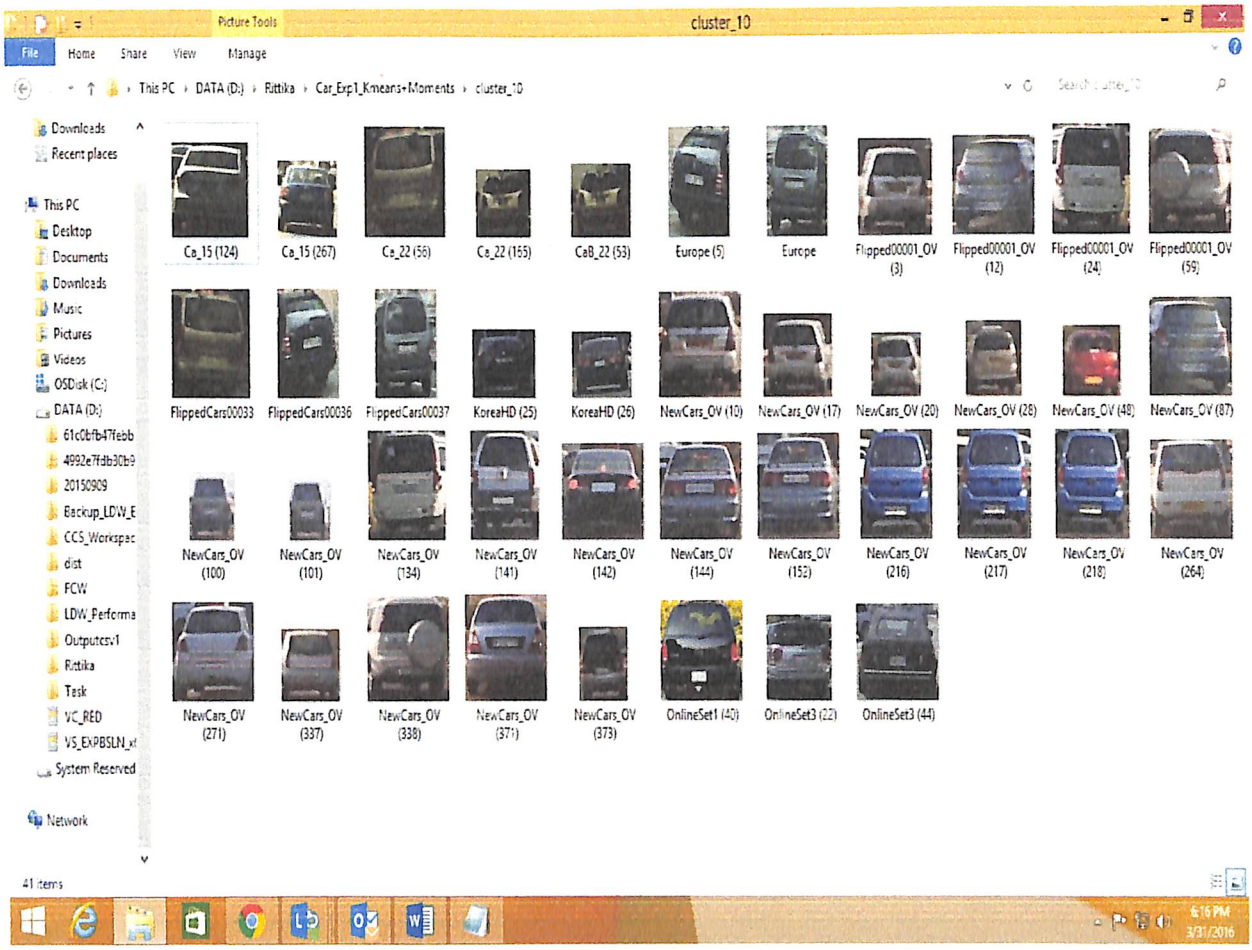


Fig. 6.14. Output Screen of clustered images using K-Means in the cluster number 10.

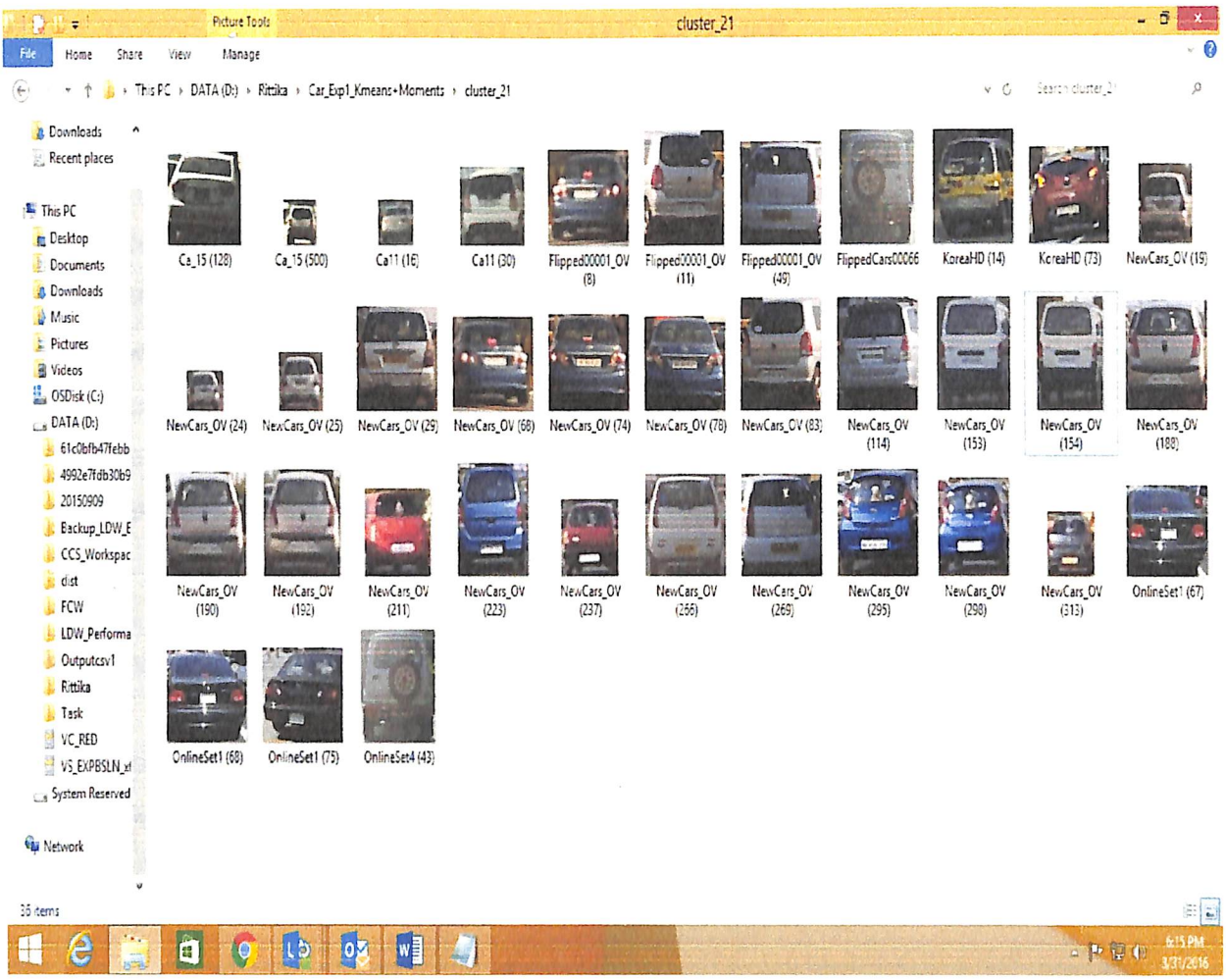


Fig. 6.15. Output Screen of clustered images using K-Means in the cluster number 21

7. LIMITATIONS AND FUTURE SCOPE

7.1 Limitations

- Precise clustering is difficult to achieve as the clustering completely depends on feature extraction method.
- The feature extraction method may differ for different datasets.
- It is difficult to achieve a completely automated system as little human intervention is needed while removing redundant images from the clusters as one needs to make sure that the clustering was properly performed.

7.2 Future Scope

- A completely automated system can be achieved with no amount of human intervention.
- Performing feature extraction such that it is applicable on most of the datasets.

8. CONCLUSION

Supervised learning, in Machine learning is a complex process as the algorithm needs to be trained. For this purpose, a training dataset is used to train the algorithm. So to provide a clean dataset for training purpose, in this study a small experiment was conducted where in the redundancy that occurs in image datasets was detected using clustering technique. Here in this study it was achieved by using moments invariance method and SOM Clustering. At the end clustered images were obtained.

Reference

1. DanasinghAsir Antony Gnana Singh, Subramanian Appavu Alias Balamurugan, Epiphany JebamalarLeavline, **An Unsupervised Feature Selection Algorithm with Feature Ranking for Maximizing Performance of the Classifiers**, International Journal of Automation and Computing, Springer [Volume 12, Issue 5, October 2015].
2. Tian Zhang, Raghu Ramakrishnan, MironLivny, **BIRCH: An Efficient Data Clustering Method for Very Large Databases**, Published in, SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data [Pages 103-114, 1996].
3. Jun Jie Foo, Justin Zobel, Ranjan Sinha, **Clustering Near- Duplicate Images in Large Collections**, Published in Proceeding MIR '07-Proceedings of the International Workshop on Multimedia Information Retrieval [Pages 21-30, 2007].
4. Winn Voravuthikunchai, Bruno Cremilleux, Frederic Jurie, **Finding Groups of Duplicate Images in Very Large Datasets**, Proceedings of the British Machine Vision Conference (BMVC 2012), Confrence Paper [pp.105.1--105.12, September 2012].
5. SanthanaKrishnamachari, Mohamed Abdel-Mottaleb, **Hierarchical clustering algorithm for fast image retrieval**, Proc. SPIE 3656, Storage and Retrieval for Image and Video Databases VII, 427 [December 17, 1998].
6. Yixin Chen, James Z. Wang, Robert Krovetz, **CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning**, Published in IEEE Transactions On Image Processing [Volume:14 , Issue: 8, August 2005].
7. Santosh Kumar Rai, Nishchol Mishra, **DBCSVM: Density Based Clustering Using Support Vector Machines**, Published in International Journal of Computer Science Issues, IJCSI [Volume:9 , Issue:4, July 2012].
8. Ji Zhang, Wynne Hsu, Mong Li Lee, **Image Mining: Issues, Frameworks And Techniques**, 2nd ACM SIGKDD International Workshop on Multimedia Data Mining [2001].
9. R. J. Ramteke, S. C. Mehrotra, **Feature Extraction Based on Moment Invariants for Handwriting Recognition**, Published in Cybernetics and Intelligent Systems, 2006 IEEE Conference [pages 1-6, June 2006].
10. Dian Pratiwi, **The Use of Self Organizing Map Method and Feature Selection in Image Database Classification System**, Published in International Journal of Computer Science Issues [June 2012].
11. Kyung Ah Han, Jong Chan Lee, Chi Jung Hwang, **Image Clustering using Self-organizing feature map with Refinement**, Published in Neural Networks, 1995. Proceedings., IEEE International Conference [Volume 1, December 1995].
12. JuliRejito, RetantyoWardoyo, Sri Hartati, AgusHarjoko, **Optimization CBIR using K-**

Means Clustering for Image Database, Published in International Journal of Computer Science and Information Technologies (IJCSIT) [Volume 3, Issue 4, 2012].

13. Jayashree S. Pillai, Annamma Abraham, **Content based Public Key Watermarking Scheme for Image Verification and Authentication**, Published in International Journal of Computer Applications [Volume: 93, Issue: 2, May 2014].
14. Fernando Bacao, Victor Lobo, Marco Painho, **Self-organizing Maps as Substitutes for K-Means Clustering**, Published in Computational Science, 5th International Conference, Atlanta, GA, USA, Proceedings, Part III- ICCS [pp 476-483, 2005].
15. H. Ming-Kuei, **Visual Pattern Recognition by Moment Invariants**, Published in Information Theory, IRE Transactions [volume. 8, pp. 179-187, 1962].
16. Cho-huakTeh, Roland T. Chin, **On image analysis by the methods of moments**, Published in IEEE Transactions on Pattern Analysis and machine Intelligence [volume. 10, Issue: 04, 1988].
17. NavneetDalal, Bill Triggs, **Histograms of Oriented Gradients for Human Detection**, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. [volume. 1, 2005].