RESEARCH ARTICLE OPEN ACCESS

Adaptive Coal Classification Using Centroid Contour Distance Object Recognition Method Using Deep Learing

K.E.Eswari M.C.A.,M.E.,[1] , K.Dhivyapriya[2]

[1]Associate Processor, Department of Computer Applications, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.

[2]Final MCA, Department of Computer Applications, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.

Email: [1]eswarisaravanan2001@gmail.com, [2]dhivyak508@gmail.com

Abstract. The primary effort in learning coal detail is observing coal features. This project developed a coal classification system which allows researchers to do a search by classification even when they don't know coal name by simply observing their characteristics. The system consists of coal-features, finds the features according to input features, and returns selected clusters. Nowadays, coal classification utilizes machine vision to grab and analyze color, shape, size and surface texture. However, the newly proposed extraction margin method only carries out roughly and there is a difference between the margin of the extracted shape, polygon, and the shape margin of the original image still. This project groups i.e., clusters the coal using image area size, pixel color values similarity, based on image's brightness values and coal shapes. In addition, the study aims in finding the gangue in coal. Based on the pixels count of gangue colors, total gangue percent in the coal is calculated and displayed. This assists in evaluating coal quality. If future, researchers were to expand to other features, coal gangue quantity, etc., even those that are hard to quantify, can also be quantified. ANN is used to classify the coal dataset.

Keywords: Coal Classification, Centroid Contour, Deep Learning, Object Recognition.

I. INTRODUCTION

Today, coal is the most required energy source for modern society. Its has a complex recovery processis, and it can be mixed with considerable amount of Silicon Dioxide Sio2 as coal gangue. The coal gangue's main components are both Al_2o_3 and Sio_2, which are sulfur rich and great quantities of heavy metals like cadmium, arsenic, chromium, copper, etc. Burning coal gangue results in emission of hazardous substances which cause environmental pollution.

In addition, when comparing with coal, the combustion value of coal gangueis lower, which can minimze the total energyfor coal mixed with coal gangue. Therefore, sorting coal gangue from its main coal is an extremely important link and it has two traditional sorting formula/methods: a) human sorting and b) wet cleaning methods.

A sieving machine sorts pure/raw coal into coal equal to or greater than 100 mm and also less than 100 mm; a new transportation system is used to transport the coal from underground to basement ground; and coal = or >=100 mm is transported to sorting workshop where skilled workers can sort coal gangue from its coal according to gray values and texture differences. In addition to these above traditional methods, representative research methods cover radar detection,ray casting, color separation, mechanical vibration, etc., which has good detection methods/properties, but has high requirements also for runtime environments and can impact human health.

With the computer technology development, ImageNet is also combined with convolutional neural networks (CNNs), and deep learning developed rapidly. Compared with old traditional sorting methods, latest object detection algorithms learn from sample images using a CNN, which extract features of coal and coal gangue and has most significant advantages, like high identification speed and high precision.

Recognizing coal from coal gangue is an important working part of the coal industry and is mainly conducted using human sorting at present. Consequently, considerable man power is required, which adds a risk burden to companies and results in poor efficiency. As a main and important branch of artificial intelligence, deep learning has widely applied in many fields, main in machine vision/voice recognition, its performance is improved greatly compared with performances of traditional learning methods, and it has good a transfer learning ability also.

This project proposed an improved ANN algorithm as classic deep learning method for intelligent and highly accurate recognition for coal and coal gangue. Compared to YOLO algorithms, ANN has a better anchor value using cluster analysis application to different data sets, a good anti-interference ability to minimize the impacts from mine dust/shock and acquires more richer detailed information by adding number of layers of the feature pyramid.

a) Auto-inspecting and grading system for machine vision for extracting and analyzing size, color, shape, and surface texture. The proposed extraction margin

method can be carried out roughly and there is still a variatin between both margin of the extracted shape, polygon, and margin of the shape of the original image. Therefore, for improving the method to capture the coal outline, this project proposes a Centroid-Contour distance for capturing outline of the coal features and the distance from center point to each margin point to more accurately quantify coal features of original image.

As a result, that captured image can be consistent with original coal image. Since image recognition technology to quantify three dimensional features, is difficult, and accuracy of the quantified value can not be verified, accuracy of the feature search query is definitely impacted and so cannot be performed. Therefore, this project applied Association Rule method to those coal features that cannot be accurately quantified.

The associating similarity method with association rule analysis effectively improves fault-tolerance and accuracy of overall systematic coal search query. This study tried to develop a new coal classification system with high tolerance and accuracy, as well as prove that Association Rule will be effectually complement the shortcomings of the inability for quantifying features and further improve fault-tolerance and accuracy of the search query system. The system's advantage is that it allows users to easily find gangue information too. Furthermore, science curriculum combines the teaching strategy of inquiry-based learning with this study's coal system for improving elementary school students' coal observation ability. Students can thus pursue learning about coal through an easy approach based on self-observation. In this study, coal research was implemented based on coal features using approach of Centroid-Contour distance (for object recognition) in addition to Fuzzy Function calculation.

However, coal quantification is difficult with various features which leads to biased tolerances. That's why this study has incorporated Association Rule analysis using manual feedback collecting on similarities among coal features, and apply Association Rule for identifying similarity rules by coal as the supplemental calculation for the feature similarity and also increase the overall tolerance and accuracy of system research. In addition, three different methods were applied to identify three different accuracy rates in the study.

Association Rule analysis will not only effectively increase tolerance / accuracy of the system but also help users to find unknown coal based on observed features of the targeted coals. After giving features input, the system would calculate / screen out Top 10 coals with highest similarities to get their names and related information. In this study, a coal search query system was originally developed to do three different analyses of similarity, namely Fuzzy Function, Association Rule, and similarity from combined Fuzzy Function and Association Rule and performed for single feature error conditions and 'n' feature error conditions.

The rest of this paper is organized as follows: Section 2 reviews the existing security approaches under recent studies and explains previous works and their drawbacks. Section 3 provides proposed methodology of the study. Section 4 provides finds and Section 5 is conclusion of the study.

II. LITERATURE REVIEW

In this paper [1] the authors stated that to solve problems of difficult poor feature credibility, feature extraction, and low recognition accuracy of coal and gangue, so this paper utilizes a difference in the coal's dielectric properties and gangue and combining with a SVM (support vector machine) for proposing a recognition method based on dielectric characteristics of coal and gangue. The influence rule of edge effect of the electrode plate on capacitance value was analyzed when the thickness of electrode plate changes. By changing the frequency and voltage of excitation source, curves of dielectric constant of coal and gangue in it versus frequency and voltage are obtained.

Combined with Kalman filter, adaptive noise complete set empirical mode decomposition (CEEMDAN) denoising method was improved, which results in signal with higher signal-to-noise ratio / lower root mean square error after the denoising. An effective value and frequency of denoised response signal were extracted to construct the feature vector sets to form training set / test set. Data of the training set are given as input into SVM for training the intelligent classification model, test set is used for testing the SVM classification effect, and make classification accuracy to 100%.

Unlike those of probabilistic neural network (PNN), intelligent classification model and learning vector quantization (LVQ) NN (neural network) classification model, the recognition / classification accuracies of the three reach 100%, but classification speed of SVM is the most fastest, by taking 0.007916s, which fully reflecting the feasibility as well a efficiency of capacitance method in identifying coal with gangue. Here in this paper, capacitance method and SVM were applied for identifying coal and gangue, accurate and efficient identification results were obtained, which provided a new feasible solution in research on coal gangue identification.

Normally, coal gangue produced will be in large amount during coal mining. Coal gangue is a type of solid waste with low carbon content, accounts for 10%-15% of the raw coal. Its main components are organic molecules, hydrocarbon active while main components of coal gangue are Al2O3 / SiO2. The coal gangue is mixed with coal; it will not only reduce quality of coal combustion, but also increase emission of waste gas. For improving quality of coal combustion and reduction of the emission of poisonous and harmful gases, separation of coal gangue from the raw coal is an noteworthy problem in coal mine/ engineering.

Coal gangue recognition is one of the key technologies of gangue separation. Hou et al. [6-8] analyzed a difference data between coal with its gangue in terms of a) surface texture and b) grayscale characteristics, and combined them with the classification algorithm to recongize coal gangue. Because of coal's texture and grayscale characteristics and its gangue are greatly affected by light, recognition accuracy may be not high.

Liu et al. [9-13] found that the morphological differences between coal / coal gangue was on the basis of studying a) texture and b) gray features, and introduced multifractal for extracting geometric features of coal gangue, however multifractals geometric features' extraction process is more complex and has poor adaptability.

Alfarzaeai et al. [14-18] studied a near infrared spectrum, thermal infrared spectrum and found coa and its gnague's multispectral characteristics, and got high recognition accuracy in laboratory environment using one of the neural network algorithms. Still, this technology was not mature, and it was difficult to apply in practical because of influence of the ambient temperature as well as light. Zhao et al. studied the radiation and attenuation characteristics of X-rays , γ -rays in the coal and coal gangue.

Coal gangues are identified in essence through their attenuation characteristics of two: X-rays and γ-rays; but , radiation produced by the rays will cause physical harm to the workers, and equipment's maintenance cost will also bw high. Wang et al. discussed a method of measuring volume using 3D laser scanning technology, which is combined with dynamic weighing technology for identifying coal gangue. The volume is an estimated value; the measurement error is large relatively.

Yang et al. studied the vibration signals of coal and coal gangue particles colliding with metal plates, and extracted the eigenvalues of the signals in combination with machine learning algorithms for identifying coal gangue, damage identification, reduces the quality of coal. Finding a recognition feature with high reliability, easy extraction and few side effects has become a difficult task in the current recognition of coal and coal gangue. Nelson et al. studied the dielectric properties of pulverized coal, and found that the dielectric constant of pulverized coal decreases regularly with increasing frequency, which provides a reference for studying dielectric properties of coal with coal gangue.

Muhammad et al. as a team conducted research in cutting-edge and pioneering studies in the signal desiccating, signal decomposition and the machine learning, with good references. In this paper, a difference between dielectric properties of coal gangue are studied, and a recognition method of coal and its gangue based on the dielectric properties was proposed. This method realizes nondestructive testing in coal and coal gangue. X-ray and γ-ray identification equipments, have high radiation intensity; so the internal features of coal and gangue that cannot be perceived just by image recognition could be obtained. Here in this study, coal and gangue collection were obtained from Huainan mining area, and SVM intelligent classification models were trained by combining coal and gangue's dielectric constant characteristics with the SVM.

Test results showed that capacitance method had high accuracy and strong timeliness in finding coal and gangue, and had great research prospects. They also concluded that according to differences in the dielectric properties coal and its gangue, dielectric constants of coal and gangue are proposed first as the identification characteristics, which provided a new method for identification of coal and gangue.

Also in this paper, a new capacitance identification method usingcoal and gangue with regular shapes were conceived, and remarkable recognition, classification results were obtained by combining SVM intelligent classification models. The main contributions of that paper were as follows: 1) The capacitance identification model of coal and gangue was established, and finite element simulation analysis of capacitor model was carried out. Influence of the edge effect produced by plate thickness on the calculation of capacitance value was obtained, the calculation formula of the capacitance value was modified for accurately calculating the capacitance value of the capacitor when medium changes.

In this paper [3] the authors stated that bounding box regression is a crucial step in detecting object. In existing methods, while `n-norm loss were adopted widely for bounding box regression, it was not tailored to valuation metric, i.e., IntersectionoverUnion (IoU). Recently, IoU loss and generalized IoU (GIoU) loss have been introduced to benefit IoU metric, however, still suffer from a problem of i) slow convergence and ii) inaccurate regression. The authors proposed in this paper, a Distance-IoU (DIoU) loss by taking the normalized distance between the predicted box and target box, which converges faster in training than a) IoU and b) GIoU losses.

Furthermore, this paper summarized three geometric factors in the bounding box regression,i.e., a) overlap area, b) central point distance and c) aspect ratio, based on which the Complete IoU (CIoU) loss was proposed, thereby leading for faster convergence and better performance. Using DIoU and CIoU losses into the state-of-the-art object detection algorithms, e.g., SSD, YOLO v3 and Faster RCNN, they achieved notable performance gains not only in IoU metric but also GIoU metric. In additin, DIoU could be easily adopted into non-maximum suppression (NMS) for acting as a criterion for further boosting performance improvement.

Object detection is one of the key issues in machine vision tasks, and has received considerable research attention for recent decades (Redmonetal.2016; RedmonandFarhadi 2018; Wang et al. 2019; 2018, Ren et al. 2015; Yang et al. 2018; He et al. 2017). Generally, existing object detection methods are categorized as: one-

stage detection, like YOLOseries (Redmonetal.2016) and SSD (Liu et al. 2016), a two-stage detection, such as R-CNN series (Girshick et al. 2014; Renetal.2015;Heetal.2017), and also even multistage detection, like Cascade R-CNN (Caiand Vasconcelos 2018). Despite of these different detection frameworks, bounding box regression is the most crucial step for predicting a rectangular box to locate the target object.

They concluded that in that paper, that they proposed two losses, i.e., i) DIoU loss and ii) CIoU loss, for bounding box regression with DIoUNMS to suppress redundant detection boxes. Using direct minimization of the normalized distance of two central points, DIoU loss achieved faster convergence than GIoU loss CIoU loss take three geometric properties into account, i.e., i) overlap area, ii) central point distance and iii) aspect ratio, and leads to faster convergence and better performance. The proposed losses and DIoU-NMS could be easily incorporated to any of the object detection pipelines, and could achieve superior results on benchmarks.

In this paper [2-3] the authors stated that lately, maximum modern-day item detection structures undertake anchor field mechanism to simplify the detection model. Neural networks most effective need to regress the mapping members of the family from anchor containers to ground fact containers, then prediction containers can be calculated the usage of records from outputs of networks and default anchor bins.

But, whilst the hassle turns into complicated, the quantity of default anchor boxes will increase with big risk of over-becoming during education. On this paper, they adopted an adaptiveanchorbox mechanism that one anchor box can cover more ground reality containers. So networks only need a few adoptive anchor bins to resolve the same hassle and the version may be more robust. The sizes of adaptive anchor bins might be adjusted mechanically in line with the depth accrued via a Time of Flight (TOF) digital camera.

The network adjusts the aspect ratios of anchor boxes to get final prediction boxes. The experimental consequences exhibit that the proposed method can get more correct detection consequences. Specially, the use of the proposed adaptive anchor field mechanism, the mean common Precision (mAP) of YOLO-v2 and YOLO-v3 networks increases glaringly on open public datasets and their self-built battery photograph dataset. Furthermore, the visible outcomes of prediction comparisons also illustrate that the proposed adaptive anchor container mechanism can obtain better performance than unique anchor box mechanism.

In this paper [4] the authors stated that organizing information into practical groupings is one of the maximum essential modes of knowledge and getting to know. As an instance, a not unusual scheme of clinical classification places organisms into a system of ranked taxa: area, nation, phylum, class, and many others. Cluster evaluation is the formal study of strategies and algorithms for grouping, or clustering, gadgets in keeping with measured or perceived intrinsic traits or similarity. Cluster evaluation does no longer use class labels that tag items with earlier identifiers, i.E., magnificence labels.

The absence of class statistics distinguishes records clustering (unsupervised gaining knowledge of) from classification or discriminant evaluation (supervised mastering). The intention of clustering is to find structure in information and is therefore exploratory in nature. Clustering has a protracted and wealthy history in a selection of clinical fields.

One of the maximum famous and simple clustering algorithms, k-method, was first posted in 1955. In spite of the reality that k-method become proposed over 50 years ago and lots of clustering algorithms were posted due to the fact that then, ok-manner remains extensively used. This speaks to the issue of designing a generalpurpose clustering set of rules and the ill-posed problem of clustering.

The author supplied a brief evaluation of clustering, summarize widely known clustering techniques, talk the important demanding situations and key problems in designing clustering algorithms, and point out some of the emerging and useful studies guidelines, which include semi-supervised clustering, ensemble clustering, simultaneous characteristic choice throughout records clustering and big scale data clustering. Advances in sensing and storage generation and dramatic increase in packages inclusive of net seek, digital imaging, and video surveillance have created many excessive-extent, high-dimensional information units. It's miles estimated that the virtual universe fed on approximately 281 exabytes in 2007, and it's miles projected to be 10 times that size by way of 2011. (One exabyte is ~1018 bytes or a million terabytes) [Gantz, 2008]. Maximum of the records is stored digitally in electronic media, accordingly providing massive potential for the improvement of automatic records analysis, category, and retrieval techniques. Further to the growth in the quantity of statistics, the type of available information (text, image, and video) has additionally extended. Inexpensive digital and video cameras have made to be had big documents of images and motion pictures. The prevalence of RFID tags or transponders because of their low fee and small length has resulted in the deployment of tens of millions of sensors that transmit statistics regularly. E-mails, blogs, transaction facts, and billions of net pages create terabytes of new records every day. A lot of those facts streams are unstructured, adding to the difficulty in analyzing them. The growth in each the extent and the kind of statistics requires advances in method to robotically recognize, system, and summarize the facts.

Information evaluation techniques are extensively categorised into important types [Tukey, 1977]: (a) exploratory and/or descriptive, that means that investigator may now not have pre-specified fashions and/or hypotheses but wants to understand overall characteristics or shape of high dimensional information, and (b) confirmatory or inferential, which means that the investigator desires to

verify the validity of a hypothesis/version or a fixed of assumptions given the available information.

More statistical techniques are proposed now to analyze data, like variance analysis, discriminant analysis, linear regression, principal component analysis, multidimensional scaling, canonical correlation analysis,factor analysis, and cluster analysis to name a few. A useful overview of these are given in [Tabachnick&Fidell, 2007].

In pattern recognition methods, data analysis is dealt with predictive modeling: by giving some training data, an author wants to predict behavior of the unseen test data. This task is referred to as learning. Often, the clear distinction is made among learning problems that are (a) supervised (classification) or (b) unsupervised (clustering), first involving labeled data only (training patterns with well known category labels) but latter involving only unlabeled data [Duda et al. 2001]. Clustering is the more difficult and challenging problem compaed to classification. There is a growing interest in a hybrid setting, called semisupervised learning [Chapelle et al., 2006]; in semi-supervised classification, labels of only a small portion of the training data set are available.

The unlabeled data, instead of being eliminated, are used in learning process. In semi-supervised clustering, instead of specifying class labels, the pair-wise constraints are specified, which is one of the weaker ways of encoding the prior knowledge.

Pair-wise must-link constraint points to the requirement that those two objects should be assigned same cluster label, whereas cluster labels of two objects participating in a cannot-link constraint must be different. Constraints are particularly beneficial in data clustering [Basu et al., 2008, Lange et al., 2005], but precise definitions of underlying clusters were not found. In the searching of good models, anybody would like to add the entire available information, no matter it is unlabeled data, data with constraints, and/or labeled data. Figure 1 illustrates a spectrum of different types in learning problems of interest in machine learning and pattern.

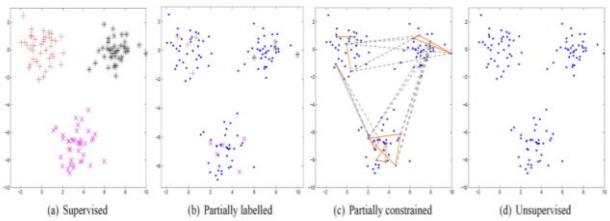

Figure 2.1: Learning problems: dots correspond to points without any labels. Points with labels are denoted by plus signs, asterisks, and crosses. In (c), the must-link and cannot-link constraints are denoted by solid and dashed lines, respectively (figure taken from [Lange et al., 2005]).

Cluster analysis is one of the main goals of data clustering, and is to find natural grouping(s) of group or set of points, patterns, and/or objects. Webster [Merriam-Webster Online Dictionary, 2008] defines cluster analysis as “a classification technique of statistical to discover whether the individuals of the population fall into various groups by making quantitative comparisons in multiple characteristics.”

A clustering example is shown in Figure 2. The main objective is developing an automatic algorithm that discovers the natural groupings (Figure 2 (b)) in unlabeled data (Figure 2 (a)). Another definition of clustering can be stated as follows:

Given a representation of ‘n‘ objects, find ’K‘ groups based on the measure of similarity such that similarities between objects in same group are high while similarities between objects in the different groups are low. But, what is a notion of similarity? What is definition of the cluster? Figure 2 explains that clusters are differ in terms of their size, shape, and density.

The presence of noise in the data makes the detection of the clusters even more difficult. An ideal cluster can be defined as a set of points that is compact and isolated. In reality, a cluster is a subjective entity that is in the eye of the beholder and whose significance and interpretation requires domain knowledge. But, while humans are excellent cluster seekers in two and possibly three dimensions, we need automatic algorithms for high dimensional data.

It is a challenge along with unknown number of clusters for given data that has resulted in the thousands of clustering algorithms which have been published and that continue to appear.

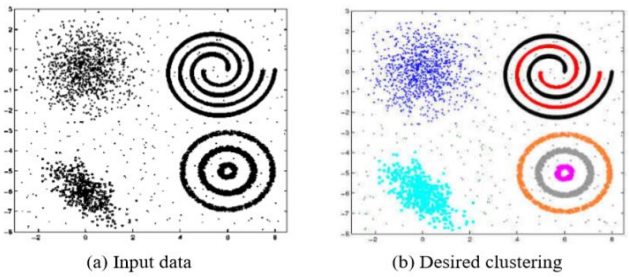

Figure 2: Diversity of clusters. The seven clusters in (a) (denoted by seven different colors in 1(b)) differ in shape, size, and density. Although these clusters are apparent to a data analyst, none of the available clustering algorithms can detect all these clusters.

Clustering algorithms are broadly divided into two groups: a) hierarchical and b) partitional. Hierarchical clustering algorithms find nested clusters recursively either in agglomerative mode starting with each data point in its own cluster and then merging most similar pair of clusters successively to form a cluster hierarchy; otherwise in divisive (topdown) mode starting with all the data points in one cluster and then recursively dividing each cluster into smaller clusters.

Comparing with hierarchical clustering algorithms, the partitional clustering algorithm finds all the clusters simultaneously as the partition of the data and don‘t impose the hierarchical structure. Input to a hierarchical algorithm is an n*n similarity matrix, where ‘n‘ is the number of objects taken for clustereing clustered. But on the other hand, the partitional algorithm could be used either an nxd pattern matrix, where ‘n‘ objects are embedded in the ddimensional feature space, or an nxn similarity matrix. Note that the similarity matrix is easily derived from the pattern matrix, however ordination methods such as MDS (multidimensional scaling) are needed to derive a pattern matrix from a similarity matrix.

The well-known hierarchical algorithms are single-link and complete-link; K-means is the most popular and simplest partitional algorithm. Since partitional algorithms prefer in pattern recognition due to the nature of available data, they coveraged here is focused on these algorithms. Thousands of clustering algorithms are being proposed in literature in various scientific disciplines. This makes it difficult to review the entire published approaches. Still, many clustering methods differ on the choice of objective function, probabilistic generative models, and heuristics. The authors briefly reviewed some of the important major approaches.

Clusters are defined as high density regions in feature space separated by low density regions. Algorithms following this clusters directly search for connected dense regions in feature space. Different algorithms use various definitions of connectedness. Jarvis-Patrick algorithm defines similarity between the pair of points as a number of common neighbors they share, where neighbors are points present in the region of a pre-specified radius around the point [Frank and Todeschini, 1994]. Ester et al. [Ester et al. 1996] proposed a DBSCAN clustering algorithm, which is similar to Jarvis-Patrick algorithm.

It directly tracks for connected dense regions in feature space by estimating density using the Parzen window method. The Jarvis-Patrick algorithm and DBSCAN performane depend on two parameters: a) neighborhood size in terms of distance, and b) the minimum number of points in a neighborhood for its inclusion in a cluster. Moreover, a number of probabilistic models are being developed for data clustering which models the density function by the probabilistic mixture model.

These approaches are being assumed that data are generated from the mixture distribution, in which each cluster is described by more mixture components [McLachlan and Basford, 1987]. The EM algorithm [Dempster et al. 1977] is used to infer parameters in the mixture models.

Several clustering algorithms include the information in theoretic formulation. Take for example, minimum entropy method presented by [Roberts et al. 2001] assumed that a data is generated using the mixture model and each cluster(s) is modeled using the semi-parametric probability density. The parameters are then estimated by maximizing KL-divergence between unconditional density and conditional density of the data points conditioned over the cluster.

This minimizes a overlap among the conditional and unconditional densities, and thereby separating the clusters between each other. In other words, the formulation results in an approach which minimizes the expected entropy of partitions over the observed data.

The information bottleneck method [Tishby et al. 1999] was being proposed as a generalization to rate-distortion theory and adopts the lossy data compression view. In simple words, given the joint distribution over two different random variables, Information Bottleneck compresses one of the available variables while retaining maximum amount of mutual information with respect to other variable. An application of this for document clustering is displayed in [Slonim and Tishby, 2000] where two random variables are the words and the documents. The words are clustered first, such that mutual information with respect to documents is retained maximally, and then using clustered words, documents are clustered such that mutual information between clustered words and clustered documents is retained maximally.

The authors concluded that organizing data into sensible groupings may arise naturally in different scientific fields. It was so not surprising to see the continued popularity of the data clustering. It is important to remember that the cluster analysis is an exploratory tool; output of clustering algorithms only suggest the hypotheses. Numerous clustering algorithms have been published and new ones will continue to appear;but there is no particular clustering algorithm has been shown to dominate eveery other algorithms across the entire application domains. Most algorithms, including simple K-means, are also admissible algorithms.

Since new applications have emerged, it is increasingly clear that task of seeking the best clustering principle may indeed be futile. For example, consider an application domain of the enterprise knowledge management. Given same set of document corpus, the different user groups (e.g., marketing, legal, management, etc) might be interested in generating partitions of the documents based on their respective requirement. The clustering method Which satisfies needs for one group may violate needs of another.

As mentioned earlier "clustering is in eye of beholder" - so indeed data clustering involve the user or the application needs. Clustering has its numerous success stories in the data analysis. In spite of this, the machine learning and the pattern recognition communities need to address for a number of issues to improve their understanding of data clustering. Below is the list of problems and their research directions that are most worth focusing in this regard.

(a) There needs to be the suite of benchmark data (with ground truth) available for research community to

test and then evaluate clustering methods. The benchmark includes data sets from different domains (documents, images, customer transactions, time series, biological sequences, social networks, etc). Benchmark should also include both both static and dynamic data (latter would be useful in analyzing the clusters that change over time), the quantitative and/or the qualitative attributes, linked and non-linked things/objects, etc. Though the idea of providing the benchmark data is not novel (e.g., KDD and UCI ML repository), current benchmarks are now limited to small, static data sets.

(b) They need to achieve the tighter integration between clustering algorithm and application needs. For example, some of the applications may require generating only the few cohesive clusters (the less cohesive clusters can be ignored), while the others may require best partition of entire data. In most of the applications, it may not necessarily be the best clustering algorithm which really matters. But, it is the more crucial to choose right feature extraction method which identifies underlying clustering structure of the data.

(c) Regardless of the objectie (or principle), most clustering methods are eventually made into combinatorial optimization problems which aim to find and analyze the partitioning of data which optimizes the objective. So, as a result, computational issue becomes more critical when an application inviolves large scale data. For instance, finding global optimal solution for K-means is practically NP hard. Hence, it is most important to select clustering principles which lead for computational efficient solutions.

(d) A fundamental issue related to the clustering is its consistency or stability. A good clustering principle shall result in the data partitioning which is stable with respect to the perturbations in data. They need to develop the clustering methods that lead to the stable solutions.

(e) Choosing the clustering principles according to their satisfiability of stated axioms. Despite Kleinberg's theorem of impossibility, various studies have shown that it could be overcomed by relaxing some of the axioms. Thus, it may be one way for evaluating the clustering principle is to decide to what degree will it satisfies the axioms.

(f) Given the clustering inherent difficulty, it makes more sense to the develop semisupervised clustering techniques in which labeled data and (user specified) pairwise constraints are used to decide both (a) data representation and (b) appropriate objective function for data clustering.

III. PROPOSED METHODOLOGY

In existing system, among the coal features, the coal shape feature refers to the marginal outline of the coal. This system groups leaves into various clusters based on Centroid-Contour distance, i.e, sum of distance of all border points from the midpoint to the edges of the coal. In addition, images are clusters based on area, pixel color values and brightness factor.

- The scheme will be not helpful in the diagnosis of coal gauge.
- Extracting gangue features of the cola is not implemented.
- The proposed system detects and classifies the examined gangue with high accuracy.
- Statistical data about the number of gangue and similarity between objects is not made.

Automatic detection of coal gangue is an essential research topic as it may prove benefits in monitoring and thus automatically detect the symptoms of gangue as soon as they appear on coal. The proposed system is a software solution for automatic detection and classification of coal. The developed processing scheme consists, color transformation structure for the input RGB image is created, then the gangue area is detected using specific threshold value followed by segmentation process, the texture statistics are being computed for useful segments, finally the extracted features are grouped as small, medium and big. Process is carried out to find the diseased region using color values and the coal gangue are graded by calculating the quotient of disease spot and coal areas.

IV. FINDINGS

- The proposed scheme will be helpful in the diagnosis of coal gangue.
- The proposed method was successfully applied in the coal image with very high precision.
- Extracting gangue features of the coal is implemented.
- The proposed system detects and classifies the examined gangue with high accuracy.

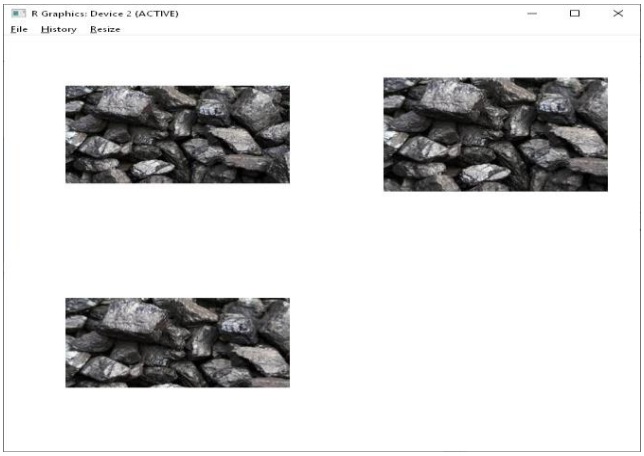

Figure 4.1 Coal image in cluster 1

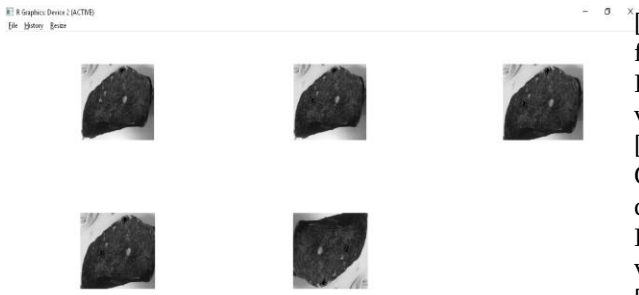

Figure 4.2 Coal images in cluster 2

V. CONCLUSION

This project developed a coal search system that applied some features and cluster them using adaptive approach. The first method used was the Centroid-Contour distance to quantify some features and combine the Fuzzy Functiontheory;thesecondmethodusedwas similarity based on area, brightness values, and pixel color similarities. The method with high efficiency can be applied for other search query systems besides coals and can be easily applied to other systems with the concept of quantifying feature similarities for applications. ANN works better in the given dataset and accuracy percent is above ninety. In future, this project may find the similarity using gangue present in the coal.

REFERENCES

[1] Guo Y, WangX, Wang S, Hu K, Wang W. (2021).“Identificationmethodof coal and coal ganguebased on dielectriccharacteristics,”IEEE Access.,vol. 9, pp. 9845-9854,Jan. 2021

[2] K.E.Eswari, P.Sivanandham “Enhancement of market data partitioning scalability and high diamentionality management using Deep Learning“, Ïnternational Journal of Computer Techniques“, Vol 8 March 2021.

[3] ZhaohuiZheng,Ping Wang,Wei Liu, Jinze Li, RongguangYe,DongweiRen. “Distance-IoULoss: Fasterand BetterLearning for Bounding Box Regression,” Presentedat AAAIConference on Artificial Intelligence 2020.
[Online].Available:https://ojs.aaai.org//index.php/AAAI/article/view/6999.
[3] Gao, Mingyu, et al. “Adaptive anchor box mechanism to improve the accuracy in the object detection system,” Multimedia Tools and Applications., vol. 78, no. 19, pp. 27383-27402, Oct. 2019.
[4] Jain, Anil K. . “Data clustering:50 years beyondK-means.” PatternRecognitionLetters., vol. 31, no. 8, pp. 651-666,June. 2010.
[5] S. Ren, K. He, R. Girshickand J. Sun, “Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence.,vol. 39,no. 6, pp. 1137-1149,June. 2017.
[6] Hou, Wei. Identification of Coal and Gangue by Feed-forward Neural Network Based on Data Analysis [J]. International Journal of Coal Preparation and Utilization., vol. 39, no. 1, pp. 33-43, Jan. 2019.
[7] Dou, Dongyang; Zhou, Deyang; Yang, Jianguo, et al. Coal and gangue recognition under four operating conditions by using image analysis and Relief-SVM [J]. International Journal of Coal Preparation and Utilization., vol. 40, no. 7, pp.473-482, Jul. 2020.
[8] Li, Man; Duan, Yong; He, Xianli, et al. Image positioning and identification method and system for coal and gangue sorting robot [J]. International Journal of Coal Preparation and Utilization., May. 2020. DOI.10.1080/19392699.2020.1760855.

[9] Jian Song. “Research on Identification Algorithm of Coal and Gangue Based on Image Feature,” M.S. thesis, Dept. Computer science and technology., Hebei University of Engineering., Hebei, China, 2016.
[10] Xianmin Ma, Xiaoru ng. “Coal Gangues Automation Selection System Based on ARM Core and CAN Bus, ”Chinese Journal of Scientific Instrument.,vol. 26, no. 8, pp. 305-307, 318, August.2005.
[11] Hailing Zhang, Jialin Wang, Jiansheng Wu, Rong Shi. “Application of Rough Sets Theory to Image Enhancement Processing,” Journal of Tongji University (Natural Science)., vol. 36, no. 2, pp. 254-257, Feb. 2008.
[12] Wanzhi Wang, Zengcai Wang. “Characteristic Analysis and Recognition of Coal and Rock Based on Visual Technology,” Coal Technology., vol. 33, no. 10, pp. 272-274,Oct. 2014.
[13] Yunxia Wu, Yimin Tian. “Methodof coal-rock image feature extraction and recognition based on dictionary learning,” Journal of China Coal Society., vol. 41, no. 12, pp. 3190-3196, Dec. 2016.
[14] J. Redmon, S. Divvala,R. Girshickand A. Farhadi, “You only look once: Unified real-time object detection,” in Proc. CVPR, Las Vegas, NV, United States, 2016, pp. 779-788.
[15] Ping Huo, Hanlin Zeng, KeyanHuo. “Research on density identification system of coal and refuse based on image processing technology, ”Goal Preparation Technology., no. 2, pp. 69-73, Apr.2015.
[16] Guo Y, Wang, X, Wang S, Hu K, Wang W. (2021).“Identification method of coal and coal gangue based on dielectric characteristics,” IEEE Access., vol. 9, pp. 9845-9854,Jan. 2021.
[17] Xijie Dou, Shibo Wang, Yang Xie, Tong Xuan. “Coal and gangue identification based on IMF energy moment and SVM, ” Journal of Vibration and Shock, vol. 39, no. 24, pp. 39-45, Dec. 2020.
[18] Wei Hou. “Identification of Coal and Gangue by Feed-forward Neural Network Based on Data Analysis,” International Journal of Coal Preparation and Utilization, vol. 39, no. 1, pp.33-43, Jan. 2019.