International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April, 2020

# Character Recognition Based on Euclidean Distance Calculation Over Predictive and Centroid Coordinates

**S. Ponmaniraj, Sanjay Sharma, R. Vijay, Gokul Rajan. V**

***Abstract:*** ***Now a day people are living with internet technology but those technologies brings many problems to the people through many hacking techniques. Image spam is the one among them. In the earlier stages, hackers used to annoy targeted victims with their fabricated text called spam text. Hackers are passing their bogus information on many ways such as advertising, spam emails, buttons, query distributions etc. From which spam emails are very specific to attack and they are filtered by text based filter. Then attackers nurtured their attacks on new way i.e., spreading spam mails by images. Those images are non related content to the concerned users on their corresponding mails or any web pages. Because of those spam images, text based filter couldn't identify spam texts. On the basis of an image's features, Attackers used to embed their spam text or mischief coded links into some of the attracted images. To identify spam contents from an image, security functions of a system must be able to recognize the characters imbedding on any images. This research paper is going to present views on image spam, Data mining approaches for dataset analysis, proposed optical character recognizer model and implementation of character recognition from images using Euclidean distance values.***

***Keywords:*** ***Centroid coordinates, Euclidean distance, Glyptic art, Image analysis, Image spam, Optical Character Recognizer, Pattern recognition, Spam analysis.***

## I. INTRODUCTION

Spam contents are irrelevant information to the users who received on their emails and browsed or visited link pages. Some time these spam contents leads to phishing or fraudulent activities to the targeted one [13]. In the earlier stages, attackers used to hack systems or the victim's sensitive data via passing spam based text contents on internet medium. After many researches, Internet authorities found some frequently used spam texts and patterns used to exploit user's data. Based on those patterns and characters they developed few anti spam applications for avoiding transaction vulnerabilities over email/internet [10]. Few well-known patterns or characters used for text based spam are 1) Congratulations, 2) Dear friend, 3) Check your order now, 4) Special promotion, 5) This is not spam, 6) You won and etc.,

This digital era holds many anti spam techniques on server towards protecting connected end devices from spam phenomenon. Recently many filtering techniques are applied to prevent the above mentioned text based spam at the server side [1]. Those modules efficiently detect spam emails based on attractive key spam texts. Server side filters working under classification techniques for the "state of art" on malicious text information. After introduction to these text filter modules and classification techniques the spammers moved on to embedding their spam text/code into attached images and those images takes some illegitimate actions when the user clicks on that images or does any activity on that images. These unwanted images are called spam images.

Emails are bulky and batches to remove tagging as spam and those emails are distributed through some illegal or unauthorized and non reliable sharing servers. If one of the servers is being compromised to perform malfunctions then aggressor starts to penetrate their malicious functions on that particular distributed server [14] without bothering of broadcast links from it. Still many technologies are trying to control and monitoring the sharing information from where they are fetching sensitive data from server. Once security system identified wrongful transactions on any servicing links then they used to update their block lists web sites on server database with those identified system account details to stop their further attacks based on network communications [9].

This paper is going to analyze and implement a method to capture characters from an image for text based filter which works on spam filtering process. Prior steps dwell into abstraction of spam texts a victim needs to understand text abstraction and pattern recognition using data mining for knowledge discovery. In the data mining process, abstraction of data will be done through classifying and refining input queries [2]. Fig.1 shows that the stages of data processing by means of data mining progresses.

Revised Manuscript Received on April 25, 2020.
* Correspondence Author
**S. Ponmaniraj*,** Asst. Professor, School of Computing Science and Engineering, Galgotias University, Greater Noida. Email: ponmaniraj@gmail.com
**Sanjay Sharma,** Asst. Professor, School of Computing Science and Engineering, Galgotias University, Greater Noida. Email: sanjay.sharma@galgotiasuniversity.edu.in
**R. Vijay,** Asst. Professor, School of Computing Science and Engineering, Galgotias University, Greater Noida. Email: r.vijay@galgotiasuniversity.edu.in
**Gokul Rajan V,** Asst. Professor, School of Computing Science and Engineering, Galgotias University, Greater Noida. Email: gokulrajan.v@gmail.com

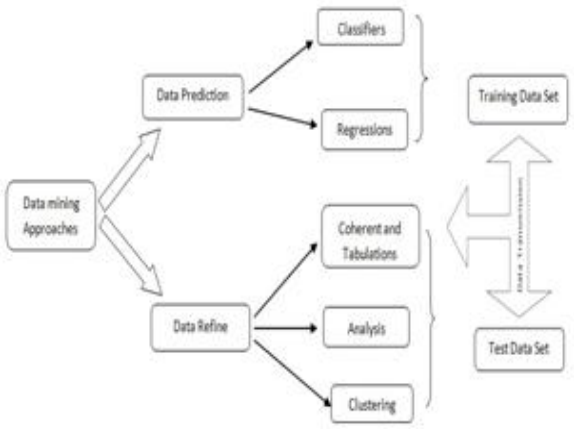


**Fig.1. Data mining approaches for data abstraction**

Data mining process contains the following methods of functions to process the given input datasets [11],

- Prediction
- Classification
- Identification and
- Optimization

## II. OCR MODEL STRUCTURE FOR PROPOSED SYSTEM

To perform data extraction on the given input data sets using KDD, analyzing the multiple dataset clusters is a mandatory function. For detecting texts in image, Support Vector Machine (SVM) method is used along with OCR model. These OCR applications are building with system components and applications for analyzing text which are embedded into images and this is used to build many more computer vision algorithms such as image search, document analysis and text recognition etc., Main phases of this OCR are preprocessing, segmentation, feature extraction, normalization and post processing. Fig.2 outlines the proposed model of OCR for recognizing characters from images.

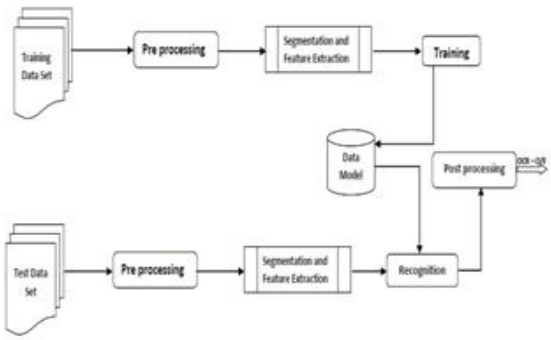


**Fig.2. Proposed model for character recognition from images**

To recognize any characters from a given image is not so easy task, instead many challenges needed to face for better performance. To produce an enhanced reliability of recognition, system has to access high definition and well structured images. The given images are scanned and converted in to grayscale image from RGB to analyze the back ground properties of the image. Segregation of background color and the image color along with it's some other features are playing a vital role in Optical Character Recognition. If the image is not considerable on the basis of low qualities such as device faults, human errors or lighting conditions etc., are causes to bring high false positive rates.

### A. Challenges of reading characters from images

Scene Complexities: Manmade objects captured by camera such as buildings, paintings and symbols are very hard to read the text from non text area of the images. It is a challenging process to read text from the manmade objects. The complexity of dissolved text in images makes OCR to get failed in some images surrounded by paintings and symbols.

Uneven lighting conditions: Less accuracy and segmentation faulting are the result from uneven lighting effects of captured images. Lighting and Shadows are the important parameters distinguish text from images. On some circumstances of lighting effect, the flash will be used but it leads to a new challenge from lighting illumination character.

Slanting: In character recognition, the angle of view is important to read characters. The images taken from various handheld devices may have different directions and it challenges to read and recognize any characters bounded by taken image. To recognize those characters researchers has implemented many innovative ideas such as projection profiles, Rapid Annotation using Subsystem Technology (RAST) algorithm for curled texts [5], Hough Transformation etc.,

Vague Impression: Blurring happens at the time of capturing images from short or long distances and object movement time. Those two categories are obscures the image quality. This blur character affects the image quality to recognize its items properly [6]. During the object movements or capturing any object while user's carelessness is the main reason for resulting to image blurriness. Highly functioned sensors needs to be presented to take each and every frames of a moving object with good accuracy.

Lettering: Different styles fonts and scripts overlapping each characters of a text in the images are some of the reasons to classifying them clearly. Font style like Italy is looking like handwritten contents of texts and this leads to give another challenging task to separation and identification of characters. Segmenting is the main problem for overlapped characters. Pattern and sub spaces of fonts from various classes are very difficult to recognize.

Multilingual: Multi languages are playing vital role in recognition of characters using OCR. Chinese and Japan languages are contains more symbolic like notations on their languages. Hindi and Tamil languages are using more classes to differentiate every character. Identifying those classes from different languages are the major intricacies of text recognition using OCR [7].

Above said challenges are removed by many image processing techniques. In most of the recognition algorithms, Image binarization is the main idea for pre processing of an image before classification [8].

*Retrieval Number: D8733049420/2020©BEIESP*
*DOI: 10.35940/ijeat.D8733.049420*
*Journal Website: www.ijeat.org*



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

### B. Phases of optical character recognition

Reading characters from the image is a vital role in spam image detection. There are three phases used in OCR to recognize texts in images. The high true positive rates for reading texts in any images depends on i) Sharpness in character's border ii) High contrast level iii) Well arranged characters and iv) Less noise in character's pixels [10].

Pre processing: Pre processing the image focuses on rescaling, blurring, averaging and setting of threshold values. In rescaling process the taken image has been scale down to fixed size. So that easily system can process text area and pixels of rest places. Blurring used to remove pixel noises such as salt and pepper noises. Specific range of threshold values must be set for all the regions of an image for better analysis.

Character Recognition: This process is working with glyptic art methodology. This glyptic art contains the 3D projections of the given image features like lines, curves, loops and intersections. These projections will be matching with concerned patterns on pixel by pixel basis. Those methods conjointly known as pattern matching, pattern recognition or image correlations. Now K- Nearest Neighbor (KNN) classifiers are used to hold glyptography arts to find the closest match of the cluster dataset for searching character.

Post processing: An accuracy of the predictive solution is made by lexicon values and lexicons are the squared meanings of a language. This squared mean value yields the good accuracy in result. In OCR, the post processing is finding lexicon of a characters from a given image.

## III. IMPLEMENTATION OF CHARACTER RECOGNITION FROM IMAGE(S)

**Feature generation**, in which 64 dimension vectors created to generate visual features of an image such as gradients of local colors, size, scaling and orientations etc.,

Img=imread('ex1.jpg'); [a b] = size(Img)

K = imresize(Img, [256,256]); [a b] = size(K)

**Calculate** relative predicted and ground truth coordinates of an anchor place for bounding box (Vertical axis),

$$v_c = (c_y - c_y^a)/h^a,\ v_h = \log(h/h^a) \quad (1)$$

$$v_c^* = (c_y^* - c_y^a)/h^a, v_h^* = \log(h^*/h^a) \quad (2)$$

Where,

$\{v_c, v_h\}$ and $\{v_c^*, v_h^*\}$ are relative predict and ground truth coordinators.

$\{C_y^a, h^a\}$ are center and height of the anchor box (Y-axis)

**Predict horizontal axis** or horizontal proposals for bounding box's relative offset,

$$Rel(O) = (x_{side} - c_x^a)/w^a \quad (3)$$

$$Rel(O)^* = (x_{side}{}^* - c_x^a)/w^a \quad (4)$$

Where,

$x_{side}$, $x_{side}{}^*$ are horizantal nearest next horizontal coordinates and pre computed ground truth bounding box anchor location. $c_x^a$ is a center of x-axis anchor location.

**Calculate correlation coefficient** for identifying symbols in an image,

$$Co(r) = \frac{\sum_m \sum_n (Img_{mn} - \overline{Img})(Temp_{mn} - \overline{Temp})}{\sqrt{\sum_m \sum_n (Img_{mn} - \overline{Img}) \sum_m \sum_n (Temp_{mn} - \overline{Temp})}} \quad (5)$$

Where,

Imgmn = Input image coordinates

Tempmn = Template coordinates for the existing Symbols

**Create nearby cluster** visual word data sets by using K-NN classifiers,

Test_set = knnAlg(img_test_data, img_train_data,img_train_set,k);

**Create histogram** for creating centroids of the words. Then need to find Euclidean distance between descriptors and centroids. Numbers of centroids are used to find the frequencies of words.

func [vector levels] = hierarchicalCentroidn(img, depth,ImgplotFlag)

Centroid coordinates for an object of a bounding image or box is denoted by,

$$C_x = \sum \frac{C_{ix} A_{ix}}{A_{ix}}, \quad C_y = \sum \frac{C_{iy} A_{iy}}{A_{iy}} \quad (6)$$

It will be simply denoted as follows,

$$C(\overline{x,y}) = \frac{Total\ movements\ in\ (x,y) directions}{Total\ area} = \frac{1}{A}\int_a^b \int_c^d xf(x)yf(y)dxdy \quad (7)$$

**Calculate K means** for clusters to identify average distance between trained and test data patterns (a,b) via Euclidean distance($d_E$), which is defined as,

$$d_E(\vec{a},\vec{b}) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (8)$$

**Calculate dissimilarity** between objects and cluster means by means of variations.

$$Cluster_{variation} = \frac{1}{m_k}\sum_{i=1}^{m_k} \left\| x^i - \mu_c k \right\|^2 \quad (9)$$

Where,

$x^i$=Data point value

$\mu_c$k=Centroid of the cluster sets

$m_k$=Previous step value

**Assign cluster** object which means is closest to the near object($m_k$) step.

$$m_k = 2\sum_{i=1}^{m} \omega_{ik}(x^i - \mu_k) \quad (10)$$

Where,

$m_k$= Cluster assignment update (M-Step)

$\mu_k$=Centroid of the cluster(x)

$\omega_{ik}$=data point belongs to cluster(x)

**Recalculate dissimilarity** accuracy associated with ($m_k$) step.

$$Re(m_k) = \frac{\sum_{i=1}^{m} \omega_{ik} x^i}{\sum_{i=1}^{m} \omega_{ik}} \quad (11)$$

## IV. RESULT AND DISCUSSION

Many methods are there to recognize character from images by distance calculations. It is found that, among those methods Euclidean distance calculation using predictive and centroid coordinates is giving efficient throughput on optimized latency time duration.

This method provides more than 80% of accuracy values with many datasets and still looking for its improvements from all aspects so that it reaches utmost 99% of accuracy. Table – I, shows that the comparison of multiple distance calculation methods and average values of precision, recall and F-score on different image's dataset.


**Fig.3.Sample output – Characters with bounding box**

**Table – I. Comparison of Distance Calculation Methods**

| Distance Calculation Methods | Average Computation Time | Precision | Recall | F -Score |
|---|---|---|---|---|
| Euclidean distance | 0.0186 | 80 | 81 | 80 |
| City block distance | 0.0189 | 75 | 78 | 77 |
| Canberra distance | 0.0193 | 73 | 78 | 75 |
| Sum of Squared of Absolute Differences (SSAD) | 0.0191 | 71 | 77 | 74 |
| Maximum value distance | 0.0191 | 70 | 76 | 73 |
| Minkowski distance | 0.0192 | 61 | 75 | 67 |
| Sum of absolute difference (SAD) | 0.019 | 61 | 70 | 65 |
| Average | 0.019 | 70 | 76 | 73 |

Fig.4, shows that the average time taken for comparing test image dataset with the existed cluster datasets for character identifications using various methods of distance calculations.

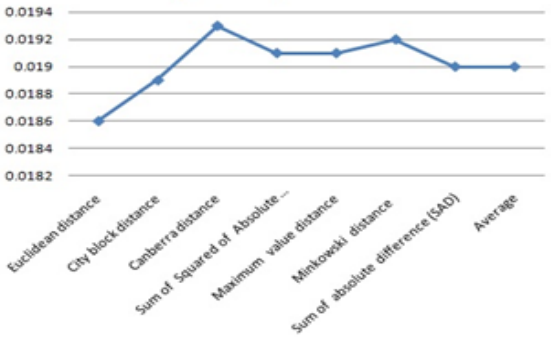


**Fig.4.Comparision of Distance algorithms and Time duration**

## V. CONCLUSION

Character recognition from an image is not as much as easy task because of its quality and feature extractions. Still many researches are going on this topic with help of many new tools which contains more features and components for easy accessing libraries. The main idea behind this proposed model for recognizing text contents from an image is for identifying spam contents or the spam images to protect targeted victims sensitive data and systems against intrusion types of attacks via security breaches and port exploitation over network communications. In future spam (image) based URL navigation will be monitored and controlled by analyzing the image features and embedded contents.

## REFERENCES


1. A.Androutsopoulos, J. Koutsias, K.V. Cbandrinos and C.D. Spyropoulos, An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In Proceedings of the 23rd ACM International Conference on Research and Developments in Information Retrieval, pages 160–167, Athens, Greece, 2000.
2. Harjot Kaur* , Er. Prince Verma, Survey On E-Mail Spam Detection Using Supervised Approach With Feature Selection, ISSN: 2277-9655 [Kaur* et al., 6(4): April, 2017] Impact Factor: 4.116 IC™ Value: 3.00 CODEN: IJESS7.
3. Karez Abdulwahhab Hamad, Mehmet Kaya, A Detailed Analysis of Optical Character Recognition Technology, September 2016, ISSN: 2147-82282, DOI: 10.18100/ijamec.270374.
4. Ye Q, Doermann D. Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence. 2015 Jul 1;37(7):1480-500.
5. Christoph H. Lampert, Thomas Breuel,Document Image Dewarping using Robust Estimation of Curled Text Lines, DOI: 10.1109/ICDAR.2005.90 · Source: DBLP.
6. Jain A, Dubey A, Gupta R, Jain N, Tripathi P. Fundamental challenges to mobile based ocr. vol. 2013 May;2:86-101.
7. Smith R, Antonova D, Lee DS. Adapting the Tesseract open source OCR engine for multilingual OCR. In Proceedings of the International Workshop on Multilingual OCR 2009 Jul 25 (p. 1). ACM.
8. Maninder Kaur , Miss. Manjeet Kaur, A Brief Review on Optical Character Recognition Techniques, IJCSMC, Vol. 6, Issue. 2, February 2017, pg.95 – 100, ISSN 2320–088X.
9. https://www.qualtrics.com/blog/how-to-keep-your-surveys-out-of-the-spam-folder/
10. Karez Abdulwahhab Hamad, Mehmet Kaya, A Detailed Analysis of Optical Character Recognition Technology, IJAMEC, ISSN: 2147-82282, 2016.
11. Ong, Veronica, and Derwin Suhartono. "Using K-Nearest Neighbor in Optical Character recognition", ComTech, vol. 7, no. 1, 2016, pp. 53-65.
12. D. K. Patel, T. Som, S. K. Yadav, and M. K. Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric," Journal of Signal and Information Processing, vol. 3, no. 2, pp. 208-214, 2012.
13. Nawaf Hazim, Mohanad Hazim Nsaif Al-Mayyahi, Mohammed Faiz Aboalmaaly, An Efficient Character Recognition Technique Using K-Nearest Neighbor Classifier, IJET, 7 (4) (2018) 3148-3153
14. https://www.saedsayad.com/k_nearest_ neighbors.htm
15. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and- drawbacks-aa03e644b48a
16. Fazal Malik, Baharum Baharudin, Analysis of distance metrics in content-based imageretrieval using statistical quantized histogramtexture features in the DCT domain, JKSUCIS, 2013, 207-218.


## AUTHORS PROFILE

**S. Ponmaniraj**, is an Assistant Professor in School of Computing Science and Engineering at Galgotias University, Greater Noida. He got his Master Degree from CSE at Sri Muthukumaran Institute of Technology, affiliated to Anna University, Chennai. Currently he is perusing Ph.D. in the area of "Network Security" for the concept of "Intrusion Detection Mechanism Based on Web Images and Sub Links" under the guidance of Prof. Dr. Tapas Kumar. He has 12 years of academic experiences from various educational institutions.

**Mr Sanjay Sharma** completed his Bachelor of Engineering in the stream of Electronics from Bangalore University, Bangalore, (Karnataka) in year 1995 and Master of technology in Computer Science and Engineering from Punjabi University, Patiala (Punjab) in the year 1999. He worked as Executive Engineer for around three years with Punwire Mobile Communications Limited and later was involved for around eight years in software and hardware development. in the year 2008 he came intoeducation for sharing his experience in both Software and Hardware development with new comers in the field of engineering and technology. Till now he has worked with couple of Universities and institutes and currently he is associated with School of Computing Sciences and Engineering, Galgotias University, Greater Noida, (Uttar Pradesh) as Assistant Professor since 2013. His area of interest are Networks, Wireless Sensor Networks and IOT.

**Mr. Vijay Ramalingam** pursed Bachelor ofEngineering in the stream of Computer Science and Engineering from Anna University Chennai Tamil Nadu in 2010. Master of Engineering in Computer Science andEngineering from Annamalai University Chidambaram Tamil Nadu in 2014. He is currently pursuing Ph.D. in Annamalai University Chidambaram. He is currently working as Assistant Professor in School of Computing Science and Engineering, GalgotiasUniversity, Greater Noida, Uttar Pradesh since 2018. He has published 1 research papers in reputed international journals and it's also available online. His main research work focuses on Image Processing, Biometrics, Image Segmentation, Feature Extraction. He has 2 years of teaching experience.

**Mr Gokul Rajan V** pursed Bachelor of Engineering in the stream of Computer Science and Engineering from Anna University of Coimbatore, Tamil Nadu in 2011 and Master Engineering in Computer Science and Engineering from Anna University Chennai in 2013. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computing Sciences and Engineering, Galgotias University of Greater Noida, Uttar pradesh since 2017. He has published 7 research papers in reputed international journals and it's also available online. His main research work focuses on Image Processing, Biometrics, Image Segmentation, Feature Extraction. He has 5 years of teaching experience.