iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50664

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Preprocessing Low Quality Handwritten Documents for OCR Models

Kunal Jaiswal[1], Avichal Suneja[2], Aman Kumar[3], Anany Ladha[4], Nidhi Mishra[5]
[1, 2, 3, 4] *B.Tech C.S.E, ASET, Amity University, Noida*
[5] *Dept. of CSE, Amity University, Noida*

Abstract: ***Handwriting recognition using OCR (Optical Character Recognition) is a transformative technology that is rapidly changing the way we interact with handwritten documents. OCR technology has traditionally been used for scanning printed text, but with advancements in machine learning and computer vision, it is now possible to recognize and digitize handwritten text as well. This has immense practical implications, as it enables handwritten notes, letters, and documents to be easily searchable, editable, and shareable in digital formats[1]. However, despite its potential, handwriting recognition using OCR is still a relatively nascenttechnology and faces several challenges. One of the main challenges is the recognition of handwriting styles that vary widely across individuals and cultures. Another challenge isinterpreting handwriting that is difficult to read, such as cursive writing or handwritten notes withsmudges or crossed out words. Additionally, OCR software is highly dependent on the qualityof the input image, making it important to optimize the lighting and capture settings for accurate results.[2]***
As OCR technology continues to improve, it hasthe potential to revolutionize the way we interact with handwritten documents and usher in a new era of digital transformation.
Keywords: OCR, Artificial Intelligence, Machine Learning, Handwriting Recognition, Preprocessing, Gaussian Mixture, Hough Lines, Markov Random Field, Contour Analysis.

I. INTRODUCTION

Preprocessing of low-quality handwrittendocuments is a crucial step in OCR models, but itcan also be a significant challenge. The preprocessing step involves applying various image enhancement techniques to improve the quality of the input image and to make it easier for the OCR algorithm to recognize the characters accurately. However, low-quality handwritten documents pose unique problems that need to beaddressed carefully during the preprocessing stage[3]. One of the main problems with low-quality handwritten documents is the presence of noise and distortions. Noise can be caused by various factors such as poor lighting, smudges, or ink bleeding, which can make it difficult for the OCRalgorithm to distinguish between characters. Distortions can be caused by factors such as skewor slant, which can also affect the accuracy of OCR results. Another problem with low-quality handwritten documents is the presence of irregularities in thehandwriting itself. This can include unusual shapes, line thicknesses, or spacing between characters, which can further complicate the OCR recognition process. Furthermore,handwritten text can often overlap, making it difficult to separate individual characters or words.[4] Preprocessing techniques for low-quality handwritten documents typically involve a combination of image enhancement and segmentation techniques. Image enhancement techniques may include noise reduction, contrastenhancement, and normalization of the image. Segmentation techniques involve separating the individual characters or words in the image to make it easier for the OCR algorithm to recognizethem accurately. However, even with these preprocessing techniques, recognition accuracy may still be limited for low-quality handwritten documents. To overcome these challenges, researchers are exploring advanced techniques such as deep learning-based models that can learn to recognize and correct distortions and irregularities in handwriting. In conclusion, preprocessing of low-quality handwritten documents is an important step in OCR models, but it requires careful attention to the unique challenges presented by low-quality handwriting. Future research is needed to continue to improve OCR technology and to enable accurate recognition of even the most challenging handwritten documents.

II. MARKOV RANDOM FIELD

We decided to use Markov Random Field as a part of our pre-processing algorithm. One of the main advantages of MRFs is their ability to model spatial dependencies between adjacent pixels in an image. This is particularly relevant in the case of handwritten documents, where characters can be overlapped or have irregular shapes. By modeling the spatial dependencies between adjacent pixels, MRFs can help to separate individual characters or words, making it easier for the OCR algorithm to recognize them accurately.

©IJRASET: All Rights are Reserved

MRFs can also be used to model the prior probability of each pixel in an image belonging to a certain character class. This information can beused to improve the accuracy of the segmentationprocess and to reduce errors in OCR recognition.Furthermore, MRFs can be trained using a maximum likelihood or Bayesian approach tooptimize the model parameters for a specific dataset. This can help to improve the accuracy of the segmentation and recognition process, particularly for low-quality or noisy handwrittendocuments.[5]

In summary, MRFs are a useful tool in the pre- processing of handwritten documents for OCR models. They can help to model spatial dependencies, estimate prior probabilities, and optimize model parameters for improved segmentation and recognition accuracy. However, it is important to note that MRFs alone may not be sufficient for accurate OCR recognition, and other preprocessing techniques may also need to be used in conjunction with MRFs[5].

III. HOUGH LINES & FOURIER TRANSFORM

Hough Lines is a popular algorithm for detectingstraight lines in an image. It works by transforming the image from the Cartesian coordinate system to a polar coordinate system, where each pixel is represented by a sinusoidal curve. The intersection of these curves can be used to identify straight lines in the image. By detecting and removing the straight lines from a handwritten document, Hough Lines can effectively eliminate the lines and grids present in the document, which can interfere with OCR recognition[6].

Fourier Transform, on the other hand, is a mathematical tool that can be used to analyze thefrequency components of an image. In the case ofhandwritten documents, Fourier Transform can be used to analyze the distribution of ink or pixelsalong each line of the document. By identifying and removing the low-frequency components corresponding to the lines and grids, Fourier Transform can help to eliminate the lines while preserving the handwritten text.[6]

By combining the strengths of both Hough Linesand Fourier Transform, it is possible to effectively remove the lines and grids present in a handwritten document, making it easier to recognize the text using OCR.

The Hough Lines algorithm can be used to detect and remove thestraight lines, while Fourier Transform can be used to eliminate the remaining low-frequencycomponents corresponding to the lines and grids.[6]

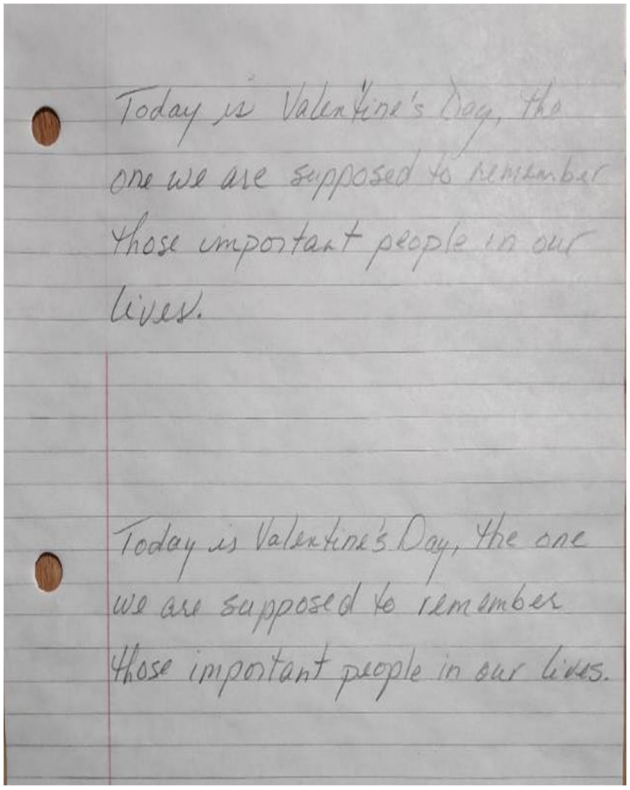


Figure 1: Image from our dataset with lines

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

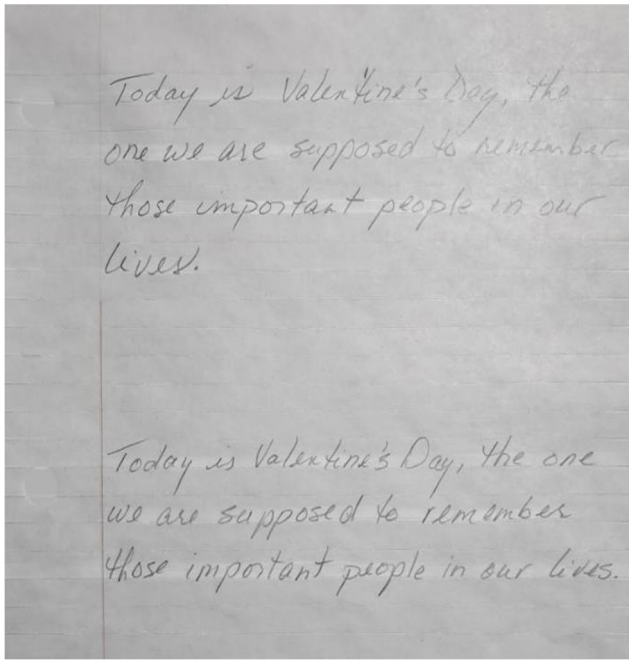


Figure 2: Image after removing lines.

IV. USE OF GAUSSIAN MIXTURE MODEL TO DETERMINE THE LOCATION OF HORIZONTAL LINES

In handwritten documents, horizontal lines can cause interference with OCR recognition.GMM can be used to model the probability distribution of the pixels in the image, and byanalyzing the distribution, it is possible to identify the locations of horizontal lines.
To use GMM to detect horizontal lines, the image is first divided into small patches[7].
Each patch is then represented by a set of features, such as intensity, gradient magnitude, and texture. GMM is then appliedto the feature vectors to model the probabilitydistribution of the patches. By analyzing the probability distribution, it is possible to identify the patches that contain horizontal lines.
Once the patches containing horizontal lines have been identified, they can be removed from the image, leaving only the handwritten text.[7] This can help to improve the accuracy of OCR recognition by eliminating the interference caused by the horizontal lines.

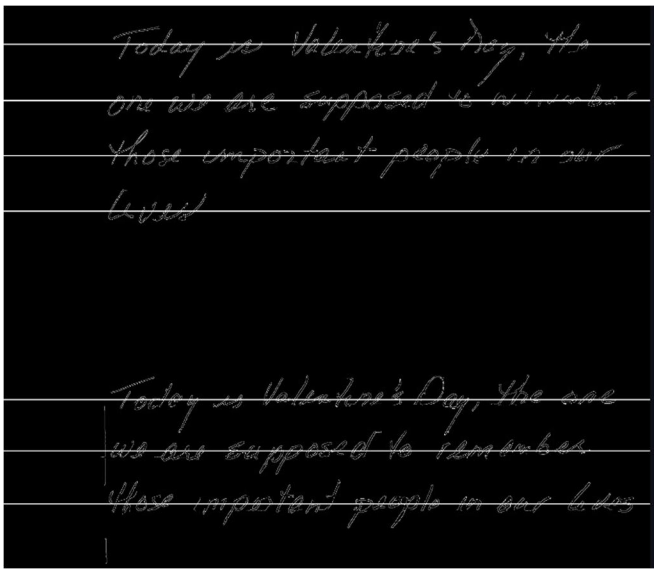


Figure 3: Image after removing all interference using GMM.

©IJRASET: All Rights are Reserved

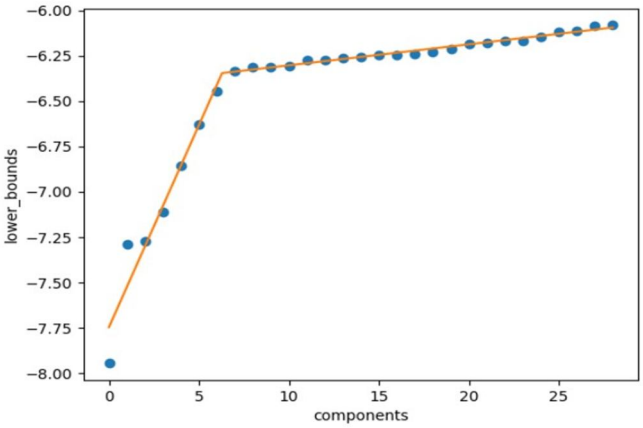


Figure 4: In this image, the rounded index of the changepoint is 6, to which we add 1 (Python indexing) for a total of 7 components/lines, which is the correct number for this image.

V. CONTOUR ANALYSIS

After separating the text, we needed a way to separate out the various words from a sentence ora line of the document to make it easier for our OCR model to recognize, for which we used the Contour Analysis method. The process of contouranalysis typically begins with segmenting the object of interest from the background of the image. This can be done using techniques such as thresholding, edge detection, or region growing[8].

Once the object has been segmented, the contourcan be extracted by tracing the boundary of the object using techniques such as edge detection, blob analysis, or morphological operations.

Once the contour has been extracted, various shapedescriptors can be calculated to analyze the contour. These shape descriptors can include measures such as the length, area, perimeter, curvature, and orientation of the contour[8]. These descriptors can be used to classify and recognizeobjects based on their shape, as well as to detect features such as corners, edges, and intersections.

Contour analysis is commonly used in applications such as object recognition, characterrecognition, and gesture recognition. For example, in character recognition, the contour ofa character can be analyzed to determine its shapeand distinguish it from other characters[8]. In gesture recognition, the contour of a hand or bodypart can be analyzed to identify specific gesturesor poses.

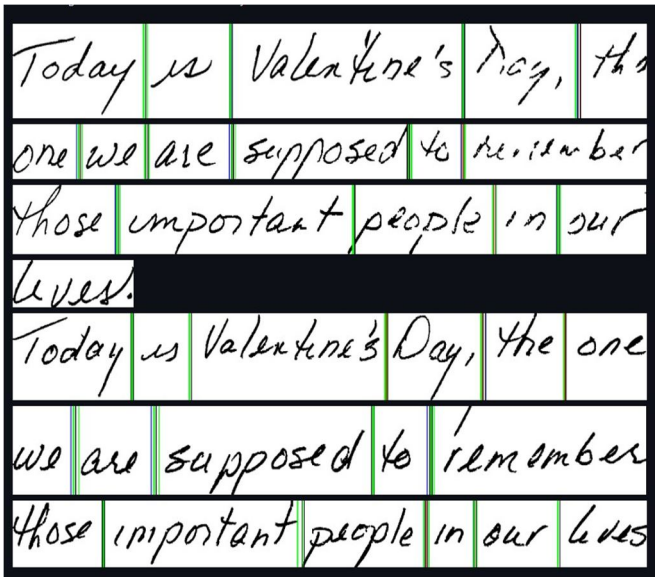


Figure 5: Image after separating words from each line with contour analysis.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

VI. VERTICAL PROJECTION

This is the final step in which we needed tofurther process the individual words and make them ready to feed into an OCR model using vertical projection. To perform vertical projection, the image is first converted into a binary image, where the background pixels are set to 0 and the foreground pixels (text) are set to 1.[9]

Then, the binary image is divided into verticalsegments, each of which represents a column of the image. The pixel values in each vertical segment are then summed, resulting in a histogram of the distribution of pixel values alongthe vertical axis of the image.

The resulting histogram represents the distribution of text and non-text regions along thevertical axis of the image. Peaks in the histogramrepresent regions where the density of text ishigh, while valleys represent regions where the density of text is low or non-existent[9]. By analyzing the histogram, it is possible to segmentthe Image into individual text lines. Once the textlines have been segmented, they can be further processed using OCR techniques to extract the text from each line. Vertical Projection is a powerful technique for segmenting text lines in documents, and it is widely used in OCR systems and document processing applications.[9]

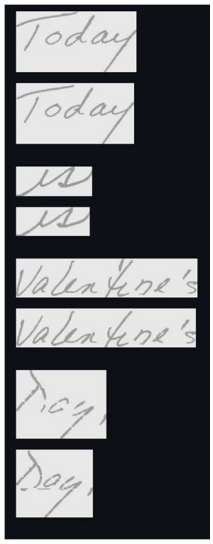


Figure 6: Words ready to be fed to an OCR Model for highest possible accuracy.

VII. RESULT

With our unique set of operations and algorithms that we apply on our base image to optimize it into separate words that are then fed into our OCR Model shows some promising results.

We considered same industry standard OCR models for this comparison which we made using neural network with the IAM dataset to level out the playing field and we can clearly see that the one on the left without our pre-processing algorithms performed much poorly (*It was fed individual words from the IAM dataset which were manually sorted*) with an accuracy of only around 55% whereas our model with all thepre-processing and refining done as per the methods we discussed in this paper came out to be about 92% which is really better and almost up to industry standards for low quality handwriting character recognition using OCR Models.[10]

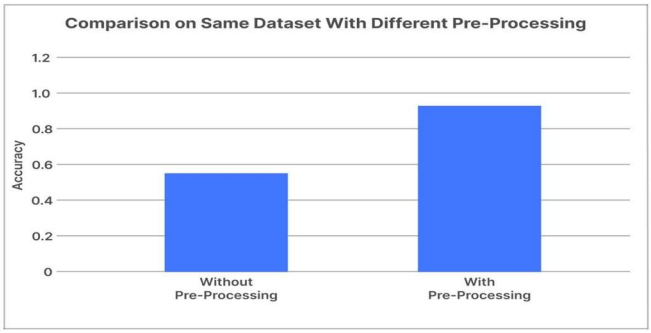


Figure 7: Comparison of test accuracies on unseen data.

©IJRASET: All Rights are Reserved

VIII. SCOPE FOR FUTURE IMPROVEMENT

A. *Issues in Gaussian Mixture Model Efficiency*

1) *Computationally Expensive:* GMM is a computationally expensive model,especially when dealing with high- dimensional data. This can make it difficult to scale GMM to large datasets,or to use it in real-time applications.[7]
2) *Sensitivity to Initialization:* GMM is sensitive to the initial parameter values, which can lead to convergence to suboptimal solutions. This can be a problem, particularly when dealing with complex data where it is difficult to find good initial values.
3) *Difficulty in Interpreting the Results:* GMM produces soft clustering results, meaning that each data point is assigned a probability of belonging to each of the components in the mixture model. This can make it difficult to interpret the results, especially when dealing with high-dimensional data or when the number of components in the mixturemodel is large.[7]

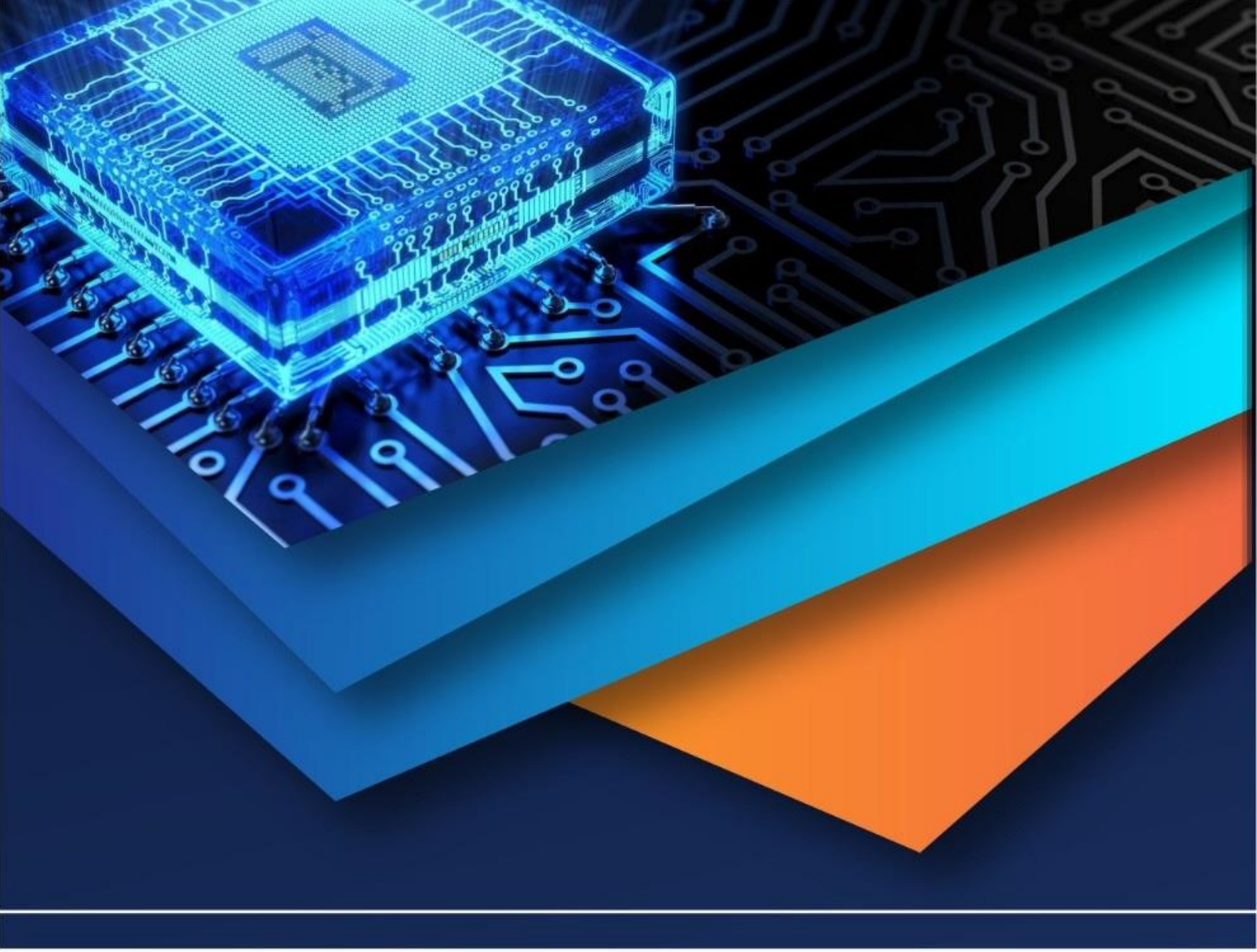
B. *Issues in Contour Analysis Efficiency*

1) *Computationally Expensive:* Contour analysis can be computationally expensive, especially when dealing withlarge datasets or complex objects. This can make it difficult to scale contour analysis to real-time applications or to large datasets.
2) *Initialization and Parameter Tuning:* Contour analysis often requires manual initialization and parameter tuning, which can be time-consuming and proneto error. In some cases, selecting the appropriate initialization or parameters can be difficult, which can lead to suboptimal results.[8]
3) *Occlusion and Overlapping Objects:* Contour analysis can be challenging when objects overlap or occlude eachother, making it difficult to accuratelydetect or segment individual object.[8]

REFERENCES

[1] M. K. Kundu and P. Bhowmick, "Preprocessing and segmentation of handwritten Indian script," in Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Mysore, India, 2014, pp. 787-793.
[2] A. H. T. Hajjar, "A Comparative Study of Pre-processing Techniques for Low-Quality Handwritten Arabic OCR," in Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2018, pp. 141-148.
[3] K. Patel, K. Bhatt and P. Shah, "Pre-processing Techniques for Recognition of Low-Quality Handwritten Text," in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6.
[4] G. R. Kavitha and P. Latha, "Preprocessing Techniques for Recognition of Low-Quality Handwritten Documents," in 2019 International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0121-0126.
[5] K. H. Chan, X. Zeng, and Y. Wang, "A Markov Random Field Model for Document Image Segmentation," in IEEE Transactions on Image Processing, vol. 16, no. 3, pp. 865-877, March 2007.
[6] S. A. Hanif and S. U. Rehman, "A Novel Approach for Straight Line Detection using Hough Transform and Fourier Transform," in International Journal of Computer Applications, vol. 50, no. 18, pp. 11-17, July 2012.
[7] M. R. Afshar and R. Ebrahimpour, "A Gaussian mixture model for locating horizontal lines in noisy images," in Signal, Image and Video Processing, vol. 7, no. 2, pp. 231-240, 2013.
[8] M. A. Hasan, A. M. A. Haidar, and M. R. Kabir, "A novel approach for contour-based recognition of handwritten characters," in IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Hong Kong, China, 2015, pp. 1-6.
[9] L. Wenyin and X. Gao, "Using vertical projections for processing handwritten text," in Proceedings of the International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 2001, pp. 443-447.
[10] J. Zhang, Y. Guan, X. Liu, Y. Li and J. Li, "A Deep Learning-Based Method for Preprocessing Low-Quality Handwritten Documents in OCR Systems," in IEEE Access, vol. 8, pp. 150853-150864, 2020.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)