

**PREPROCESSING AND SEGMENTATION TECHNIQUE
FOR AN OFFLINE NOVEL GURMUKHI HANDWRITTEN
DATASET**

Sukhdeep Kaur ¹, Ganesh Kumar Sethi ²

Research Scholar¹ Department of Computer Science, Punjabi University,
Patiala, India.

Associate Professor² Department of Computer Science, M. M. Modi
College Patiala, India.

Abstract

Writer identification systems are advanced technology solutions used in forensic investigation, authentication of historical document and the identification of writer that are crucial for identifying handwritten text with specific individuals. This study presents a novel dataset called GUR-PRIMTA of 2100 handwritten text images in the Gurmukhi script which is often used in Punjab, India. The experimental setup involved many crucial phases, primarily segmentation and preprocessing. Using deep convolutional neural networks (CNNs) to reduce and eliminate noise from images, VGG16 model is used to preprocess the handwritten text images. A contour curving-based segmentation approach has been applied to segment handwritten text words for further analysis. The proposed methodology shows effectiveness to enhance the accuracy and speed in the tasks of writer identification, with implications for various real-world applications that required accurate features of handwritten text.

Keywords: Preprocessing, Word Segmentation, GUR-PRATIMA Dataset, Contour Curve.

1. Introduction

In today's technologically evolved society, handwriting identification is a requirement across various sectors. The identification of writers is one of the most challenging tasks in pattern recognition. For writer identification, a text is compared with a training dataset in order to determine who wrote it. Because different people have different writing styles, this procedure is very complicated. Identification depends on several factors, including writing style, emotions, perceptions, behavior, and distinct cognitive processes. With the rapid development of pattern recognition and artificial intelligence, handwriting analysis has emerged as a prominent research topic. Biometric identification, a field within computer science, focuses on identifying individuals from a group by using various identifiers such as signatures, retinal scans, handwriting, and fingerprints. Compared to non-Indic scripts, less work has been done on writer identification systems with excellent efficiency rates in Gurmukhi, an Indic script. Many preprocessing method and text word segmentation techniques for handwritten text images have been performed in recent decades. Various techniques work directly on grayscale images, while others work in binary images. Still others combine binary and grayscale images

to achieve text word segmentation. The objective of the paper is to create a model that preprocess and segments the words, lines etc. on the handwritten images of Novel GUR PRATIMA dataset which is based on Gurmukhi script. The Gurmukhi script, an alphabetic writing system primarily employed for the Punjabi language, was pioneered by the Sikh Guru, Guru Angad Dev Ji, during the 16th century. In addition to religious scriptures, Gurmukhi is utilized in Punjab, India, and other Punjabi-speaking areas of the world for literature, education, and government purposes.

1.1 General Architecture of handwritten identification system

Figure 1 depicts the generalized architecture of the handwriting identification system which consists of various phases such as:

Image Acquisition: An image is captured and digitalized using a sensor, such as a scanner or camera.

Image Preprocessing: The initial stage of getting an image ready for further processing. The objective is to extract desirable characteristics that represent important information in an image and to eliminate unsuitable features that will interfere with later processing. At this point, the computer eliminates noise and sometimes enhances specific object attributes that are essential when interpreting the image.

Image Segmentation This is fundamental task in computer vision that involves partitioning an image into multiple segments or regions based on certain criteria, such as color, texture, or intensity. The goal of image segmentation is to simplify the representation of an image for easier analysis and interpretation.

Feature Extraction: Feature extraction step includes extraction of salient features, capture curvature and roundness, texture based and spectral features of handwriting. The off line feature extraction includes document feature, paragraph feature, line feature, word feature and character feature

Classification: Classification of writer is the last step of the handwriting identification process. The extracted features are to be fed into some classifier to assign writer-id. We will use machine learning with suitable classifier.

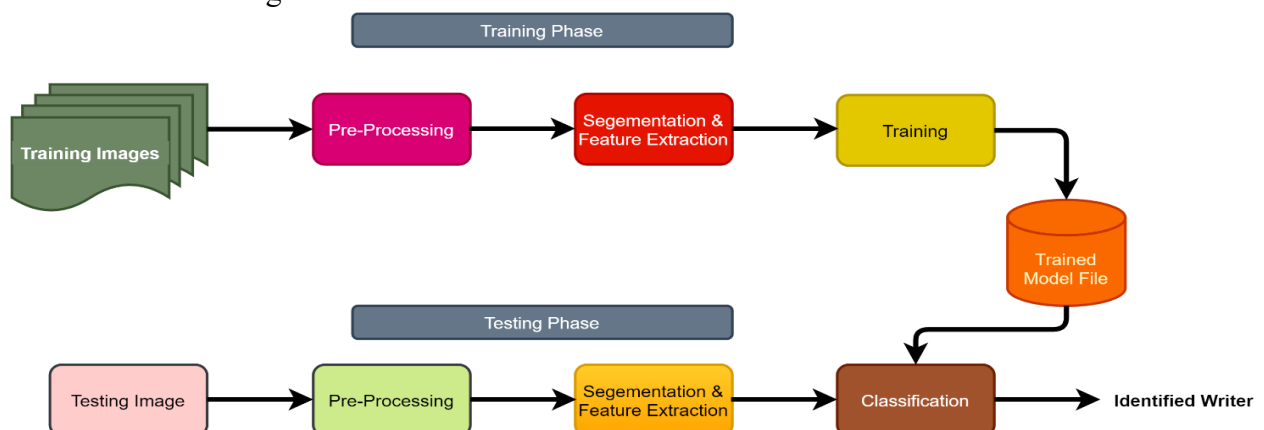


Fig 1: General Architecture of Writer Identification and Classification

1.2 Motivation and Contribution

No work is available the writer identification of handwritten documents of Gurmukhi Script. So, it is of great motivation to, process such handwritten documents with a large variability. Writer identification system has been used for the digitization of documents where segmentation of documents is a necessary and inherent part of the process. Thus, this research work is a modest attempt to segment the handwritten documents having text written in Gurmukhi scripts at the word. It contributes in the manner as described below:

1. Large variations in handwriting have been captured by taking samples from persons belonging to different domain and content of text from Academic domain, etc.
2. There is limit on the number of words or lines written on the document for sample 1 and 2 as print- printed pangram is mentioned, no restriction of line, word and character size and spacing between the lines. An effort has been made to capture straight, skewed as well as the curved text which is a complex task in handwritten identification. Skew can be present at the document, line or word level.
3. There is no limit of word and lines for sample 3. Writer can write the data of own interest
3. The approach contour curve tracking has been implemented for segmentation.
4. A novel GUR-PRATIMA dataset has been constructed from the various domain of 2100 documents with handwritten text collected from writers of various age groups. The documents considered for the purpose include Punjabi and Alphanumeric characters data.
5. A preprocessing technique VGG16 model has been applied to handwritten images of GUR-PRATIMA dataset.
6. The contour curve approach is capable of segmenting documents with various structures like application format, notes to be displayed, text starting from the middle of page, document with paragraphs etc.
7. Later, the performance of the applied segmentation technique has been computed on the basis of the accuracy

1.3 Challenges

It is hard to segment the text in handwritten documents due to variable line spacing and curved text lines. In Punjabi, characters are connected by a header line; and modifiers are present in the upper and lower zones of text, and overlap with lower and upper text lines. To segment such text we need certain methods which can segment the text without losing the meaning of characters like modifiers etc. Segmenting text with header line has more challenges as there are some characters which use header line while some characters have empty space in the header line (eg. Sassa and khakha).. Gurumukhi text is written in three zones and it has many modifiers present in the upper and lower zones. Further some modifiers like sihari and bihari are present in upper and middle zone. Segmentation of such text requires special techniques to reduce the over and under segmentation errors. Taking into consideration above mentioned all points, we have implemented curve contouring method on Gurumukhi and script. Some of the other challenges for text segmentation are as follows:

1. The segmentation algorithms need to be evaluated on the basis of number of over, mixed and bad as the over segmentation rates result in high computation power.
2. The present state-of-the-art segmentation algorithms produced different results after their application on the same dataset for various iterations.
3. In Gurumukhi, characters are connected through header line, and are overlapped in the upper and lower zone due to the modifiers. It makes the segmentation process more difficult as the approaches used in segmenting the cursive script cannot be applied in segmenting the Gurumukhi script.
4. It is necessary to focus on increasing the diversity of training data to process variety of documents as one training set can handle similar kind of documents. This fact can be seen by the large difference between volumes of both training as well as testing data set.
5. Some writing styles have variable inter and intra word gap which challenges the word segmentation task.
6. Irregular page layout cannot be handled using a single approach.

1.4.Format of the manuscript

The subsequent sections outline the reported work (Section 2), elaborate Material and methods (Section 3), discuss and analyze experimental outcomes (Section 4), and conclude by summarizing findings and suggesting future directions (Section 5).

2. Reported Work

Bal A. et al. (2016) [1] used the IAM database and proposed modified horizontal and vertical projection, baseline recognition and normalization, and writing pressure for detecting handwritten manuscripts in cursive. The accuracy of the segmentation was 92.56% for words and 95.65% for lines. *Zuo L. et al. (2002) [2]* used PCA for writer identification on the basis of their handwriting. The experiment encompassed 400 pages of handwritten texts, comprising approximately 16,000 Chinese characters written by 40 distinct writers. This extensive dataset was employed to assess the efficacy of the method. The outcomes demonstrated a remarkable average identification accuracy of 97.5%, coupled with rapid identification processing. *Suteddy W. et al. (2023) [3]* presented evaluation of several deep learning architectures, including VGG16, ResNet50, MobileNet, Xception, and EfficientNet, for offline handwriting writer identification using the IAM and CVL databases has been accomplished. ResNet50 and VGG16. When the accuracy of the training and validation metrics for each architecture was evaluated, ResNet50 produced the best train accuracy, at 98.86%. *Suteddy W. et al. (2022) [4]* proposed CNN model, which was trained, evaluated, and analyzed using Adam and the stochastic gradient descent (SGD) optimizer with batch sizes of 2, 4, and 8 and applied learning rates (LR) of 0.001, 0.01, and 0.1. and contained different classes of 10 writers, each included 400 images of handwritten city names in Gurmukhi script, created the dataset which was utilized for the implementation task. By comparing the metrics, it had been analyzed that ResNet50 had the best training accuracy of 98.86% and 99.13% validation accuracy, for the CNN model with an Adam optimizer, batch size of 4 with LR of 0.001. *Rehman A. et al.*

(2019) [5] QUWI dataset consist of English and Arabic language, was utilized to train a deep transfer convolutional neural network (CNN) to identify writers. Techniques like contours, negatives, and sharpness implemented on text-line images on the target data-set resulting in reduced the possibility of over-fitting. For English, 92.78%, and Arabic, 92.20%, and the combination of Arabic and English, 88.11%, were the accuracy rates utilizing the frozen Conv5 layer. *Purohit N. et al. (2023) [6]* proposed an offline writer identification system using a multi-input hybrid convolution neural network (CNN) which worked on multilingual and monolingual handwritten texts. By merging local CNN features with the traditional handcrafted features of HOG (Histogram of Oriented Gradients), the model generated global features for classification. The model achieved an accuracy rate of 98.23% using the English script benchmark dataset CVL; however, the accuracy rate was 96.93% with an in-house bilingual dataset consisting of handwritten English and Hindi text. *Roy P. P et al. (2016) [7]* provided information that identifying a writer of Indic scripts—such as Devanagari, Bangla, Gurumukhi, and others—required a difficult process. A model based on HMM sequence has been utilized in these script and the reported accuracy for Bangla and Devanagari was 92.82% and 94.51%, respectively. *Ahmed B.Q. et al. (2023) [8]* developed a technique for offline text-independent writer identification, by using handwritten samples in Arabic and English. The contour points curve angle (CPCA) and contour point concavity/convexity (CON3) have two characteristics which has been applied for analysis of handwritten ink strokes. The reported accuracy was 89.5% for the KHATT dataset while 98.2%. for IAM. *Dargan S. et al. (2020) [9]* implemented technique on handwritten Gurumukhi text for identifying writers. A dataset of 100 authors which contained $100 \times 53 \times 10 = 53000$ Gurumukhi characters, has been acquired for experiment. Feature extraction approaches such as Zoning, transition, diagonal, and peak extent were employed, further Multi Layered Perceptron (MLP), Artificial Neural Network (ANN), and Random Forest (RF) classifiers were trained to accomplish the classification phase. A 93.06% accuracy rate for writer identification system has been reported. The proposed study analysis the performance metrics with a 0.39% False Positive Rate (FPR) and a 93.2% True Positive Rate (TPR). *Dargan S. et al. (2020) [10]* a system that implemented feature extraction with a mixture of zoning, diagonal, transitions, and a linear SVM classifier, in addition to preprocessing and digitization. An accuracy rate of 91.53% was reported on 24,500 samples of Devanagari characters written by 100 distinct authors. *Longjam T. et al. (2023) [11]* implemented an adaptable CNN-based system to verify offline multi-scripted signatures by using two publically accessible databases (BH-Sig260 and CEDAR) which include English, Bengali, and Hindi signatures. The reported method has been tested with individual script signature classes as well as a combination of these scripts, and the attained accuracy of 93.33%, 95%, 98.33%, and 90%, respectively. *Dos Santos et al. (2009) [12]* introduced a text line segmentation technique implemented on the IAM database, and demonstrated its effectiveness, speed, and safety. It extracted the picture characteristics using a morphological operator. *He J. et al. (2003) [13]* introduced a configurable archive document image analysis system for the construction of digital libraries, which has been designed using top-down iterative development methods and rapid prototyping. The system achieved

approximately 93% correct segmentation of the required archive card fields when tested on set of 2000 images. *Nguyen K. C., et al. (2016) [14]* discovered text-line segmentation by utilizing a morphological approach, zone projection. Every text-line segmented by vertical projection, Stroke Width Transform (SWT), bridge discovery, and Voronoi diagrams. The technique divides intricate touching characters into curving, touching, and skew text-lines. Using the HIT-MW database, 83.93% accuracy has been reported during the trial. *Arivazhagan M. et al. (2007) [15]* demonstrated a method for segment a handwritten document into separate text lines. Based on experimental results, 97.31% of lines in 720 papers (including Arabic, English, and handwritten by children) with a total of 11, 581 lines have been accurately segmented. *Su T. H. et al. (2007) [16]* proposed a technique for skew detection in handwritten Chinese documents. It utilized the horizontal stroke histogram after examining the features of Chinese characters. The approach has 8.44% better than the Wigner-Ville Distribution (WVD) technique, identifying 98.83% of the skew angles within a degree. *Weliwitage C. et al. (2005) [17]* proposed a technique Cut text minimization (CTM) for text segmentation into lines of text, produced accuracy levels of 96% with fewer than five foreground cuts and 60% of text lines with no front cuts when applied to the NIST data set. *Yin F. et al. (2009) [18]* reported that Using a database of 803 unconstrained handwritten Chinese document images with a total of 8,169 lines, a text line segmentation algorithm based on minimal spanning tree (MST) clustering with distance metric learning has been developed. The suggested algorithm achieved a correct rate of 98.02% of line detection and outperformed other competing algorithms. *Basu S. et al. (2007) [19]* demonstrated a text line extraction approach for multi-skewed document picture of handwritten English or Bengali text, and stated accuracy of 90.34% and 91.44%, respectively. *O'Gorman et al. (1993) [20]* proposed the docstrum as a page layout analysis technique. It formed text lines and blocks, automatically estimates text orientation and spacing settings, and created k nearest-neighbor pairings between picture components in a bottom-up way. Examples have been provided to demonstrate how to use the docstrum on various page types. *Kise K. et al. (1998) [21]* provided a page segmentation technique by utilizing the estimated area from the Voronoi diagram. The approach extracted body text regions from 128 images featuring with non-Manhattan layouts and skew angles ranging from 0° to 45°, as well as 98 images with Manhattan layouts. It attained an accuracy of 97.9%. *Simon A. et al. (1997) [22]* proposed methodology used for the CLIDE (Chemical Literature Data Extraction) system, however the approach outlined here worked with a wider variety of documents. The physical page structure was built using a unique distance metric between the components, which was based on Kruskal's method. *Li Y et al. (2008) [23]* proposed script-independent text line segmentation in freestyle handwritten documents, and the obtained accuracy on different scripts were 85.6% on Arabic, 92% for Hindi, 92% for Chinese, and 96% for Korean. *Doermann D. et al. (2006) [24]* developed a method for text line detection for many scripts, including Arabic, English, Chinese, Hindi, and Korean. According to statistical findings, algorithm consistently yields accurate results with a stated accuracy of 85.6%, even when skew angles, character sizes, and noise levels were reasonably varied. *Louloudis G. et al. (2006) [25]* applied text line identification approach

based on linked component on unrestricted handwritten Greek manuscripts, and obtained 96.87%-line detection accuracy. *Saha S. et al. (2010) [26]* developed a Hough transform based technique for line and word segmentation. This technique was applied on public domain datasets published on the CMATER, Jadavpur University website, as well as datasets for business card reader systems and license plate recognition systems, in order to standardize the performance of the system. Experimental verification showed that the word segmentation approach has 85.7% accurate for document images, 94.6% accurate for business card images, and 88% accurate for security camera images. In addition to this, a comprehensive overview of the diverse preprocessing methodologies and datasets employed in recent studies on handwritten images had been also compared and analyzed in Table 1. By analyzing and synthesizing these approaches, we aim to gain insights into the current state-of-the-art techniques and identify potential avenues for future research in this field.

Table I. Comparison of Different Preprocessing Techniques

Ref	Author Name	Dataset	Techniques
[1]	Bal A. et al.	IAM	Salt and Pepper noise using Median filter technique. Image binarization using Otsu Thresholding
[2]	Zuo L. et al.	16000 Chinese words	Linear normalization Skewness correction
[3]	Suteddy W. et al.	IAM and CVL	Noise Removal using dilation and erosion operations. Gaussian Filter Otsu Thresholding
[4]	Sharma S. et al.	4000 samples of the handwritten Images	Normalization Brightness Adjustment Contrast Adjustment Intensity level Adjustment
[5]	Rehman A. et al.	QUWI	Skew Detection and Correction Normalization The Sliding Window Patching Strategy Contours Extraction Sharp Images and Negatives.
[6]	Purohit N. et al.	CVL and a bilingual dataset	Otsu Thresholding Gaussian Filter
[7]	Roy P. et al.	Bangla/Devanagari scripts	Global histogram-based Otsu binarization Water reservoir for skew correction
[8]	Ahmed B.Q. et al.	KHATT and IAM	Average Filter Otsu Thresholding
[9]	Dargan S. et al.		PCA

		53000 Gurumukhi characters	LDA
[10]	Dargan S. et al.	24,500 Devanagari Character sample	The nearest neighbor interpolation method is used to normalize character image. Bitmap Image. Image thinned with a parallel thinning technique.
[11]	Longjam T et al.	CEDAR and BH-Sig260	Image cropping and resizing: 256 x 256. Noise removal: The bilateral filter A Color image converted to grayscale by averaging the pixel values across all three channels.

Table II highlights the diversity of methodologies put forth by various researchers aimed at segmenting text lines, word and character within document images.

Table II. Comparison of Different Segmentation Techniques

Ref	Author	Dataset	Brief description
[12]	dos Santos et al.	IAM	In order to divide characters and lines in a document image, they use both horizontal and vertical projection histograms.
[13]	He J. et al.	Archive card	This technique cannot be extended to narrow lines as it assumes parallel text lines with a wide space between them.
[14]	Nguyen K. C.	Japanese	Text is divided into vertical strips, and each strip's local projection profile is calculated to identify lines.
[15]	Arivazhagan M. et al.	CEDARFOX, ARABIC, Exam Essay	Every stripe is thought to consist of straight lines that are divided by a horizontal projection histogram and connected by heuristic rules between adjacent components.
[16]	Su T. H. et al.	HIT-MW	Along with the determined skew angle, Chinese text is provided using a skew detection approach based on the segmentation of handwritten text lines and the horizontal stroke histogram.
[17]	Weliwitag e C. et al.	NIST	Suggested extending the projection-based method to estimate the border of adjacent lines while utilizing optimization techniques for cut text minimization (CTM), assuming that the

			initial points of lines are identified using projection.
[18]	Yin F. et al.	HIT-MW	a minimum spanning tree (MST) is produced using a certain distance metric in line segmentation based on connected components.
[19]	Basu S. et al.	Bengali and English document images	A theoretical approach that predicts the spaces between lines using an hypothetical stream of water is presented, originating from both ends of the text document.
[20]	O'Gorman et al.	Page formats (printed documents)	A method that uses the criteria of the k closest neighbour unit to combine nearby linked components. The technique works well with printed materials and slightly curved handwritten documents.
[21]	Kise K. et al.	Voronoi diagram.	This method integrates CCs into text lines by using a Voronoi diagram and heuristic criteria.
[22]	Simon A. et al.	Images from the Journal of Organic Chemistry, the Journal of Medicinal chemistry, and the Journal of the American Chemical Society	A method is provided that assembles MST using a hand-crafted distance metric; The effectiveness relies on both the distance metric and heuristic rules.
[23]	Li Y. et al.	Arabic, Hindi, Chinese, and Korean	The method for segmenting freely written documents combines the density estimation with the level-set approach. The initial measure defines the boundary between adjacent lines.
[24]	Doermann D. et al.	Arabic, Hindi, and Chinese	The technique uses the Gaussian window to enhance the text-line architecture and maintain the level-set in order to determine the distance.
[25]	Louloudis G. et al.	Unconstrained Handwritten Greek Documents	This method divides CCs into three groups that are evenly spaced, each of which is then subjected to a distinct method and votes in the Hough domain. Once probable text lines have been identified, it fixes possible errors.
[26]	Saha S. et al.	Handwritten and printed text lines with multi script variations	HT-based method for digital picture segmentation into words and lines. The method is used on a variety of datasets,

saved in .tif format. The storage structure of dataset contains 700 folders; each folder contains 3 images stored in it.

Sample No. 1 Writer No. 31

WRITER NAME: Jasleen Kaur

AGE: 16 Years

GENDER: Female

QUALIFICATION: 10th

PROFESSION: Student

DESIGNATION: Student

CITY: Katangarh (Bareilly)

STATE: Punjab

WRITING INSTRUMENT/COLOR: Black ball Pen

LEFT/RIGHT HANDED: Right Hand

MOBILE NUMBER: 96539-03857

WRITER NUMBER: 31

I give my consent to use this data for research/ publication.

(In English) Jasleen Kaur (In Punjabi) ਜਸਲੀਨ ਕੌਰ

Signature

Fig 3: Writer's Information

Sample No. 2 Writer No. 31

ਪੰਜਾਬ ਵਿਧਾਨ ਸਭਾ ਨੇ 1961 ਦੇ ਐਕਟ ਨੰ. 35 ਅਧੀਨ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਦੀ ਸਥਾਪਨਾ 30 ਅਪ੍ਰੈਲ 1962 ਈ: ਨੂੰ ਕੀਤੀ। ਇਸਰਾਈਲ ਦੀ ਹਿਬਰੂ ਯੂਨੀਵਰਸਿਟੀ ਤੋਂ ਬਾਅਦ ਸ਼ਾਹੀ ਸ਼ਹਿਰ ਪਟਿਆਲਾ ਵਿੱਚ ਸਥਾਪਤ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਦੁਨੀਆ ਦੀ ਅਜਿਹੀ ਯੂਨੀਵਰਸਿਟੀ ਹੈ, ਜੋ ਭਾਸ਼ਾ ਦੇ ਨਾਮ 'ਤੇ ਸਥਾਪਿਤ ਕੀਤੀ ਗਈ ਹੈ। ਇਸ ਯੂਨੀਵਰਸਿਟੀ ਨੇ ਵੱਖ-ਵੱਖ ਖੇਤਰਾਂ, ਜਿਵੇਂ ਕਿ ਫਿਲਮ, ਸੰਗੀਤ, ਖੇਡਾਂ, ਰਾਜਨੀਤੀ, ਕਲਾ-ਜਗਤ ਨਾਲ ਸੰਬੰਧਿਤ, ਉਹ ਨਾਮਵਰ ਚਿਹਰੇ ਪੈਦਾ ਕੀਤੇ ਹਨ, ਜਿਨ੍ਹਾਂ ਨੇ ਵਿਸ਼ਵ-ਪ੍ਰਸਿੱਧੀ ਪ੍ਰਾਪਤ ਕੀਤੀ। ਯੂਨੀਵਰਸਿਟੀ ਵਿੱਚ ਪਾਠਕਾਂ ਵਾਸਤੇ ਵਿਸ਼ਾਲ ਲਾਇਬਰੇਰੀ ਅਤੇ ਸੰਦਰਭ ਸੈਕਸ਼ਨ 22 ਘੰਟੇ ਖੁੱਲ੍ਹਾ ਰਹਿੰਦਾ ਹੈ, ਜਿਥੇ ਵਿਦਿਆਰਥੀ ਨੂੰ-ਅਧਿਐਨ ਕਰ ਸਕਦੇ ਹਨ। ਬੋਜ-ਕਾਰਜ ਦੇ ਖੇਤਰ ਵਿੱਚ ਯੂਨੀਵਰਸਿਟੀ ਕਾਨਫਰੰਸਾਂ ਦੇ ਨਾਲ ਹੋਰ ਖੋਜਕਾਰੀ ਸੰਮੇਲਨਾਂ ਮੁਹੱਈਆ ਕਰਵਾ ਰਹੀ ਹੈ। ਯੂਨੀਵਰਸਿਟੀ ਦੇ ਸਭਿਆਚਾਰਕ ਪ੍ਰੋਗਰਾਮਾਂ ਦੌਰਾਨ ਵੱਖ-ਵੱਖ ਰੰਗਾਂ ਦੇ ਪੰਜਾਬੀ ਪਹਿਰਾਵੇ ਵਿੱਚ ਸਜੇ ਗੱਡਰੂ-ਮੁਟਿਆਰਾਂ ਦੇ ਉਚਮਟ ਅਤੇ ਕੱਛਣਾਂ ਦੀ ਆਵਾਜ਼ ਪ੍ਰਗਟ ਸਮੇਂ ਦੀ ਜੋੜ ਵਰਗਾ ਮਾਹੌਲ ਬਣਾ ਦਿੰਦੇ ਹਨ। ਯੂਨੀਵਰਸਿਟੀ ਦਾ ਖੂਬਸੂਰਤ ਫੁੱਲਾਂ ਦੇ ਬਗੀਚਿਆਂ ਨਾਲ ਸਜਿਆ ਵਿਹੜਾ ਜਿਸ ਨਾਲ ਆਲਾ-ਦੁਆਲਾ ਮਹਿਕਾਂ ਵੰਡਦਾ ਪ੍ਰਤੀਤ ਹੁੰਦਾ ਹੈ। 2020 ਵਿੱਚ ਖੋਜਿਤ ਹੋਈਆਂ ਐਨ ਆਈ ਆਰ ਐਫ ਰੈਂਕਿੰਗ ਵਿੱਚ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਨੇ 100 ਤੋਂ ਵੱਧ ਸਥਾਨਾਂ ਦੀ ਫ਼ਲਾਂਗ ਲਾਉਂਦੇ ਹੋਏ 64ਵਾਂ ਸਥਾਨ ਹਾਸਿਲ ਕੀਤਾ ਹੈ।

Kindly read the instruction carefully before filling the sample.

- Do not fold the sheets.
- Fill the sample carefully.
- Kindly write the paragraph within the boundary.
- On page No. 3 write the given sample within the lines.
- On page no 4 write the paragraph with in boundary.
- One sample must be filled by one writer.
- Don't cross or mark anything else without the given paragraph.
- Don't over write and cross the character or word.

Fig 4: Pre-printed Gurmukhi Pangram

Sample No. 3 Writer No. 31

ਪੰਜਾਬ ਵਿਧਾਨ ਸਭਾ ਨੇ 1961 ਦੇ ਐਕਟ ਨੰ. 35 ਅਧੀਨ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਦੀ ਸਥਾਪਨਾ 30 ਅਪ੍ਰੈਲ 1962 ਈ: ਨੂੰ ਕੀਤੀ। ਇਸਰਾਈਲ ਦੀ ਹਿਬਰੂ ਯੂਨੀਵਰਸਿਟੀ ਤੋਂ ਬਾਅਦ ਸ਼ਾਹੀ ਸ਼ਹਿਰ ਪਟਿਆਲਾ ਵਿੱਚ ਸਥਾਪਤ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਦੁਨੀਆ ਦੀ ਅਜਿਹੀ ਯੂਨੀਵਰਸਿਟੀ ਹੈ, ਜੋ ਭਾਸ਼ਾ ਦੇ ਨਾਮ 'ਤੇ ਸਥਾਪਿਤ ਕੀਤੀ ਗਈ ਹੈ। ਇਸ ਯੂਨੀਵਰਸਿਟੀ ਨੇ ਵੱਖ-ਵੱਖ ਖੇਤਰਾਂ, ਜਿਵੇਂ ਕਿ ਫਿਲਮ, ਸੰਗੀਤ, ਖੇਡਾਂ, ਰਾਜਨੀਤੀ, ਕਲਾ-ਜਗਤ ਨਾਲ ਸੰਬੰਧਿਤ, ਉਹ ਨਾਮਵਰ ਚਿਹਰੇ ਪੈਦਾ ਕੀਤੇ ਹਨ, ਜਿਨ੍ਹਾਂ ਨੇ ਵਿਸ਼ਵ-ਪ੍ਰਸਿੱਧੀ ਪ੍ਰਾਪਤ ਕੀਤੀ। ਯੂਨੀਵਰਸਿਟੀ ਵਿੱਚ ਪਾਠਕਾਂ ਵਾਸਤੇ ਵਿਸ਼ਾਲ ਲਾਇਬਰੇਰੀ ਅਤੇ ਸੰਦਰਭ ਸੈਕਸ਼ਨ 22 ਘੰਟੇ ਖੁੱਲ੍ਹਾ ਰਹਿੰਦਾ ਹੈ, ਜਿਥੇ ਵਿਦਿਆਰਥੀ ਨੂੰ-ਅਧਿਐਨ ਕਰ ਸਕਦੇ ਹਨ। ਬੋਜ-ਕਾਰਜ ਦੇ ਖੇਤਰ ਵਿੱਚ ਯੂਨੀਵਰਸਿਟੀ ਕਾਨਫਰੰਸਾਂ ਦੇ ਨਾਲ ਹੋਰ ਖੋਜਕਾਰੀ ਸੰਮੇਲਨਾਂ ਮੁਹੱਈਆ ਕਰਵਾ ਰਹੀ ਹੈ। ਯੂਨੀਵਰਸਿਟੀ ਦੇ ਸਭਿਆਚਾਰਕ ਪ੍ਰੋਗਰਾਮਾਂ ਦੌਰਾਨ ਵੱਖ-ਵੱਖ ਰੰਗਾਂ ਦੇ ਪੰਜਾਬੀ ਪਹਿਰਾਵੇ ਵਿੱਚ ਸਜੇ ਗੱਡਰੂ-ਮੁਟਿਆਰਾਂ ਦੇ ਉਚਮਟ ਅਤੇ ਕੱਛਣਾਂ ਦੀ ਆਵਾਜ਼ ਪ੍ਰਗਟ ਸਮੇਂ ਦੀ ਜੋੜ ਵਰਗਾ ਮਾਹੌਲ ਬਣਾ ਦਿੰਦੇ ਹਨ। ਯੂਨੀਵਰਸਿਟੀ ਦਾ ਖੂਬਸੂਰਤ ਫੁੱਲਾਂ ਦੇ ਬਗੀਚਿਆਂ ਨਾਲ ਸਜਿਆ ਵਿਹੜਾ ਜਿਸ ਨਾਲ ਆਲਾ-ਦੁਆਲਾ ਮਹਿਕਾਂ ਵੰਡਦਾ ਪ੍ਰਤੀਤ ਹੁੰਦਾ ਹੈ। 2020 ਵਿੱਚ ਖੋਜਿਤ ਹੋਈਆਂ ਐਨ ਆਈ ਆਰ ਐਫ ਰੈਂਕਿੰਗ ਵਿੱਚ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਨੇ 100 ਤੋਂ ਵੱਧ ਸਥਾਨਾਂ ਦੀ ਫ਼ਲਾਂਗ ਲਾਉਂਦੇ ਹੋਏ 64ਵਾਂ ਸਥਾਨ ਹਾਸਿਲ ਕੀਤਾ ਹੈ।

Sample No. 3 Writer No. 31

ਪੰਜਾਬ ਵਿਧਾਨ ਸਭਾ ਨੇ 1961 ਦੇ ਐਕਟ ਨੰ. 35 ਅਧੀਨ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਦੀ ਸਥਾਪਨਾ 30 ਅਪ੍ਰੈਲ 1962 ਈ: ਨੂੰ ਕੀਤੀ। ਇਸਰਾਈਲ ਦੀ ਹਿਬਰੂ ਯੂਨੀਵਰਸਿਟੀ ਤੋਂ ਬਾਅਦ ਸ਼ਾਹੀ ਸ਼ਹਿਰ ਪਟਿਆਲਾ ਵਿੱਚ ਸਥਾਪਤ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਦੁਨੀਆ ਦੀ ਅਜਿਹੀ ਯੂਨੀਵਰਸਿਟੀ ਹੈ, ਜੋ ਭਾਸ਼ਾ ਦੇ ਨਾਮ 'ਤੇ ਸਥਾਪਿਤ ਕੀਤੀ ਗਈ ਹੈ। ਇਸ ਯੂਨੀਵਰਸਿਟੀ ਨੇ ਵੱਖ-ਵੱਖ ਖੇਤਰਾਂ, ਜਿਵੇਂ ਕਿ ਫਿਲਮ, ਸੰਗੀਤ, ਖੇਡਾਂ, ਰਾਜਨੀਤੀ, ਕਲਾ-ਜਗਤ ਨਾਲ ਸੰਬੰਧਿਤ, ਉਹ ਨਾਮਵਰ ਚਿਹਰੇ ਪੈਦਾ ਕੀਤੇ ਹਨ, ਜਿਨ੍ਹਾਂ ਨੇ ਵਿਸ਼ਵ-ਪ੍ਰਸਿੱਧੀ ਪ੍ਰਾਪਤ ਕੀਤੀ। ਯੂਨੀਵਰਸਿਟੀ ਵਿੱਚ ਪਾਠਕਾਂ ਵਾਸਤੇ ਵਿਸ਼ਾਲ ਲਾਇਬਰੇਰੀ ਅਤੇ ਸੰਦਰਭ ਸੈਕਸ਼ਨ 22 ਘੰਟੇ ਖੁੱਲ੍ਹਾ ਰਹਿੰਦਾ ਹੈ, ਜਿਥੇ ਵਿਦਿਆਰਥੀ ਨੂੰ-ਅਧਿਐਨ ਕਰ ਸਕਦੇ ਹਨ। ਬੋਜ-ਕਾਰਜ ਦੇ ਖੇਤਰ ਵਿੱਚ ਯੂਨੀਵਰਸਿਟੀ ਕਾਨਫਰੰਸਾਂ ਦੇ ਨਾਲ ਹੋਰ ਖੋਜਕਾਰੀ ਸੰਮੇਲਨਾਂ ਮੁਹੱਈਆ ਕਰਵਾ ਰਹੀ ਹੈ। ਯੂਨੀਵਰਸਿਟੀ ਦੇ ਸਭਿਆਚਾਰਕ ਪ੍ਰੋਗਰਾਮਾਂ ਦੌਰਾਨ ਵੱਖ-ਵੱਖ ਰੰਗਾਂ ਦੇ ਪੰਜਾਬੀ ਪਹਿਰਾਵੇ ਵਿੱਚ ਸਜੇ ਗੱਡਰੂ-ਮੁਟਿਆਰਾਂ ਦੇ ਉਚਮਟ ਅਤੇ ਕੱਛਣਾਂ ਦੀ ਆਵਾਜ਼ ਪ੍ਰਗਟ ਸਮੇਂ ਦੀ ਜੋੜ ਵਰਗਾ ਮਾਹੌਲ ਬਣਾ ਦਿੰਦੇ ਹਨ। ਯੂਨੀਵਰਸਿਟੀ ਦਾ ਖੂਬਸੂਰਤ ਫੁੱਲਾਂ ਦੇ ਬਗੀਚਿਆਂ ਨਾਲ ਸਜਿਆ ਵਿਹੜਾ ਜਿਸ ਨਾਲ ਆਲਾ-ਦੁਆਲਾ ਮਹਿਕਾਂ ਵੰਡਦਾ ਪ੍ਰਤੀਤ ਹੁੰਦਾ ਹੈ। 2020 ਵਿੱਚ ਖੋਜਿਤ ਹੋਈਆਂ ਐਨ ਆਈ ਆਰ ਐਫ ਰੈਂਕਿੰਗ ਵਿੱਚ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਨੇ 100 ਤੋਂ ਵੱਧ ਸਥਾਨਾਂ ਦੀ ਫ਼ਲਾਂਗ ਲਾਉਂਦੇ ਹੋਏ 64ਵਾਂ ਸਥਾਨ ਹਾਸਿਲ ਕੀਤਾ ਹੈ।

Fig 5: Sample of Restricted Environment

Fig 6: Sample of Unrestricted Environment

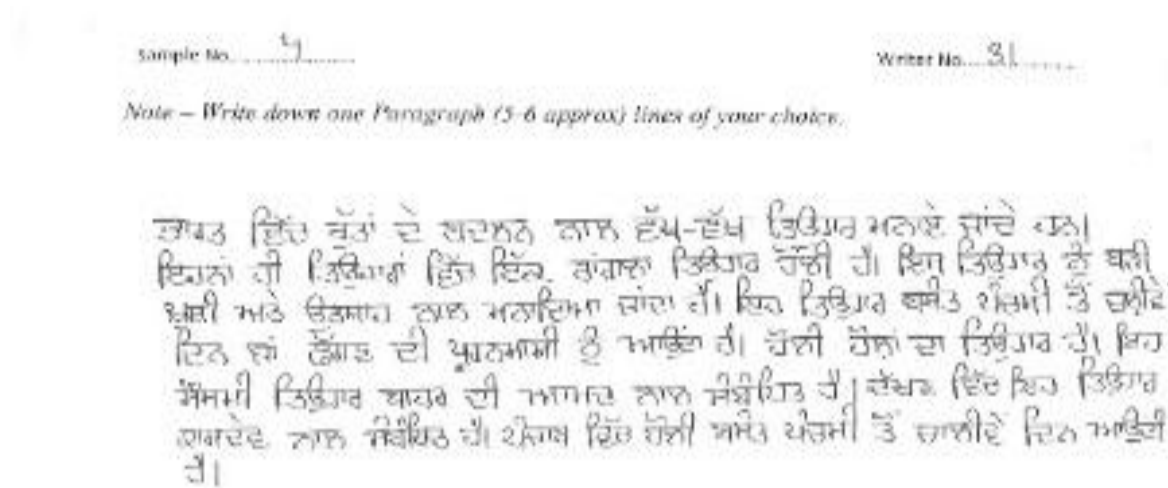


Fig 7: Sample of free paragraph

3.2 Preprocessing

Data preprocessing is the most important task before training any model. Hence in this research work, VGG16 model have used to preprocess the handwriting image. The reason behind to choose this model is that it is a pre trained model which means that this model has been already trained on millions on images so its accuracy of pre-processing the images is much efficient than any traditional preprocessing techniques. VGG model stands for the Visual Geometry Group. Here 16 layers have used of VGG model in which thirteen are convolutional layers, five are Max Pooling layers, and three are Dense layers. No doubt they sum up to 21 layers but it has only sixteen weight layers which are also called as learnable parameters layers. The architecture of VGG16 is shown in Fig 8. The results of preprocessed images in which the original image has been resized to enhance the image quality are shown in the following section. In this report, the three samples have shown of one writer where the dimensions of their original image were 5100 x 6600 for each set of image. After applying VGG16 for preprocessing to the original images as shown in Fig: 9(a), (b), and (c), the size of preprocessed images shown images Fig: 10(a), (b) and (c) have been reduced to 448*448.

Image Resizing

The original image size is $M \times N$, resize it to 224×224 .

$$I_{\text{resized}}(x, y) = I(M/224 \cdot x, N/224 \cdot y) \\ \text{where } x, y \in \{0, 1, \dots, 223\}$$

Mean Subtraction and Normalization

VGG16 was trained on images with certain mean values subtracted and normalized.

Normalization Equation:

$$\text{For an image with channels R, G, B } I_{\text{norm}}(x, y, c) = I_{\text{resized}}(x, y, c) - \text{mean}_c \\ \text{where } c \in \{R, G, B\}.$$

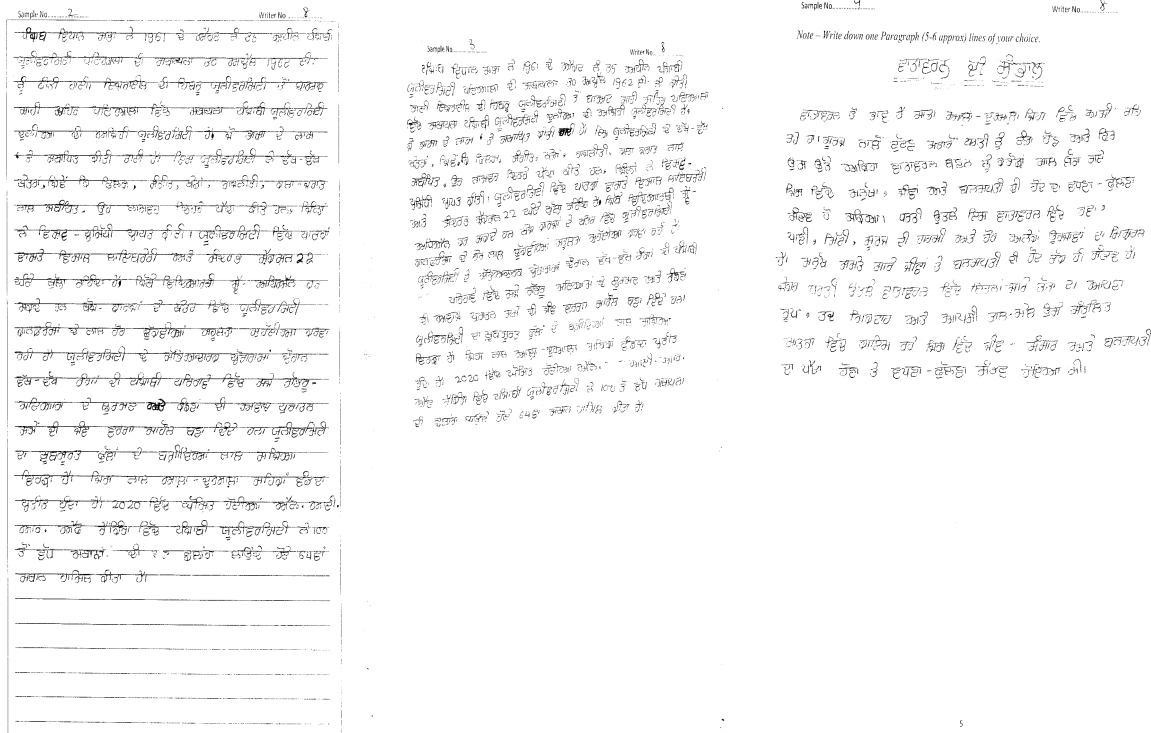


Fig 9: Original Images of Writer Number (a) Restricted (b) Unrestricted (c) Free paragraph Gurmukhi Handwriting Sample

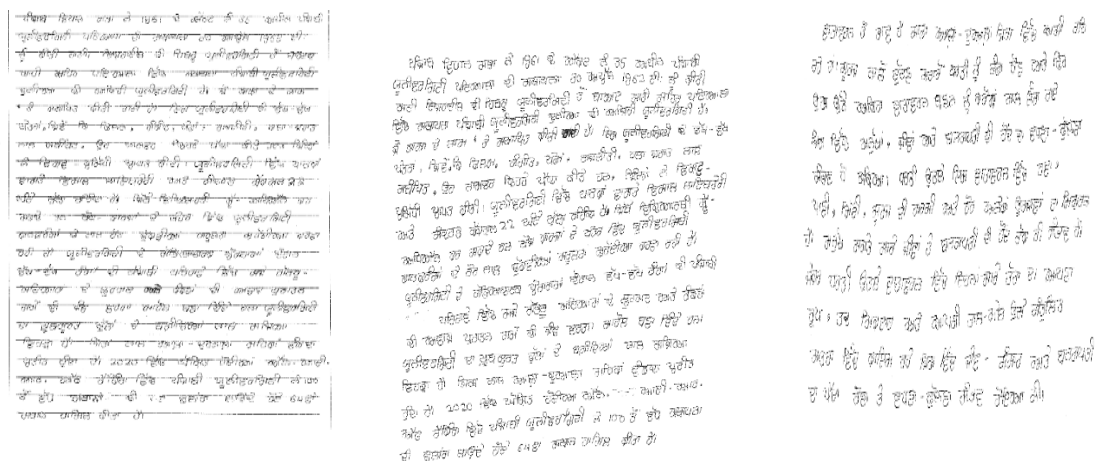


Fig 10: Preprocessed Images of Writer Number (a) Restricted (b) Unrestricted (c) Free paragraph Gurmukhi Handwriting Sample

3.3 Segmentation

The Word segmentation technique for handwritten documents based on contour curve tracking involves identifying and isolating individual words from handwritten text by tracking the contours or outlines of the characters. The step-by-step approach to achieve this includes Contour Detection, Contour Curve Tracking and Word Segmentation. Contour detection algorithm is used to identify the contours of individual characters in the preprocessed image.

Filter out small contours that are likely noise or artifacts. In the next phase Contour Curve Tracking performed by identifying the outer contours of words, which typically form the boundaries of the text regions. For each contour representing a word boundary, track the contour curve to extract the word's outline. This involves finding the bounding box or convex hull of the tracked contour to isolate the word region. The final step is Word Segmentation: Once the word boundaries are identified using contour curve tracking, extract the corresponding regions from the original image. The result of segmentation method depicted in Fig: 11 (a), (b) and (c).

Contour Detection and Tracking

Track the contours to identify the boundaries of the words. Each contour can be represented as a sequence of points: $C = \{(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$

where C represents the contour with n points.

Calculate properties like bounding boxes for each contour:

$$\text{BoundingBox}(C) = \{x_{\min}, y_{\min}, x_{\max}, y_{\max}\}$$

where $x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$ are the minimum and maximum x and y coordinates of the contour points.

Word Segmentation

Segment words based on bounding box properties and heuristics (e.g., size, aspect ratio):

$$\text{SegmentedWord}_i = I_{\text{preprocessed}}(x_{\min} : x_{\max}, y_{\min} : y_{\max})$$

Here, SegmentedWord_i represents the image of the segmented word i, extracted using the bounding box coordinates $(x_{\min} : x_{\max}, y_{\min} : y_{\max})$.

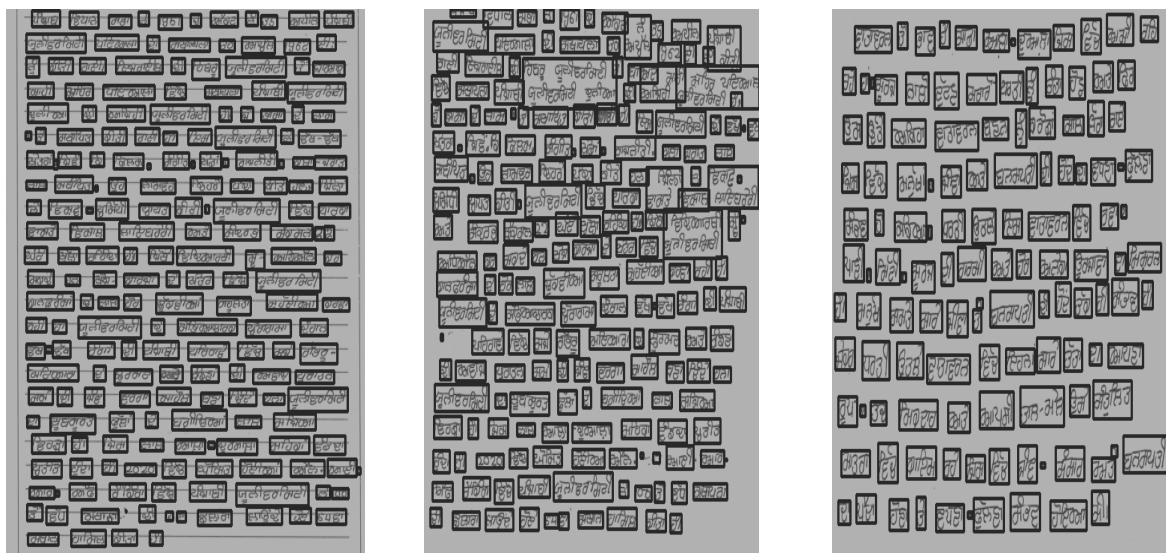


Fig 11: Segmented Images of Writer Number (a) Restricted (b) Unrestricted(c) Free paragraph Gurmukhi Handwriting Sample

The major preprocessing and segmentation steps are shown in following algorithm.

Step 1: Import necessary libraries including **cv2**, **numpy**, **matplotlib**, **seaborn**, **skimage**, **keras**, and others.

Step 2: Define functions:

- a. **showImg(img, cmap)** - Display images using Matplotlib.
- b. **prepareImg(img)** - Convert the given image to grayscale.
- c. **createKernel(kernelSize, sigma, theta)** - Create anisotropic filter kernel.
- d. **find_score(arr, angle)** - Evaluate skew correction score.
- e. **main_processing(ds_path, cd_path)** - Main processing function:
 - i. Set **ds_path** to the dataset directory.
 - ii. Set **cd_path** to the code directory.
 - iii. Create a 'Segmented' directory in **cd_path**.
 - iv. Change the working directory to **ds_path**.
 - v. Loop through each directory and file in the dataset:
 - Call **process_image(directory, file)** function.
- f. **process_image (directory, file)** - Process each image in the dataset:
 - i. Read an image file using **cv2.imread()**.
 - ii. Convert the image to grayscale using **prepareImg function**.
 - iii. Apply binary thresholding using Otsu's method to obtain a binary image.
 - iv. Perform edge detection using the Sobel filter to get output image.
 - v. Correct skew in the output image to obtain new image using the **find_score function**.
 - vi. Resize new image to **448x448 pixels**.
 - vii. Preprocess img4 using **VGG16's preprocess_input** function.
 - viii. Save the preprocessed image using **cv2.imwrite()**.
 - ix. Apply anisotropic filtering to img4 and binary thresholding to get imgThres.
 - x. Find contours in imgThres using **cv2.findContours()**.
 - xi. Create a directory with the same name as the original image in 'Segmented'.
 - xii. Iterate through each contour:
 - Extract the segmented region from img4.
 - Save the segmented region as a separate file in the created directory using **cv2.imwrite()**.
 - Draw a rectangle around the segmented region on img4 using **cv2.rectangle()**.
 - xiii. Save img4 with rectangles drawn around segmented regions using **cv2.imwrite()**.

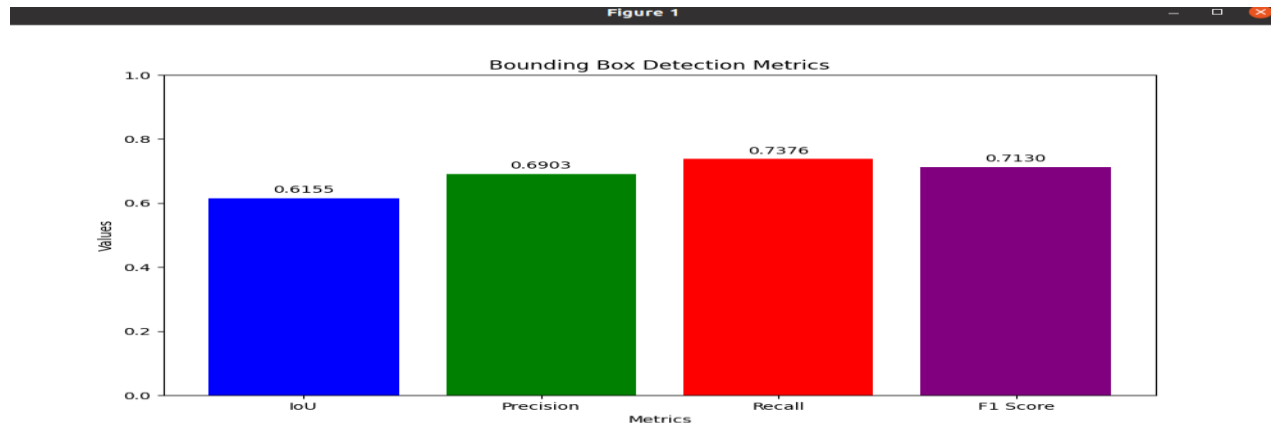
Step 3: Call **main_processing(ds_path, cd_path)** to initiate the processing.

Step 4: End of the program.

3.4 Performance Evaluation

The segmentation technique's performance is evaluated primarily based on its accuracy in a handwriting identification system. Accuracy here pertains to the segmentation technique's capability to precisely identify and isolate individual words within a handwritten document. The analysis of this vital measurement, which is usually presented as a percentage, involves dividing the total words in the document (total words segment) by the number of successfully segmented words (obtained words segment). A higher accuracy score indicates how effectively

the segmentation algorithm divides the handwritten text into words, facilitating a quicker to do further analysis or identification.



On the other hand, a decreased precision score indicates that word segmentation can be difficult, which could lead to errors in subsequently processing phases. Consequently, attaining high segmentation accuracy is essential to ensuring a handwriting identification system's reliability and efficacy since it has a direct impact on the system's ability to accurately analyses and interpret handwritten information. Using the following method, we test our implemented approach on 2100 handwritten Gurmukhi script document images dataset by manual comparison.

$$Accuracy = \frac{\text{Obtained words segment}}{\text{Total words segment}}$$

4. Results and Discussion

This section offers a thorough analysis of the metrics utilized in segmentation based on handwriting of different writers. Table III shows the segmentation task's accuracy of each writer. Every row represents a distinct writer (Writer ID), and the accuracy (%) of their samples (Samples 1,2 and 3) is shown in three columns. The accuracy results show how well the segmentation process worked; higher percentages correspond to better results. The wide range of accuracy differences across writer's points to differences in writing, styles, patterns and perhaps even handwritten sample quality. This table III is an invaluable tool for evaluating and contrasting the handwriting identification skills of different writer, which may be essential for enhancing identification rate and comprehending the difficulties associated with handwritten text analysis. The highest accuracy rate of implemented segmentation technique for our dataset is 100% for various documents.

Reference	Technique	Dataset	Accuracy
Implemented Method	Contour Curve Tracking	2100 handwritten Gurmukhi Script images	71.30%

Jindal P.[28]	Mid-Point Detection Technique	2 handwritten Gurmukhi Script images	65.89%
Tofiq T. A[27]	Projection Based	63000 Kurdish word	64.60%

Table III. Segmentation Outcome of GUR-PRATIMA dataset

5. Conclusion

In this study, a newly curated dataset named GUR-PRATIMA, comprising 2100 handwritten text images utilized for experimental task. VGG16 model has employed for preprocessing tasks, encompassing quality enhancement, resizing, and reduction of image dimensions. Subsequently, a contour tracking-based segmentation method to approximate solutions for word segmentation within the images has been applied. Table III outlines the percentage distribution of each sample attributed to corresponding writers and reports an overall segmentation accuracy of 80.38% for Gurmukhi words. However, a limitation of our study is the inability to determine the accuracy for sample 3 (Free paragraph) due to the absence of a specific text length constraint provided by the writer. In future endeavors, we aim to enhance word extraction by employing advanced segmentation techniques.

Competing Interests: No Conflict of Interest

Funding Information: Not Applicable

Author contribution: The authors have contributed for collection of dataset used in the present research.

Data Availability Statement: Collected data is available with the users and will be made available to the researchers. Research Involving Human and /or Animals: Not Applicable

Informed Consent: The consent is taken from all the subjects for using the data for research.

Reference

- [1] Bal A. & Saha R. (2016). An improved method for handwritten document analysis using segmentation, baseline recognition and writing pressure detection. *Procedia Computer Science*, 93, 403-415.
- [2] Zuo L. Wang Y. & Tan T. (2002, July). Personal handwriting identification based on PCA. In *Second International Conference on Image and Graphics* (Vol. 4875, pp. 766-771). SPIE.
- [3] Suteddy W. Agustini D. A. R. Adiwilaga A. & Atmanto D. A. (2023). End-To-End Evaluation of Deep Learning Architectures for Off-Line Handwriting Writer Identification: A Comparative Study. *JOIV: International Journal on Informatics Visualization*, 7(1), 178-185.

- [4] Sharma S. Gupta S. Gupta, D. Juneja S. Singal, G. Dhiman G. & Kautish S. (2022). Recognition of gurmukhi handwritten city names using deep learning and cloud computing. *Scientific Programming*, 2022, 1-16.
- [5] Rehman A. Naz S. Razzak M. I. & Hameed I. A. (2019). Automatic visual features for writer identification: a deep learning approach. *IEEE access*, 7, 17149-17157.
- [6] Purohit N. & Panwar S. (2023). A Novel Multi-Input Hybrid Deep Convolution Neural Network Methodology for Offline Writer Recognition. *Tuijin Jishu/Journal of Propulsion Technology*, 2023, 1612-1624.
- [7] Roy P. P., Bhunia A. K. Das, A. Dey P. & Pal U. (2016). HMM-based Indic handwritten word recognition using zone segmentation. *Pattern recognition*, 60, 1057-1075.
- [8] Ahmed B.Q., Hassan Y.F., Elsayed A.S. (2023). Offline text-independent writer identification using a codebook with structural features. *PLoS ONE* 18(4): e0284680.
- [9] Dargan S., & Kumar M. (2020, November). Writer identification system based on offline handwritten text in Gurumukhi script. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 544-549). IEEE.
- [10] Dargan S., Kumar M., Garg A., & Thakur K. (2020). Writer identification system for pre-segmented offline handwritten Devanagari characters using k-NN and SVM. *Soft Computing*, 24, 10111-10122.
- [11] Longjam T., Kisku D. R., & Gupta P. (2023). Multi-scripted Writer Independent Off-line Signature Verification using Convolutional Neural Network. *Multimedia Tools and Applications*, 82(4), 5839-5856.3
- [12] dos Santos, R. P., Clemente, G. S., Ren, T. I., & Cavalcanti, G. D. (2009, July). Text line segmentation based on morphology and histogram projection. In *2009 10th International Conference on Document Analysis and Recognition* (pp. 651-655). IEEE.
- [13] He J., & Downton A. C. (2003, August). User-assisted archive document image analysis for digital library construction. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* (pp. 498-502). IEEE.
- [14] Nguyen K. C., & Nakagawa M. (2016). Text-line and character segmentation for offline recognition of handwritten Japanese text. *IEICE technical report*, 115(517), 53-58.
- [15] Arivazhagan M., Srinivasan H., & Srihari S. (2007, January). A statistical approach to line segmentation in handwritten documents. In *Document recognition and retrieval XIV* (Vol. 6500, pp. 245-255). SPIE.
- [16] Su T. H., Zhang T. W., Huang H. J., & Zhou Y. (2007, September). Skew detection for Chinese handwriting by horizontal stroke histogram. In *Ninth international conference on document analysis and recognition (ICDAR 2007)* (Vol. 2, pp. 899-903). IEEE.

- [17] Weliwitage C., Harvey A. L., & Jennings A. B. (2005, December). Handwritten document offline text line segmentation. In *Digital Image Computing: Techniques and Applications (DICTA'05)* (pp. 27-27). IEEE.
- [18] Yin F., & Liu C. L. (2009). Handwritten Chinese text line segmentation by clustering with distance metric learning. *Pattern Recognition*, 42(12), 3146-3157.
- [19] Basu S., Chaudhuri C., Kundu M., Nasipuri M., & Basu D. K. (2007). Text line extraction from multi-skewed handwritten documents. *Pattern Recognition*, 40(6), 1825-1839.
- [20] O'Gorman, L. (1993). The document spectrum for page layout analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 15(11), 1162-1173.
- [21] Kise K., Sato A., & Iwata M. (1998). Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding*, 70(3), 370-382.
- [22] Simon A., Pret J. C., & Johnson A. P. (1997). A fast algorithm for bottom-up document layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3), 273-27.
- [23] Li Y., Zheng Y., Doermann D., & Jaeger, S. (2008). Script-independent text line segmentation in freestyle handwritten documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8), 1313-1329.
- [24] Doermann D., & Jaeger S. (2006, October). A new algorithm for detecting text line in handwritten documents. In *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft.
- [25] Louloudis G., Gatos B., Pratikakis I., & Halatsis K. (2006, October). A block-based Hough transform mapping for text line detection in handwritten documents. In *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft.
- [26] Saha S., Basu S., Nasipuri M., & Basu D. K. (2010). A Hough transform based technique for text segmentation. *arXiv preprint arXiv:1002.4048*.
- [27] Tofiq T. A., & Hussein J. A. (2021). Kurdish Text Segmentation using projection-based approaches. *UHD Journal of Science and Technology*, 5(1), 56-65.
- [28] Jindal P., & Jindal B. (2015, December). Line and word segmentation of handwritten text documents written in Gurmukhi script using mid point detection technique. In *2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS)* (pp. 1-6). IEEE.