Advances in Science, Technology and Engineering Systems Journal Vol. 10, No. 1, 48-59 (2025)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Text Line Segmentation on Myanmar Handwritten Document using Average Linkage Clustering Algorithm

Nilar Phyo Wai [1*], Nu War [2]

[1]*Image and Signal Processing Lab, University of Computer Studies, Mandalay, Mandalay, 05071, Myanmar*

[2]*Faculty of Computer Systems and Technologies, Myanmar Institute of Information Technology, Mandalay, 05071, Myanmar*

ARTICLE INFO

Article history:
Received: 06 January, 2025
Revised: 22 January, 2025
Accepted: 23 January, 2025
Online: 09 February, 2025

Keywords:
Myanmar Handwritten Document
Text Line Extraction
Text Line Segmentation
Connected Component Analysis
Average Linkage Clustering

ABSTRACT

Text line segmentation from document images is a significant challenge in the field of document image analysis. It involves extracting individual text lines from Myanmar handwritten document images to enable text recognition. This task becomes particularly challenging in Myanmar handwritten documents, especially those with irregular or cursive writing styles, due to variations in line spacing, and touching and overlapping characters in Myanmar handwritten documents. This paper proposes a text line extraction method based on an average linkage clustering algorithm for handwritten document images to address segmentation errors caused by characters with inconsistent spacing, different writing styles, and line overlaps due to ascenders and descenders. In this paper, Connected Components (CCs) are extracted by using Connected Component Analysis (CCA) and Anisotropic Gaussian multiscale technique. And then convex-hull computation based on the divide and conquer method is used to re-segment the irregular touching components. Then the text lines are extracted by the proposed system based on an average linkage clustering algorithm that consider both the smaller and larger within-cluster variance. The performance of the proposed method is evaluated using the Pixel and Line Intersection over Union (IU) values, which are found to be 93.27% of Pixel IU and 95.09% of Line IU on dataset I and 92.61% of Pixel IU and 89.90% of Line IU on dataset II, respectively. According to the experimental results based on the existing dataset and their own data set, the proposed system can give a better result than the Density-Based Spatial Clustering and Application with Noise (DBSCAN) clustering algorithm.

1. Introduction

Handwritten Text Recognition (HTR) is a crucial technology that converts handwritten content into digital text, typically for further processing such as information retrieval, classification, or translation. HTR continues to pose significant challenges and remains the focus on the research community's attention. Text line segmentation is a critical task for enhancing the performance of the handwriting recognition process. Despite advancements in text line segmentation methods for handwritten documents in various languages such as English, Chinese, Arabic, and Japanese, a significant need persists to segment text lines of handwritten documents written in the Myanmar language.

Myanmar has a lot of characters and is written in a left-to-right direction. Although it lacks uppercase and lowercase characters, it includes ascenders, descenders, and vowel diacritics. The challenges of Myanmar handwritten documents include cursive writing, diverse writing styles, diacritics, and overlapping and touching characters between text lines. Therefore, the development HTR of Myanmar handwritten document images for the Myanmar language has not been yet. The work presented in this paper extracts the text lines from handwritten document images by considering different styles in the language. Text line segmentation is generally seen as a preprocessing step for tasks

* Corresponding Author: Nilar Phyo Wai, University of Computer Studies, Mandalay, Mandalay, 05071, Myanmar, +959253150565, nilarphyowai@ucsm.edu.mm

https://dx.doi.org/10.25046/aj100106

such as document structure extraction, printed characters or handwritten recognition.

This research focuses on text line segmentation of handwritten documents for undergraduate students in Myanmar. The study works to achieve the segmentation of both long-length and short-length lines in handwritten documents and address the challenges of segmenting overlapping and touching lines.

The contributions of this research paper are summarized in the following:

- Creating a new dataset specific to Myanmar handwritten with detailed annotation and developing an average linkage clustering method tailored to Myanmar handwritten documents.
- Distinct blobs representing characters are obtained using Connected Component Analysis (CCA) and Gradient Magnitude, and then the convex hull refines the segmentation of the extracted connected characters.
- A proposed clustering algorithm extracts the detected components into text lines, handling overlapping and variable-length text lines.

The paper is organized as follows: section 2 describes the literature review of text line segmentation approaches. Section 3 explains the proposed method and then datasets are described in section 4. Section 5 demonstrates the experimental results and section 6 summarizes the research work.

2. Related Work

There are various techniques for text line segmentation in handwritten documents, including top-down and bottom-up approaches, scale-space representation and connected component analysis, and deep learning-based methods. This section discusses a representative of various text line segmentation trends in recent years.

Top-down approaches are analyze the entire document and partition it into smaller segments, such as text lines. In contrast, bottom-up methods begin by grouping smaller elements, such as characters or connected components, to construct text lines. Scale-space with Anisotropic Gaussian was used to obtain blob lines in handwritten images and extract text lines by analyzing connected characters on the detected blob lines. In deep learning techniques, neural networks extract relevant heterogeneous features.

In the paper [1], a text line and word segmentation system are presented using horizontal projection histogram (HPH) for unconstrained handwritten documents. The authors estimated the midpoint from the histogram and filled the gap between two consecutive lines. This algorithm is tested on two datasets which are Meitei Mayek handwritten and English languages. However, this approach varies segmentation accuracy that depends on constraints such as close and touching text line in handwritten documents. In [2], the authors utilized grid projection profiles to detect text areas in horizontal and vertical directions. Text areas in historical Tibetan images were extracted by applying specific rules to connected characters with corner points. The method focused on the straight text line of the Tibetan images.

In [3], the authors employed to construct graphs for the nearest neighbors based on pixel points, and then computed eigenvectors to normalized vectors of neighbors. Spectral clustering to group clusters as text lines based on the vectors. The authors [4] proposed that modified DBSCAN clustering was used to detect the baseline of Arabic handwritten documents using interest points which are obtained from the Extremely Randomized Trees (ERT) method.

The authors [5] proposed an innovative method for segmenting text lines in historical handwritten documents. The approach utilizes a scale-space representation with anisotropic Gaussian filters to identify blob lines. Subsequently, a component tree algorithm is employed to binarize the detected blobs. The final step involves extracting text lines through energy minimization using graph cuts. However, that needs to be the segmentation of multi-skewed text lines. In [6], an anisotropic Gaussian kernel was used to extract blob lines from a historical Manchu handwritten document, and broken blob lines were merged using morphological operations. Connected Component Analysis was performed to extract the text and segment the connected components based on overlapping set theory.

The authors [7] proposed a text lines extraction technique for handwritten images based on four aspects: horizontal and vertical projection, height of the text lines, size of the characters and curved nature of the alphabet. The authors consider the variation in character size caused by different handwriting styles. The problems of segmentation due to spaces of the character or overlapping text lines on account of the ascenders and descenders are reduced in the research work. To separate overlapping lines within the extracted segments, different thresholds are performed based on the average character height and width. The proposed method achieves 85.507% of the MatchScore value. This algorithm improves the performance of text line extraction and attains 99.39% of accuracy, 85.5% of detection rate and 91.92% of F-measure values. However, it is needed to consider touching and skew of text lines.

The authors developed a baseline detection algorithm using upper edges set and a text line segmentation method based on connected region analysis and baseline location [8]. The baselines were defined by expressions of two features, the number of upper edges and the duty ratio in the horizontal position. After that, the connected region with a height exceeding a specified threshold was identified as an adhesion region, which was then further truncated. Regions associated with the same baseline were then grouped to form a text line. The results indicated that segmentation accuracy was high, with strong resistance to distortion, and text line adhesion could be effectively managed. In this paper, distortion and line adhesion present significant but ongoing challenges, and future efforts could address these issues in two ways: learning-based segmentation methods and incorporating contextual semantics into text line segmentation.

In [9], author introduced VML-UTLS, an unsupervised convolutional network designed to learn proximity and similarity features within Arabic handwritten documents. This approach labels text lines at the pixel level by applying an energy minimization framework combined with detected blob lines.

The author proposed a method for predicting text line structure using Fully Convolutional Networks (FCNs). Their approach employed a line adjacency graph (LAG) to effectively segment components that intersect multiple text lines. While this model demonstrates robustness across various languages and can handle skewed text lines, it still needs to process complex natures involving touching components [10].

The Segmentation of text lines in heterogenous handwritten documents using deep learning techniques is robust and high-performance that requires large amounts of annotated label data which expensive and demands significant resources and time. The authors [11], presented text line extraction from historical documents using neural network techniques based on the Mask-RCNN network. This network is used to perform instance segmentation and to avoid misalignment between the region of interest and the extracted features. It was trained with the DRoSB dataset, French documents. To compare the performance of Mask-RCNN with U-Net networks, the cBAD 2017 database is used. The pixel and object level metrics are assessed for the performance evaluation. In hyperparameter turning, it is found that small scales and horizontal ratios gave better performance. In this paper, many experiments are conducted on Mask-RCNN, U-Net, and Doc-UFCN networks on different datasets. The results showed a better performance for Mask-RCNN that delivered with manageable computational demands, making it a practical choice for real-world text segmentation.

In [12], a neural network-based open-source Tesseract OCR engine is applied to recognize Myanmar handwritten characters. To convert binary images and reduce noises, Otsu's binarization algorithm and median filter method are used for pre-processing. The connected components labelling method is used for detecting text regions. Each connected component is bounded by a block of equal size. Despite some research on Myanmar character recognition, it remains in its early stages, with only limited literature available.

The text line segmentation system is presented for text line segmentation of challenging handwritten document images. The manuscript images contain narrow interline spaces with touching components, interpenetrating vowel signs and inconsistent font types and sizes. In addition, they contain curved, multi-skewed and multi-directed side note lines within a complex page layout. Therefore, bounding polygon labelling would be very difficult and time consuming. Instead of relying on line masks that connect the components on the same text line, these line masks are predicted using a Fully Convolutional Network (FCN). In the literature, FCN has been successfully used for text line segmentation of regular handwritten document images. This paper demonstrates that FCN is useful for challenging manuscript images as well. Using a new evaluation metric that is sensitive to over segmentation as well as under segmentation, testing results on a publicly available challenging handwritten dataset are comparable with the results of a previous work on the same dataset [13].

The authors [14] proposed an accurate text line segmentation in the presence of both cleanly written and struck-out text by addressing a crucial challenge in handwritten document processing. The approach effectively integrates stroke width estimation, noise filtering, morphological operations, and a DenseNet-based deep learning model to distinguish struck-out components from clean text. Additionally, spatial features and component directionality are utilized to form text lines progressively. The study's strength lies in its novel dataset, which specifically includes struck-out text, making it more representative of real-world scenarios. The experiments conducted on both the new dataset and standard benchmarks (ICDAR2013 and ICDAR2019 HDRC) demonstrate superior performance compared to existing methods. However, the approach's reliance on deep learning models may introduce computational overhead, which could be a limitation in resource-constrained environments. Furthermore, while the proposed methodology is effectiveness, its adaptability to various handwriting styles, languages, and extreme document degradation remains an area for further exploration.

Previous research works focus on different handwriting styles, overlapping text lines, multi-touching text lines, skewed text lines on handwritten document images. In this paper, an unsupervised clustering algorithm with average linkage is designed for text line segmentation in Myanmar handwritten document images, which contain cursive writing, overlapping and skewed text lines, and characters of varying scales. This approach enhances segmentation performance by effectively handling overlapping and skewed text lines. The proposed system obtains the higher performance on the segmentation of different line spacing, both short and long-length text lines, overlapping and skewed text lines.

3. Proposed Methodology

The steps of the proposed system are summarized as pre-processing, scale-space representation and gradient magnitude calculation, CCs extraction with CCA, text line extraction using a proposed algorithm, and then segments touching CCs. The design of the proposed system is shown in Figure 1.

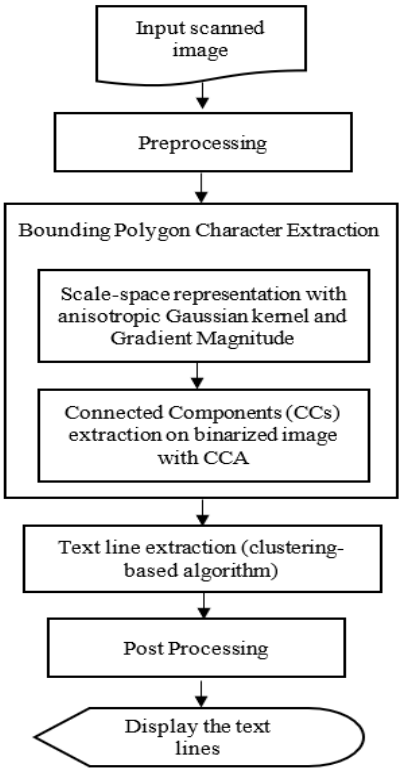


Figure 1: The design of the proposed system

3.1. Preprocessing

Pre-processing steps are essential for enhancing the performance of text line segmentation in scanned handwritten images of undergraduate students in Myanmar. The approach

incorporates several pre-processing techniques, including removing lines, red marks, noise, and binarization on and cropping the input handwritten image.

Firstly, the input scanned image is converted to the Hue, Saturation, and Value (HSV) color space is better isolate specific colors., saturation, and value (HSV) color thresholding, which isolates red hues within defined HSV ranges and then removes them with inpainting, which is filled the detected regions with surrounding pixel values to restore the original appearance of the input image. After red mark removal, morphological closing operations are applied using two line-shaped kernels: a horizontal kernel to remove underlines and a vertical kernel to remove margin lines. Next, median filtering addresses bleed-through, stains, and salt-and-pepper noise. The input scanned handwritten images of undergraduate Myanmar students are shown in Figures 2 and 3. The pre-processed images are described in Figures 4 and 5.

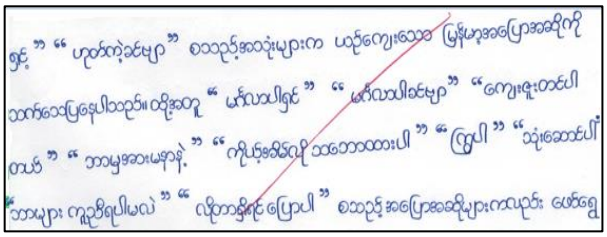


Figure 2: The scanned handwritten image of Dataset I

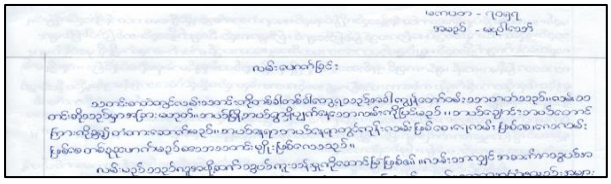


Figure 3: The scanned handwritten image of Dataset II

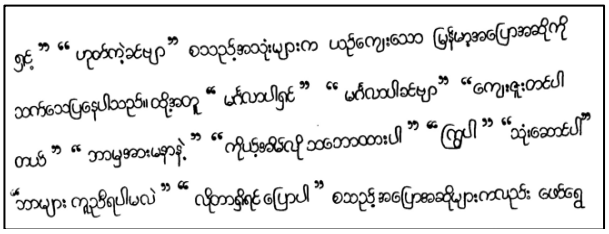


Figure 4: Preprocessed image of Dataset I

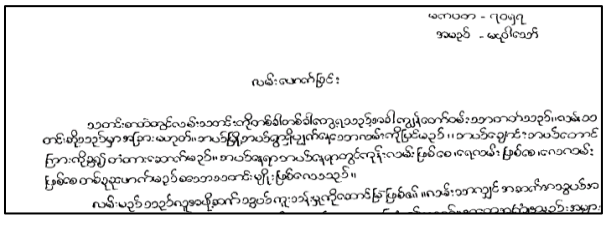


Figure 5: Preprocessed image of Dataset II

3.2. Scale-space Representation with Directional Gaussian Kernel

Scale-space representation is widely used multiscale technique for analyzing unknown image structures at various levels of detail. The scale-space representation of the image uses a directional Gaussian kernel with appropriate scale aims to obtain coarse and fine image structuring. It can also retain the significant features of the image. The choice of scale crucially influences the resulting scale-space representation. An anisotropic Gaussian kernel, as described in (1), is applied to the input handwritten image to produce the scale-space image..

$$G(x, y; \sigma) = \frac{1}{2\pi\sigma_x\sigma_y}\, e^{-\left(\frac{x^2}{2\sigma_x{}^2} + \frac{y^2}{2\sigma_y{}^2}\right)} \tag{1}$$

The image scaled in the horizontal direction by convolving it with a Gaussian kernel $G(x, y; \sigma_x)$ in (2).

$$I_x = G(x, y; \sigma_x) * I(x, y) \tag{2}$$

where σ_x represents a standard deviation along the x-axis. To scale the image vertically, it is convolved with a Gaussian kernel $G(x, y; \sigma_y)$, as specified by (3).

$$I_y = G(x, y; \sigma_y) * I(x, y) \tag{3}$$

where σ_y represents a standard deviation along the y-axis.

The concepts of gradient orientation and gradient magnitude is essential for extraction information from digital images, enabling various applications such as object recognition, image segmentation, and edge detection. The gradient magnitude is computed for the scale-space representation in (4), emphasizing regions of significant intensity changes are essential for identifying structural features in both directions.

$$I(x, y) = \sqrt{I_x{}^2 + I_y{}^2} \tag{4}$$

3.3. Coarse Segmentation

Connected Component Analysis (CCA) is a fundamental technique in image processing used to identify and label distinct regions within a binary image. It leverages gradient magnitude from scale-space images to accurately identify and group pixels of connected components (CCs), such as text characters, based on their intensity and boundary features. This approach ensures the effective extraction of components to overcome the challenges of faint characters, touching characters, or noisy backgrounds. Binarization of the gradient image uses the adaptive thresholding method to identify foregrounds (text) and backgrounds. In Figure 6, CCA is applied to the binarized image to detect and extract a group of connected pixels (x_{min}, y_{min}, x_{max} , x_{max}) into distinct CCs using 8-connectivity to ensure accuracy.

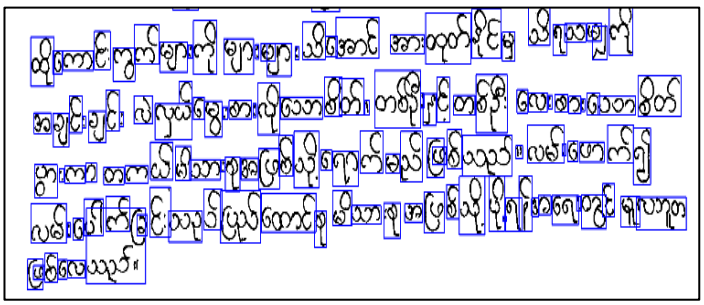


Figure 6: CCs extraction on the preprocessed image

3.*4 Fine Segmentation*

After extracting connected characters from distinct blobs in handwritten document images, some bounding boxes may be incorrect due to overlapping (almost touching) components, as illustrated in Figures 7 (a) and 8 (a).

3.4.1 Correction of CCs

The median value of CCs is calculated to combine (◌့, ◌်, and ◌း) and detect the irregular bounding box by the following equation.

M=median {all height of connected characters on the entire document}
H=height of each connected character

Two sets are defined to combine (◌့, ◌်, and ◌း) from the connected characters extraction.

The set of small CCs is defined as:

$$S = \{s_1, s_2, s_3 \ldots, s_a | \, \mathrm{H}\,(s_i) < \frac{M}{2}\} \quad (5)$$

where: a is number of small connected components

The set of large CCs is defined as larger than equal of M

$$L = \{l_1, l_2, l_3 \ldots, l_b | \, \mathrm{H}\,(l_i) \geq \mathrm{M}\} \quad (6)$$

where: b is the number of large connected components

Find the horizontal proximity and vertical proximity (between CCs from S set and L set between 10 and 20 to combine CCs as Figures 7(b).

(a) (b)

Figure 7: (a) Before and (b) After combination of some CCs

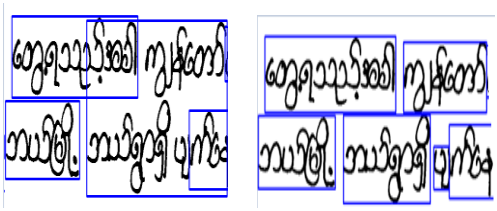


Figure 8: (a) Before and (b) After applying convex hull to segment the almost-touching CCs

And in this system, the irregular bounding boxes are considered by using the following three conditions

(i) Cursive writing case if (H> 2 M and M >=60)
(ii) Overlapping writing case if (H> 2.5 M and M <60)
(iii) Untouched Consonants Case (Ascenders & Descenders) if (H >3 M and M <50)

Connected Components (CCs) fit the above three conditions which must be considered irregular and continued.

To correct the bounding box each CCs from the irregular Connected Component list, the system extracts the boundaries of connected characters or the external contours (endpoints of horizontal, vertical, or diagonal segments). For finding the convex hull of a set of points in irregular list, the divide-and-conquer algorithms is used to compute the convex hull and works by recursively dividing the set of points into two halves, computing the convex hull for each half, and then merging the two hulls together [15]. The correct bounding boxes is shown in Figure 8 (b).

3.4.2 Finding Distance Threshold

After correction of bounding box, the system extract height of CC from the updated Connected Component list of each document as below:

$$\mathrm{H} = \{h_1, h_2, h_3, \ldots \ldots \ldots, h_n\}$$

threshold = Median (H)

The procedure of proposed text line extraction based on average linkage clustering method is described as following:

Proposed Algorithm

Result: Text line of hierarchical clustering records C_{new_k} where k is the number of text lines

Input: The bounding box of the connected component b_i, where i=1, 2, …. …, n; number of bounding boxes of handwritten document image

$$b_i = \{x_{i_{start}}, x_{i_{end}}, y_{i_{start}}, y_{i_{end}}\}$$

begin Initialize each cluster of handwritten document image and threshold

$$mc_i = \frac{(y_{i_{start}} + y_{i_{end}})}{2}, where: \; i = 1, 2, \ldots, n$$

C = {C₁, C₂, …, Cn} = {{ mc_1 },{ mc_2 }, …., { mc_n }}
where n is the number of clusters
Minimum distance: $d_{min} = \infty$
k=1

while length (C) > 1 **do**
for i=1 to (length (C) - 1) **do**
for j= (i+1) to length (C) **do**
Calculate distances using average linkage between C_i *and* C_j :

$$d = \frac{1}{|C_i| \times |C_j|} \sum_{mc_i \in C_i} \sum_{mc_j \in C_j} D(mc_i, mc_j)$$

if $d < d_{min}$ **then** $d_{min} = d$
if $d_{min} < threshold$ **then**
Record the cluster index for merging:
Merge_i= i, Merge_j=j
Merge cluster $C_{\mathrm{Merge_i}}$ and $C_{\mathrm{Merge_j}}$: $C_{new} = C_{\mathrm{Merge_i}} \cup C_{\mathrm{Merge_j}}$
Remove $C_{\mathrm{Merge_i}}$ and $C_{\mathrm{Merge_j}}$ from C.
Append C_{new} to C.
end
end
end
end
Record $\{V(b_i) | b_i \in C_{new_k}\}$: $C_{new_k} = C_{new}$
k=k+1
end

Final Merging for larger and smaller within-cluster variance (Very small handwritten style or high line skew):

find y_{min}, y_{max} *and* y_{mid} for C_{new_k} cluster

$y_{mid_m} = \frac{y_{min_m} + y_{max_m}}{2}$ where: m=1, 2, 3, ……, k

for i= 1 to k-1

j=i+1

$distance = \left| y_{mid_i} - y_{mid_j} \right|$

if $(distance \leq threshold)$ **then**

merge cluster $C_{\text{new\_i}}$ and $C_{\text{new\_j}}$

end

end

update C_{new_k}

end

The inputs of proposed algorithm are $x_{start}, x_{end}, y_{start,}, y_{end}$ coordinate points of the detected characters. The algorithm initializes the center point of the detected characters as its own cluster. In the matrix, the different values of all clusters are saved and then found the minimum values of distance are found. If the minimum values are smaller than the threshold, they will form as new cluster. To update the distance between clusters, average linkage is used. Then, minimum distance is found again in the clusters and forming clusters are combined with other clusters until the clusters are smaller than the threshold. In the combining merge clusters, the midpoint is calculated and merged if it is smaller than the threshold. Finally, the algorithm provides the clusters of text lines from handwritten document images.

The proposed algorithm employs an unsupervised hierarchical clustering approach with average linkage. This clustering method begins by treating each data point as its own cluster and iteratively merges the closest clusters within the range of the distance threshold. The hierarchical clustering is influenced by the choice of linkage criteria, with average linkage being a popular and effective method for clustering data points. Average linkage is the method that involves considering the distances between all pairs of points in the two clusters and averages all of these distances. It is applied to cluster similarity of spatial constraints on time series [16] and to determine elements in coal [17].

The proposed algorithm merges the extracted CCs as text lines using average distance linkage to efficiently extraction on text line of Myanmar and Malayalam handwritten document images [7] shown in Figures 9 and 10.

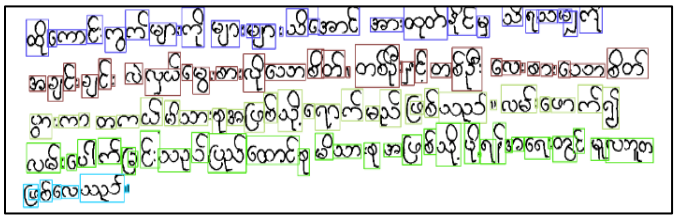


Figure 9: Text line extraction with the proposed algorithm

3.4. Post Processing

After text line extraction, the irregular bounding boxes, defined by the above three conditions, are segmented to improve the performance of text line segmentation based on the height of CCs of nearest clusters, as shown in Figure 11.

Each color represents a specific text line in the clustering results, providing a clear visual distinction for segmenting text lines in the handwritten document.

Figure 10: Text line extraction on the Malayalam handwritten document

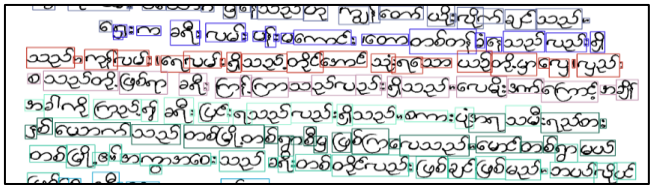


Figure 11: Text line segmentation on the Myanmar handwritten document

4. Myanmar Handwritten Document Images Dataset

4.1. Characteristics of Myanmar language

Myanmar language consists of 33 consonants, 12 vowels, 4 medials, 10 special characters, double-layered (Pali) characters, digits, and punctuation marks as shown in Figure 12. Some fundamental consonants in the Myanmar language can be standalone words. One must combine a consonant with one or more additional characters to create new words. It is written from left to right and read from top to bottom. Although it does not distinguish between uppercase and lowercase characters, it includes ascenders, descenders, and diacritics as shown in Figures 13 and 14.

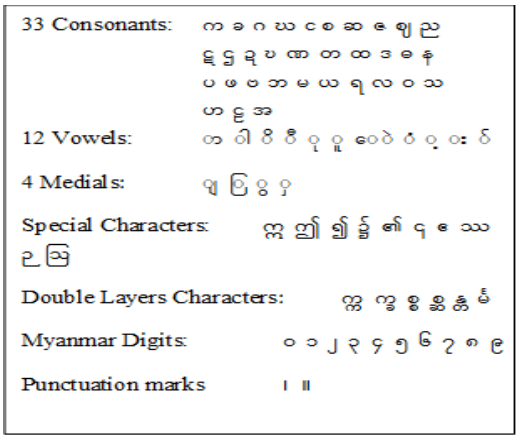
4.2. Dataset

There is no benchmark dataset for segmenting text lines for Myanmar, even though many other languages have datasets of heterogeneous handwritten documents for document analysis. Therefore, in this work, Myanmar Handwritten documents are gathered by undergraduate students to create datasets. These documents were scanned at 300 dpi resolution. The text lines in the dataset are annotated using the Aletheia ground-truth tool [18] to obtain tight polygons of x and y coordinate points for each text line in PAGE-XML format. The Myanmar handwritten dataset contains Myanmar's alphabets, proverbs, prose, essays, numbers, adopted words, open and closed quotes, and some Pali words.

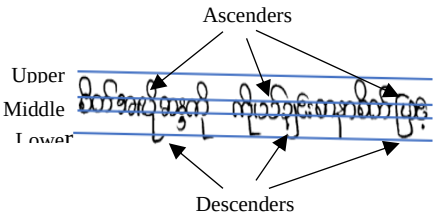
There are two formats of datasets: Dataset I) and Dataset II. The Dataset I includes simple handwritten documents and Dataset II contains challenging handwritten documents.

- Dataset I, contains 4161 text lines across 287 documents (images) covering various topics. The documents were written by 10 students from Myitkyina Education College. This

dataset includes the overlapping 50 lines and 39 skewed text lines.



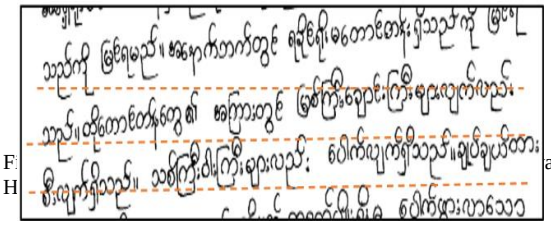
- Dataset II comprises 2664 text lines across 118 documents (images) on the same topics, written by 32 students from



UCSM. This dataset includes 654 overlapping text lines, 79 skewed text lines, and 20 touching lines, and narrow line spacing which causes segmentation errors.

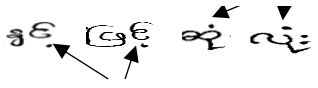
Figure 12: Myanmar Alphabets

Figure 14: Sample of Vowel diacritics on Myanmar Handwritten Word



F anmar H

Figure 15: Inconsistent Spacing, Overlapping and Skewed Text Lines on Myanmar Handwritten Document



Vowel diacritic (outkamyint)

The characters, ascenders, descenders, and diacritics, narrow spacing are shown in Figures 13 and 14. These features introduce vertical complexity, making accurate segmenting of individual text lines more challenging, especially in Dataset II as shown in Figure 15.

The next section will discuss about evaluation metric and experimental details using the proposed approach. After that, the findings and discussion will be described.

5. Experiments

In this section, the performance of the proposed method and Density-Based Spatial Clustering and Application with Noise (DBSCAN) can be evaluated on the 405 documents for Myanmar Datasets and 293 Malayalam handwritten document images of LIPI-Database [7].

DBSCAN [19] groups data points are close to each other based on a density criterion, making it particularly effective for clustering irregularly shaped clusters and those with varying densities. The text line extraction of handwritten documents can adapt well to skewed or curved lines, commonly found in natural or cursive handwriting.

The DBSCAN algorithm clusters the center point of the y coordinate of each CC based on epsilon (ε) and minPts to extract text lines of the input image.

- Epsilon (ε): Half of the distance threshold for clustering (distance threshold is the median height of the connected characters of each handwritten document).
- minPts: this parameter requires at least two points to form a cluster.

The proposed method and DBSCAN function as unsupervised learning algorithms that do not require prior knowledge of the number of clusters.

5.1. Evaluation Metrics

Among many evaluation metrics for text line segmentation of handwritten documents, Pixel and Line Intersection over Union (IU) [20] are used to evaluate text line segmentation on the Myanmar handwritten datasets.

Pixel IU defines the accuracy of line detection at the pixel level in (7). In this case, TP represents the correctly detected pixels, FP refers to additional pixels erroneously identified, and FN indicates missed pixels.

$$Pixel\ IU\ = \frac{TP}{TP+FP+FN} \tag{7}$$

Line IU is a metric to evaluate how accurately individual lines have been detected in (8). Specifically, it measures the proportion of correctly identified lines. A threshold value of 0.75 is used to evaluate the experimental results.

$$Line\ IU = \frac{Intersection}{Union} \tag{8}$$

Intersection: Intersection is the number of correctly detected lines (true positives, TP).

Union: Union is the total number of lines in both the detected and ground truth sets, which include true positive (TP), false positive (FP) is extra text lines that were wrongly detected, and false negative (FN) indicates missed text lines. Precision, recall and F-measure are defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

$$\text{F-Measure} = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{11}$$

5.2. Experimental Results

The distance threshold for each handwritten document is defined by the median height of the document's CCs. The threshold values are different based on individual handwriting styles, which impact on the performance of text line extraction with the clustering algorithms.

The following tables present the comparative analysis of different scale-space representation methods used for text line extraction: Gradient Magnitude with Anisotropic Gaussian, Laplacian of Anisotropic Gaussian, Laplacian of Gaussian (LoG), and Difference of Anisotropic Gaussian. The three methods are evaluated with two scales, (2,1) and (4,2), representing the sigma values applied in the x and y directions, respectively. LoG with automatic scale selection uses a minimum sigma of 1 and a maximum scale of 2.

Gradient Magnitude with Anisotropic Gaussian method on the clustering algorithms demonstrates good performance compared to other scale-space representation methods, particularly in challenging Myanmar handwritten document images, as shown in Tables 4 and 5. It also provides good result for Malayalam handwritten document images of LIPI Database in Table 6. Additionally, for simpler handwritten documents, the LoG yields more precise results than alternative approaches, as indicated in Tables 1 and 2. The difference of Anisotropic Gaussian is faster than other approaches in execution time in Tables 3 and 6. The execution time in Gradient Magnitude scale-space is faster than other methods in Table 9.

The proposed method uses average linkage to cluster the midpoints of CCs, so the method can perform on overlapping and inconsistent spacing text lines well. The proposed method on the Gradient Magnitude scale-space outperforms the results of text line segmentation with 95.09%, 89.90%, and 93.24% of Line IU on two datasets of Myanmar and LIPI database in Tables 2, 5, and 8 respectively. The results of DBSCAN are in 97.77%, 81.51%, and 87.68% for the evaluation of Line IU on those datasets in these three tables. Although DBSCAN has performed well in simple handwritten documents, the performance of challenging datasets is less efficient than the proposed clustering algorithm.

In Tables 1,4 and 7 are compared the Pixel IU of proposed method is compared with DBSCAN of Myanmar and Malayalam handwritten document images (LIPI) dataset. The results of Pixel IU of proposed method on the Gradient Magnitude scale-space are obtained 93.38%, 92.65% and 93.24% respectively. The DBSCAN method are found 96.01%, 89.63 % and 95.40% in Pixel IU in these three tables. The proposed method solves the problem overlapping text lines and variable length of text lines by using the integration of scale-space representation with average linkage clustering approach. The idea improves the segmentation results of different handwriting styles, varying line spacing and extracts different scales in handwriting.

The proposed method achieves higher recall by detecting more lines, however it introduces more false positives. Overall, text line segmentation of the proposed method is more effective, as indicated by its higher F-measure, balancing both precision and recall. The reliance of DBSCAN on fixed density threshold limits its ability to adapt to the variability of text line spacing and the presence of ascenders, descenders, and diacritics like Myanmar and Malayalam script. In comparison, the proposed method enables handling the varying structure of text lines, resulting in precision, recall, and F-measure values, particularly in complex datasets in Table 10. Experiments have been conducted on the HP Core i7 (12th-gen) system, using PyCharm Community 2020 and Python 3.9.

Table 1: Comparison results of Pixel IU on Myanmar Dataset I

Scale-Space Representation Methods	DBSCAN (%)	Proposed (%)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	74.50	66.07
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2) (σ_{x2}=2, σ_{y2}=1)	85.26	78.70
Laplacian of Gaussian (LoG) $minimum_\sigma = 1$ and $maximum_\sigma$=2	**96.01**	**96.80**
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	89.69	93.38
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	90.06	93.27

Table 2: Comparison results of Line IU on Myanmar Dataset I

Scale-Space Representation Methods	DBSCAN (%)	Proposed (%)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	84.61	69.94
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2) (σ_{x2}=2, σ_{y2}=1)	89.22	80.04
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	**97.77**	**97.82**
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	94.83	94.59
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	93.31	95.09

Table 3: Comparison results of execution time in second on Myanmar Dataset I

Scale-Space Representation Methods	DBSCAN (sec)	Proposed (sec)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	684.15	5614.66
Difference of Anisotropic Gaussian ($\sigma_{x1} = 4, \sigma_{y1}$=2) ($\sigma_{x2}$=2, σ_{y2}=1)	396.93	**324.68**
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	376.93	399.70
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	362.50	369.49
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	356.78	453.91

Table 4: Comparison results of Pixel IU on Myanmar Dataset II

Scale-Space Representation Methods	DBSCAN (%)	Proposed (%)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	76.55	62.09
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2) (σ_{x2}=2, σ_{y2}=1)	86.59	73.96
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	83.23	90.31
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	89.26	**92.65**
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	**89.63**	92.61

Table 5: Comparison results of Line IU on Myanmar Dataset II

Scale-Space Representation Methods	DBSCAN (%)	Proposed (%)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	68.33	55.16
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2) (σ_{x2}=2, σ_{y2}=1)	77.63	67.34
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	76.95	87.89
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	81.00	89.60
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	**81.51**	**89.90**

Table 6: Comparison results of execution time in second on Myanmar Dataset II

Scale-Space Representation Methods	DBSCAN (sec)	Proposed (sec)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	395.57	417.35
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2)(σ_{x2}=2, σ_{y2} = 1)	**185.00**	187.06
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	203.32	210.27
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	200.54	211.08
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	206.07	207.85

Table 7: Comparison results of Pixel IU on LIPI Database

Scale-Space Representation Methods	DBSCAN (%)	Proposed (%)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	89.73	70.87
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2)(σ_{x2}=2, σ_{y2} = 1)	84.78	86.92
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	94.21	**95.69**
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	95.21	94.56
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	**95.40**	94.85

Table 8: Comparison results of Line IU on LIPI Database

Scale-Space Representation Methods	DBSCAN (%)	Proposed (%)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	84.18	47.17
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2)(σ_{x2}=2, σ_{y2} = 1)	92.29	74.02
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	**92.85**	82.59
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	87.28	92.67
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	87.68	**93.24**

Table 9: Comparison results of execution time in second on LIPI Database

Scale-Space Representation Methods	DBSCAN (sec)	Proposed (sec)
Laplacian of Anisotropic Gaussian (σ_x=2, σ_y=1)	6013.87	913.82
Difference of Anisotropic Gaussian (σ_{x1}=4, σ_{y1}=2)(σ_{x2}=2, σ_{y2} = 1)	706.10	684.93
Laplacian of Gaussian (LoG) $minimum_\sigma$=1 and $maximum_\sigma$=2	342.31	285.64
Proposed Scale-Space Gradient Magnitude (σ_x=2, σ_y=1)	**261.03**	268.29
Proposed Scale-Space Gradient Magnitude (σ_x =4, σ_y=2)	1730.14	913.82

Table 10: Evaluation results of text line segmentation for proposed scale-space representation on Precision (P), Recall (R) and F-Measure (FM) on scale σ_x =4, σ_y=2

Dataset	Methods	No. of Text Lines	Correctly Text Lines	P (%)	R (%)	FM (%)
Myanmar Dataset I with 287 documents	DBSCAN	4161	4067	97.45	97.76	97.61
	Proposed Algorithm		4083	95.55	98.14	96.83
Myanmar Dataset II with 118 documents	DBSCAN	2664	2204	93.46	82.73	87.77
	Proposed Algorithm		2500	90.67	93.84	92.23
LIPI Database with 293 documents [7]	DBSCAN	5658	4937	91.52	92.95	92.23
	Proposed Algorithm		5285	97.97	93.40	95.63

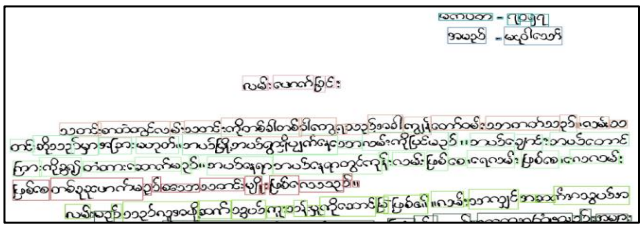


Figure 16: Text line segmentation of Gradient Magnitude with Anisotropic Gaussian on scale (2, 1)

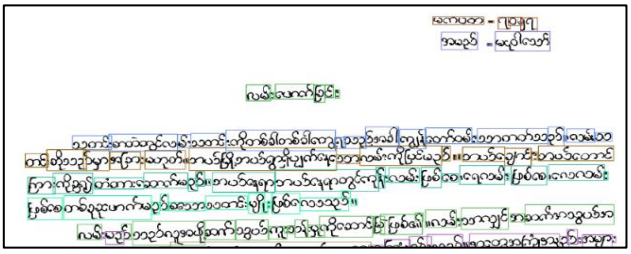


Figure 17: Text line segmentation of Gradient Magnitude with Anisotropic Gaussian on scale (4, 2)

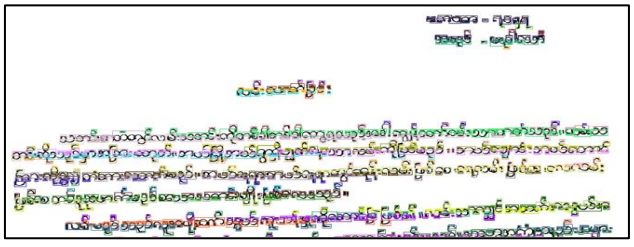


Figure 18: Text line segmentation of Laplacian of Anisotropic Gaussian on scale (2,1)

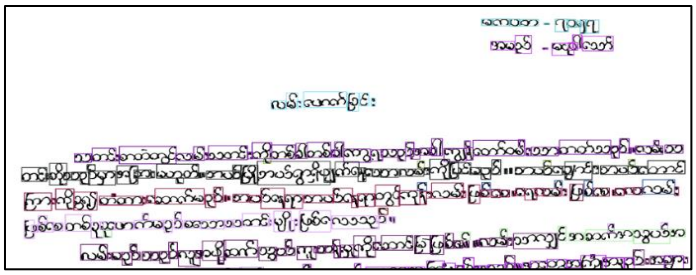


Figure 19: Text line segmentation of Difference of Gaussian with Anisotropic Gaussian on scales (4,2) and (2,1)

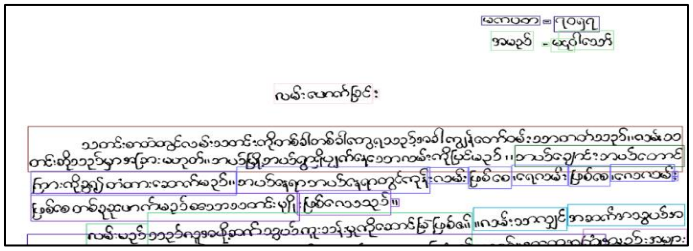


Figure 20: Text line segmentation of LoG with a minimum scale of 1 and maximum scale of 2

5.3. Finding and Discussion

Scale selection plays a crucial role in extracting connected components (CCs) of handwritten text using scale-space representation methods. The choice of scale significantly impacts the detection of CCs in handwritten document images. The selection of scale and the combination of certain characters significantly influence the determination of threshold value for clustering CCs as text lines.

Ascenders, descenders, open or closed quotes, and writing styles often present challenges in clustering text lines within handwritten document images.

Although Dataset I is relatively simple, the presence of open and closed quotes, along with a small writing style (SWS), causes segmentation errors. Oversegmentation of text lines typically occurs under two conditions: first, when ascenders and descenders are written separately with varying spacing; and second, when the small writing style (SWS) introduces inconsistencies oversegmentation. The cursive wrting style (CSW) and touching characters cause undersegmentation. The proposed method solves overlapping lines, but there is still work to be done for highly skewed lines and mutiple touching lines. In the future, it will be necessary to address the challenges related to highly skewed text lines. multiple touching characters. The text line segmentation of different scale-space representation methods with the proposed algorithm are demonstrated in Figures 16, 17, 18, 19 and 20.

The proposed method is designed for Myanmar and, due to the nature of its features, can be successfully applied to languages of a similar nature. However, it does not work well for other languages. The accuracy of text line extraction algorithms on the Malayalam handwritten document image database, LIPI, is presented in Table 11. It achieves a higher accuracy of 97.97%, compared to 58.19%, 26.7%, and 87.7% for the A* Path Planning algorithm [21], the piecewise painting algorithm [22], and the

horizontal and vertical projection method, respectively, on the LIPI database. The proposed algorithm is also effective in extracting overlapping and short text lines from Malayalam handwritten document images in the LIPI database.

Table 11: Comparison of accuracy obtained for different existing text line methods and the proposed method on LIPI database

No	Algorithm for Text Line Extraction	No. of Correctly Segmented Text Lines	Accuracy (%)
1	A* Path Planning	3258	58.19
2.	Piecewise Painting	1495	26.7
3	Horizontal and Vertical Projection	4912	87.7
4	Proposed Method	5285	97.97

6. Conclusion

Text line extraction from handwritten documents is challenging because of the different handwriting styles. The proposed system addresses overlapping of text lines in extracted segments and varying gaps between characters due to the different handwriting styles, particularly when descenders and ascenders are written below one another, which can lead to incorrect segmentation of a single line into two. The input handwritten image is extracted as connected characters with scale-space technique and CCA. These connected characters are merged into text line using the proposed clustering algorithm. The system resolves these issues, achieving 93.27% and 95.09% match between extracted text lines and ground truth lines from Dataset I and 96.21% and 89.90% on Dataset II, as evaluated using the Pixel IU and Line IU metrics. Myanmar handwritten datasets for text line segmentation have created 405 images, and ground truth images for 6825 text lines. The proposed method is the first phase in digitizing handwritten document images of undergraduate Myanmar students. In the future, text recognition of the handwritten image will be performed on deep learning techniques.

Conflict of Interest

The authors declare no conflict of interest.

Author Contribution

The major portion of the work presented in this paper was carried out by the first author, Nilar Phyo Wai, under the supervision of the second author, Nu War also performed the data analysis, implementation, validation, and preparation of the manuscript.

Acknowledgment

I would like to thank Dr. Nu War, Professor of the Faculty of Computer Systems and Technologies at Myanmar Institute of Information Technology (MIIT), Mandalay, for her continuous guidance, support, and suggestions.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] I. Sanasam, P. Choudhary, K.M. Singh, "Line and word segmentation of handwritten text document by mid-point detection and gap trailing," Multimedia Tools and Applications, **79**(41–42), 30135–30150, 2020, doi:10.1007/s11042-020-09416-1.

[2] X. Zhang, L. Duan, L. Ma, J. Wu, "Text extraction for historical tibetan document images based on connected component analysis and corner point detection," Communications in Computer and Information Science, **772**, 545–555, 2017, doi:10.1007/978-981-10-7302-1_45.

[3] X. Han, H. Yao, G. Zhong, "Handwritten text line segmentation by spectral clustering," Eighth International Conference on Graphic and Image Processing (ICGIP 2016), **10225**(Icgip 2016), 102251A, 2017, doi:10.1117/12.2266982.

[4] A. Fawzi, M. Pastor, C.D. Martínez-Hinarejos, "Baseline detection on Arabic handwritten documents," DocEng 2017 - Proceedings of the 2017 ACM Symposium on Document Engineering, 193–196, 2017, doi:10.1145/3103010.3121037.

[5] B.K. Barakat, R. Cohen, A. Droby, I. Rabaev, J. El-Sana, "Learning-free text line segmentation for historical handwritten documents," Applied Sciences (Switzerland), **10**(22), 1–19, 2020, doi:10.3390/app10228276.

[6] K. Sun, T. Liu, L. Zhang, M. Hao, "Handwritten Manchu Historical Document Segmentation with Anisotropic Gaussian Kernel," Proceedings - 2022 Chinese Automation Congress, CAC 2022, **2022**-Janua, 727–731, 2022, doi:10.1109/CAC57257.2022.10055732.

[7] P. P V, D. Sankar, "Handwriting-Based Text Line Segmentation from Malayalam Documents," Applied Sciences (Switzerland), 13(**17**), 2023, doi:10.3390/app13179712.

[8] Z. Li, W. Wang, Y. Chen, Y. Hao, "A novel method of text line segmentation for historical document image of the uchen Tibetan," Journal of Visual Communication and Image Representation, **61**, 23–32, 2019, doi:10.1016/j.jvcir.2019.01.021.

[9] B.K. Barakat, A. Droby, R. Alaasam, B. Madi, I. Rabaev, R. Shammes, J. El-Sana, "Unsupervised deep learning for text line segmentation," Proceedings - International Conference on Pattern Recognition, (d), 2304–2311, 2020, doi:10.1109/ICPR48806.2021.9413308.

[10] Q.N. Vo, S.H. Kim, H.J. Yang, G.S. Lee, "Text line segmentation using a fully convolutional network in handwritten document images," IET Image Processing, **12**(3), 438–446, 2018, doi:10.1049/iet-ipr.2017.0083.

[11] F.C. Fizaine, P. Bard, M. Paindavoine, C. Robin, E. Bouyé, R. Lefèvre, A. Vinter, "Historical Text Line Segmentation Using Deep Learning Algorithms: Mask-RCNN against U-Net Networks," Journal of Imaging, **10**(3), 2024, doi:10.3390/jimaging10030065.

[12] A. Nyein, H. Khaung Tin, "Handwritten Myanmar Character Recognition System using the Otsu's Binarization Algorithm," 2021, doi:10.4108/eai.27-2-2020.2303219.

[13] B. Barakat, A. Droby, M. Kassis, J. El-Sana, "Text line segmentation for challenging handwritten document images using fully convolutional network," Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR, 2018-August, 374–379, 2018, doi:10.1109/ICFHR-2018.2018.00072.

[14] P. Shivakumara, T. Jain, U. Pal, N. Surana, A. Antonacopoulos, T. Lu, "Text line segmentation from struck-out handwritten document images," Expert Systems with Applications, 210(July 2021), 118266, 2022, doi:10.1016/j.eswa.2022.118266.

[15] J. Sklansky, "Finding the convex hull of a simple polygon," Pattern Recognition Letters, **1**(2), 79–83, 1982, doi:10.1016/0167-8655(82)90016-2.

[16] A. Benevento, F. Durante, "Correlation-based hierarchical clustering of time series with spatial constraints," Spatial Statistics, **59**(April 2023),

100797, 2024, doi:10.1016/j.spasta.2023.100797.

[17] N. Xu, R.B. Finkelman, S. Dai, C. Xu, M. Peng, "Average Linkage Hierarchical Clustering Algorithm for Determining the Relationships between Elements in Coal," ACS Omega, 6(**9**), 6206–6217, 2021, doi:10.1021/acsomega.0c05758.

[18] C. Clausner, S. Pletschacher, A. Antonacopoulos, "Aletheia - An advanced document layout and text ground-truthing system for production environments," Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 48–52, 2011, doi:10.1109/ICDAR.2011.19.

[19] M. Daszykowski, B. Walczak, "2.26 - Density-Based Clustering Methods," Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Second Edition: Four Volume Set, **2**, 565–580, 2020, doi:10.1016/B978-0-444-64165-6.03005-6.

[20] F. Simistira, M. Bouillon, M. Seuret, M. Wursch, M. Alberti, R. Ingold, M. Liwicki, "ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts," Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, **1**, 1361–1370, 2017, doi:10.1109/ICDAR.2017.223.

[21] O. Surinta, M. Holtkamp, F. Karabaa, J.P. Van Oosten, L. Schomaker, M. Wiering, "A Path Planning for Line Segmentation of Handwritten Documents," Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR, **2014-Decem**, 175–180, 2014, doi:10.1109/ICFHR.2014.37.

[22] A. Alaei, U. Pal, P. Nagabhushan, "A new scheme for unconstrained handwritten text-line segmentation," Pattern Recognition, 44(**4**), 917–928, 2011, doi:10.1016/j.patcog.2010.10.014.