International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

# Euclidean Distance Based Text Line Extraction and Skew Correction

**Neha[1], Apoorva Arora[2]**

[1,2]Department of Computer Science and Technology, Chandigarh Engineering College, Landran, Mohali

**Abstract:** ***There are many organizations like cultural, educational, governmental; a commercial that manages a wide range of handwritten text documents. In handwritten documents, Text line extraction remains a challenging problem in document image analysis due to multi-skew in handwritten documents. Skew detection and correction of extracting text line becomes a crucial step in document image analysis. In this paper, we intended an innovative technique for text line extraction followed by skew correction by using Euclidean Distance suitable for handwritten documents. We aim to handle single and multi-skew of handwritten text of various writers (single and multiple). Precisely, the problem is stated as energy minimization which affects the accuracy of text line extraction. Additionally, it is necessary to correct the skew of lower baseline and fluctuations of these text lines. Afterwards, text lines are extracted one by one on the proximity of joining of words and align the skew of each text line towards horizontal line. This innovative technique was implemented over 90 documents of various scripts, font size and font style written by single/ multiple writers and multi-skew of text. Till now, huge research has been done to develop the handwritten recognition systems so as to recognize and classify the Characters with the highest possible accuracy and within a shortest period of time. But all the existing systems according to my research extracts the text line features individually using the different technique for each feature which leads to the large amount of processing. So, I am trying to classify the text line documents with the highest possible accuracy and shortest possible time constraints by extracting the text lines rather than segmenting the text lines and words.***

**Keywords:** Text Line Extraction, Skew Detection and Correction, Euclidean Distance, Gradient of a Line

## 1. Introduction

Document Analysis and Recognition (DAR) aims to extract information from documents and also give results according to human apprehension. Text line extraction is an utmost effective competitive layout structure of the text of a distinctive collection of handwritten document scanned images. Text line extraction is a crucial stage in handwritten document images for several image processing tasks like layout analysis and OCR. Extraction of text lines in handwritten documents is much more difficult than machine printed documents. Subsequently, handwritten text line extraction is a considerable challenge in document image analysis.

A considerable amount of research on text line extraction and skew correction has been done over the past few years. Euclidean Distance has the ability of exacting lines from scanned image as well as corrects the skew of each line.
The text lines provided by an extraction process, extract isolated lines from document image. The extraction process is relatively easy when text lines are straight, adequately spaced, and located in the same direction. Although, Variation of the skew angle along the same text line or between text lines, existence of touching lines or overlapping, variable word size are the challenges of text line extraction. For humans, skewed texts are troublesome for visualization and introduce more obstacles in text reading. For machine processing, skewed text brings a number of obstacles that range from requiring extra space for storage to making more failure level the recognition and replica of the text by automatic OCR tools. They are treated as noise. Likewise, handwritten text lines are represented by fluctuations. Variations in the baseline position exist along the text line due to correspondent writing flow. The lower part of character bodies are followed and join by fictitious line or baseline. Therefore, it is essential to consider the entire situations for efficient text line extraction.

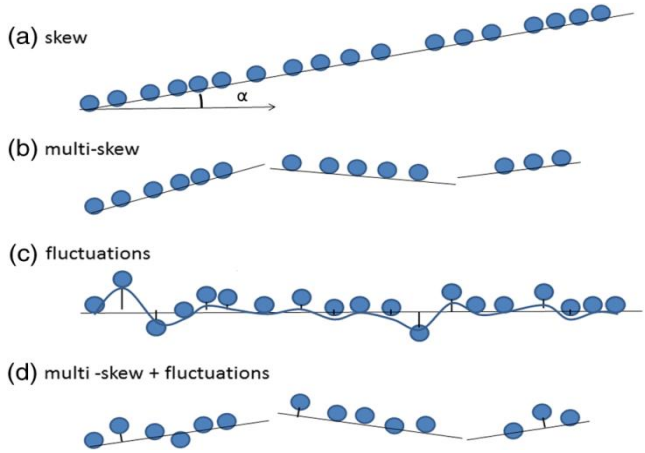


**Figure 1(a)** Skewed baseline: bottom positions of writing components lay on a line with angle ! ≠ 0. (b) Multiskewed baseline. (c) Fluctuating baseline: bottom positions of writing components deviate from a horizontal line. (d) Multiskewed and fluctuating baseline. [3]

This work represents the text line extraction and skew correction method for handwritten text documents by employing Euclidean Distance. The objective is to enhance the accuracy of text line extraction and skew correction. The main steps of the proposed approach are: Noise removal by using DCT filtering, Document binarization by applying K means color Quantization, Joining of words employed by Weighted Sum, Euclidean Distance based Text Line Extraction and Skew Correction.

This paper is organized as follows. Section II justifies the detail about Euclidean Distance and the gradient of a line. Section III details the proposed work and methodology.

Section IV experimentally compares the proposed Work with Existing Work. Section V concludes the Paper.

## 2. Euclidean Distance and Gradient of a Line

In mathematics, the Euclidean Metric or Distance is the ordinary distance (i.e. Straight line) between two points in Euclidean space. It can be defined as the sum of the square roots of the difference between corresponding points. In this distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean space**.** In older literature refers to the metric as Pythagorean metric.

In Euclidean geometry, the Euclidean Distance between two points in the plane with coordinates $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$ in Euclidean space, then the distance (d) from **p** to **q** is given by the Pythagorean formula.

$$D(p, q) = \sqrt{(p_2 - p_1)^2 + (q_2 - q_1)^2}$$

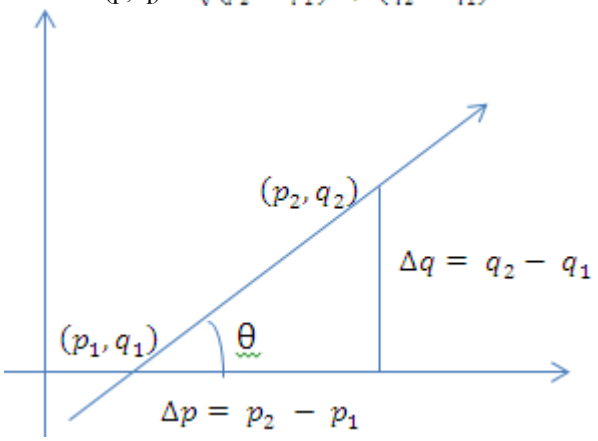


**Figure 2:** Euclidean Distance and Gradient of a Line

**Gradient of a Line**

The **gradient** (slope) of a line is the ratio of the amount that *y* increases as *x* increases in some amount. Slope tells you how much *y* increases as *x* increases. [4]

In mathematical language, the slope *m* of the line is

$$m = \frac{q_2 - q_1}{p_2 - p_1}$$

In trigonometric, gradient or slope of a line is related to its angle to incline θ by its tan function.

$$m = \tan\theta$$

And inverse function of tan function is

$$\theta = \tan^{-1} m$$

## 3. Proposed System and Methodology

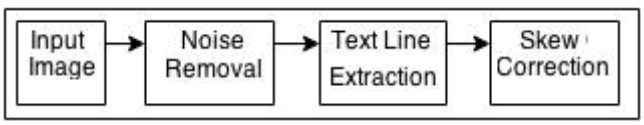


**Figure 3:** Block Diagram of Proposed System

Figure 3 gives the survey of the proposed system. Handwritten script documents go through noise removal where all nonessential noise is removed. Output of noise removal is given to the text line extraction where text lines of handwritten documents are extracted. Afterwards, the skew is detected and corrected in skew correction stage. In this stage obtained text lines are free from skew.

The method works on handwritten document images with single and multiple skews. Extracting the text lines from document images is a crucial step in order to identify the various lines written in different orientations. After the extraction it would be easy to estimate the skewed lines.

Figure4 represents the various steps of the proposed work.

- In the initial step we read the scanned handwritten document image in Matlab and do needful alternations.

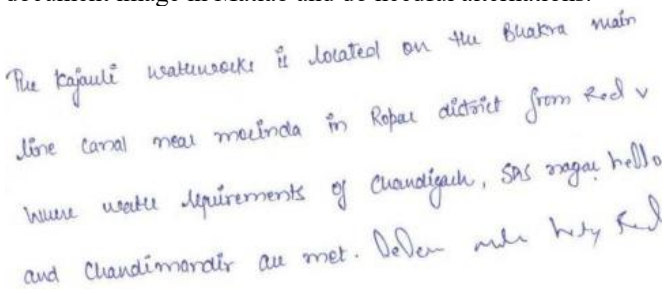

**Figure 4 (a):** Original Image

- Then extract R, G, B components individually from the scanned original image.

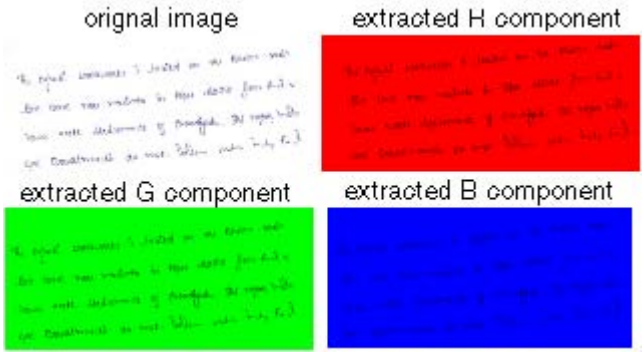


**Figure4(b):**Extracted R, G, B Components

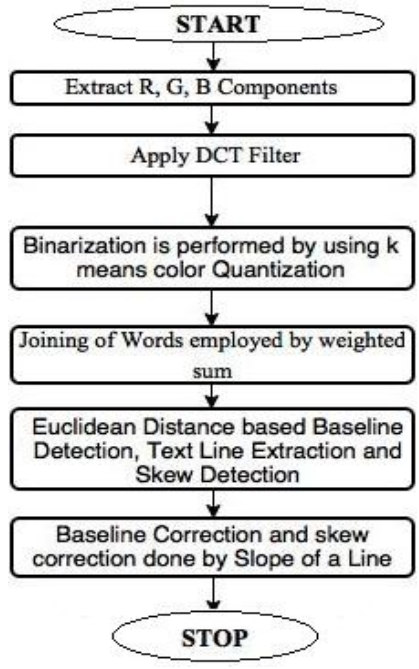


**Figure 4:** Flow Chart of Proposed System

- Apply DCT filter for R, G, B separately to remove noise from each R, G, B matrix as shown in figure.

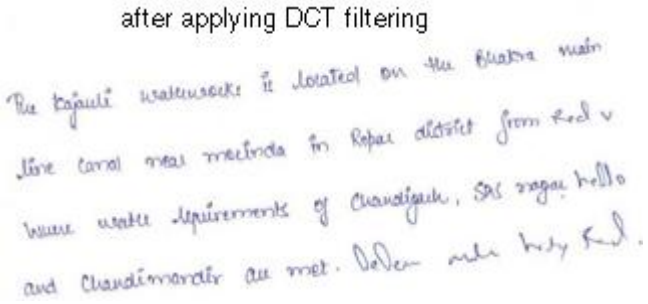


**Figure 4(c):** After applying DCT Filter

- In the third step, binarization is performed on the basis of k means color quantization. We employed a K means color quantization or clustering to remove the distinct colors from an input image.

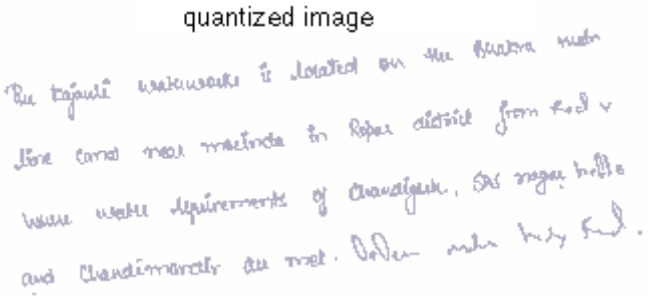


**Figure 4(d):** Quantized Of Text Image

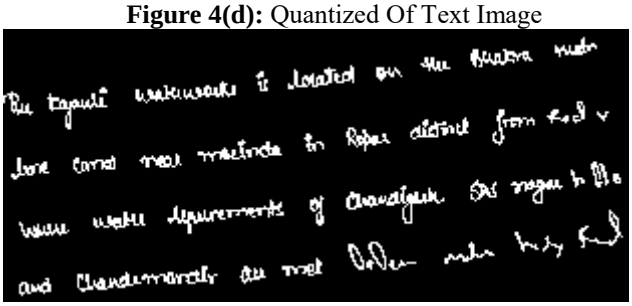

**Figure 4(e):** Binarization of Text Image

- In the fourth step, Joining of words are employed by using weighted sum. The weighted sum model is popular method,a technique for MCDM (multi-criteria decision making) to decide various alternatives with reference to decision criteria.
-

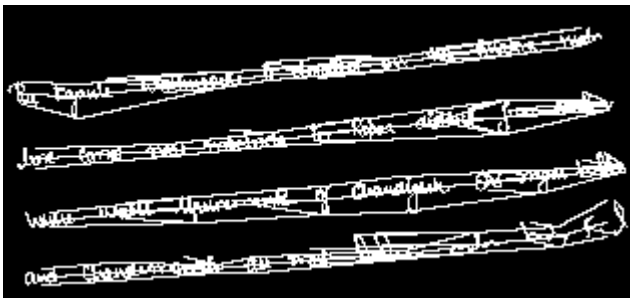

**Figure4(f).** Joining of Words

- After joining of words, morphological commands are used to fill the holes/ regions between these test lines.

**Figure 4(g):** Filled text line area by using Morphological Commands

- Furthermore, in the proposed work Euclidean distance used for baseline detection, text line extraction one by one from document image and skew detection of each extracted line as shown in figure.

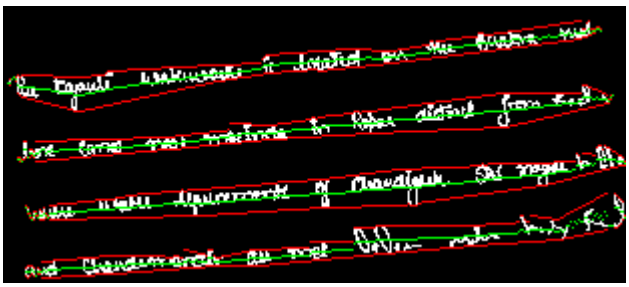

**Figure4(h).** Baseline Detection of Text Image

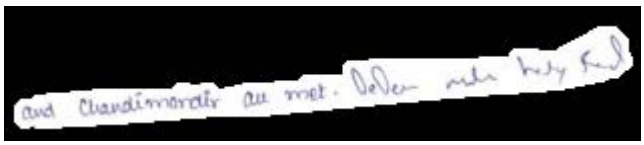

**Figure4(i).** Text Line Extraction Of Text Image

- At the end, baseline correction and skew correction of each line done is by using gradient of a line and afterwards showing the final image of the document.

**Figure 4(j):** Skew corrected Image

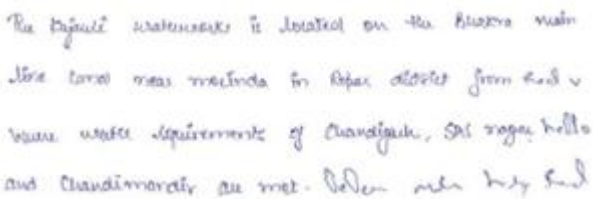

**Figure4(k):** Text Document After Skew Correction Stage

## 4. Algorithm

***Input:*** In this we consider Scanned handwritten document Image

***Output:*** Skew Corrected Image

1. For i=1 to i=x
2. For j=1 to j= $[0]_i$

3. $d = \sqrt{(mean\ [0]_{jx} - R_{xi}) + (mean\ [0]_{jy} - R_{yi})}$
4. $m = {mean\ [0]_{jy} - R_{yi}} / {(mean\ [0]_{jy} - R_{yi})}$
5. $\theta = \tan^{-1}(m)$
6. $if \begin{cases} 0 > d < 110\ and\ 0 > \theta < 45 & j \in CN \\ otherwise & j \notin CN \end{cases}$
7. End for
8. For j=1 to j=k
9. $WCN_j = (CN\ (\theta)_j \times w_1 + CN\ (dis)_j \times w_2)$
10. End for
11. $CN_{sel} = \min[WCN]$
12. End for

- Where **x** are the number of detected objects
- $[0]_{jx}$ & $[0]_{jy}$ are the x and y pixel positions of $j^{th}$ object detection.
- $R_{xi}$ And $R_{yi}$ is the mean centered position of $i^{th}$ object.
- CN is the pool of selected connected objects
- $WCN_j$ are the weighted criteria of the $j^{th}$ object.
- $w_1$ & $w_2$ are weights of (theta & dis) respectively.
- $CN_{sel}$ Is the selected neighboring object.

## 5. Results and Discussions

In order to show the results, we have created our own database of 90 handwritten documents using various scripts like English, Hindi, Punjabi, Arabic and Latin. It showed the best results over English, Punjabi, Hindi and Latin. In proposed work, baseline detection, text line extraction and skew detection are evaluated with the help of only one method called Euclidean Distance. We used the Euclidean Distance to find a minimum word distance form all its neighboring words and on the basis of this distance, joined words as connected components by using a weighted sum for text line extraction. We are extracting the line and calculate and correct the scans of the document images and check the accuracy of the final image.

**Table 1:** Performance Evaluation Of Proposed System

| | *Existing Work* | *Proposed Work* |
|---|---|---|
| Number of Documents | 80 | 90 |
| Average Number of lines in a document | 20 | 32 |
| Total number of lines analyzed | 1600 | 2880 |
| Number of lines Extracted | 1596 | 2864 |
| Number of Skew Corrected Lines | 1562 | 2857 |
| Accuracy of Extraction | 99.13% | 99.4% |
| Accuracy of Skew Correction | 97.63% | 99.2% |

## 6. Conclusion

In the proposed work, we represent a novel approach for Euclidean Distance for text line extraction and skew detection and correction over scanned handwritten document images. For the noise removal, DCT employed over the handwritten document images. Afterwards, Binarization is performed on the basis of using K means Color Quantization rather than using thresholding methods. Euclidean Distance is used to calculate the distance between words and this calculated distance helps in the extraction of text lines. This distance also helps in to find the lowest baseline of words. Thus, the skew is corrected on the basis of distance calculated. Baseline correction is done by gradient or slope of a line. This proposed system shows best accuracy of 99.2% over 90 different script documents.

## References

[1] S. Dixit, S. H. Narayan, M. Belur, "Kannada Text Line Extraction Based on Energy Minimization and Skew Correction", IEEE International Advance Computing Conference (IACC), pp. 62-67, 2014.
[2] Hyung IL Koo and Nam Ik Cho, 'Text-Line Extraction in Handwritten Chinese Documents Based on an Energy Minimization Framework', *IEEE Trans. On Image Process.*, vol. 21, no. 3, pp. 1169-1175, 2012.
[3] Olivier Morillot, Laurrence Likforman Sulem and Emmanuele Grosicki, " New Baseline Correction algorithm for text line recognition with bidirectional recurrent neural networks," Journal of Electronic Imaging 22(2), 023028 (Apr- Jun 2013).
[4] Wikipedia, 'Euclidean distance', 2015. https://en.wikipedia.org/wiki/Euclidean_distance.
[5] 'Slope of a line', 2015. http://study.com/academy/lesson/what-is-slope-definition-formulas-quiz.html.
[6] Wikipedia, 'Slope', 2015. https://en.wikipedia.org/wiki/Slope.
[7] M. M. Blumenstein, C. K. Cheng and X. Y. Liu, " New Preprocessing Techniques for Handwritten Word Recognition," in The 10th IASTED International conference for Visualization, Imaging and Image Processing, pp. 480-484, Acta Press (2002).