A MODEL TO CONVERT WAVE–FORM-TEXT TO LINEAR-FORM-TEXT FOR BETTER READABILITY BY OCRS

C.S. Vijayashree[1] and Vasudev T[2]

[1]P.E.T Research Foundation,
P.E.S College of Engineering, Mandya-571401, India
[2]Maharaja Research Foundation,
Maharaja Institute of Technology, Mysore-571438, India

ABSTRACT

The existing Optical Character Readers (OCRs) are capable of reading linear form text and have limitations to read artistic and non-linear form text. Wave-form-text is an artistic-text which is quite common in several documents such as certificates, advertisements and history documents. OCRs fail to read such wave-form-text and it is necessary to transform the same to linear-form-text at preprocessing stage. In this paper, we present a transformation model for better readability by OCRs. This method takes the wave-form-text as input, extracts each character. The characters are subjected to tilt correction to correct any tilt present. Then the characters are subjected to alignment correction and are finally concatenated together to form linear text. The proposed method is implemented on several wave-form-text inputs and the readability of the transformed text is analyzed with an OCR.

KEYWORDS

Artistic text, wave-form-text, linear-form-text, Tilt correction, OCR

1. INTRODUCTION

A significant area in the field of Digital Image Processing is Document Image Analysis (DIA). DIA is very important in applications like document identification/recognition, language identification, automatic reading from document etc. Many researchers are working on different problems on document images starting from image acquisition to image understanding [1,2]. Processing activities in DIA can be divided into Pre-processing, Segmentation, Script Identification, Page Layout Analysis (PLA) and Classification, Character Recognition etc [3], which open up multiple areas of research. The research in this field is focusing to come out with generic approaches to accomplish automation in document reading, extracting contents from documents and these have lead into many vibrant research problems [2]. The results of the research on the above problems are converging towards the generic solutions to major issues in DIA.

DOI : 10.5121/ijma.2015.7301

In spite of considerable research work in the area of DIA, a major issue which is not sufficiently addressed is, reading or extracting the contents of the text which appear in artistic-form in a document. Many documents, especially certificates, marks cards, sign boards, logos, etc., have artistic text. In addition, many official seals on the documents for authentication are also artistic in nature. The contents of such artistic-text definitely have some valuable information that has to be processed. Most of the graduation certificates issued by the Universities contain the name of the university in artistic form. If such document has to be processed by an Optical Character Reader (OCR), the OCR should be able to read such artistic-text or proper pre-processing is required to make that text readable by OCR. Few such artistic-texts in documents are, text appearing in triangular-form, arc-form, circular-form, wave form. Samples of such artistic-texts are shown in Figure 1. The contents of such text normally convey the identity information like company's name, type of document, etc., which is the main source for classification of the document.

Figure 1. Samples of artistic-form text in document

Documents containing artistic-text, when subjected to reading by OCRs, fail to be read, as the OCRs are developed to read linear texts. Hence, it is necessary to transform artistic-text to linear text such that OCRs are able to read the contents efficiently. Approaches developed for general skew detection and correction are not suitable to transform such artistic-text documents into linear form. Hence, it is required to come out with different approaches that can transform artistic-form text into linear form text and make the same suitable for reading by an OCR.

In this research work, we propose a methodology to transform wave-form-text to linear-form-text suitable for OCR processing. Some samples of wave-form-text is shown in Figure 2. The wave-form-text exhibit two distinct characteristics which are tilted characters and characters are at multiple horizontal line levels. Tilt is the angular slant to the baseline introduced in the character. Tilted characters are mainly noticed in many artistic texts. We notice that there are tilted characters in the wave-form-text. Such tilted characters hinder the investigation of generic methods of recognition and the efficiency of recognition drops relatively. Hence tilt in characters contributes a major share in affecting the efficiency of the recognition algorithms. In addition to tilted characters, the characters in wave-form-text are not placed in the same horizontal line. This also affects the efficiency of the recognition algorithms. In the proposed work these two features are addressed in obtaining the solution.

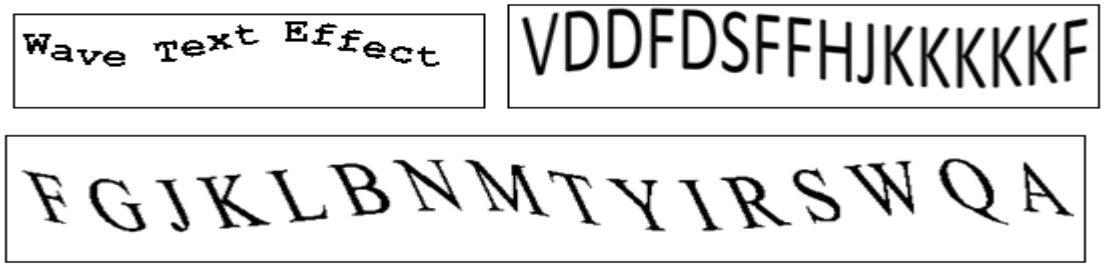

Figure 2. Samples of Wave-Form-Text

2. RELATED WORK

One of the major problems encountered in DIA is implicit/inherent skew noticed in document images [4,5]. Inherent skew, is due to the natural inclinations of text lines in the document. Considerable amount of work is reported in literature on explicit skew detection [6-16]. Each of the approaches reported in literature on explicit skew detection has its own advantages and limitations, and these approaches are not extendable for detecting inherent skew. Since artistic texts also have inherent orientation in the document, artistic-texts are said to have implicit skew. To the best of our efforts while surveying for literature in the direction of implicit skew detection and correction, we could find the work of Pal et al in detecting multiple implicit skewed lines within a document[3], i.e., detecting lines within the document having different orientations. Given below are details on some of the related work in this area.

Vasudev et al[18] have proposed transformation of arc-form-text to linear-form[18]. The work proposed by Vasudev et al[18] performs transformation to considerable extent but suffers from tilt deformation and an additional stage is required for tilt corrections in the model. The proposed work assumes that the arc-form-text has been segmented out from the document and is free from noise. Further, it is assumed that the arc-form-text in document is either circular or elliptical in shape and is limited to the upper half circle or ellipse. The arc-form-text is enclosed in two arcs an inner arc and an outer arc. The principle adopted to perform transformation is a point processing technique[19], where a set of points representing an ellipse is transformed to represent a line. After enclosing the arc-text between two suitable arcs, it is required to transform all the points on this elliptical band into a linear band of points. The transformation process introduces tilt to the characters. An additional stage is required to correct the tilt in the characters. Vasudev et al[20] have proposed a model where initially the direction of skew is identified in character at a macro level with the help of a knowledgebase. The amount of inclination of tilt to its base line is detected with the support of line drawing algorithm. In the final stage, the detected tilt in the character is corrected by transformation process. Further, the readability efficiency after transformation is claimed as 84% in this method.

Vishwanath et al[21] have proposed connected component technique for character extraction from document image having artistic-form-text. The technique starts with artistic form text as input and transforms the same to linear form. First, the characters in artistic text are segmented and extracted using Connected Component Analysis technique. Due to the intrinsic nature of artistic text, the extracted characters exhibit skew. In the next stage, such implicit skew in extracted characters is detected using Hough Transform and corrected. Further, skew corrected characters are concatenated to put in linear form. Experimental results of the proposed method show an average 80% of readability by OCR as efficiency.

Further, Vishwanath et al[22] have proposed Radon transform for the detection of implicit skew in the characters and its correction. The technique starts with artistic form text as input and transforms the same to linear form. First, the characters in artistic text are segmented and extracted using Connected Component Analysis technique.. In the next stage, the implicit skew in the extracted characters is detected using Radon Transform and corrected. Further, skew corrected characters are concatenated to put in linear form. Experimental results of the proposed method show an average 85% of readability by OCR as efficiency.

The work proposed by Vijayashree et al[23] presents a simple technique to estimate and correct the tilt present in artistic text. Initially, the direction of tilt of the characters is detected using a heuristically constructed knowledgebase. Next, the inclination of the character to its base is

estimated using line drawing algorithm. Finally, the estimated tilt is corrected through rotation in the counter direction of the tilt. The method has been tested with sufficient samples and readability analysis is performed with an OCR. Experimental results show an average improvement in readability by OCR from 20% before tilt correction to 82% after the tilt correction. The current work incorporates this technique for tilt correction to take care in wave-form-text conversions.

Vijayashree et al[24] present an alternate technique to transform arc-form-text to linear form. The proposed model has two stages. The initial stage is to estimate two concentric imaginary ellipses to enclose the arc-form-text. After enclosing the arc- text between two imaginary suitable arcs, it is required to transform all the points on this elliptical band into a linear band of points. In this transformation model, a set of points representing line in one orientation is transformed to represent a line of points in another orientation. An arc-form text can be considered as a set of n consecutive lines in different orientations, where n being the distinct points on surface of the outer arc. These n lines with different orientations are transformed to n vertical lines, which results in the text appearing horizontally linear. The results of experiments establish an overall readability between 73% - 100% in transformed and tilt corrected text.

The current proposed work considers as input wave-form-text. It assumes that the wave-form-text has been segmented out from the document and is free from noise. The proposed model has three stages. The first stage extracts each character from the wave-form-text and is described in section 3. The second stage performs tilt correction as proposed in [21] and the same is briefed in section 4. The third stage performs character alignment correction and is described in section 5. Then the corrected characters are concatenated such that they are all aligned together to be on the same horizontal line. The flow in the procedure is illustrated as block diagram in Figure 3. Experimental results are described in section 6. The conclusion of the work is described in section 7.

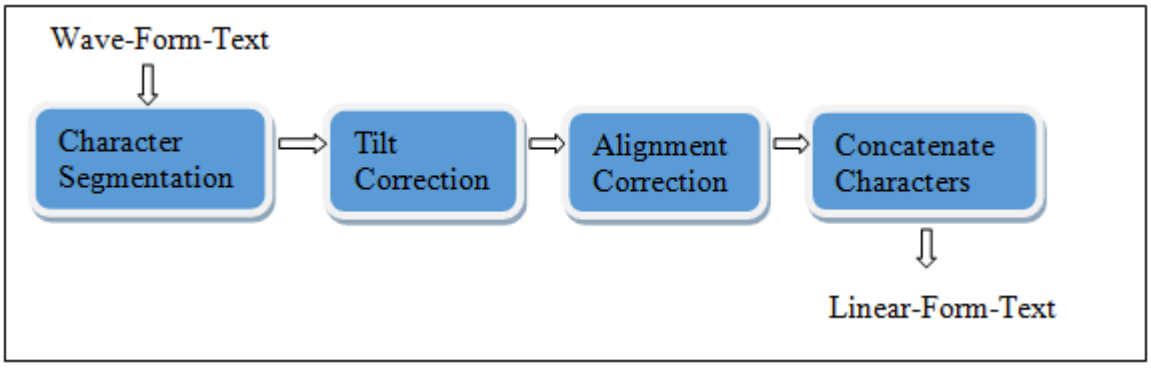

Figure 3. Sequence of Stages in the Proposed Work

3. CHARACTER SEGMENTATION

The characters in the wave-form-text are typically at different horizontal levels and different tilt inclinations. These characters need to be brought to a single horizontal level and their angle of inclination needs to be corrected. The wave-form-text which is segmented out from the document is taken as input. Segmentation of each character is done through searching suitable pair of vertical lines which enclose each character. For each character, pair of horizontal lines are

searched in the top and bottom of the character so that the complete character is enclosed in a box consisting of vertical and horizontal lines. The character enclosed in the box is extracted out for tilt correction and alignment correction. The methodology is illustrated in the form of algorithm subsequently.

1. Input the wave-form-text segmented from the document

2. While not end of text

    a. Search a vertical line till the line touches the character in sequence and is the left boundary line for the character
    b. From the left boundary line of the character, search another vertical line till the edge of the character is found and the line is right boundary line for the character
    c. Search a horizontal line from the top between the two vertical lines that touches the character. Similarly, search another horizontal line from bottom that touches the character between the two vertical lines.
    d. Steps a to c ensure that the character is enclosed in a box defined by the pair of vertical and horizontal lines.

3. Extract each character enclosed in the box for subjecting the same to tilt and alignment correction.

The Figures 4, 5 and 6 show the input text, text with separated characters and extracted characters respectively.

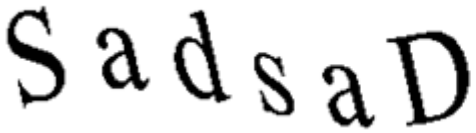

Figure 4. Input Text

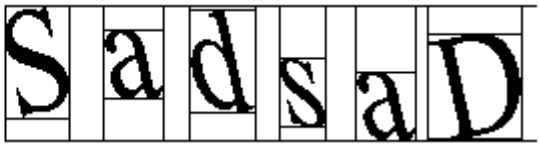

Figure 5. Enclosing Characters in box

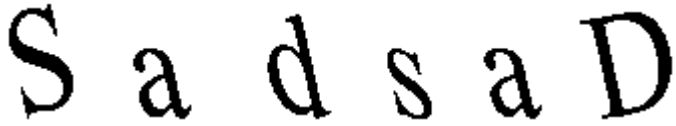

Figure 6. Extracted characters

4. TILT CORRECTION

Each extracted character of wave text is next subjected for tilt correction procedure [21]. The method initially performs a macro level decision to find the direction of tilt i.e. detection of tilt is towards left or right to baseline. The proposed tilt detection algorithm detects the direction of tilt in the input character with the support of a heuristic knowledgebase. Next, the degree of tilt to the baseline is estimated using line drawing algorithm. Finally the character is rotated by the estimated tilt in the counter direction of the tilt. The Figure 7 illustrates the tilt corrected characters.

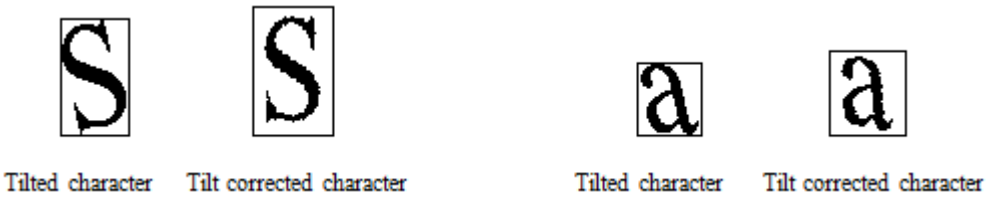

Figure 7. Character Tilt Correction

5. CHARACTER ALIGNMENT CORRECTION

Each tilt corrected character may need correction for the proper alignment of the character. The correction is applied either at the top of the character or at the bottom of the character depending on the type of the character. In the box enclosing the character, at the bottom or the top horizontal line, a line is drawn from the edge of the character touching the horizontal line to the furthest opposite vertical edge at the same horizontal level. If no part of the character is touched by the line, the line is drawn from the character edge to the adjacent pixel on the vertical line. This is continued till the drawn line touches any part of the character. The difference in height from each pixel of the drawn line to the horizontal line is calculated. All the pixels of the wave-form-text in the same column as the pixel of the line are moved by the difference calculated so that the character is aligned correctly.

The methodology is illustrated as algorithm subsequently.

1. Search a horizontal line in the box enclosing the character where the character edge touches the point A on the line furthest from the center. This could be the bottom horizontal line or the top horizontal line.
2. From the point A where the edge of the character touches the horizontal line, search a line to the point(B) on the vertical edge furthest from A such that the line AB creates a free zone corner.
3. For every pixel on line AB, find the distance of the pixel from the horizontal line. Move all pixels in that column of the wave-form-text by the distance calculated.

Figure 8 illustrates the line AB drawn in the box enclosing the character. Figure 9. shows the alignment corrected character.

Figure 8. Line drawn from the edge of the character

Figure 9. Alignment corrected characters

6. EXPERIMENTAL RESULTS

Experiments are conducted on the wave-form-text with different sizes and different wave shapes for English texts considering about 200 wave-form-texts. The results of experiments establish readability ranging between 93% - 100% with an overall average readability of 98% which is considerably a quite encouraging initial result. The result of the transformation provides a better and more suitable pre-processed input for OCRs. The main reason for readability failures in certain cases are due to error in tilt and alignment corrections. Few experimental results are illustrated in Table-1.

Analysis of readability by an OCR of the text before transformation and after transformation is performed with respect to English text using the OCR "Readiris Pro 9"[23]. In this process, first, the samples of wave-form-text are taken as input to OCR and subjected to be read by the OCR and the result is tabulated in Table 1. Column "Input Wave-Form-Text" shows the input document. Column "Readiris Output1" displays the readability output of OCR for the input document. The transformed text which is a result of the proposed method is subjected to be read by OCR and the result is tabulated in Table 1. Column "Corrected Linear-Form-Text" tabulates the transformed document. Column "Readiris Output2" displays the readability output of OCR for the transformed document. We can see a marked improvement in the readability of the OCR with the proposed transformation.

Table 1. Experimental Results

Input Wave-Form-Text	**Readiris Output1**	**Corrected Linear-Form-Text**	**Readiris Output2**
GHTYUIOLKJHFDD	GHTYUIOIKJHrDD	GHTYUIOLKJHFDD	GHTYUIOLKJHFDD
ABCDEFGHIJ	ABCDEFOHIJ	ABCDEFGHI J	ABCDEFGHI J
SadsaD	$adsaJ}	SadsaD	Sad saD
SadsaD	Sa\\~aU	SadsaD	SadsaD
HGTYUIOPDFAWX	\\GT{\J\U111)rI\ ~1	HGTYUIOPDFAWX	HGTYUIOPDFA~X
ABCDEFGHIJKLMN	[illegible]	ABCDEFGHIJKLMN	t\BCDEFGHIJKLMN
FGJKLBNMTYIRSWQA	FGJKLBNMrYIRS WQA	FGJKLBNMTYIRSWQA	FGJKLBNMTYIRSWQA
N V F G HJKLX	N V f G RJKLX	NVFGHJKLX	NVFGRJKLX

6. CONCLUSION

The proposed approach efficiently transforms a wave-form-text into linear-form-text. The results do not show much of tilt deformations and very little variations in the character alignment in the transformed characters. The transformed text serves as better pre-processed input to the OCR for better readability. OCR shows an average readability of 98% after transformation which is

reasonably good performance achieved at initial attempt. The proposed method is claimed as a simple and less complex approach to transform wave-form-text to linear-form. The method does not require any special arrangements to acquire the document image or to perform the transformation.

Failures in reading certain cases are due to persistence of error in tilt and alignment of the corrected characters. The performance can be further increased through exploring efficient tilt and alignment corrections which are under investigation. The method can be extended suitably to transform wave-form-text of other languages including other Indian languages. The current tilt correction is based on knowledgebase constructed for English language. This can be extended to cover Indian languages as well. Further, there is scope to produce a much better distortion free, smooth and neatly aligned text in the transformation which is under investigation.

REFERENCES

[1] O'Gorman, Lawrence, Kasturi, Rangachar. Executive Briefing: Document Image Analysis. IEEE Computer Society Press, 1998.

[2] Nagabhushan, P. Document Image Processing, in : Proc. National Pre-Conf. workshop on Document Processing, India, pp 114, 2001.

[3] Vasudev T, Hemanthakumar G H, Nagabhushan P., 2005, Segmentation of characters in an arc-form, 7th Int. Conf. on Cognitive Systems (ICCS 2005), India.

[4] Pal U, Mitra M, Choudhri B B. Multi-Skew Detection of Indian Script Documents, in: Proc. Int. Conf. on Document Analysis and Recognition (ICDAR 2001), 2001.

[5] Vasudev T, Hemanthkumar G, Nagabhushan P. Detection and Correction of Vertical Skew in Characters, in: Proc. 3rd Int. Conf. on Innovative Applications of Information Technology for Developing World(AACC 2005) CD version, Nepal, 2005.

[6] Zheng Zhang. Restoration of Curved Document Images Throug 3D Shape Modeling, in: Conf. on Computer Vision and Pattern Recognition(CVPR2004), 2004.

[7] Amin A, Fischer A. A Document Skew Detection Method Using the Hough Transform. J. Pattern Aual. Applicat. 3,243-253, 2000.

[8] Kavallieratou E, Fakotakis N, Kokkinakis G. Skew Angle Estimation for Printed and Handwritten Documents Using the Wigner-Ville Distribution, Image Vis. Comput. 20, 813-824, 2002.

[9] Liolios N, Fakotakis N, Kokkinakis G. On the Generalization of the Form Identification and Skew Detection Problem. Pattern Recognition(35), 243-264, 2003.

[10] Murali S, Vasudev T, Hemanthkumar G, Nagabhushan P. Language Independent Skew Detection and Correction of Printed Text Document Images: A Non-rotational Approach. VIVEK – Int. J. Artif. Intell. 16(2), 08-15,2006.

[11] Shivakumar P, Nagabhushan P, Hemanthkumar G, Manjunath. Skew Estimation by Improved Boundary Growing for Text Document in South Indian Languages. VIVEK –Int. J. Artif. Intell. 16(2), 15-21, 2006.

[12] Lu Yue, Tan, Chew Lim. A Nearest Chained Aproach to Skew Estimation in Document Images. Pattern Recognition Lett. 24, 2315-2323, 2003.

[13] Cao,Yang, Wang, Shuhua, Li, Heng. Skew Detection and Correction in Document Image Based On Straight Line Fitting, Pattern Recognition Lett. 24(12), 1871-1879, 2003.

[14] Vasudev T, Hemanthkumar G, Nagabhushan, P. Segmentation of Characters in Arc-form-text, in: Proc. on Cognitive System(ICCS 2005), CD version, India, 2005.

[15] Breuel T. The Future of Document Imaging in the Era of Electronic Documents , in: Proc. Int. Workshop on Document Analysis, India, pp.275-296, 2005.

[16] Kennedy L M, Basu M. Image Enhancement Using a Human Visual System Model, Pattern Recognition 30(12), 2001-2014, 1997.

[17] Vasudev T, Hemanthkumar G, Nagabhushan, P. An Elliptical Approximation Model for Removal of Text-Line Bending Deformation at Page Borders in a Document Image, in: Proc. Int. Conf. on Cognition and Recognition, India, pp.645-654, 2005.

[18] Vasudev T, Hemanthkumar G, Nagabhushan P. Transformation of Arc-form-text to Linear-form-text Suitable for OCR, Pattern Recognition Letter 28 (2008) 2343-2351, 2008.

[19] Rafael C Gonzales & Richard E Woods, 2002, Digital Image Processing, 2nd Edition, Pearson Education Publication.

[20] Vasudev T, Hemanthkumar G, Nagabhushan P. Detection and Correction of Vertical Skew in Characters, in: Proc. 3rd Int. Conf. on Innovative Applications of Information Technology for Developing World(AACC 2005) CD version, Nepal, 2005.

[21] Vishwanath C. Kagawade, Vijayashree C.S., Vasudev T. Transformation of Artistic Form Text to Linear Form Text for OCR Systems, International Conference on Advances in Computing (ICAdc2012)

[22] Vishwanath C. Kagawade, Vijayashree C.S., Vasudev T. Transformation of Artistic Form Text to Linear Form Text for OCR Systems using Radon Transform, International Conference on Emerging Research in Electronics, Computer Science and Technology(ICERECT-12)

[23] C. S. Vijayashree, Vishwanath C. Kagawade, Vasudev T. Estimation of Tilt in Characters and Correction for better readability by OCR systems, International Journal of Computer Applications (0975 – 8887). Volume 90 – No 13, March 2014

[24] Vijayashree C S, Shrithi C V, Vasudev T. Modified Approach To Transform ArcForm-Text to Linear-form-text : A Preprocessing Stage for OCR, SIPJ, Vol 5, No 4, 67-75, Aug 2014.

[25] http://www.irislink.com/readiris.

AUTHORS

Vasudev T is currently Professor in the Department of Computer Applications, at Maharaja Institute of Technology, Mysore. He obtained his Bachelor of Science and Post-Graduate Diploma in Computer Programming with two Masters Degrees, one in Computer Applications and the other one in Computer Science and Technology. He was awarded Ph.D. Degree in Computer Science from University of Mysore. He has 30 years of experience in academics. He has published over 30 articles in reputed journals in his area of research Digital Image Processing, specifically Document Image Processing.

C. S. Vijayashree obtained her B.E. Degree in Computer Science from B.I.T, Bangalore and M.E. Degree in Computer Science from U.V.C.E, Bangalore. She is pursuing research towards her Ph. D. Degree in Computer Science of University of Mysore, Mysore, at P.E.S. College of Engineering, Mandya.