Reference Text Line Identification Based on "Water Flow" Algorithm

Darko Brodić[1] and Zoran Milivojević[2]

Abstract – **In this paper a robust and efficient algorithm for detecting reference text line at different skew angles is presented. Algorithm called "water flow" assumes hypothetical water flows under different specified angles of the image frame from left to right and vice versa. As a result of algorithm, unwetted image frames is extracted as area needed for text line extraction. From extracted text line by the least square method, reference text line is identified. Examples using different water flow angles are examined and inspected. Results are analysed, elaborated and summarized. Proposed algorithm showed robustness and should be used for all sorts of skewness.**

Keywords – **Reference text line, Reference text line extraction, OCR**

I. INTRODUCTION

Handwritten text is fully or partially cursive text. It can be composed of discrete characters, broken words or its combination. But handwritten text tend to be multi-oriented and skewed. Appearance of different orientation skewed lines i.e. multi-skewed lines are made the handwritten text to be less readable. Hence, reference text line extraction from optically scanned documents is primary goal of optical character recognition (OCR) [1].

Various reasons exist for appearance of multi-skewed lines in text, but two are most common [1]. Firstly, during scanning process a degrees of misalignment of the document made is unavoidable. But, all text lines in the scanned document are uniformly skewed i.e. reference text line are almost parallel. Secondly, text lines in original document are differently skewed due to specific individual handwriting.

All handwriting text lines are made under different orientation i.e. multi-skewed. To enhance the ability of document analysis system robust algorithm for reference line extraction in multi-skewed text is needed.

In ideal situation extraction of reference text line is simple. A reference text line is described by relation

$$y = ax + b\,. \tag{1}$$

It means parameters a and b define slope and y-intersection, respectively. Due to uniformly skewed text, one reference line extraction led to extraction of other reference lines.

In real situation, all proposed techniques of reference line extraction based on identifying valleys of horizontal pixel density histogram quoted in [1] failed due to multi-skewed text lines.

[1]Darko Brodić is with the Technical Faculty Bor at University of Belgrade, Vojske Jugoslavije 12, Bor 19210, Serbia, E-mail: dbrodic@tf.bor.ac.rs

[2]Zoran Milivojević is with the Technical College Niš, Aleksandra Medvedeva 20, Niš 18000, Serbia, E-mail: zoran.milivojevic@ vtsnis.edu.rs

Method of identifying words contour area as a start of detecting baseline point proposed in [2]. But the assumption made on the definition of elements of word are too specific.

Another method proposed in [3] deal with "simple" multi-skewed text. It uses as a basis simple type of Hough transform for straight lines. This approach is again specific.

Algorithm proposed by [4] model text line detection as an image segmentation problem by enhancing text line structure using a Gaussian window and adopting the level set method to evolve text line boundaries. Authors specified method as robust, but rotating text by an angle of 10° has an significant impact on reference line hit rate.

In this paper proposed method by [1] is modified as in [5] and more examined as well. Actually method [1] hypothetically assumed a flow of water in a particular direction across image frame in a way that it faces obstruction from the characters of the text lines. Definition of unwetted area is corner stone for calculating text line extraction. But in [1] water flow angle is locked up by four values: 45°, 26.6°, 18.4° and 14°. Our modified water flow method works with water flow angle from 0° to 90°. As can be seen different water flow angle region is proposed for the best referent line hit rate. Text under investigation is computer printed sample text rotated around x-axis from -20° to 20°.

Organization of paper is following. In Sect. II. we give an information on proposed algorithm. Every detail of modified algorithm is described. In Sect. III. experimental results are given. Results are analysed and elaborated as well. In Sect. IV. conclusion is made.

II. PROPOSED ALGORITHM

Text line detection procedure consists of three elements as shown in Fig. 1.

Fig. 1. Text line detection procedure

In preprocessing stage, algorithm for line distinction is used. After preprocessing all text lines are separated. During the processing stage, algorithm for identification of skew and reference text line is employed. After all, in postprocessing stage reference text, based on skew and stroke angle, is straightened and repaired.

In this paper we deal with element of processing. Some assumption should be made before defining algorithm. We suppose that there is an element of preprocessing. It had been

made before applying our algorithm. It's job is to split up text lines due to white area between every neighbor text lines. After preprocessing, every text line is separated. So, it initiates distinct entity consists of group of words.

Split up and extracted text line represents digitalized image dimension MxN. Each word in an image contains black points i.e. pixels. Every point is represented by number of coordinate pairs $X_{i,j}$. Algorithm needs to extract bounding areas and define pixel type from situation in Fig.2.

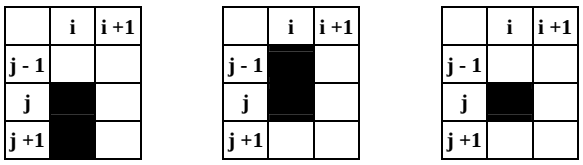

Fig. 2. Boundary pixel type
a. Upper boundary pixel, b. Lower boundary pixel,
c. Candidate for additional investigation of boundary type

Due to making unwetted areas under angle those regions are mark out by lines defined by:

$$y = kx \ , \tag{2}$$

where slope k = tan(α) from Fig.3.

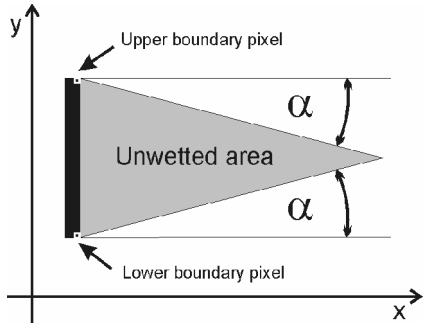

Fig. 3. Unwetted area definition

First, proposed algorithm have to verify boundary pixel type in document image. After verification it makes unwetted areas around the words. Due to pixel type, i.e. upper or lower, slope is α or –α. Lines defined by slope make connection in specific pixel creating unwetted area defined as grey region in Fig. 3.

Pseudo code for pixel type detection (See Fig.2. for reference) can be expressed as follows:

```
begin
   for I = 1 to M step 1
   begin
      for J = 1 to N step 1
if x[i,j]=black and x[i,j+1]=black and x[i,  j-1] =
white and P
      then
         pixel=upper_boundary
         slope=alpha
elseif x[i,j]=black and x[i,j-1]=black x[i,j+1]=
white and P
      then
         pixel=lower_boundary
         slope=-alpha
elseif  x[i,j]=black and x[i,j-1]=white and
x[i,j+1]=white and P
      then
         additional_investigation
   end
end
```

where P is defined as:

```
x[i,j]=white and x[i,j+1]=white and x[i,j-1]=white
```

Additional investigation is made on pixel without complete location. It can be lower, upper or no boundary pixel. It depends on neighbor area of pixels. Apart from [6] and [7] enlarged window RxS pixels is defined as a basis. In this paper R=5 and S=7 is proposed and analysed. Position of window is backwards from pixel candidate for additional investigation. Pixels occurence and position in proposed window led to identification on pixel boundary type.

After additional investigation pixel type is completely located. Throughout previous decision making, algorithm for unwetted areas simply draw area under specified angles. As a result words are bounded by unwetted dark stripes. This situation is given in Fig. 4.

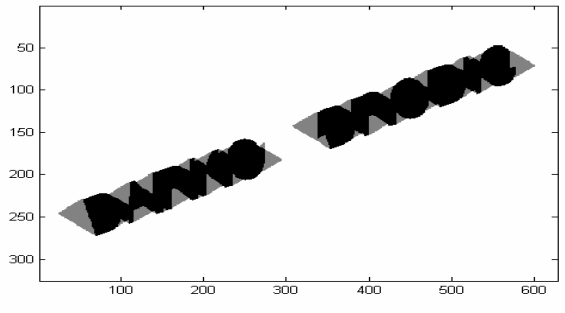

Fig. 4. Document text with unwetted areas

The skew angle detection is based on information obtained from presented algorithm. Defining reference text line means calculating specific average position of every image column. That specific average position is average position of only black pixels in every column of image.

Relation for calculating reference text line is:

$$x_i = \frac{\sum_{j=1}^{L} y_j}{L} \qquad i = 1,\ldots, K \ , \tag{3}$$

where y_j is position of black pixel in column j and L is sum of black pixel in specified column j of an image.

After calculation image matrix with only one black pixel per column is obtained. That black pixel per column defines calculated reference text line and text line skewness. Fig. 5. presented line obtained by reference text line calculation.

Calculated "reference text line" forms discontinuous line partly representing reference text line. To form continuous reference text line from point's collection some numerical method could be used. Candidate methods are interpolation, extrapolation, and combination of interpolation and extrapolation or least squares method [8].

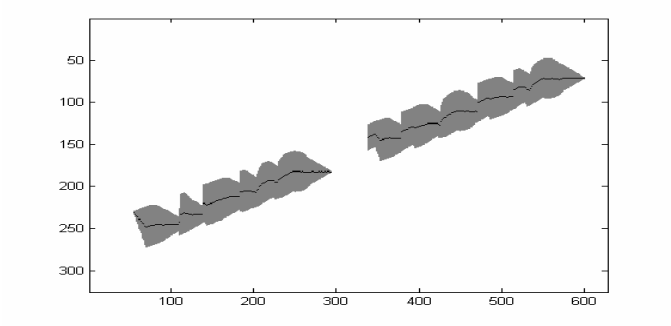

Fig. 5. Calculated "reference text line"

The goal is more common due to linear aspect of reference text line representation from original sample text. But, fundamentally method of fitting data is needed. Least squares method can be interpreted as that. It is an alternative to interpolation for fitting a function to a set of points. Unlike interpolation, it does not require the fitted function to intersect each point. The method of least squares is probably best known for its use in statistical regression. Hence, least square method is used for achieving continuous reference text line from point's collection. The common computational procedure to find a first degree polynomial function approximation in a situation like this is as follows [8]:

1. Use n for the number of data points.
2. Find the four sums:

$$\sum x \ , \ \sum x^2 \ , \sum y \ , \sum xy \ . \tag{4}$$

3. The calculations for the slope, a', and the y-intercept, b', are as follows:

$$a' = \frac{(\sum y)(\sum xy) - n\,(\sum xy)}{(\sum x)^2 - n\,(\sum x^2)} \ , \tag{5}$$

$$b' = \frac{(\sum x)(\sum xy) - (\sum y)(\sum x^2)}{(\sum x)^2 - n\,(\sum x^2)} \ . \tag{6}$$

As a result, continuous reference text line is obtained by:

$$y = a'x + b' \ . \tag{7}$$

III. Experimental results

For the sake of experiment, reference computer sample text no.1 rotated from -20° to +20° by step of 5° around x-axis is used. It is represented in Fig.6.

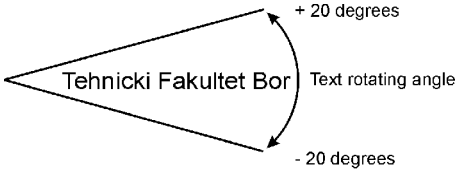

Fig. 6. Sample text rotating from -20° to +20° for the robustness investigation of the algorithm

As a reference, sample text no.2 is introduced to represent handwritten text. It is given in Fig.7.

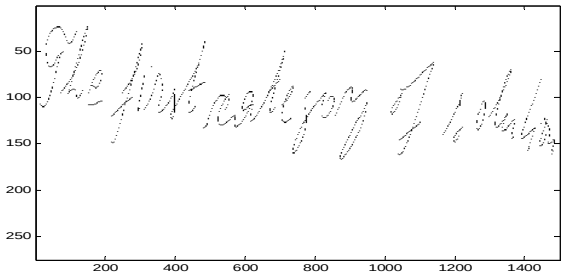

Fig. 7. Sample text no.2 representing handwritten text

Our modified water flow method works with water flow angle from 0° to 90°. Referent line hit rate (RLHR) is defined by:

$$RLHR = 1 - \frac{\beta_{ref} - \beta_{est}}{\beta_{ref}} \ , \tag{8}$$

where β_{ref} = arctan(a) from eq.(1) and β_{est} = arctan(a') from eq.(7).

The best referent line hit rate depends on different water flow angle region. At the beginning, referent line hit rate for "water flow" small angles, i.e. from 0° to 20°, are investigated (see Table I and II). "Water flow" angles smaller than 15° aren't enough robust. Further investigation on varying full range "water flow" angle from 0° to 90° by step 5° gave results presented in Table I and II. In Table II mean value i.e. X, and RMS values are calculated by:

$$\overline{X} = \frac{1}{L}\sum_{i=1}^{L} X_{est} \ , \tag{9}$$

$$RMS = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(X_{ref} - X_{est})^2} \ , \tag{10}$$

where in (9) and (10) L is number of examined text rotating angles range from -20° to +20° by step 5°, X_{ref} is RLHR for β_{est} equal to β_{ref} i.e. due to normalization equal to 1, and X_{est} is RLHR.

Inspecting given results led to conclusion that any "water flow" angle bigger than 20° has enough robustness. This instance is presented in Fig.8.

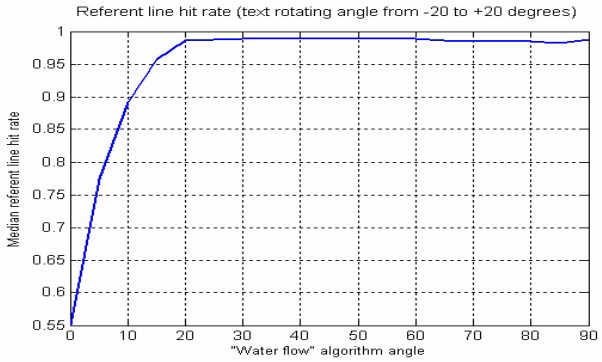

Fig. 8. Referent line hit rate for "water flow" angle from 0 to 90°

Further inspecting Fig.8., the best referent line hit rate is obtained for ”water flow” angle region from 20° to 60°. Specified ”water flow” angle region should be used on further investigation of complicated reference text samples.

Proposed algorithm is suited for printed as well as for handwritten text. Applying algorithm to sample text no.2 proved those claims. Visual results of calculated “reference text line” from sample text no.2 is represented in Fig.9.

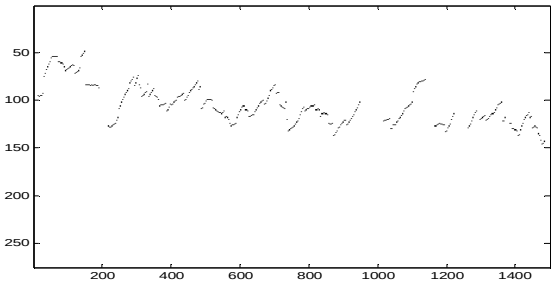

Fig. 9. Calculated ”reference text line” from sample text no.2

As can be seen from Fig.9. algorithm is suited for multi-skew text common in handwritten text. Further investigation will be made on calculating skew rate and reference text line in handwritten text.

IV. Conclusion

In this paper we presented modified algorithm for extraction of reference text line in multi-skewed text image. It assumes hypothetical water flows under different specified angles of the image frame from left to right and vice versa. As a result of algorithm, unwetted image regions defined. Those regions are corner stone needed for reference text line calculation.

Using least square method on calculated reference text line, reference text line is extracted and identified. Robustness of algorithm is examined and validated using computer sample text rotating in region of ±20° around x-axis. Obtained results led to conclusion that ”water flow” angle in region from 20° to 60° is the best suited for high-quality reference line hit rate.

At the end, calculated ”reference text line” from hand-written sample text visually prove that algorithm is suited for multi-skew handwritten text.

References

[1] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri, D.K. Basu, "Text Line Extraction from Multi-Skewed Handwritten Documents", Pattern Recognition, Vol.40, pp. 1825-1839, 2006

[2] Jiren Wang, Mazlor K.H. Leung, Siu Cheung Hui, "Cursive Word Reference Line Detection", Pattern Recognition, Vol.30, No.3, pp. 503-511, 1997

[3] G. Louloudis, B.Gatos, I.Pratikakis, C.Halatsis, "Text Line Detection in Handwritten Documents", Pattern Recognition, Vol.41, pp. 3758-3772, 2008

[4] Yi Li, Yefeng Zheng, D. Doermann, S. Jaeger, "A New Algorithm for Detecting Text Line in Handwritten Documents", 18th International Conference on Pattern Recognition, Vol.2, pp. 1030-1033, Hong Kong, 2006

[5] D. Brodić, Z. Milivojević, “Uniform Modified Method for Handwritten Text Reference Line Detection”, MIPRO 2009, CIS Section, paper no.26, Opatija, Croatia, 2009

[6] R.C. Gonzalez, R.E. Woods, “Digital Image Procesing”, 2nd ed., New Jersey: Prentice-Hall, 2002, pp. 67-70

[7] M. Sonka, V. Hlavac, R. Boyle, “Image Processing, Analysis and Machine Vision”, Toronto: Thomson, 2008, pp. 174-177

[8] W.M. Bolstad, “Introduction to Bayesian Statistics”, New Jersey: John Wiley & Sons, 2004, pp. 40-44, 235-240

TABLE I
REFERENCE LINE HIT RATE FOR WATER FLOW ANGLE VARYING FROM 0° TO 90° AND FOR TEXT ROTATING FROM -20° TO 20°

Text	“Water Flow” Angle																		
Rotation	0°	5°	10°	15°	20°	25°	30°	35°	40°	45°	50°	55°	60°	65°	70°	75°	80°	85°	90°
-20°	0.4988	0.6164	0.7362	0.8607	0.9860	0.9879	0.9885	0.9882	0.9879	0.9868	0.9863	0.9893	0.9887	0.9871	0.9865	0.9860	0.9854	0.9860	0.9860
-15°	0.4983	0.6588	0.8223	0.9888	0.9907	0.9907	0.9910	0.9907	0.9899	0.9899	0.9888	0.9862	0.9888	0.9873	0.9851	0.9843	0.9847	0.9851	0.9851
-10°	0.5048	0.7408	0.9870	0.9892	0.9904	0.9915	0.9915	0.9915	0.9909	0.9904	0.9887	0.9858	0.9881	0.9824	0.9796	0.9796	0.9779	0.9779	0.9779
-5°	0.4851	0.9737	0.9771	0.9783	0.9817	0.9840	0.9828	0.9828	0.9828	0.9817	0.9794	0.9817	0.9657	0.9668	0.9634	0.9622	0.9622	0.9565	0.9600
0°	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900	0.9900
5°	0.4874	0.9760	0.9794	0.9805	0.9817	0.9851	0.9863	0.9874	0.9874	0.9886	0.9931	0.9989	1.0023	1.0023	1.0046	1.0080	1.0092	1.0103	1.0126
10°	0.4980	0.7368	0.9813	0.9841	0.9858	0.9875	0.9887	0.9898	0.9909	0.9926	0.9949	0.9949	1.0000	0.9960	0.9938	0.9960	0.9977	0.9994	1.0011
15°	0.4976	0.6566	0.8163	0.9828	0.9851	0.9858	0.9862	0.9869	0.9877	0.9881	0.9899	0.9903	0.9922	0.9910	0.9881	0.9873	0.9884	0.9903	0.9914
20°	0.4993	0.6169	0.7376	0.8579	0.9868	0.9887	0.9887	0.9893	0.9901	0.9907	0.9901	0.9912	0.9854	0.9780	0.9706	0.9703	0.9722	0.9437	0.9893

TABLE II
REFERENCE LINE HIT RATE MEAN AND RMS FOR WATER FLOW ANGLE VARYING FROM 0° TO 90° AND FOR TEXT ROTATING FROM -20° TO 20°

Text	“Water Flow” Angle																		
Rotation	0°	5°	10°	15°	20°	25°	30°	35°	40°	45°	50°	55°	60°	65°	70°	75°	80°	85°	90°
Mean	0.5510	0.7740	0.8919	0.9569	0.9865	0.9879	0.9882	0.9885	0.9886	0.9887	0.9890	0.9898	0.9890	0.9868	0.9846	0.9849	0.9853	0.9821	0.9881
RMS	0.1647	0.1606	0.1119	0.0555	0.0034	0.0026	0.0027	0.0026	0.0026	0.0031	0.0044	0.0051	0.0104	0.0103	0.0122	0.0134	0.0138	0.0206	0.0146