International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:4, 2009

Word base line detection in handwritten text recognition systems

Kamil R. Aida-zade and Jamaladdin Z. Hasanov

Abstract—An approach is offered for more precise definition of base lines' borders in handwritten cursive text and general problems of handwritten text segmentation have also been analyzed. An offered method tries to solve problems arose in handwritten recognition with specific slant or in other words, where the letters of the words are not on the same vertical line. As an informative features, some recognition systems use ascending and descending parts of the letters, found after the word's baseline detection. In such recognition systems, problems in baseline detection, impacts the quality of the recognition and decreases the rate of the recognition. Despite other methods, here borders are found by small pieces containing segmentation elements and defined as a set of linear functions. In this method, separate borders for top and bottom border lines are found. At the end of the paper, as a result, azerbaijani cursive handwritten texts written in Latin alphabet by different authors has been analyzed.

Keywords—azeri, azerbaijani, latin, segmentation, cursive, HWR, handwritten, recognition, baseline, ascender, descender, symbols.

I. Introduction

Text recognition is being one of the most researched areas for last 20 years. Generally, it is divided into 3 parts.

- Printed text recognition. Mainly used for reformatting, changing the content of the printed documents. Mostly used in publishing houses, universities, libraries, offices, etc.
- Hand printed text recognition. This text recognition approach is used for automatic processing of forms, surveys, etc. Main applied areas are post offices, service companies, banks. In this approach, document to be filled usually has particular design and handprinted letters are written inside the predefined boundaries.
- Handwritten text recognition (HWR).

Handwritten text recognition is used for reducing the manual text input time. In this case, operator's time is spent only for the error correction, which percentage depends on recognition system itself. From the counted 3 recognition models, the HWR is being the most complicated by the structure and functionality. In printed text recognition, all letters of the same font do not vary in shapes, which makes the recognition system's work easier, avoiding the complicated symbol classification. In handprinted text recognition systems, texts usually vary depending authors but here text boundaries are previously defined, which helps the recognition system in segmentation part. In HWR, shapes, thickness, accuracy of all letters or symbols vary depending an author, and additionally, different writers write in different manners (using connecting ligatures, word spacing, distance between the lines, etc) which impacts the accuracy of the recognition result and adds complexity to a recognition system. Handwritten text recognition can be subdivided into 2 categories: off-line recognition [1], [2] and on-line recognition [3], [4]. On-line recognition deals with real-time data processing and has ability to integrate pen movement and pressure information. Off-line recognition, however, is based on a static input of the data and relies only on pixel information for the recognition of each word [5]. Both on-line and off-line recognition approaches use various methods like segmentation, feature extraction, neural network or HMM based learning/recognition for script recognition. For off-line handwritten cursive recognition, as features, the ascending and descending parts of the letters or words are also being used. These recognition systems mainly based on human reading model which classifies and recognizes the words by their shape. Various shape detection algorithms for different languages and alphabets were offered and in this paper, we describe a baseline detection method for cursive handwritten words.

II. PROBLEM STATEMENT

In contemporary handwritten recognition systems, as an informative parameters for learning and recognition, various feature extraction methods are being used. This feature extraction methods vary depending on handwritten script and recognition method. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

In cursive handwritten text recognition as a features, the word shape and ascending/descending parts of the letters are also used. Ascending/descending parts of the letters give additional information about the letters which helps to classify them and rise recognition result [7]. The general idea is that we see words as a complete patterns rather than the sum of letter parts. Some claim that the information used to recognize a word is the pattern of ascending, descending, and neutral characters. Another formulation is to use the envelope created by the outline of the word. The word patterns are recognizable to us as an image because we have seen each of the patterns many times before. James Cattell (1886) was the first psychologist to propose this as a model of word recognition. Cattell is recognized as an influential founder of the field of psycholinguistics, which includes the scientific study of reading.

In this paper, we offer a method for precise identification of the ascending/descending parts of the words. The ascending/descending parts of the words are found after the base line detection - the lines or ligatures above or below the word base lines are supposed to be ascenders and descenders respectively. Some border line detection methods are based on vertical histogram analysis and find these border lines as a one solid line [6]. Unfortunately, this method doesn't satisfy the requirements of HWR systems which process scripts with various kinds of slant and slope parameters or words where letters are not located on the same horizontal line. In this paper, a base line detection method for these kind of words is reviewed.

Ideally, the borders of the body part of the words should be continuous curve lines. In Fig. 1, the ideal baseline borders of "riyaziyyat" word is shown. As shown from this picture,

Fig. 1. Ideal base line borders for "riyaziyyat" word.

the ideal borders are curved lines instead of lines as offered in [6]. By defining the horizontal axis as x, the length of the word as X, the ideal borders of the base line shown in Fig. 1 could be defined as next polynomial:

$$\begin{aligned} u(x) &= a_1x^n + a_2x^{n-1} + ... + a_{n-1}x + a_n, \\ v(x) &= b_1x^n + b_2x^{n-1} + ... + b_{n-1}x + b_n, \\ x &\in [0, X]. \end{aligned} \quad (1)$$

Here, $u(x)$ and $v(x)$ are the functions of the upper and bottom borders of the base line, respectively. Definition of baseline borders' functions as described in (1) might require complicated and high-cost calculations. Therefore we suggest to define these borders as l-piece piecewise linear functions.

$$\begin{aligned} u(x) &= a_{0i}x + a_{1i}, \ x \in [x_{i-1}, x_i), \ i = 1, .., l, \\ v(x) &= b_{0i}x + b_{1i}, \ x \in [x_{i-1}, x_i), \ i = 1, .., l. \end{aligned} \quad (2)$$

Whereas:

x_{i-1}- start of i-th part, $i = 1, .., l, x_0 = 0, x_l = X$;

$a_{0i} = \tan(\alpha_i)$, where α_i - the angle of border line in i-th part;

a_{1i} - vertical shift of the border line in i-th part.

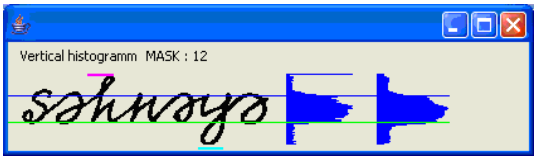
Note. The polygonal lines described in equation (2) might not be continuous:

$$\begin{aligned} a_{0i}x + a_{1i} &\neq a_{0i+1}x + a_{1i+1}, \\ b_{0i}x + b_{1i} &\neq b_{0i+1}x + b_{1i+1}. \end{aligned} \quad (3)$$

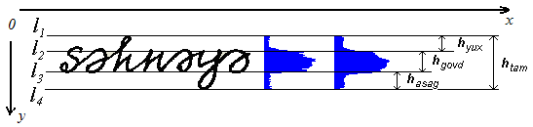
For word baseline's border determination, as described in (2) equation, the words are analyzed by parts. In other words, the border lines for each slice of the word is defined and merged together.

III. BASELINE DETECTION METHOD

Recall from the previous part, to find the handwritten word's baseline borders, the vertical density histogram of the word is used [6], [9]. In Fig. 2(a), the baseline and border line detection by mentioned method for "səhnəyə" word is shown.



(a) pixel density histogram



(b) baselines of the word

Fig. 2. Histogram based baseline and border line detection

At the start of the handwritten text recognition, the text lines are extracted. After that, each text line is divided into words. For word's baseline detection, a line extraction method offered by Srihari and Bozinovic [9] is used and some changes considering handwritten Latin azerbaijani text specifics were applied.

After baseline detection, 4 horizontal border lines are defined [9] : top border line of word; top border line of baseline; bottom border line of the baseline; bottom border line of word.

These lines are defined as l_1, l_2, l_3, l_4 which are located in top to bottom order on vertical y axis (Fig. 2(b)). The calculation algorithm for finding these border lines goes through these steps:

1) $h_0(y)$ vertical density histogram of the word is calculated, $y \in [0, h_{tam}]$, h_{tam} - height of the word and equals to histogram width.
2) Smoothed $s_0(y)$ histogram is calculated:

$$s_0(y) = \sum_{i=-\Delta y}^{i=\Delta y} = h_0(y + i).$$

Here, Δy is a smoothing parameter.In their paper, Srihari and Bozinovic define Δy value as 5 pixels. This parameter may vary depending the size of the word image. In other words, it depends on parameters like author script, scanner, paper and pen quality. As a result of experiments on Latin handwritten texts, for azerbaijani handwritten text, the value of Δy is defined to be as $0.1h_{tam}$. In this step, raw histogram is smoothed around Δy pixels.
3) Starting from the top, the fist position where $s_0(y)$ doesn't equal to zero, marked as l_1 point.
4) Starting from bottom, the first position where $s_0(y)$ doesn't equal to zero defined as l_4 point.
5) Starting from the center of the l_1 and l_4 towards the l_1 , position where $s_0(y)$ is closer to zero, is redefined as l_1.

6) Starting from the center of the l_1 and l_4 towards the l_4, position where $s_0(y)$ is closer to zero, is redefined as l_4. Note. In steps 5 and 6, Srihari and Bozinovic check for equality with zero, whereas we try to find the closest value to zero. This closeness is defined as:

$$s_0(y) < \varepsilon;\ \varepsilon = 0.014X.$$

Where, X - is the pixel width of the word. The value of ε is found experimentally for azerbaijani handwritten text.

7) The peak point where $s_0(y)$ has its maximum value is defined as p.
8) 1_3 point is defined as a d_0 local minimum value between p and l_4.
9) l_2 point is defined as a d_0 local minimum value between l_1 and p.
10) If l_1 and l_2 points are very close to each others, then

$$l_1 = l_2 - max((l_4 - l_3), (l_3 - l_2)) \quad |l_1 - l_2| < dy.$$

11) If l_3 and l_4 points are very close to each others, then

$$l_4 = l_3 - max((l_2 - l_1), (l_3 - l_2)) \quad |l_3 - l_4| < dy.$$

The redefinition of l_1 and l_4 in steps 10 and 11 is used to fix these values for words with no ascending or descending parts. Otherwise, due to noisy elements on the top or the bottom parts of the word, these parts might be assumed as a ascending or descending ligatures, respectively. For prevention of these cases, the value of l_1 and l_4 changed in steps 10 and 11. In other words, the values of h_{yux} and h_{asag} parameters are increased and all noisy elements on top and bottom parts are ignored.

In Fig. 2 the baseline detection shown for normal word with no slant or slope, whereas in practice, it's hard to find handwritten words where all letters are ideally written on the same horizontal line (Fig. 3).

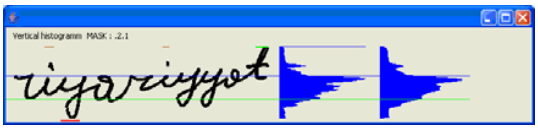
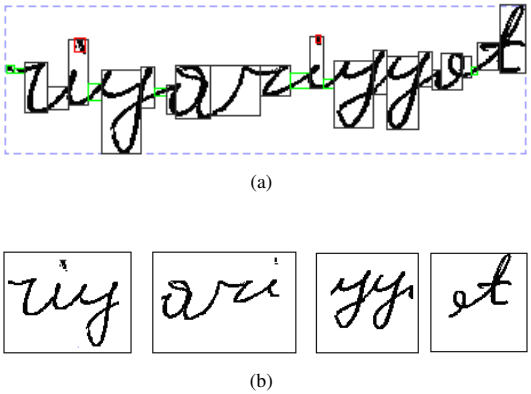


Fig. 3. Baseline detection by standard method.

In this case, due to wrong baseline determination, the ascending or descending parts defined by distance from the top and bottom borders of the body area, might also be found incorrectly [10]. As shown in Fig. 3, the first "y" letter's lower part is enough below from the bottom border and that's why it's correctly detected as a letter with a descending part (a short horizontal line under or upper the letter shows the found ascending or descending parts, respectively). At the same time, the second and third "y" letters in the same word don't satisfy the described rule and thereafter weren't found as a letter with descending parts.

For defining the baseline borders as shown in equation (2), this method splits word into parts and analyzes each part. At the beginning, the word is segmented and the borders of word elements (letters, symbols, ligatures) are found (Fig. 4(a)). Afterwards, each word is split into parts with $N = \frac{X}{kh_{tam}}$ segmentation elements on each (Fig. 4(b)), where k is an author dependent coefficient. In this paper for simplicity we use $k = 3$ value. The word could be split by pixel length also but in this case, there's a risk where one letter might be split into 2 or more parts. In segmentation part based split, all letters in split words remain whole. After split, vertical density histogram for each part is calculated and baseline border by Srihari and Bazinovic classical method is determined (Fig. 5). As shown from the picture, the baseline borders for word parts seems to be found correctly (in 3rd part, last two "y" letters) but there're still deficiencies.



(a)

(b)

Fig. 4. Segmentation and splitting the "riyaziyyat" word.

In paper [11] a very useful research on HWR is done, slope texts specifics analuzed. According to this paper, maximum slope angle is around 20°. Taking into account this fact, we take vertical density histogram around 20° for each word piece and the optimal baseline borders are considered as a final result. As a criteria for optimum, we get the maximum histogram value difference for top and bottom borders among the all variants (Fig. 5):

$$F_u(\alpha) = \max_{y\in[l_1, l_2+\frac{l_3-l_2}{2})} |s_0(y;\alpha) - s_0(y-\Delta s;\alpha)|$$
$$F_v(\alpha) = \max_{y\in[l_2+\frac{l_3-l_2}{2}, l_4)} |s_0(y;\alpha) - s_0(y-\Delta s;\alpha)|$$

Where α - rotate angle, $\alpha \in [-20°, 20°]$;

$s_0(y;\alpha)$- smoothed vertical density histogram function of rotated image with α angle;

Δs - step used for threshold detection and its experimentally calculated value vary around 1-4 pixels. In this work, we set $\Delta s = 1$;

$F_u(\alpha), F_v(\alpha)$ - maximum threshold value functions respectively for top and bottom borders.

α_u^* and α_v^* values for top and bottom borders of the baselines for given word pieces are found by equation (4). After defining the α_u^* and α_v^* values, the piecewise linear functions for borders are defined.

$$\alpha_u^* = \arg\max_{\alpha\in[-20°,20°]} F_u(\alpha)$$
$$\alpha_v^* = \arg\max_{\alpha\in[-20°,20°]} F_v(\alpha) \qquad (4)$$

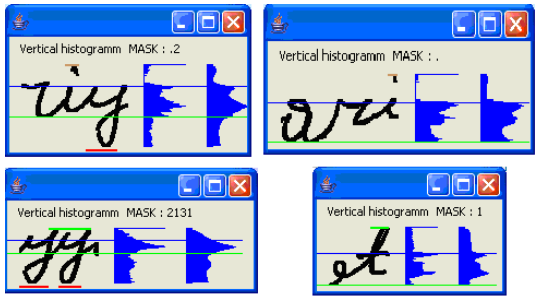
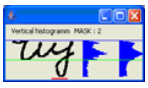
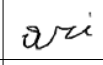
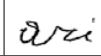
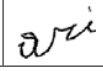


Fig. 5. Vertical density histogram analysis for each piece of the "riyaziyyat" word.

TABLE I
VERTICAL DENSITY HISTOGRAM ANALYSIS FOR "RIY" AND "AZI" WORDS UNDER 0, 10°, -10° ANGLES

	image	histogam	image	histogram
0°				
10°				
-10°				

By definition of the values of α_u^* and α_v^*, the values of l_2 and l_3 parameters are also defined, which correspond to a_{1i}, b_{1i} coefficients in equation (2).

Note. By analyzing the values of α_{ui}^* and α_{vi}^* for each i-th segment, there is a possibility to take a decision about the accuracy of the author: if the values of α_{ui}^* and α_{vi}^* don't differ for all word pieces, then this word might be supposed to be written on the same horizontal line. Otherwise the handwritten script assumed as inaccurate. Different methods for accuracy measurement might be used. One of them could be written like:

$$A_u = \sum_{i=1}^{L} (\alpha_{ui}^*)^2, \;\; A_v = \sum_{i=1}^{L} (\alpha_{vi}^*)^2 \qquad (5)$$

Where

A_u and A_v - accuracy parameter for top and bottom border line, respectively. For sign avoidance, the square power operator is used in this formula which also creates a sensitivity to higher angles.

For correcting or smoothing the border lines, the value of k can be changed. After each processing, the accuracy could be measured and the best result saved. There should be minimal value for k which should prevent the oversplit part problem, where event each letter might be analyzed.

In Fig. 6, the vertical density histogram analysis for rotated images are shown. As shown from this picture, the highest jump is on the second picture which is the vertical density histogram of the 10° rotated image. The optimal border line angle might differ for top and bottom line.

In table I, histograms of "riy" and "azi" subwords taken under 0°, -10°, 10° angles and baselines are shown. As shown from figures, for the same subword, best upper and bottom border lines could be under different angles. For example, for "riy" subword, the best upper border line is taken under 0° angle, whereas best bottom border line is calculated under 10° angle.

As a result of this method, the border lines of the "riyaziyyat" word is determined and shown in Fig. 7. Compared to Fig. 3, here the borders are linear, but close to original.

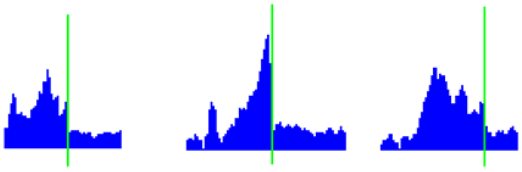


Fig. 6. Defining the $v(x)$ bottom border line for "riy" word piece under 0°, 10°, -10° rotation.

IV. RESULTS

For measuring the effectiveness of the offered method we analyzed the recognition results of the handwritten documents written by 6 authors. As a recognition system, 3-layer Azerbaijani cursive HWR system developed by K.R. Aida-zade and J.Z. Hasanov [8] is used. In this system, as a final step the lexicon based search is performed which ensures the recognition result at the output of the neural network. In this final step, an ascending/descending part sequence mask for each word is found and lexicon search is done by both mask and recognition results. In our experiment, all authors have

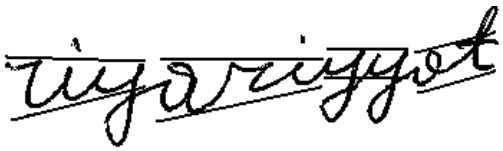


Fig. 7. Baseline borders of "riyaziyyat" word found by offered method.

been instructed about writing with a slope for challenging the described recognition system and offered method. In table II the recognition results are shown. In first column the index of the author takes place, the second column stores the words written by the corresponding author. The third and last columns show the recognition rate (count of correctly recognized words to the word count) for recognition without (before) and with offered baseline detection method (after), accordingly. As shown from table II, recognition rates are not satisfactory in the "before" column, which can be explained by over-sloped text samples. Even in this situation, the offered method helped to increase (somewhere more than twice) the recognition rates.

As mentioned at the start of this paper, different approaches and methods for cursive HWR were offered. An offered

TABLE II
RECOGNITION RESULTS

author	words	before	after
1	208	35 %	77 %
1	208	35 %	77 %
2	122	47 %	71 %
3	123	33 %	69 %
4	201	58 %	74 %
5	90	19 %	55 %
6	104	28 %	58 %

method might be useful not only for azerbaijani Latin text but for other Latin alphabet languages. The accuracy of this method depends on segmentation results.

REFERENCES

[1] C.Faure and E.Lecolinet. OCR: Handwriting. In R.A.Cole et al, editor, "Survey of the State of the Artin Human Language Technology", *Center for Spoken Language Understanding*, Oregon Graduate Institute, pp 86-89, 1995.

[2] W. Senior. "Off-line handwritten recognition: A review and experiments". Technical Report CUED/F-INFENG/TR105, Cambridge University Engineering Department, Dec. 1992.

[3] C.Higgins and P.Bramall. "A non-line cursive script recognition system". *Handwriting and Drawing Research - Basic and Applied Issues*, IOS Press, pp. 285-298, 1996.

[4] R. K. Powalka, N. Sherkat, L. J. Evett and R. J. Whitrow. "Dynamic cursive script recognition: A hybrid approach". In Advances in Handwriting and Drawing: A multidisciplinary approach, 1994.

[5] S. Wesolkowski. "Cursive script recognition: A survey". *Handwriting and Drawing Research -Basic and Advanced Issues*, IOS Press, pp. 267-284, 1996.

[6] R.G. Casey and E. Lecolinet. "Strategies in character segmentation: A survey". *Proc. of the 3rd International Conference on Document Analysis and Recognition*, Montreal, Canada, pp. 1028-1033, 1995.

[7] K.R. Ayda-zadə. C.Z. Həsənov. Latin əlifbali əlyazmalari tanima sistemlərində sozlərin seqmentlənməsi ucun usul. *AMEA XEBERLERI. F.-t. və r. elmləri seriyasi*. XXVI cild. 3. 2006.

[8] K.R.Aida-zade, J.Z. Hasanov. "Handwritten recognition system for azerbaijani latin text". *Proc. of PCI 2008 International conference*, pp. xx-yy, 2008.

[9] S. Srihari and R.Bozinovic. "Multi-level perceptron approach to reading cursive script". *Artificial Intelligence*, vol. 33, pp. 217-255, 1987.

[10] A.W. Senior. "Off-line cursive handwritting recognition using recurrent neural networks", Trinity Hall, Cambridge, England, 1994.

[11] B. Yanikoglu and P. A. Sandon. "Segmentation of off-line cursive handwriting using linear programming". *Pattern Recognition*, Vol. 31, No. 12, pp. 1825-1833, 1998.