Technical Sciences 18(4), 2015, 313–321

MATHEMATICAL PROBLEM OF STRAIGHTENING TEXT LINES

***Oleksandr Tymchenko*[1], *Irena Kulczycka*[2], *Karolina Szturo*[1]**

[1] Department of Safety Engineering
University of Warmia and Mazury
[2] Ukrainian Academy of Printing, Lviv

Received 19 February 2015; accepted 31 July 2015; available on line 17 August 2015.

K e y w o r d s: recognition, optical recognition system, text document image, image distortion.

A b s t r a c t

The rapid distribution of digital cameras has caused several new problems related to text recognition. Based on experimental studies, was revealed that existing OCR systems cannot cope with complex perspective and geometric distortions that arise when photographing text document. Therefore it is necessary to apply text documents pre-processing so that the text lines were straight and horizontal. This article briefly consider existing methods pre-processing documents and found that it depend on the type of distortion and not universal. Proposed new method involving the mathematical raising of straightened text lines on the image and heterogeneous distortion correction based on a page surface transformation model. This method is better than others because it is universal and corrects any type of distortion, including a combination of several types of distortion.

Introduction

Very often there is a need to convert a paper text document or book into an electronic form (ARMS 2000). The process of translating paper documents into a digital form is carried out by a scan or photography. Because it is difficult to work with text document images (it is sometimes necessary to edit some part of the text), it is more convenient to present this image in a text editor. Optical recognition systems are often sufficient to deal with this task (such as FineReader, Omnipage, Readlris). The optical text recognition system acquires digital representation of the scanned or pictured document and has to form a text which is contained in this image, in a form suitable for saving in an electronic text document format.

Correspondence: Oleksandr Tymchenko, Katedra Inżynierii Bezpieczeństwa, Uniwersytet Warmińsko-Mazurski w Olsztynie, ul. Oczapowskiego 11, 10-719 Olsztyn, phone: 89 524 61 25

Optical recognition systems can generally recognize high quality images with high enough precision. However, if an image has some distortions (this problem is very often found in photographed text documents), the quality of recognition is considerably worsened and, sometimes, the process of recognition becomes generally impossible. Such images need previous geometrical correction of existing distortions to ensure that the lines of text on the image are direct and horizontal.

The existing methods (MASALOVICH 2007, FU et al. 2007, YIN et al. 2007) are based on certain models distortion document and depend on the type of distortion (distortion types discussed below). First out some text lines and text lines distortion function is constructed on the image. Then, based on this information, there is straightening the image. If only geometric distortions present in the document (when scanning thick books), then the total distortion function documents are taking from information of distortions of two text lines and build a linear approximation between them (FU et al. 2007). If there are only perspective distortions on the image, then it is enough to find the point of intersection of the lines on the image (YIN et al. 2007). To find the point of intersection is possible to construct a linear approximation for each word in the image and find the point at which intersect the continuation of all received segments. All of analyzes methods have their drawbacks, are not universal and can be used in the case when only one of all the types of distortion are on the image. Therefore, recognition of distorted test documents remains an actual problem.

Types of distortions. Maximal possibilities of the optical character recognition systems

To allow this recognition system to recognize a text document image without errors, all text lines must be straight and horizontal. However, after scanning or photographing there are often problems which can result in a worsening of image quality and the recognition will become impossible. Several algorithm is a result of recognition:

1) A skewed page which is characterized by the turning of all of the text lines towards a corner. Such curvature can appear as a result of unequal position of a document in a scanner.

2) Geometrical distortions can appear during the scanning of thick books in the area of the bend of a book.

If an image is acquired by a digital photocamera, except for previous problems, there can be perspective distortions and more difficult geometrical

distortions, related to inequality of the initial document (incurving, concavity, etc.) which is hard to predict.

As shown in Figure 1*a*, an image defect usually takes a place when the image plane of a digital camera (R) is parallel with the document (plane D). If the image plane of camera R is not parallel to plane (D), perspective distortions appear, as shown in Figure 1*a*. Geometrical distortions appear when scanning or photographing thick books (in the area of the bend), when a text is placed on a smoothly curved and not flat surface (Fig 1*b*).

Note that there are not only damaged documents considered in this paper; distortion in scanning documents related with damage of a document eg. crease or tear are not analyzed.

From OCR Software rating (OCR Software Review, on line) for test we selected such OCR programs as OmniPage, ABBY FineReader and Rediris. Based on experimental studies, the maximal recognition possibilities of the disfigured texts in this system are:

1) Distortion of image. ABBY FineReader and Rediris recognizes images, where the angle between the horizontal line of image and the line of text is less than 25 degrees. If this angle is more than 25 degrees, the system does not recognize an image and reports an error (it did not find text characters in the image); pretreatment of the image is needed. OmniPage corrects documents with any angle of text strings and recognizes characters with almost no mistakes.

2) Perspective and geometrical deformation. All systems recognizes an image with a great amount of errors. The quality of recognition is poor.

None of this programs cannot correct non-linear distortion. So, it is obvious that for high-quality character recognition of a text document, pretreatment of such images is necessary.

Definition of the mathematical problem of straightening text lines

The task of straightening text lines in the picture is formulated as follows (MASALOVICH 2007):

If an image I is given on rectangular area $m \times n$, the image can belong to the plural of images with the distorted lines $£_{\text{damaged}}$ or to the plural of images without distortions of lines $£_{\text{normal}}$. It is necessary to build a reflection ϕ of source image I, to get such an image $\bar{I}$, that will satisfy the following terms:

1) if the initial image lines were distorted, on a regenerated image, where lines are practically level, the quality of recognition must be better by far, than on the initial;

2) if there were direct lines on an initial image, the regenerated image quality of recognition must be not worse than on the initial image.

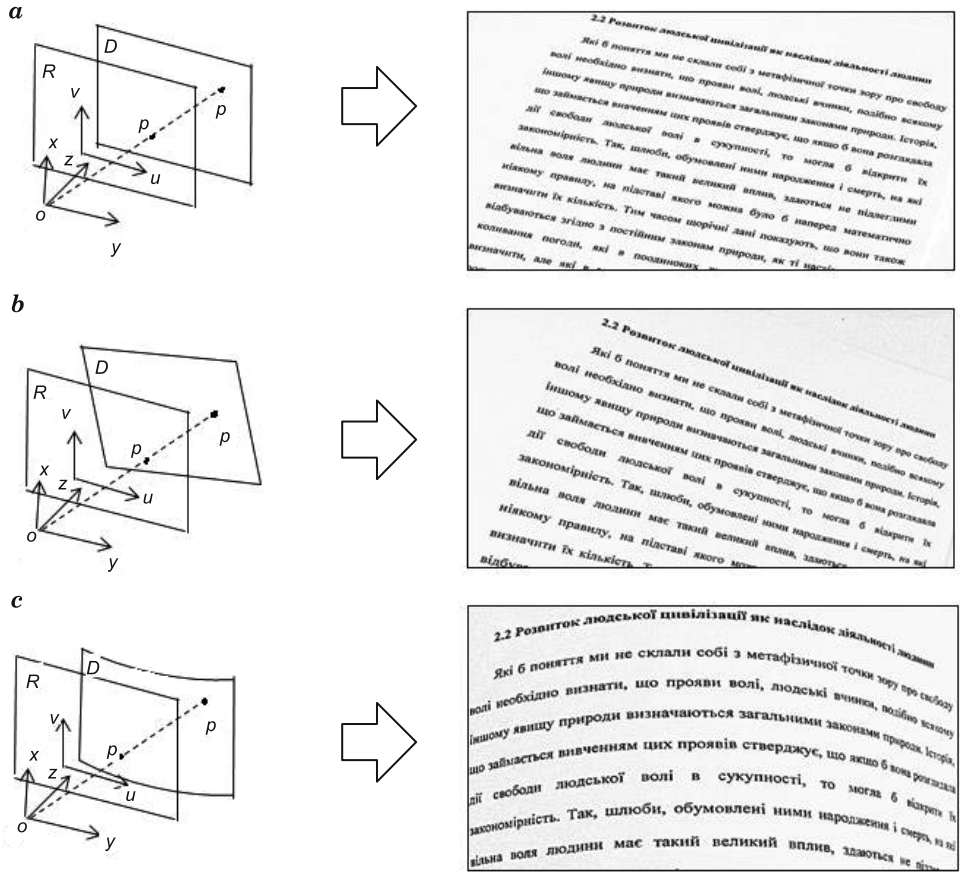

Fig. 1. Issues of the image projections

The digital image can be represented as a rectangular matrix of size $m \times n$. Each element of the matrix (pixel) a_{ij}, $i = \{1, ..., m\}, j = \{1, ..., n\}$ is a color in the corresponding image point. If the image is binary, then $a_{ij} = \{1,0\}$ [6,7].

The task of straightening the text lines can be described by the following mathematical formula:

$$\bar{I} = \Phi^{-1}(I) = \{\bar{C}(x,y) = C(\Phi_x(x,y), \Phi_y(x,y))\} \quad (1)$$

$C{:}R^2 \rightarrow \{0; 1\}$ – presentation of binary image I is as a two – dimensional function of color

$$C(x,y) = \begin{cases} 1, & if\ x < 0 \wedge x > m \wedge y < 0 \wedge y > n \\ 0, & if a_{ij} = \text{white};\ i = [x], j = [y] \\ 1, & if a_{ij} = \text{black};\ i = [x], j = [y] \end{cases} \quad (2)$$

$\Phi(t, u)$: $R^2 \to R^2$ – function of image transformation:

$$\Phi(t, u) = \begin{cases} \Phi_x(t, u) \\ \Phi_y(t, u) \end{cases};\ \Phi_x(t, u): R^2 \to R,\ \Phi_y(t, u): R^2 \to R \quad (3)$$

The ultimate goal of every text document image distortion correction algorithm is improvement of the results of recognition of the corrected image. As it is impossible to create an ideal initial document, the efficiency of the straightening limits are estimated with segments, by algorithm will accept a value which equals the difference between a value which determines the quality of text recognition before the straightening of text lines and a value which determines the quality of text recognition after straightening text lines.

Correction method by the construction of a page surface transformation model

It is necessary to build a model of transformation to represent the projection of an extended surface in a two-dimensional rectangular area. To find the projection of an extended surface it is necessary to estimate the text limits of the document. The left and right text limits are estimated using segments by the least-squares method, which base on of all of the discovered most left/right points, except for those points of text lines which do not begin from the beginning of document (titles, cross-headings). For the estimation of the highest and lowest bound of text, the polynomial of the third degree least-squares method is also utilized.

Thus, we will have two segments-off: AD, which relies on the left text bound

$$y = a_l x + b_l \quad (4)$$

and BC, that relies on the right limit of text

$$y = a_r x + b_r \quad (5)$$

the equation of the curved highest bound of text AB will look like:

$$y = a_{u1}x^3 + a_{u2}x^2 + a_{u3}x + a_{u4} \tag{6}$$

the same for curve BC:

$$y = a_{l1}x^3 + a_{l2}x^2 + a_{l3}x + a_{l4} \tag{7}$$

It is necessary to perform a transformation to represent the projection of the curved surface, limited by curves AB, DC and by lines AD, BC in a two-dimensional rectangular area. Let the $A'(x'_1,y'_1)$, $B'(x'_2,y'_2)$, $C'(x'_3,y'_3)$, $D'(x'_4,y'_4)$ – angular points of rectangular area (Fig. 2). Let $|\widehat{AB}|$ be length of arc between points A and B, $|AB|$ – Euclidean distance [8] between points A and B.

Width W of the rectangular area determined as:

$$W = \min\left(|\widehat{AB}|, |\widehat{DC}|\right) \tag{8}$$

Height H of the rectangular area equals:

$$H = \min\left(|AD|, |BC|\right) \tag{9}$$

In our example, (Fig. 2) $W = |\widehat{AB}|$, $H = |AD|$.

The angular points of the rectangular area are calculated as follows:

$$\left.\begin{array}{ll} x'_1 = x_1, & y'_1 = y_1 \\ x'_2 = x'_1 + W, & y'_2 = y'_1 \\ x'_3 = x'_2, & y'_3 = y'_2 + H \\ x'_4 = x'_1, & y'_4 = y'_3 \end{array}\right\} \tag{10}$$

The function Φ now creates the accordance between curves AB and DC.

$$\Phi(E(x_u,y_u)) = G(x_l,y_l), \text{ if } \frac{|\widehat{AE}|}{|\widehat{AB}|} = \frac{|\widehat{DG}|}{|\widehat{DC}|} \tag{11}$$

Where $E(x_u,y_u)$ – is a point-on curve AB, $G(x_l,y_l)$ – is a point-on-curve DC

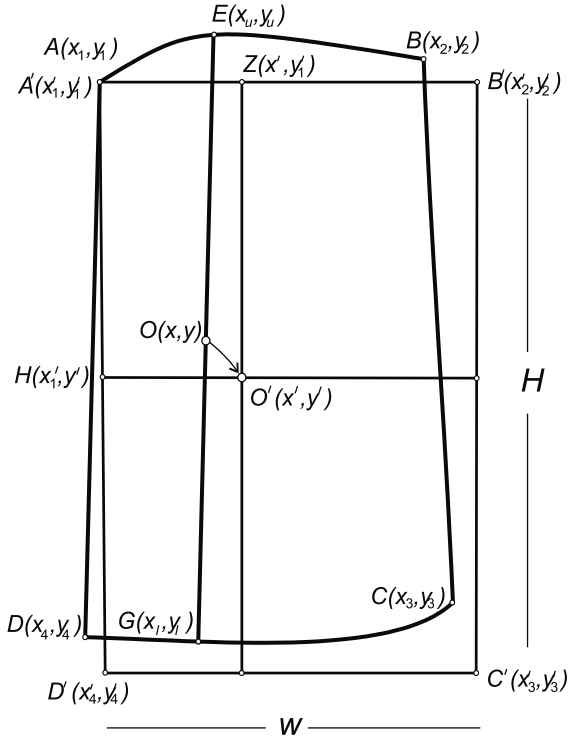

Fig. 2. Projection of curved surface

All points from the projection of curved surface determine the new position. Let $O(x,y)$ be a point on the projection of a curved surface. Our task is to define the new position $O'(x',y')$ of point $O(x,y)$ (Fig. 2).

First, we will define a line EG, which satisfies the next terms:

1. It crosses points $E(x_u,y_u)$, $G(x_l,y_l)$ that lie on curves AB and DC, accordingly.
2. $\Phi\,(E(x_u,y_u)) = G(x_l,y_l)$
3. Point $O(x,y)$ belongs to the line EG.

Farther calculate position $O'(x',y')$:

$$x' = x'_1 + |A'Z| \tag{12}$$

$$y' = y'_1 + |A'H| \tag{13}$$

where H – is a point $H(x'_1,y')$ Z – is a point $Z(x',y'_1)$. The lengths of segments $|A'Z|$, $|A'H|$ are calculated as follows:

$$\frac{\widehat{|AB|}}{\widehat{|AE|}} = \frac{W}{|A'Z|} \Rightarrow |A'Z| = \frac{W}{\widehat{|AB|}} \widehat{|AE|} \tag{14}$$

$$\frac{|EG|}{|EO|} = \frac{H}{|A'H|} \Rightarrow |A'H| = \frac{H}{|EG|} |EO| \tag{15}$$

Repeat this sequence of executions for all points from the projection of a curved surface. If a point is out of an area, it accepts transformation of the nearest point.

Figure 3 shows the experimental application of this method to images taken with a digital camera (with geometric distortion).

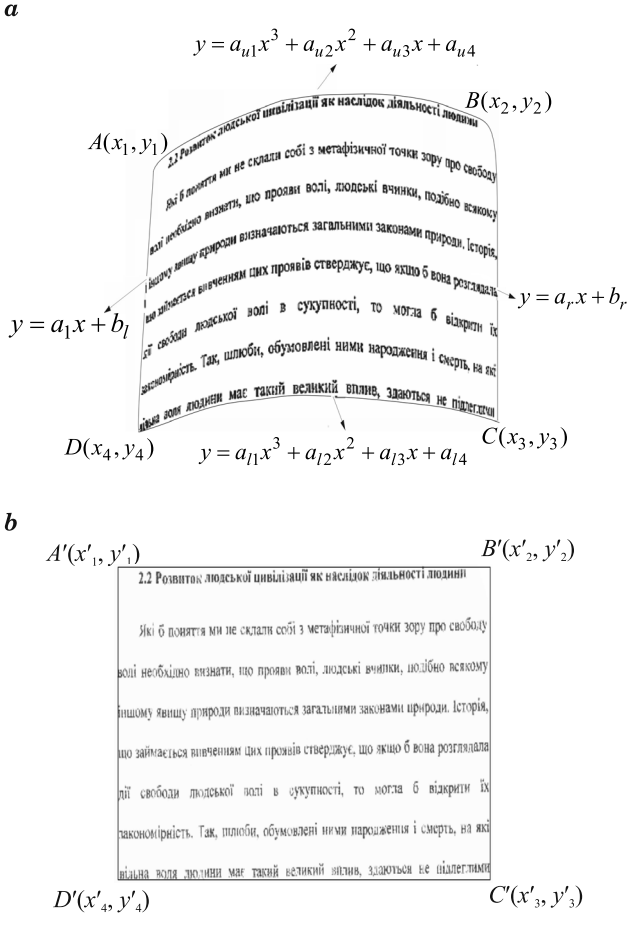

Fig. 3. Example of dewarping image: *a* – extraction of curved surface projection, *b* – corrected image

Conclusion

All of types of distortions which can arise on a text document image are analyzed in this article. If there are only perspective distortions or the problem of a page defect on the image, after a prognostication model of the curvature of all of lines of phototypography it is possible to align text strings (YIN et al. 2007). However, various heterogeneous distortions can appear in the photography of a text image. They can carry an arbitrary form and differ within the bounds of one page, or one line. For example, a page can be curved from one side, and curved inwards from other, with the different angles of slope and yet, despite this, there can be perspective distortions. Thus, it is hard to predict the model of distortion. The described method is universal and does not depend on the type of distortion. It is necessary to develop software on the basis of the described algorithm to define the efficiency of its application to the images with arbitrary heterogeneous line curvatures.

References

ARMS W. 2000. *Digital Libraries*. M.I.T. Press. On line: http://www.cs.cornell.edu/wya/diglib/ (access: October 2014).

FU B., WU M., LI R., LI W., XU Z., YANG CH., 2007. *A model based book dewarping method using text line detection*. Proceedings of the Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR-2007), Curitiba, pp. 63–70. On line: http://imlab.jp/cbdar2007/proceedings/papers/P1.pdf (access: October 2014).

JÄHNE B. *Digital image processing*. 2002. Springer, Berlin, p. 29. On line: http://cgrava.webhost.uoradea.ro/teaching/PAI/documentatie/Jahne_Digital_Image-Processing.pdf (access: October 2014).

MARTI U.V. 2001. *Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system*. Int. Journal of Pattern Recognition and Artifical Intelligence, 15(1): 65–90.

MASALOVICH A.A. 2007. *Use of patch Bez;e for approximation of distortion of text documents*. In: // Labours 17 International Conference on Computer Graphics and Sight. Ed. A.A.Masalovich, L.M. Mesteckiy. Grafikon, Moscow. On line: http://www.machinelearning.ru/wiki/images/a/a2/Gr–2007-Masalovich.pdf (access: October 2014).

OCR Software Review. On line: http://ocr-software-review.toptenreviews.com (access: October 2014).

PRATT W. 2007. *Digital image processing*. Fourth edition. PixelSoft, Inc. Los Altos, California. A Wiley-Interscience Publication, pp. 91–99. On line: https://docs.google.com/file/d/0B30qTepNcnnAOGhlOUQ4Z1o1Sm8/edit (access: October 2014).

YIN X.-C., SUN J., NAOI S. 2007. *Perspective rectification for mobile phone camera-based documents using a hybrid approach to vanishing point detection*. Proceedings of the Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR-2007), Curitiba, pp. 37–44.