Document Image De-warping for Text/Graphics Recognition

Changhua Wu and Gady Agam

Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
{agam,wuchang}@iit.edu

Abstract. Document analysis and graphics recognition algorithms are normally applied to the processing of images of 2D documents scanned when flattened against a planar surface. Technological advancements in recent years have led to a situation in which digital cameras with high resolution are widely available. Consequently, traditional graphics recognition tasks may be updated to accommodate document images captured through a hand-held camera in an uncontrolled environment. In this paper the problem of perspective and geometric deformations correction in document images is discussed. The proposed approach uses the texture of a document image so as to infer the document structure distortion. A two-pass image warping algorithm is then used to correct the images. In addition to being language independent, the proposed approach may handle document images that include multiple fonts, math notations, and graphics. The de-warped images contain less distortions and so are better suited for existing text/graphics recognition techniques.

Keywords: perspective correction, document de-warping, document pre-processing, graphics recognition, document analysis, image processing.

1 Introduction

Document analysis and graphics recognition algorithms are normally applied to the processing of images of 2D documents scanned when flattened against a planar surface. Distortions to the document image in such cases may include planar rotations and additional degradations characteristic of the imaging system [3]. Skewing is by far the most common geometric distortion in such cases, and have been treated extensively [2].

Technological advancements in recent years have led to a situation in which digital cameras with high resolution are widely available. Consequently, traditional graphics recognition tasks may be updated to accommodate document images captured through a hand-held camera in an uncontrolled environment. Examples of such tasks include analysis of documents captured by a digital camera, OCR in images of books on bookshelves [12], analysis of images of outdoor

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 348–357, 2002.

signs [7], license plate recognition [13], and text identification and recognition in image sequences for video indexing [10]. Consequently, distortions characteristic of such situations should be addressed.

Capturing a document image by a camera involves perspective distortions due to the camera optical system, and may include geometric distortions due to the fact that the document is not necessarily flat. Rectifying the document in order to enable its processing by existing generic graphics recognition algorithms require the cancellation of perspective and geometric structural distortions.

A treatment of projective distortions in a scanned document image have been proposed [9] for the specific case of scanner optics and a thick bound book modeled by a two parameter geometric model. Extending this approach to more general cases requires the generation of parametric models for specific cases, and the development of specific parameter estimation techniques. A more general approach that is capable of handling different document deformations by using structured light projection in order to capture the geometric structure of the document is described independently in [4] and [5]. The disadvantage of these approaches lie in the need for additional equipment and calibration needs. In [1] a method is described for correcting geometric and perspective distortions based on structure inference from two uncalibrated views of a document. In this approach the structure recovery may be in some cases inaccurate and lead to distortions in the recovered image. Finally in [15] small deformations of a document image of a thick bound book obtained by a scanner are treated by rotating and aligning segmented words. Entities other than words such as math notations or graphics cannot be handled by this approach.

Contrary to the above described approaches, the proposed approach is not coupled to specific structural models and does not depend on external means for structure recovery. Instead, the document structure distortion is inferred from a single document image obtained by a hand held camera. In addition to being language independent, the proposed approach may handle document images that include multiple fonts, math notations, and graphics.

In the proposed approach, the restoration of the document image so as to reduce structural and perspective distortions in the acquired image, depends on a reconstruction of a target mesh which is then used to de-warp the original image. This target mesh is generated from a single image of the document based on the assumption that text lines in the original document are straight. The detection of the starting position and orientation of curved text lines is described in Section 2. The tracing of curved text lines in outlined in Section 3. The mesh reconstruction and document de-warping presented in Section 4. Section 5 concludes the paper.

2 Detecting the Starting Position and Orientation of Curved Text Lines

Let $F \equiv \{f_p\}_{p \in \mathbb{Z}_m \times \mathbb{Z}_n}$ be an $m \times n$ image in which the value of each pixel is represented by an s-dimensional color vector $f_p \in \mathbb{Z}^s$ where in this expression $\mathbb{Z}_m$ represents the set of non-negative integers $\{0, \ldots, m-1\}$. Without loss of generality, for efficiency reasons, we assume that the input image is binarized [11] so as to produce $G \equiv \{g_p\}_{p \in \mathbb{Z}_m \times \mathbb{Z}_n}$ where $g_p \in \mathbb{Z}_2$ and black pixels have intensity of 0.

In the proposed approach the user interactively specifies the four corner points p_{lt}, p_{lb}, p_{rt}, p_{rb} of a quadrilateral containing the portion of the image that has to be rectified. It is assumed that the user specification of the corner points is such that text lines in the document are in a general direction which is approximately perpendicular to the left edge $\overline{p_{lb}p_{lt}}$ of the quadrilateral. While the identification of these corners may be done automatically under certain assumptions, interactive point specification is simple and so we did not find it necessary to address this problem. Without loss of generality we assume in the rest of this paper that the orientation of the left edge $\overline{p_{lb}p_{lt}}$ is approximately vertical.

Based on the above assumptions the starting position of non-indented text lines should be approximately along the line segment $\overline{p_{lb}p_{lt}}$ and their orientation should be approximately perpendicular to that line segment. Detecting the starting point of text lines is a common problem in document analysis that is normally treated after skew correction by detecting the extremum points in the graph of a cumulative horizontal projection [8]. It should be noted, however, that the performance of this approach strongly depends on the preliminary skew distortion correction. As the distortion in our case is non-linear, skew correction will not suffice. Furthermore, the required correction is the overall target of the proposed approach and can not be used at this stage. In order to solve this problem, the cumulative projection that is used in the proposed approach is constructed from a local neighborhood of the left edge $\overline{p_{lb}p_{lt}}$ which due to is locality is assumed to be less distorted. As the text lines are not necessarily perpendicular the the left edge $\overline{p_{lb}p_{lt}}$ directions adjacent to the horizontal direction should be checked as well, and the maximal projection should be retained.

Consequent to the above description, the graph of a cumulative horizontal projection is replaced in the proposed approach by a graph of local adaptive cumulative projection. This graph is constructed by computing the local adaptive cumulative projection $\Phi(p)$ at each possible starting point $p \equiv (x, y)$ starting with p_{lb} and progressing along $\overline{p_{lb}p_{lt}}$ towards p_{lt}. The value of $\Phi(p)$ is defined by:

$$\Phi(p) \equiv \min\{\Phi_\beta(p) \quad | \quad \theta - \alpha < \beta < \theta + \alpha\} \tag{1}$$

where $\Phi_\beta(p)$ is the local cumulative projection in the direction of β at p, the angle θ if the angle that produced the minimal projection of the previous starting point $(x, y-1)$ and α is a preset angular limit (see Figure 1). The use of θ is designed to promote a smoothing constraint. Its initial value at p_{lb} is taken as 0. The angle β that produced the minimal value of $\Phi_\beta(p)$ is the estimated starting orientation of the text line emanating from p. It is stored for later use. The local cumulative projection $\Phi_\beta(p)$ is computed by the sum:

$$\Phi_\beta(p) \equiv \sum_{p \in R(p,\beta)} g_p \tag{2}$$

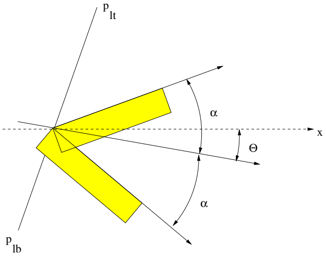


Fig. 1. Constructing the local adaptive cumulative projection

where $R(p, \beta)$ is the set of pixels contained within a rectangle emanating from p in the direction of β. Based on simple geometric considerations, the corner points of this rectangle may be computed by:

$$p_1 \equiv (x_{lt} + (x_{lb} - x_{lt})(y - y_{lt})/(y_{lb} - y_{lt}), y) \tag{3}$$

$$p_2 \equiv (x + h\cos(\beta), y + h\sin(\beta)) \tag{4}$$

$$p_3 \equiv (x + w\cos(\beta - \pi/2) + h\cos(\beta), y + w\sin(\beta - \pi/2) + h\sin(\beta)) \tag{5}$$

$$p_4 \equiv (x + w\cos(\beta - \pi/2), y + w\sin(\beta - \pi/2)) \tag{6}$$

where w and h are preset parameters corresponding to the width and height of the rectangle respectively. By using the corner points p_1, p_2, p_3, p_4, the pixels belonging to the rectangle are obtained by using a standard scan-line filling algorithm [6]

Once obtaining the local adaptive cumulative projection graph, extremum points in it may be used to separate text lines. Minimum points in particular are used to detect the beginning of text lines. In order to reduce errors due noise this graph is smoothed by using a low-pass filter prior to the extremum points detection. In addition, in cases of several detected minimum points in close proximity to each other, only the smallest one is kept. Figure 2 presents the smoothed local adaptive cumulative projection graph of the Chinese document image in Figure 3. Minimum points in this graph correspond to starting points of text lines. The identified starting points of text lines together with the estimated starting orientation are overlaid in gray on the binarized document image in Figure 3. As can be observed the fact that the text lines are curved does not mislead the outlined detection algorithm. It should be noted that in the experiments we conducted, the described algorithm was equally successful in detecting the starting position and orientation of text lines in images of English and Chinese documents in accordance with the observation in [8].

3 Tracing Curved Text Lines

After obtaining the starting position and orientation of each text line, the complete text lines may be traced in a similar way. That is, given a point and

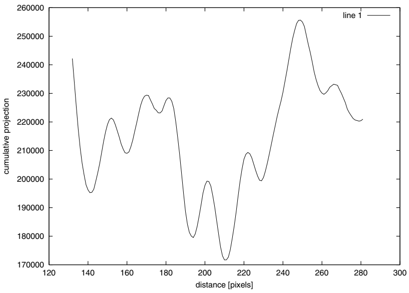


Fig. 2. The smoothed local adaptive cumulative projection graph of the Chinese document in Figure 3. Minimum points in this graph correspond to starting points of text lines. The values on the x-axis of this graph were multiplied by 255

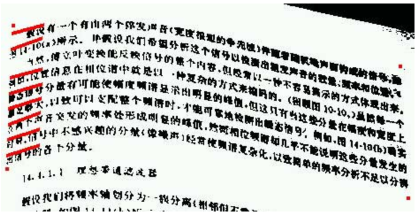


Fig. 3. Identified starting position and orientation of text lines (by using the proposed approach) overlaid in gray on the binarized image of a Chinese document. As can be observed the fact that the text lines are curved does not mislead the detection algorithm

orientation on a text line the next point on that text line is selected by evaluating the cumulative projection in a local neighborhood in a range of directions around the given direction. The next point is selected as the one for which the cumulative projection is minimal. More formally, given the j-th point on the i-th traced text line p_{ij} and the text line orientation θ_{ij} at that point (see Figure 4), the next point on that line $p_{i,j+1}$ is obtained as:

$$p_{i,j+1} \equiv p_{ij} + h \cdot s_i \cdot (\cos(\theta_{i,j+1}), \sin(\theta_{i,j+1})) \qquad (7)$$

where h is the length of the rectangular neighborhood as defined for Equation 4, the angle $\theta_{i,j+1}$ is the angle that minimizes $\Phi(p)$ in Equation 1, and s_i is a scale factor.

The scale factor s_i is introduced in order to produce a similar number of points on each text line when the specified quadrilateral $(p_{lt}, p_{lb}, p_{rt}, p_{rb})$ is not rectangular. Let $L_t \equiv ||p_{lt} - p_{rt}||$ and $L_b \equiv ||p_{lb} - p_{rb}||$ be the length of the top and bottom edges of the quadrilateral respectively. Assuming that the step length on the top edge is 1, the step length on the bottom edge is taken to be $\frac{L_b}{L_t}$.

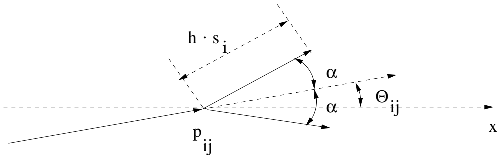


Fig. 4. Tracing a text line. Given a point p_{ij} and orientation θ_{ij} the next point on that text line is searched in an angular range of $\pm\alpha$ around θ_{ij}. The length of the step is adjusted by a scale factor s_i

The step length in an intermediate line i can be then interpolated by:

$$s_i \equiv (1 - \eta_i) + \eta_i \frac{L_b}{L_t} \tag{8}$$

where $\eta_i \equiv (y_{lt} - y_{i0})/(y_{lt} - y_{lb})$.

The tracing of a curved text line is stopped at the point p_{ij} if black pixels are not found in any of the projection rectangles of that point or if any of the projection rectangles of that point intersects the right edge $\overline{p_{rb}p_{rt}}$ of the quadrilateral. Due to the non-uniformity of characters in the document the traced lines may contain small variations. These variations are eliminated by low pass filtering the traced curves.

The angular range searched in the process of tracking a curved text line is normally smaller than the one used for the detection of the starting point of text lines. The angular range should be small enough in order to prevent possible crossings to neighboring text lines, and large enough in order to facilitate the tracing of curved text lines. In order to reduce crossings between text lines while maintaining a larger angular range search area, the local cumulative projection $\Phi_\beta(p)$ is modulated by a weight factor $W(\beta)$ which is inversely proportional to the angular deviation $(\beta - \theta)$:

$$W(\beta) \equiv 1 + |\tan(\beta - \theta)|/\mu \tag{9}$$

where μ is a constant and it is assumed that $|\beta - \theta| < \pi/4$.

The above description of modulation of the local cumulative projection assists in reducing the number of crossings between text lines, but do not eliminate them. Consequently, a consistency constraint is introduced in order to remove such crossings. For that purpose the average orientation in each column is computed by: $\overline{\theta}_j \equiv \sum_i \theta ij$, and lines containing any points with orientation deviating by more than a preset threshold τ from the average θ_j are removed. Text lines not intersecting the right edge $\overline{p_{rb}p_{rt}}$ do not contribute to the generation of a regular grid, and so they are removed as well. It should be noted that as the proposed approach does not rely on a dense grid of lines, incomplete/inaccurate traced lines may be removed instead of attempting to correct them. Figure 5 presents the result of the line removal stage where Figures 5-a and 5-b display the traced lines before and after correcting them respectively.

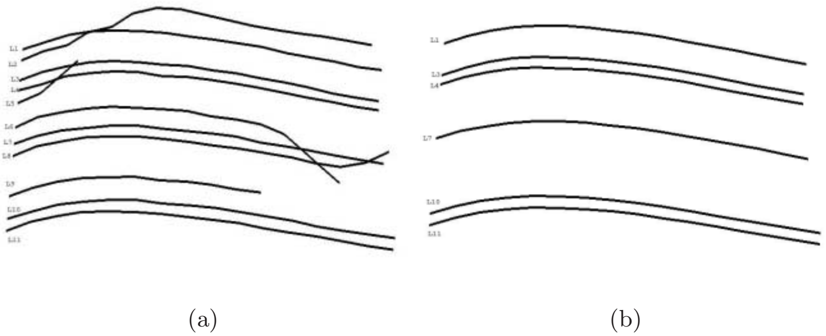


Fig. 5. Demonstration of the line removal stage. **(a)** – **(b)** The traced lines before and after correction respectively. As can be observed, lines L_2, L_5, L_6, L_8 are removed due to the orientation consistency constraint whereas lines L_5, L_9 are removed due to the fact that they do not intersect the right edge

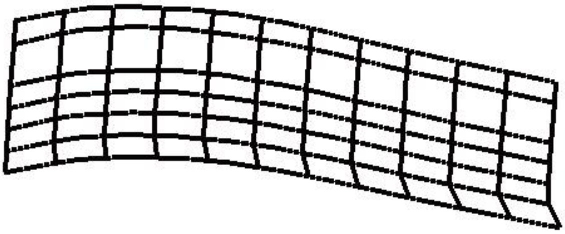
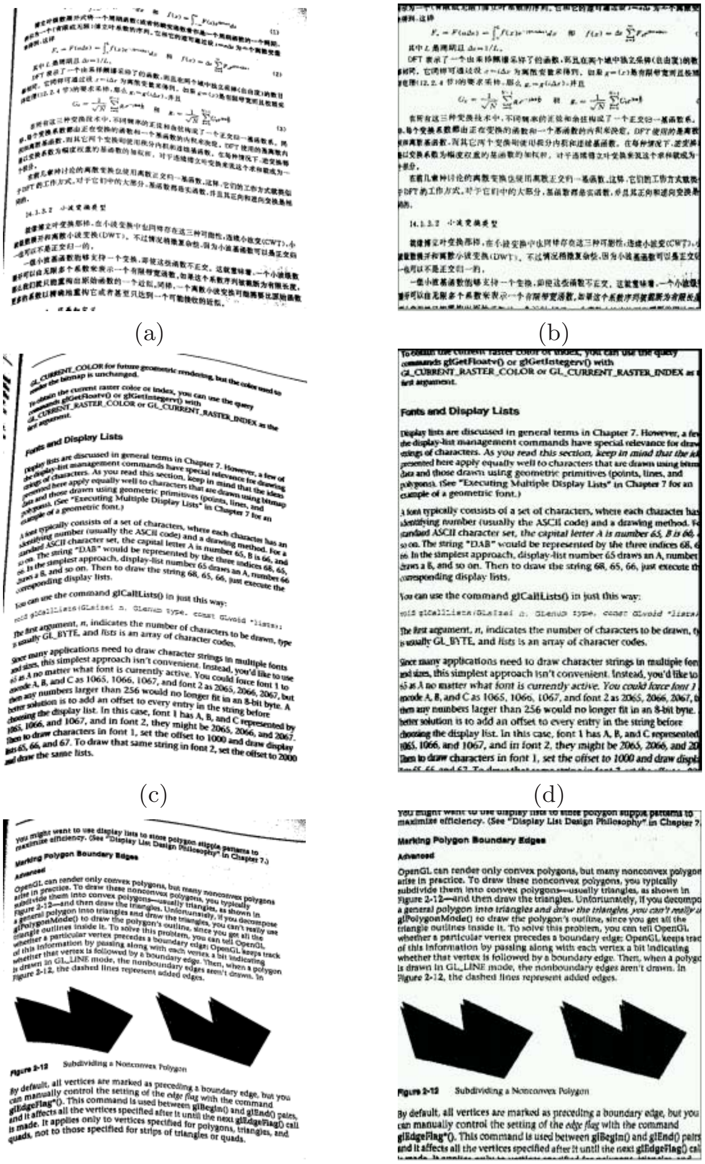


Fig. 6. The reconstructed source mesh for the document image in Figure 3. As can be observed the reconstructed mesh corresponds to the structural deformation in that document

4 De-warping the Document Image

For the purpose of de-warping the document image a source and target rectangular meshes should be produced. The source mesh contains curved lines corresponding to the structural distortion in the document image, whereas the target mesh should be rectilinear so as to represent the document structure without distortion. The source mesh is produced based on the traced lines obtained as described in the previous section. The horizontal lines of that mesh are the traced lines whereas the vertical lines are generated by subdividing each traced line into a fixed number of uniform length segments. Figure 6 presents the reconstructed source mesh for the document image in Figure 3. As can be observed the reconstructed mesh corresponds to the structural deformation in that document.

The target mesh is generated based on the source mesh and the assumption that the text lines in the document were straight before going through structural deformation. The rectilinear target mesh is generated with the same number of



(a) (b)

(c) (d)

(e) (f)

Fig. 7. Document image de-warping obtained by the proposed approach. The left column presents the input document images whereas the right column presents the rectification results obtained automatically by the proposed approach. As can be observed the proposed approach is capable of handling documents in different languages which include graphics, math notations, and different fonts

rows and columns as the source mesh. The distance between neighboring rows in the target mesh is set to the average distance between the corresponding rows in the source mesh multiplied by a uniform scale factor which is used to scale the size of the rectified image. The distance between neighboring columns in the target mesh is set to be uniform. This distance is selected as the uniform segment length on the longest row in the source mesh. It should be noted that, in general, the distance between neighboring columns in the target mesh should not be uniform due to perspective foreshortening. More specifically, the distance between columns of the document corresponding to an area of the document closer to the camera should be smaller. In future work we intend to estimate this non-uniform length based on character density estimation in each column.

Given the reconstructed source and target meshes the de-warping of the document image is done by a 2-pass image warping algorithm as described in [14]. This image warping algorithm is particularly suitable for our case as it is based on a rectangular mesh structure which is inherent to document images and as it prevents foldovers.

5 Results

The results of document image de-warping obtained by the proposed approach are presented in Figure 7. In this figure, the left column presents the input document images whereas the right column presents the rectification results obtained automatically by the proposed approach. As can be observed the proposed approach is capable of handling documents in different languages which include graphics, math notations, and different fonts. This is due to the fact that only a sparse mesh grid has to be reconstructed for the rectification.

As mentioned earlier in this work we do not yet take care of generating non-uniform columns in the target mesh. Consequent to that it is possible to observe that characters in parts of the document image which were originally closer to the camera appear to have a slightly larger width.

References

1. G. Agam. Perspective and geometric correction for graphics recognition. In *Proc. GREC'01*, pages 395–407, Kingston, Ontario, 2001. 349
2. A. Amin, S. Fischer, A.F. Parkinson, and R. Shiu. Comparative study of skew algorithms. *Journal of Electronic Imaging*, 5(4):443–451, 1996. 348
3. H. Baird. Document image defect models. In *Proc. SSPR'90*, pages 38–46, 1990. 348
4. M.S. Brown and W.B. Seales. Document restoration using 3d shape: a general deskewing algorithm for arbitrarily warped documents. In *Proc. ICCV'01*, pages 367–374, Vancouver, BC, Jul. 2001. IEEE. 349
5. A. Doncescu, A. Bouju, and V. Quillet. Former books digital processing: image warping. In *Proc. Workshop on Document Image Analysis*, pages 5–9, San Juan, Puerto Rico, Jun. 1997. 349

6. David F.Rogers. *Procedural elements for computer graphics*. McGraw-Hill, second edition, 1998. 351
7. H. Fujisawa, H. Sako, Y. Okada, and S. Lee. Information capturing camera and developmental issues. In *Proc. ICDAR'99*, pages 205–208, 1999. 349
8. D.J. Ittner and H.S. Baird. Language-free layout analysis. In *Proc. ICDAR'93*, pages 336–340, Tsukuba, Japan, 1993. 350, 351
9. T. Kanungo, R. Haralick, and I. Philips. Global and local document degradation models. In *Proc. ICDAR'93*, pages 730–734, 1993. 349
10. H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. *IEEE Trans. Image Processing*, 9(1):147–156, 2000. 349
11. J. Sauvola, T. Seppanen, S. Haapakoski, and M. Pietikainen. Adaptive document binarization. In *Proc. ICDAR'97*, pages 147–152, Ulm, Germany, Aug. 1997. 350
12. M. Sawaki, H. Murase, and N. Hagita. Character recognition in bookshelf images by automatic template selection. In *Proc. ICPR'98*, pages 1117–1120, Aug. 1998. 348
13. M. Shridhar, J.W.V. Miller, G. Houle, and L. Bijnagte. Recognition of license plate images: issues and perspectives. In *Proc. ICDAR'99*, pages 17–20, 1999. 349
14. G. Worlberg. *Digital Image Warping*. IEEE Computer Society Press, Los Alamitos, California, 1990. 356
15. Z. Zhang and C.L. Tan. Recovery of distorted document images from bound volumes. In *Proc. ICDAR'01*, pages 429–433, Seattle, WA, 2001. 349