

Rectifying the Bound Document Image Captured by the Camera: A Model Based Approach

Huaigu Cao, Xiaoqing Ding, Changsong Liu
Department of Electronic Engineering, Tsinghua University
Beijing, 100084, P. R. China
Email: caohg@ocrserv.ee.tsinghua.edu.cn

Abstract

A model based approach for rectifying the camera image of the bound document has been developed, i.e., the surface of the document is represented by a general cylindrical surface. The principle of using the model to unwrap the image is discussed. Practically, the skeleton of each horizontal text line is extracted to help estimate the parameter of the model, and rectify the images. To use the model, only a few priori is required, and no more auxiliary device is necessary. Experiment results are given to demonstrate the feasibility and the stability of the method.

1. Introduction

Our research has been motivated by the advantage of using digital camera for document image capture. On the one hand, compared with a scanner, one conventional inputting device, a camera has much more portability. And on the other hand, using a camera, one can take the image instantly in a more casual manner. However, the conveniences of using digital camera are accompanied with some image quality problems. With a scanner, one can make the documents flat simply by pressing them intentionally on the glass board. While with a camera, the natural warping existing in the document surface won't disappear, hence the image will be deformed. In this case, the process of document analysis and recognition will become complicated and unreliable.

Among a variety of warping types, the warping caused by bookbinding should be the most ordinary one, as is shown in Figure 1, one photograph of a bound document. There have been a number of literatures dealing with the rectification of this kind of warping. In [1], a method of using the combination of cylinder and plane is introduced to simulate the surface of books, but how to estimate the parameters and make use of the model is unresolved, and it can only be applied to images scanned by the scanner. In [2], a laser projector is used to project a 2D light network on the surface of the document, and then two dimensional distortions of the surface are corrected with a two pass mesh warping proposed by [3]. This method needs additional device one can not conveniently afford. The

method introduced in [4] is to get the depth of each point in the image by some stereo vision method, hence to make a depth image, and then rectify the image according to the depth image. Although it seems capable of rectifying any type of image distortions, how to map the points on the rough, noisy surface defined by the depth image to the points on the plane is still a problem. In [5], the scanned images of bound books are rectified by means of character segmentation. Characters in the shadow (where the surface is curled) are segmented, their orientations and original locations are estimated, and then characters are adjusted. Since the extent of bending is unknown by this method, it is not so accurate, and characters that are narrowed due to distortions remain narrower after the adjustment.

In this paper, we propose a novel model based approach to rectify the bound document image warping. The key advantages of our approach are: It requires very few priori about the image formation process, runs with a single image, and any stereographic device is unnecessary. In section 2, we discuss features about our cylinder model and the principle on how to use the model to rectify the image. In section 3, the detailed process of the rectification is present. And the final results are shown in section 4.

2. The cylinder model

The existence of bookbinding often prevents the document from being opened entirely. A typical camera image of one page of bound document has been shown in Figure 1. One can notice that the horizontal text lines are

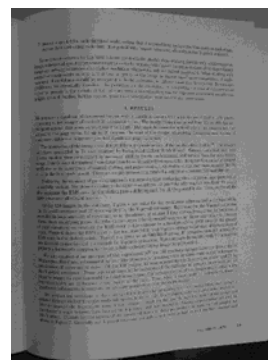


Figure 1. An example of image warp caused by bookbinding

distorted, and do not run straight and parallel any more.

Using a scanner, one may press the document on the glass so that the surface of the document is partially flat, whereas using a camera one may not. According to this, we notice that the model in [1] describing the surface as a combination of a part of cylinder and a semi-plane is no longer suit for the surface of the document in natural state. So we need a more general model. Here we adopt a general cylindrical surface to simulate the surface of the document captured by the camera, and assumed that the horizontal text lines go along the direction of the directrix. Firstly, we may locate some of these text lines, hence to get several projections of directrixes. In the following, we should see that once we are aware of the positions of these projections, with a few priori, we could get the mapping from the warping 2D image to the rectified 2D image of the 3D surface.

As we have known, the camera is a lens system, and the object distance is typically much greater than the image distance. Under this circumstance, the image formation process of a lens system can be approximately described by the perspective projection [6]. In the case that the generatrix doesn't parallel the image plane, the mapping from the points of the generatrix to the points on the image plane can be viewed as a projective transformation, generally a nonlinear one (the mid-point of a line segment in the 3D world is no longer the mid-point of corresponding line segment in the image), which makes the restoration much difficult. In order to restore the surface, we assume that the generatrix of the cylinder parallels the image plane.

Corresponding geometrical relationship of the image formation process of the surface are demonstrated in figure 2. We may choose appropriate Cartesian coordinates X and Y axes, making them perpendicular to and parallel the generatrix, respectively. From the geometrical relationship of the process of image formation shown in figure 2, we may notice two features of the process:

(1) In the image plane, each line along Y axis direction is a scaled image of a generatrix, and the mapping between them is a linear transformation.

(2) the image of any directrix projected on the image plane is a scaled image of that directrix, and the mapping between them is a linear transformation.

The first feature can be easily proved according to the

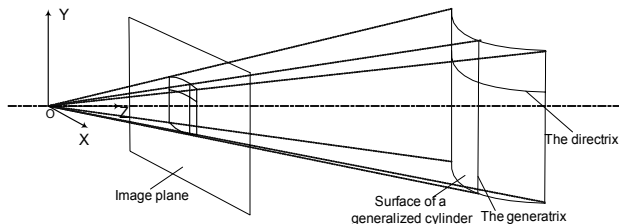


Figure 2. The process of image formation (perspective projection) of the document surface (a general cylindrical surface).

geometrical relationship. However, the second feature seems not so obvious as the first one, so we need more explanation. Just as shown in Figure 3, there is a line segment PQ in the scene, with its image $P'Q'$ on the image plane. By the geometry, we get the equation:

$$P'Q' = \frac{f}{z_0} PQ = \frac{fy_0}{z_0^2} PQ \quad (1)$$

Notice that the depth difference in one directrix is often less than 1~2 cm, while the average depth of the surface is much larger, typically 40~80cm, therefore we may assume that $PQ \ll Z_0$. In this case, equation 1 indicates that the mapping from the depth difference PQ of real 3D world to its image $P'Q'$ on the image plane is a linear map, a directly proportional transformation.

Now consider an arbitrarily selected directrix on the surface:

$$L_1 : \begin{cases} y = y_1 \\ z = D(x) \end{cases}$$

According to equation 1, we may get the image of the curve L_1 :

$$\begin{aligned} L_1' : y &= C_1 - \frac{fy_1}{(z^*)^2} D\left(\frac{z^*x}{f}\right) = C_1 - \frac{y_1}{z^*} \cdot (f/z^*) D\left(\frac{x}{f/z^*}\right) \\ &= C_1 - \frac{y_1}{z^*} \cdot D^*(x) \end{aligned} \quad (2)$$

where z^* means the approximate depth of the surface, C_1 is a constant, and $D^*(x)$ is the curve $D(x)$ zoomed by f/z^* times in scale. Since we do not concern about the zoom factor, if the value of the factor y_1/z^* is known to us, we may get the equation of the directrix immediately from the equation 2.

Although we may record the approximate distance, z^* , from the lens to the document, the estimation of the coordinate y_1 is difficult, and need camera calibration. To solve this problem, we use the following method. Take arbitrarily another directrix that different than L_1 :

$$L_2 : \begin{cases} y = y_2 \\ z = D(x) \end{cases}$$

The image of the curve L_2 can be described by

$$L_2' : y = C_2 - \frac{y_2}{z^*} \cdot D^*(x)$$

By calculating the curve L_1' minus L_2' , we get

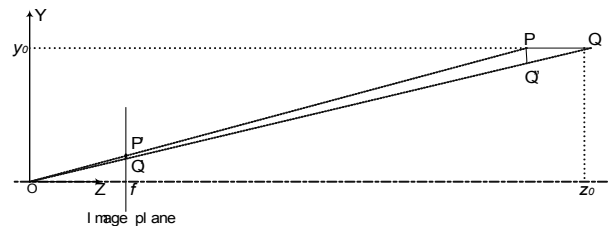


Figure 3. The geometrical relationship between the difference of depth in the scene and the distance in the image

$$L_{1-2}': y = C_1 - C_2 \frac{y_1 - y_2}{z^*} D^*(x) \quad (3).$$

Here the value $y_1 - y_2$ can be computed from the distance in the image multiplying the resolution.

By now, we have proved two features of the image formation process shown in figure 1. With these two features, we may rectify the image by the following steps:

(1) Locate two directrices $L_1: y = D_1'(x)$ and $L_2: y = D_2'(x)$ in the image, and recover the shape of the directrix ($y = D^*(x)$) according to equation 3;

(2) Decide the mapping of x coordinate from the rectified image to the original image. The mapping is U :

$$x_{\text{new}} \rightarrow x_{\text{old}}, \text{ where } \int_0^{x_{\text{old}}} \sqrt{1 + \left[\frac{dD^*(x)}{dx} \right]^2} dx = x_{\text{new}}, x_{\text{new}} = 0,$$

1, 2 ..., denoted by $U(x_{\text{new}}) = x_{\text{old}}$;

(3) Decide the mapping of y coordinate from the unwrapped image to the original image. In the original image, select arbitrarily a point (x_1, y_1) in the directrix L_1' and a point (x_2, y_2) in the directrix L_2' , respectively. We may take the y coordinates y_1 and y_2 as one kind of "standard" points, mapping the y coordinates of all the points within L_1' to y_1 , and the y coordinates of all the points within L_2' to y_2 . For each pixel with the coordinates $(x_{\text{new}}, y_{\text{new}})$ within the rectified image, we have

$$\begin{cases} x_{\text{old}} = U(x_{\text{new}}) \\ y_{\text{old}} = \frac{y_{\text{new}} - y_1}{y_2 - y_1} \cdot [D_1'(x_{\text{old}}) - D_2'(x_{\text{old}})] + D_1'(x_{\text{old}}) \end{cases} \quad (4).$$

Equation (4) is the mapping between the two coordinates. Transferring the pixels using this mapping, we may have the image rectified. Figure 4 illustrates that process.

3. Rectifying the image using cylinder model

We adopt the cylinder model mentioned in the above section to unwrap the bound document image. The key idea is that, firstly we location the horizontal text lines in the image as many as possible. We use these lines to decide the left or right boundary of the text regions. Since those

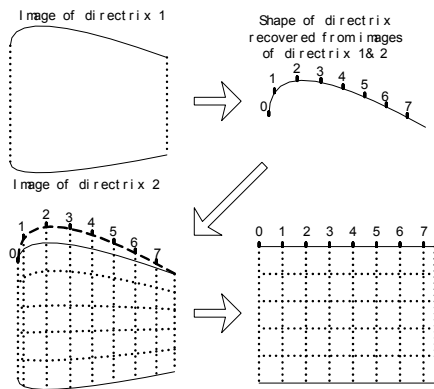


Figure 4. The process of rectification

two lines are usually vertical, we may specify the Y direction (and therefore the X direction) of figure 1 according to any of them. Among these text lines, choose two of them that are "typical" to generate two directrices according to a certain minimal error constraint. Finally, by equation 4, we may get a mapping between the coordinates of the image we are seeking and the coordinates of the original image. We perform this mapping to get the result.

3.1. Locating text lines

We may not need to locate the region of text lines indeed. Actually, what we need is that, for each text lines, find a curve which runs across the mid-height of each word in the text line. And that is enough to specify the directrices.

Firstly, we adopt the *Niblack's* algorithm [7] to threshold the image, which can effectively differentiate between foreground and background. Denote the original and the binarised image by I and I_b , respectively. In order to reduce computation, we then zoom out the binarised image by factor 2^p , hence get the image I_z . To reduce the computation effectively, and meanwhile to secure that the image are not zoomed out so much as to deteriorate the quality of text lines, we designate p as the minimal positive integer which satisfies that the width of the original image divided by 2^p is not large than 512. Suppose that in I_b , "1" or "black" denotes the foreground, and "0" or "white" denotes the background, then in our experiments, the zooming out process can be described as:

$$I_z(m, n) = \begin{cases} 1, & \text{if } \exists x, y, s.t. I_b(x, y) = 1, \\ & 2^p \cdot m \leq x < 2^p \cdot (m+1), \\ & \text{and } 2^p \cdot n \leq y < 2^p \cdot (n+1) \\ 0, & \text{otherwise.} \end{cases}$$

Based on I_z , we generate another image I_m that is the same size as I_z by the following steps:

- (1) All of the pixels in I_m are initially white;
- (2) In I_z , search such vertical line segment that all of its pixels are black, whereas the pixel just above it and the pixel just below it are white, and its length > 1 . Here we require that the length of the line segment > 1 , because in the case the length equals 1, it is possibly a noise point.
- (3) In I_m , set the pixels white at the positions corresponding to the mid-points of all such line segments.

Once we get the image I_m , we may use a connected component searching method to find points of each directrix, hence to locate those directrices. The searching strategy is that, choose a start point from one end of a directrix, and search foreground points horizontally towards another end. In our experiments, each time from a foreground pixel, search for the next foreground pixel horizontally within 16 pixels and $\pm 12^\circ$.



Figure 5 the “skeletons” of text lines extracted from image in figure 1.

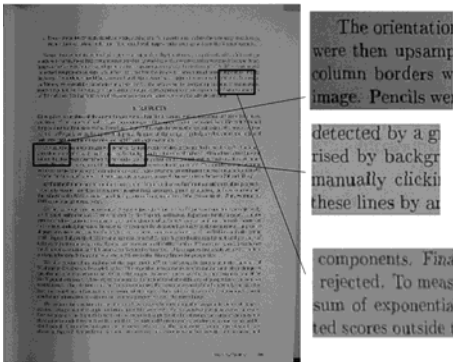


Figure 6. The rectified image of the one shown in figure 1

In reality, the text line where a paragraph is over often stops half way, and the searching process, if it is from the left to the right, sometimes does not stop, but traverses to an adjacent text line instead. To avoid this, the searching should be from right to the left.

The process mentioned above can be imagined as extraction of the “skeletons” of the text lines. The result of this process applied on figure 1 is shown in figure 5, where pixels owned by different text lines are marked by different brightness.

3.2. Estimating the directions of X, Y axes

Before we could utilize equation 4 to rectify the image, we first estimate the direction of X, Y axis. To achieve this, we make use of the pixels of directrices found in the process mentioned in section 3.1. For the pixel set of each directrix, take the left-most pixel. And all these left-most pixels form a pixel set, denoted by C_L . Similarly, all of the right-most pixels form a pixel set, denoted by C_R . C_L and C_R can be considered two groups of points on the left or right boundary of the text region, respectively. Fit a straight line L_L using points in set C_L by least mean square error criterion. The direction of L_L is the direction of Y axis. To increase accuracy, we adopt the method of iteratively finding and excluding the outliers. Each time when we fit a straight line $L_{L,k}$ with C_L , we find the point in C_L that is the farthest from $L_{L,k}$, and that point is

considered an outlier. Exclude the outlier and fit a new straight line $L_{L,k+1}$ once again, until the distance from the outlier to the straight line is not above 2 pixels, which means successful, or the number of points in C_L is less than 10, which means failed. Occasionally when it fails, we may make use of the set C_R for estimation.

3.3. Choosing two typical directrices to rectify the image

Based on the discussions above, we may now consider the question of how to choose two “typical” curves to rectify the directrix. We should decide a criterion and apply it to get the optimal result. Suppose that we have got n ($n \geq 2$) directrices that are projected on the image plane by the method in section 3.1, and have specified the directions of X and Y axes, and these n curves are represented by n point sets C_1, C_2, \dots, C_n in 2D plane. To operate conveniently, we use a smooth curve $SC_k(x)$ to fit each point set C_k , $k = 1, 2, \dots, n$. Estimate the x coordinates of the left and right boundaries a and b ($a < b$) of the text region using the average x coordinates of C_L and C_R mentioned in section 3.2 (here a and b are truncated into integers), then we get the domain of the curve, the interval $[a, b]$. For any two of the n curves, i.e. $SC_i(x)$ and $SC_j(x)$, we define the following error formulas:

$$Err(i, j, k) = \sum_{x=a}^b | [SC_j(x) - SC_i(x)] \cdot \frac{SC_k(\frac{a+b}{2}) - SC_i(\frac{a+b}{2})}{SC_j(\frac{a+b}{2}) - SC_i(\frac{a+b}{2})} + SC_i(x) - SC_k(x) |.$$

Here $Err(i, j, k)$ can be considered a measurement of the deviance of the k th curve if the i th and the j th are chosen as typical curves. It is easy to prove that $Err(i, j, k)$ equals zero when $SC_k(x)$ is a linear combination of $SC_i(x)$ and $SC_j(x)$. Here we do not use square error because of its vulnerability to the outliers. To minimize the sum of these error, we should choose the indices of the two typical curves i, j by

$$(i, j)_{opt} = \arg \min_{\substack{(i, j) \\ k=1 \\ k \neq i, j}}^n Err(i, j, k).$$

Once we have specified two typical curves, we may rectify the image by equation 4.

4. Experiment results and analysis

In our experiments, we adopt the 3rd order polynomial equation to express the smooth curve $SC_k(x)$, i.e. $SC_k(x) = a_k x^3 + b_k x^2 + c_k x + d$, $k = 1, 2, \dots, n$, because the process of polynomial regression is quite simple, and the 3rd order polynomial is fairly accurate in describing such curves. Figure 1 shows an image of 1200×1600 and proximately

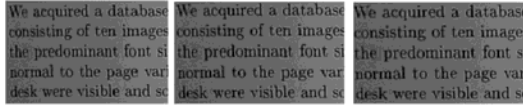


Figure 7. The compare of results in different inputs of object distance input cases. (a) The input of object distance is 50cm; (b) The input of object distance is 59cm; (c) The input of object distance is 70cm.

139dpi, and the distance from the lens to the surface of the document is about 59 cm. The rectified image is shown in figure 6, where three small regions are zoomed in. In figure 6, we may notice that the text lines are all straight, and the characters in different regions that are zoomed in have uniform width. Consequently, the image is rectified pretty well.

In another experiment, we vary the input of object distance, and compare the results. From figure 7, we may notice that there are just slight changes in the width of characters amongst different results. Using our model, sometimes it is a bit hard to estimate the object distance accurately. The experiment just mentioned indicates that even with a rough estimation of distance, one can get an acceptable result.

We use the OCR software ScanSoft OmnipagePro 11.0 to make a compare of OCR qualities between before and after rectification of the image in Figure 1. Before OCRing, we binarise the gray-scale image using *Niblack's* Algorithm. There are totally 824 English words. Before rectification, totally 764 words (92.72%) are correctly read, while after rectification, the number of words that are correctly read is 809 (98.18%). The rectification process ensures effective layout analysis and word segmentation, hence bringing about higher reading performance and less recognition errors.

We have tried our method on a number of images of different resolution ranges from about 90dpi to 180dpi, From most of them we get satisfactory results. Some of the results are shown in figure 8.

Acknowledgments

This work was supported by 863 Hi-tech Plan (project 2001AA114081) & National Natural Science Foundation of China (project 69972024).

References

- [1] T. Kanungo, R. Haralick, I. Philips, "Global and local document degradation models," in Proc. 2nd International Conference on Document Analysis and Recognition, 1993.
- [2] A. Doncescu, A. Bouju, V. Quillet, "Former books digital processing: image warping," in Proc. Workshop of Document Image Analysis, 5-9, 1997.
- [3] D. B. Smythe, "A Two-Pass Mesh Warping Algorithm for Object Transformation and Image Interpolation", ILM Technical Memo #1030, Computer Graphics Department,

Lucasfilm Ltd., 1990.

- [4] M. S. Brown, W. B. Seales, "Document restoration using 3D shape: a general deskewing algorithm for arbitrarily warped documents", in Proc. International Conference on Computer Vision, July 2001.
- [5] Z. Zhang, C. L. Tan, "Restoration of images scanned from thick bound documents" , in Proc. 6th International Conference on Document Analysis and Recognition, 2001.
- [6] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 1st Edition, pp. 725-727 Prentice Hall, 1995.
- [7] W. Niblack, *An Introduction to Digital Image Processing*, Prentice Hall, Englewood Cliffs, 1986.

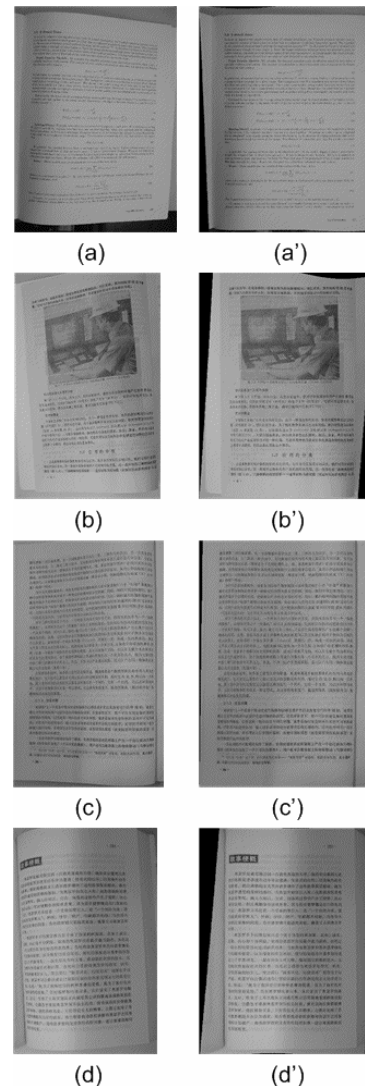


Figure 8. Some of the results: (a)-(d) are the original images, and (a')-(d') are the corresponding results.