

Rectifying Perspective Distortion into Affine Distortion Using Variants and Invariants

Masakazu Iwamura
Osaka Pref. Univ., Japan
masa@cs.osakafu-u.ac.jp

Ryo Niwa
Osaka Pref. Univ., Japan
niwa@m.cs.osakafu-u.ac.jp

Koichi Kise
Osaka Pref. Univ., Japan
kise@cs.osakafu-u.ac.jp

Seiichi Uchida
Kyushu Univ., Japan
uchida@is.kyushu-u.ac.jp

Shinichiro Omachi
Tohoku Univ., Japan
machi@ecei.tohoku.ac.jp

Abstract

For user convenience, document image processing captured with a digital camera instead of a scanner has been researched. However, existing methods of document image processing are not usable for a perspective document image captured by a digital camera because most of them are designed for the one captured by a scanner. Thus, we have to rectify the perspective of the document image and obtain the frontal image as if it was captured by a scanner. In this paper, for eliminating perspective distortion from a planar paper without any prior knowledge, we propose a new rectification method of a document image introducing variants which change according to the gradient of the paper and invariants which do not change against it. Since the proposed method does not use strong assumptions, it is widely applicable to many document images unlike other methods. We confirmed the proposed method rectifies a document image suffering from perspective distortion and acquires the one with affine distortion.

are suitable to capture an image easily and quickly. This can bring us a new application of document analysis and character recognition.

Though camera-based approach has a chance to yield excellent applications, the realization is not easy. A reason is that most existing document analysis techniques are for scanner-captured images. That is, processing a camera-captured image preliminarily requires a rectification of image to obtain a scanner-captured like image. There exists many rectification methods, however, they require strong restriction on layout and way of capturing. For example, they are not applicable to the document image shown in Fig. 1. Thus, in this paper, we propose a novel rectifying method of perspective distortion of a document image. The proposed method estimates relative depth of each region of a document without any prior knowledge by employing an area of a character as a *variant* and an area ratio as an *invariant*. Since the proposed method does not use strong restriction on layout and way of capturing, it is applicable to document images of wide variety including the document image shown in Fig. 1. The comparison with existing methods is discussed in Sec. 4.

1. Introduction

Recently, camera-based document analysis and character recognition have been researched [3, 6]. Camera-based approach is known to be more difficult than scanner-based approach since a captured image can be degraded by nonuniform lighting, out of focus, perspective distortion and so on. Despite the difficulty, the camera-based approach has advantages of portability and ease to use. For example, scanners are not easy to carry, not able to scan a very big poster and an immovable object¹. To the contrary, cameras

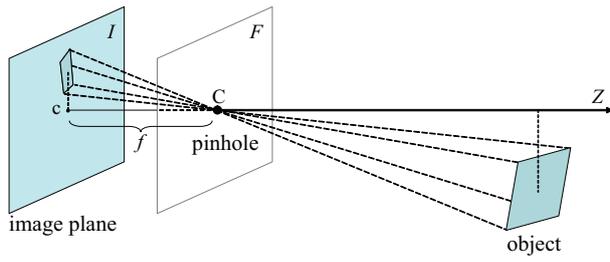
¹A portable scanner can be easily carried. However, the size of scannable paper is constrained.

2. Proposed method

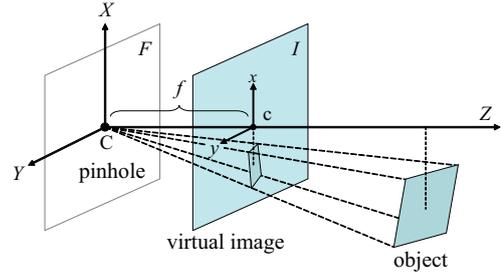
In this section, we explain the proposed rectifying method of perspective distortion. For the sake of the explanation, we begin with explaining the relationship between 3D and 2D coordinate systems.

2.1. Central perspective projection model [1, 7]

To begin with, we mention how a 2D image of a 3D object is obtained with a camera. As shown in Fig. 2(a), the



(a) The pinhole imaging model.



(b) The pinhole imaging model using virtual image. Virtual image shown in (b) is not inverted.

Figure 2. Central perspective projection model.

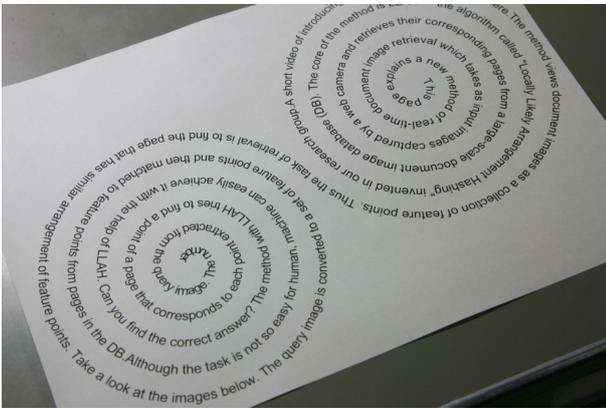


Figure 1. A document with difficulty to rectify for existing methods.

pinhole imaging model is often used. A point C is a *pinhole* at the *optical center*. Light rays passing through the pinhole C form an image of an object on the *image plane* I at a distance f from the optical center. An axis which passes through the pinhole C and is perpendicular to the image plane is called *optical axis*. Let c be the image center which is the intersection point between the image plane and the optical axis. In this model, parallel lines do not always be transformed into parallel lines. This transformation is called *perspective projection*. The distortion caused by the perspective projection is called *perspective distortion*.

In general, the image plane is rearranged as shown in Fig. 2(b). The image coordinate system is a 2D coordinate system which employs the image center c as the origin, and x - and y -axes as shown in Fig. 2(b). The camera-centered image coordinate system is the coordinate system which employs the focal point C as the origin, the optical axis as Z -axes, X - and Y -axes as x - and y -axes of the image coordinate system.

A point $(X, Y, Z)^T$ in the camera-centered image coordinate system is projected into the point $(x, y)^T$ in the 2D image coordinate system. The transformation is written as

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} X \\ Y \end{pmatrix}. \quad (1)$$

2.2. Relationship between area and depth

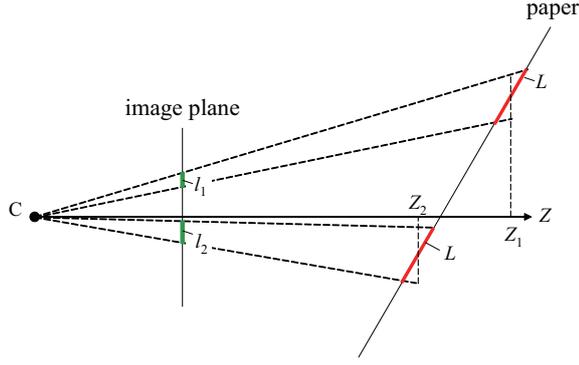
Imagine a document image suffering from perspective distortion, slanted in a certain angle. There usually exists the same characters, say “a,” printed in the same size. Due to the perspective distortion, the observed sizes of the characters vary by their positions; a character near the camera is large and that far from the camera is small. We estimate the slant angle from the changes of character sizes.

Here, we derive the relationship between the observed area and depth of a character². Fig. 3(a) illustrates two same characters in different positions on a document. The Z coordinates of them in the center are Z_1 and Z_2 , respectively. Say, $Z_1 > Z_2$.

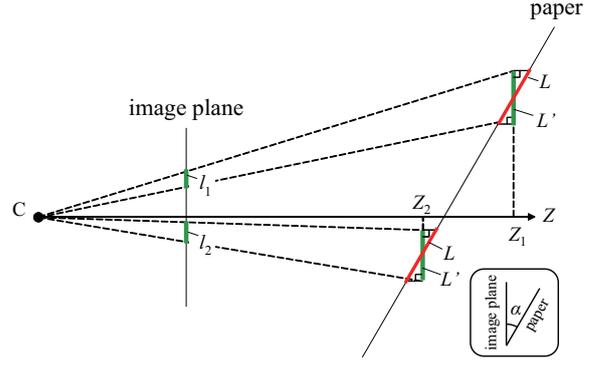
First of all, we consider a simplified problem handling a *length* (e.g., height or width) of a character instead of an area of it. Let L be the inherent length of the character. Let l_1 and l_2 be the projection lengths of the characters in the positions of Z_1 and Z_2 , respectively. To make the following calculation easier, we use an approximation as shown in Fig. 3(b). The approximation makes the slant characters standing. Let L' be their approximated length determined as

$$L' = L \cos \alpha, \quad (2)$$

²Note that the area of a character means the number of pixels in foreground colors. Actually, “area of a connected component” may be more precise expression since there exists separated characters such as “i” and “j”.



(a) Projection of characters.



(b) Projection of standing characters of approximated length L' .

Figure 3. The relationship between the inherent length L and the projection lengths l_1 and l_2 .

where α , $0 \leq \alpha \leq \pi/2$, is the angle between the paper and the image plane. Thus, by considering only x coordinate in Eq. (1), we obtain

$$l_j = \frac{f}{Z_j} L' = \frac{f}{Z_j} L \cos \alpha. \quad (3)$$

Then, we consider the area of a character. Let S and S' be the area and the approximated area of a character, which correspond to L and L' . The relationship of them is given as

$$S' = S \cos \alpha. \quad (4)$$

Then, we obtain the projection area of a character s_j as follows.

$$s_j = \left(\frac{f}{Z_j}\right)^2 S' = \left(\frac{f}{Z_j}\right)^2 S \cos \alpha. \quad (5)$$

Eq. (5) shows the relationship between an observed area and depth; a projection area is inverse proportional to the square of the Z coordinate (i.e., depth). Thus, letting Z_j and s_j be the depth and observed area of the j -th character, then we obtain the following relationship from Eq. (5):

$$Z_j = \frac{f\sqrt{S \cos \alpha}}{\sqrt{s_j}}. \quad (6)$$

Finally, we consider how to determine the angle α . Since all the characters are on a coplanar in a 3D coordinate system, the angle α will be obtained by fitting them to a coplanar. The detail of the process is discussed in Sec. 2.4. Here we mention the way to calculate a 3D coordinate from a 2D image coordinate. From Eq. (1), the coordinate of the j -th

character $(X_j, Y_j, Z_j)^T$ in the camera-centered coordinate system is denoted as

$$\begin{pmatrix} X_j \\ Y_j \\ Z_j \end{pmatrix} = \begin{pmatrix} Z_j x_j / f \\ Z_j y_j / f \\ Z_j \end{pmatrix}. \quad (7)$$

For the simplicity, we use the following expression instead:

$$\begin{pmatrix} X_j \\ Y_j \\ Z'_j \end{pmatrix} = \begin{pmatrix} Z'_j x_j \\ Z'_j y_j \\ Z'_j \end{pmatrix}, \quad (8)$$

where $Z' = Z/f$.

2.3. Clustering using area ratio

The method to estimate the slant α mentioned in Sec. 2.2 works only if all the characters in the document belong to one category. However, discussing such situation is nonsense. Thus, we discuss a method to distinguish characters into categories.

The most easily conceived method might be character recognition. However, recognizing distorted characters is not easy task. In addition, since just classification is required, recognizing (labelling) characters is unnecessary. Thus, we propose a novel classification method of characters. That is, classification by area ratios of character regions. The area ratio is known as an affine invariant. Though an affine invariant is not an invariant against perspective distortion, an affine invariant in a small region can be approximated as a projective invariant.

The area ratio has to satisfy the following two conditions. (1) The identical regions must be extracted.

Figure 4. An example of pairing of two characters. This is the case of “t” and “h.”

Since an area ratio is invariant, an (approximately) same value is calculated at the same region. However, if the same region is not extracted due to perspective distortion, the invariant cannot be obtained. To avoid the problem, the proposed method introduces the area of a convex hull which can be calculated in the same manner under linear transformation. Thus, the ratio of two areas, foreground region of a character and its convex hull, is used.

(2) Area ratios must be distinctive enough to classify characters.

Characters of different categories whose area ratios are identical can exist. In this case, they must be misclassified into a cluster. In order to avoid bad influence, we use several area ratios at the same time. This increases the classification performance since the probability that several area ratios are simultaneously identical is less than the probability that one area ratio is identical. Since the number of areas calculated from one character is limited, we calculate area ratios by pairing two adjacent characters. Fig. 4 is an example of the pairing. At most five area ratios are calculated from every pair of adjacent two characters. Fig. 5 shows the five area ratios of two characters of Fig. 4. When m out of five area ratios are used for clustering, they are used as an m -dimensional feature vector. We employ k -means clustering algorithm to distinguish the vector. Each cluster is expected to contain each character pair in ideal. Hereafter, all the processes are performed for each pair of two characters.

Here, we mention the restriction of the proposed method on the size of characters. The relationship between the area and depth derived in Sec. 2.2 bases on the assumption that characters in the same category are the same size. Thus, if characters of different sizes exist in a document, the clustering using area ratios cannot distinguish the difference in size of characters. This can cause the estimation error of depth. However, the characters in different sizes can be eliminated by noise reduction process detailed in Sec. 2.4 since most documents consist of many body text characters in the same size and few characters in headings in different sizes.

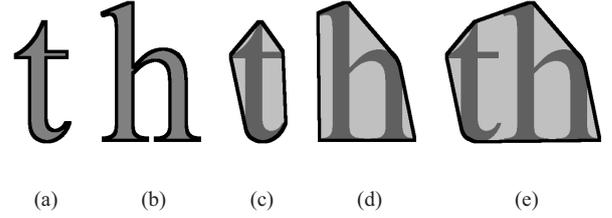


Figure 5. Five area ratios calculated from the pair of Fig. 4. (a) Area of “t,” (b) Area of “h,” (c) Area of convex hull of “t,” (d) Area of convex hull of “h,” and (e) Area of convex hull of “t” and “h.”

2.4. Fitting to plane

The clustering procedure described in Sec. 2.3 enables us to fit character pairs of each category to each plane, respectively. The estimated plane can be different since the area of a character of the same size differs by category and estimated depth varies on the size. However, these plane must be the same one. Thus, in order to estimate the slant angle α of the plane accurately, we integrate them. In Sec. 2.2, we let S be a known area of a character. However, it is actually unknown and differs by category. Thus, we estimate the area of a character S and the slant angle of only one plane α simultaneously. In order to that, we replace S , $(X_j, Y_j, Z_j)^T$ and Z'_j appeared in Sec. 2.2 with S_i , $(X_{ij}, Y_{ij}, Z_{ij})^T$ and Z'_{ij} by adding cluster number i . We also add cluster number i to from Eq. (6) to Eq. (8).

To begin with, by substituting Eq. (6) into Eq. (7), we obtain

$$\begin{pmatrix} X_{ij} \\ Y_{ij} \\ Z'_{ij} \end{pmatrix} = \sqrt{S_i} \cos \alpha \begin{pmatrix} x_{ij}/\sqrt{s_{ij}} \\ y_{ij}/\sqrt{s_{ij}} \\ 1/\sqrt{s_{ij}} \end{pmatrix}. \quad (9)$$

This means that the coordinate $(X_{ij}, Y_{ij}, Z'_{ij})$ of each character is calculated by x_{ij} , y_{ij} and s_{ij} obtained from the image. However, since the inherent area of a character S_i in Eq. (9) and the slant angle α are unknown, we let $K_i = \sqrt{S_i} \cos \alpha$ and define an error of Z coordinates between the plane and a character of a category as

$$\begin{aligned} \varepsilon_{ij} &\equiv \left| \{aX_{ij} + bY_{ij} + c\} - Z'_{ij} \right| \\ &= \left| \left\{ a \frac{K_i x_{ij}}{\sqrt{s_{ij}}} + b \frac{K_i y_{ij}}{\sqrt{s_{ij}}} + c \right\} - \frac{K_i}{\sqrt{s_{ij}}} \right|. \end{aligned} \quad (10)$$

Then, the sum of the error for all the characters of all the

categories is given as

$$E = \sum_i \sum_j \varepsilon_{ij}. \quad (11)$$

Finally, we estimate the parameters $\{K_i\}$, and a , b and c which minimize Eq. (11). Note that since there is linear ambiguity in $\{K_i\}$, we fixed as $c = 1$, and estimate $\{K_i\}$, a and b in this paper. However, to estimate the plane, we cannot avoid taking the effect of noises (outliers) into account. The noises come from failure of extracting characters from the image, misclassification, and existence of characters in different sizes mentioned in Sec. 2.3. To deal with the noises, we perform two noise removal procedures: (A) removal of outliers, and (B) removal of clusters of outliers. The former removes a character where the error ε_{ij} is not less than a threshold t_1 . The latter removes a cluster whose number of elements is not greater than t_2 , since the cluster seems to estimate a wrong plane.

2.5. Rotation

We discuss how to set the view point in right front of the paper. This is performed by rotating the paper so that the view point moves on the normal of the paper. To represent a rotation matrix, we use the roll-pitch-yaw rotation angles where rotation angles around the X -, Y - and Z -axes are represented by ψ , θ and ϕ , respectively. The rotation matrix is given as

$$\begin{aligned} \mathbf{R} &= \mathbf{R}(Z, \phi) \mathbf{R}(Y, \theta) \mathbf{R}(X, \psi) \\ &= \begin{pmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &\quad \begin{pmatrix} \cos \theta & 0 & -\sin \theta \\ 0 & 1 & 0 \\ \sin \theta & 0 & \cos \theta \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \psi & -\sin \psi \\ 0 & \sin \psi & \cos \psi \end{pmatrix}. \end{aligned} \quad (12)$$

Since we defined $Z' = Z/f$, the equation of the plane $Z' = aX + bY + 1$ is $Z = afX + bfY + f$. The normal of the plane $Z = afX + bfY + f$ is $(af \ bf \ -f)^T$. Thus, the rotation matrix \mathbf{R} should be determined so that both the X and Y coordinates become 0. Since the rotation around the Z -axis is ignored, rotation angles of \mathbf{R} are given as

$$\begin{pmatrix} \phi \\ \theta \\ \psi \end{pmatrix} = \begin{pmatrix} 0 \\ -\tan^{-1} \frac{af}{\sqrt{1+(bf)^2}} \\ -\tan^{-1}(bf) \end{pmatrix}. \quad (13)$$

Eq. (13) shows that an unknown parameter f is required for the estimation of the angles. This is in the same situation as [4] and [8]. Since f is not obtained in this paper, we simply

let $f = 1$. Due to this, affine distortion remains after the rectification. The affine distortion is the distortion where parallel lines are kept parallel after transformation.

Finally, we rectify the image by the rotation. By rotating a point $(X_{ij}, Y_{ij}, Z_{ij})^T$ in the camera-centered coordinate system, we obtain the coordinate

$$\begin{pmatrix} \tilde{X}_{ij} \\ \tilde{Y}_{ij} \\ \tilde{Z}_{ij} \end{pmatrix} = \mathbf{R} \begin{pmatrix} X_{ij} \\ Y_{ij} \\ Z_{ij} \end{pmatrix}. \quad (14)$$

Then, projecting the point into an image plane, the coordinate on the 2D image plane

$$\begin{pmatrix} \tilde{x}_{ij} \\ \tilde{y}_{ij} \end{pmatrix} = \frac{f}{\tilde{Z}_{ij}} \begin{pmatrix} \tilde{X}_{ij} \\ \tilde{Y}_{ij} \end{pmatrix} \quad (15)$$

is given.

3. Experiment

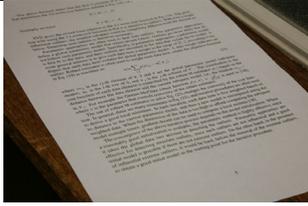
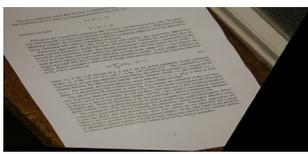
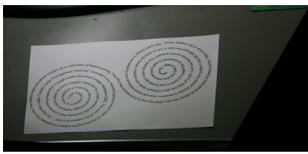
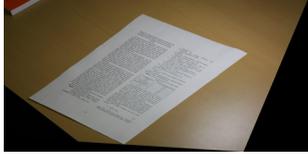
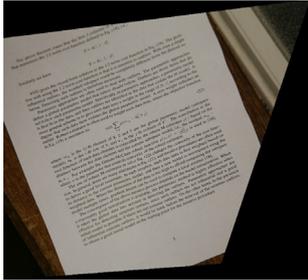
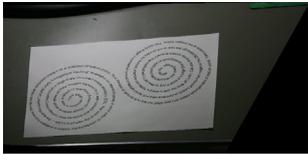
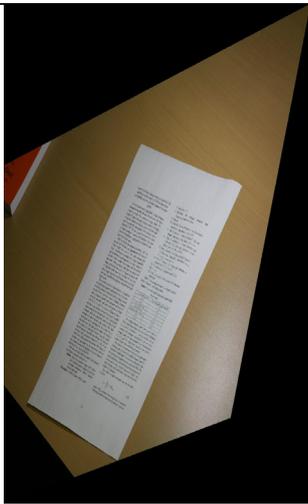
We performed experiments to evaluate the proposed method. Five combinations of area ratios we used were the following:

- i. $\frac{\text{Smaller (Fig. 5(a))}}{\text{Smaller CH (Fig. 5(c)) + Larger CH (Fig. 5(d))}$,
- ii. $\frac{\text{Larger (Fig. 5(b))}}{\text{Smaller (Fig. 5(a)) + Larger (Fig. 5(b))}$,
- iii. $\frac{\text{Smaller (Fig. 5(a))}}{\text{Larger CH (Fig. 5(d))}$,
- iv. $\frac{\text{Smaller CH (Fig. 5(b))}}{\text{Larger (Fig. 5(c))}$,
- v. $\frac{\text{Smaller (Fig. 5(a))}}{\text{Larger (Fig. 5(b))}$,

where ‘‘smaller’’ stands for the area of smaller character of the pair, ‘‘larger’’ stands for the area of larger character, and ‘‘CH’’ stands for convex hull. When the dimensionality of an invariant vector was 5, area ratios i.~v. were used. When the dimensionality of an invariant vector was 3, area ratios i.~iii. were used.

The experimental results for three images are shown in Table 1. The images were taken by Canon EOS 5D, and their sizes were $4,368 \times 2,912$. Though the proposed method does not require a paper frame, for ease of visual evaluation, document images with paper frames were used. The images in the row (A) are original images. Those in the row (B) are rectified images with $f = 1$. Parameters for clustering and noise reduction were tuned so as to obtain the best result. In theory, the parallelism of two lines

Table 1. Rectification results and parameters of the proposed method. (A) Before rectification. (B) After rectification ($f = 1$). (C) After rectification (the best value of f is selected by hand).

		Image 1	Image 2	Image 3
(A)				
	Angles of two lines in long side (L) and narrow side (S)	(L) 8.98° / (N) 5.18°	(L) 1.73° / (N) 19.2°	(L) 6.64° / (N) 7.70°
(B)				
	Angles of two lines in long side (L) and narrow side (N)	(L) 0.06° / (N) 0.73°	(L) 0.14° / (N) 2.81°	(L) 1.78° / (N) 5.43°
	No. of dim. of invariant vector	3	3	5
	No. of clusters	60	40	200
	Thresholds t_1 and t_2	0.05, 20	(No thresholding)	0.1, 20
(C)				
		$f = 11000$	$f = 2000$	$f = 15000$
	Angles of two lines in long side (L) and narrow side (N)	(L) 0.78° / (N) 0.02°	(L) 1.24° / (N) 1.09°	(L) 1.20° / (N) 3.88°
	Average difference of corner angles from right angle	0.605°	1.35°	4.16°

is rectified, though a right angle of a corner is not. In the experimental results, parallel lines of image 1 were rectified within 1° , and those of images 2 and 3 were not. The main cause is estimation errors of parameters of the plane due to outliers. The images in the row (C) are rectified images with the best f . In theory, not only the parallelism but also corner angles are rectified. In the experimental results, as the same reason as the row (B), both the parallelism and corner angles of image 1 was almost rectified, and that of images 2 and 3 were not. Therefore, improving estimation accuracy of a plane and deriving estimation method of f are required.

4. Comparison with existing methods

Since rectification of perspective distortion is a basic task on camera-based document analysis, there are many existing methods. They are roughly classified into the following three approaches: (1) using a paper frame, (2) using text lines, (3) using stereo vision. We mention the outline of the methods and discuss the difference from the proposed method.

The first approach assumes that the paper frame is a rectangle in nature and the frame can be clearly obtained. A rectangle suffering from perspective distortion becomes a quadrilateral since the parallelism of lines is lost. Using the information that the quadrilateral is originally a rectangle, we can calculate the transformation and rectify by the inverse transformation. This approach is used in [2] and some commercial products such as Ricoh Caplio R6. Though the approach is reasonable since many paper frames are rectangle, the whole document image have to be captured.

The second approach assumes the parallelism of text lines. For example, in [2], vanishing points are estimated from text lines, and then slant angles of the paper are estimated from the vanishing points³. This method first extracts text lines from a document image, and horizontal vanishing points are estimated. Next, assuming both ends of text lines are aligned vertically, vertical vanishing points are estimated by drawing three lines at right end, center and left end of the text lines. Finally, the document image is rectified by the two vanishing points and prior knowledge that horizontal text lines and vertical lines of ends are orthogonal⁴. The biggest drawback of the method is strong restriction for page layout. The method is not applicable to a document with complex layout and a document including many figures and equations since estimating both ends is difficult.

³[2] includes two methods in approaches (1) and (2).

⁴Note that the prior knowledge used here, an original angle of horizontal and vertical lines, has the same amount of information as the focal length f because both of them reduce one degree of freedom of a projection matrix. If we use such a prior knowledge, we can rectify remaining affine distortion of the experimental result (B) and obtain (C) in Table 1.

Approaches (1) and (2) mentioned above are not applicable to a document with complex layout, such that text lines are not parallel, and whole paper frame is not captured, such as the document image shown in Fig. 1.

The third approach estimates 3D shape of a paper with multiple cameras [5] or a movie capturing a document by a hand-held camera [9]. These require different devices from the proposed method.

5. Conclusions

In this paper, we proposed a method of rectifying perspective distortion into affine distortion without any prior knowledge. The proposed method estimates relative depth of each region of a document by employing an area of a character as a variant and an area ratio as an invariant. Then, 3D pose (slant angles) of the plane of the document is estimated. Since the proposed method does not use strong restriction on layout and way of capturing, it is applicable to document images of wide variety.

In the experiments, we confirmed the rectification ability of the proposed method in recovering the parallelism of lines. Though we confirmed that the rectification roughly succeeded, estimation accuracy should be improved. It can be achieved by employing the robust estimation and improving the performance of noise removing.

Due to the limitation of the proposed method, it cannot recover corner angles of a paper in principle. The cause is that there is linear ambiguity in depth estimation. This may be solved by employing other pair of variant and invariant. This is included in future work. Larger scale evaluation is also included in future work.

References

- [1] *Computer Vision: A Modern Approach*. Prentice Hall, 2002.
- [2] P. Clark and M. Mirmehdi. Recognising text in real scenes. *Int'l Journal of Document Analysis and Recognition*, 4:243–257, 2002.
- [3] D. Doermann, J. Liang, and H. Li. Progress in camera-based document image analysis. pages 606–616, 2003.
- [4] Y. ichi Ohta, K. Maenobu, and T. Sakai. Obtaining surface orientation from texels under perspective projection. In *Proc. of 7th International Conference on Artificial Intelligence*, pages 746–751, 1981.
- [5] C. H. Lampert, T. Braun, A. Ulges, D. Keysers, and T. M. Breuel. Oblivious document capture and real-time retrieval. In *Proc. First Int'l. Workshop on Camera-Based Document Analysis and Recognition*, pages 79–86, Aug. 2005.
- [6] J. Liang, D. Doermann, and H. Li. Camera-based analysis of text and documents: a survey. *International Journal of Document Analysis and Recognition*, 7:84–104, 2005.
- [7] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, June 2005.

- [8] M. Pilu. Extraction of illusory linear clues in perspective skewed documents. In *Proc. Computer Vision and Pattern Recognition, 2001 (CVPR '01)*, volume 1, pages 363–368, 2001.
- [9] T. Sato, S. Ikeda, M. Kanbara, A. Iketani, N. Nakajima, N. Yokoya, and K. Yamada. High-resolution video mosaicing for documents and photos by estimating camera motion. In *Proc. SPIE Electronic Imaging*, volume 5299, pages 246–253, Jan. 2004.