A new method for perspective correction of document images

José Rodríguez-Piñeiro[a], Pedro Comesaña-Alfaro[a,b], Fernando Pérez-González[a,b,c] and Alberto Malvido-García[d]

[a] University of Vigo, Signal Theory and Communications Dept., Vigo, Spain
[b] University of New Mexico, Electrical and Computer Engineering Dept., Albuquerque, NM
[c] Gradiant (Galician Research and Development Center in Advanced Telecommunications), Vigo, Spain
[d] Bit Oceans Research, López de Neira, 3, Office 408, 36202 Vigo, Spain

ABSTRACT

In this paper we propose a method for perspective distortion correction of rectangular documents. This scheme exploits the orthogonality of the document edges, allowing to recover the aspect ratio of the original document. The results obtained after correcting the perspective of several document images captured with a mobile phone are compared with those achieved by digitizing the same documents with several scanner models.

Keywords: Document image, perspective distortion correction

1. INTRODUCTION

Nowadays, the increasing performance and low price of portable imaging devices (e.g. mobile phones, PDAs) are boosting the usage of these devices for supplementing or even replacing traditional flat-bed scanners for document image adquisition. Unfortunately, a number of problems that traditional scanners do not have to face have arisen with this increasing use, with *perspective distortion* one of the most evident and probably most harmful for the subsequent application of document processing tools. Although this problem has already deserved some attention in previous works in the literature, most of the proposed solutions are based on rather restrictive assumptions on the nature of the captured document, so the target of this work is to follow a systematic approach with a minimum number of constraints. Indeed, the only constraint imposed on our method is that the four corners of the document are captured in the considered image.

Next, we review some of the most representative methods related to perspective distortion correction. These methods can be classified into two main categories. The methods in the first category use the text in the document for characterizing perspective distortion. The second category encompasses those algorithms that do not require that the original document includes text. A deeper discussion about both is provided below.

1.1 Methods requiring text in the document

Clark and Mirmehdi[1] proposed a perspective distortion correction method based on the use of vanishing points recovery, where these points provide the information required to correct the perspective distortion of the captured document. In this work, the recovery of the vanishing points is based on the assumption that a text paragraph must display some sort of left, right, centered or full formatting. Probably, its main drawbacks are the computational cost required by some steps of the correction method —including several image transformations and exhaustive searches on some parameters, such as the horizontal vanishing point— and the need of knowing some correspondences between imaged points and their real-world counterparts.

A different approach is followed by Lu *et al.*,[2] where a method based on applying morphological operators is proposed. This method needs neither high-contrast document boundaries nor paragraph formatting information. Nevertheless, it is constrained to deal with text documents, and, even more importantly, it is based on the use of some parameters that require much knowledge about the document contents, like the number of characters

Further author information: (Send correspondence to P. C.-A.)
P. C.-A.: E-mail: pcomesan@gts.uvigo.es, Telephone: +34 986818655

in the document image, and some parameters that are not automatically tuned, like some parameters used to discard false detections of top and bottom lines of the text; thus, small changes on the value of those parameters can make the method fail. More recently, Lu and Tan have presented an extended version of the mentioned work,[3] although it inherits the discussed drawbacks.

Morphological operations are also used in the method proposed by Miao and Peng,[4] although a smaller knowledge about the contents of the document is required in comparison with Lu *et al.*'s aproaches. Additionally, an adaptive thresholding technique is adopted to binarize the capture, which makes the method capable of dealing with lighting variations. Nevertheless, no clue is provided on how to obtain the value of some parameters —like the size of structure elements used to cluster the detected connected components into text lines—. Furthermore, the image correction process requires the use of three transformations, which are computationally expensive.

Methods that do not depend so much on the contents of the document, as they are not exclusively aimed at text-based documents, are due to Yun.[5,6] The first of them is centered on the perspective estimation problem, while the second one is focused on the rectification system design. Indeed, Yin *et al.*[5] describe a method that uses textual information —if it is available— and also other sources of information, such as document boundaries. Nevertheless, most of the method is based on the clues provided by textual information. In this case there are also some parameters that are not automatically tuned —like the thresholds used to classify a detected line as horizontal—. On the other hand, both methods are carefully designed to minimize the computational cost. In fact, a multi-stage approach to perspective correction is proposed[6] which is able to avoid some unnecesary stages. Nevertheless, the most important drawback of these methods is the fact that they are not able to recover the original aspect ratio of the document.

Iwamura *et al.*[7] proposed a method that estimates the depth of each area of the document by using measurements of its textual contents, like the variation of the area of characters with respect to their position. Nevertheless, the proposed aproach does not obtain the focal length, being only able to recover an affine distorted version of the original document.

Finally, there exist some considerations that can make the methods based on text documents undesirable for general applications. First of all, it is obvious that the requirement of dealing with text documents reduces the generality of the designed methods, and hence, the potential number of applications. In addition, most of the described methods constrain features like the size of the text, as well as its variation over the document, or parameters about the paragraphs formatting. Furthermore, according to the described algorithms, most of them could fail when restoring handwritten documents, or those written with some particular typographies, such as italics, where the tips of the characters are not vertical. Also, the presence of several columns of text can make some methods fail, as well as the consideration of different alphabets —like some kinds of writing that are not ordered from left to right and from top to bottom—. According to these facts, it is reasonable to pay attention to those methods that do not require the presence of text in the document.

1.2 Methods not requiring text in the document

A more theoretic approach than those followed in the methods presented in the previous section is introduced by Liebowitz and Zisserman.[8] This approach is not only suitable for recovering the fronto-parallel view of text documents, but also for describing the geometry, constraints and algorithmic implementations that allow metric properties of figures on a plane, like angles and length ratios, to be measured from a captured image of that plane. Perhaps the most novel contribution of this work is the presentation of different ways of providing geometrical constraints, including the availability of a known angle in the original scene, two equal but unknown angles, or a known length ratio. Unfortunately, formal proofs supporting those procedures are not provided. It must be also noted that, depending on the level of knowdelegde about the contents of the document —measured by means of the number of known pairs of orthogonal lines in the original scene—, more than a single image transformation may be required.

On the other hand, there also exist a great number of publications that deal with camera calibration. Although they are not designed for performing perspective distortion correction of captures, they can be used to estimate the camera position and its orientation in relation to the imaged document. Unfortunately, since those methods have been designed for other purposes, most of them cannot be applied to the perspective distortion correction problem.

This is not the case of Guillou *et al.*'s approach,[9] where a method is proposed for camera calibration as well as recovering fronto-parallel views of general three-dimensional scenes. Although the followed approach is similar to that described in the current work, especially in what concerns the characterization of the camera parameters, the estimation of the camera position and orientation in Guillou *et al.*'s approach is based on the use of two vanishing points, being the transformation performed by using projective coordinates. In contrast, our approach is exclusively based on Euclidean geometry, not requiring the use of vanishing points nor projective coordinates. In any case, it is interesting to remark that the limitations and requirements of both methods are very similar.

T. M. Breuel and colleagues have also published several works dealing with the rectification of document images. Of those works, the most similar one[10] to the current proposal is based on the work by Zhang and He.[11] Special attention should be devoted to the latter paper, where the authors propose a method that enhances whiteboards captures. Indeed, the approach followed by those authors to estimate the focal distance and the aspect ratio of the original object (a rectangular whiteboard) is similar to that presented here, although the specific way of performing picture points transformation is not specified (which is instead done in the current paper). Furthermore, the current work studies in detail the problematic cases, i.e. those situations where the perspective distortion parameters can not be estimated. As it happened for Guillou *et al.*'s approach,[9] both the requirements and the limitations of Zhang and He's proposal[11] are similar to those of the method introduced below.

The remainder of this paper is organized as follows: Sect. 2 introduces the considered capture process model, and the proposed perspective distortion correction method is introduced in Sect. 3. Experimental results are provided in Sect. 4, and Sect. 5 presents the main conclusions.

2. CAPTURE PROCESS MODEL

When a three-dimensional scene is photographed, a two-dimensional image is captured. Although this dimensionality reduction inherently entails ambiguity about the photographed scene, in this work it will be shown that whenever some constraints about the photographed object are met, a good estimate of the original object can be obtained (although modified by an isotropic-scaling). Specifically, a method is proposed aimed at estimating the original document-plane relative coordinates of a rectangular document.

In order to achieve this goal, in this section the capture model of a flat document will be given, and in the next section the possibility of obtaining the inverse transform of that capture process will be studied.

Three different coordinate systems will be introduced to illustrate the capture process of a flat document:

- **Original document plane**. Let us denote by $\bar{\mathbf{x}}$ the coordinates of a point of the original document in the plane defined by that document, so $\bar{\mathbf{x}} \in \mathbb{R}^2$. The origin of this two-dimensional space will be assumed to be located at a corner of the document, and its axes will be aligned with document edges.

- **Three-dimensional space**. Given that the document to be photographed is located in a three-dimensional space, $\bar{\mathbf{x}}$ will be transformed onto a different point $\bar{\mathbf{y}} \in \mathbb{R}^3$. The origin of this three-dimensional space will be located at the camera center.

- **Image plane**. When the document located in a three-dimensional space is captured, a two-dimensional image is obtained. This process needs to define an aiming direction in $\mathbb{R}^3$, and a plane orthogonal to that direction, i.e. the so-called *image plane*. Consequently, $\bar{\mathbf{y}}$ will produce a point in the two-dimensional image by intersecting the straight line defined by $\bar{\mathbf{y}}$ and the camera center with the image plane; this intersecting point will be denoted by $\bar{\mathbf{z}} \in \mathbb{R}^3$. For the sake of simplicity, the aiming direction will coincide with the third component of the three-dimensional space introduced above, so the points on the image plane will be characterized by sharing their third component, i.e. $\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2, \bar{f})^T$, where $\bar{f}$ is usually named camera's *focal length*. Finally, the origin of this image plane, i.e. $(0, 0, \bar{f})^T$, will be located at the center of the obtained rectangular image.

This notation, used for denoting the point coordinates according to the different coordinate systems, should not be confused with that followed for referring to the coordinate system components; namely, X, Y and Z

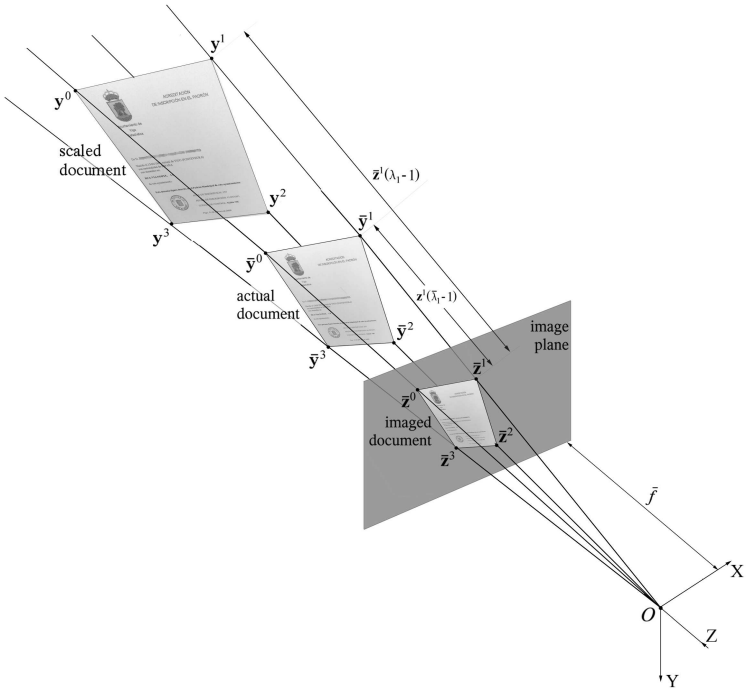

Figure 1. Pinhole model.

will be used for denoting the first, second, and (where applicable) third component of the considered coordinate system.

The used capture model is depicted in Fig. 1, and it is usually named *pinhole model of the camera*,[12] being extensively used in the literature.

Taking into account the definition of $\bar{\mathbf{z}}$ and $\bar{\mathbf{y}}$, it is straightforward to see that there exists a scalar $\bar{\lambda} > 0$ such that $\bar{\mathbf{y}} = \bar{\lambda}\bar{\mathbf{z}}$, and at the same time all the values of $\mathbf{y}$ equal to $\lambda\bar{\mathbf{z}}$ for any $\lambda > 0$ will produce the same captured point, showing the capture ambiguity mentioned at the beginning of this section. An example of this ambiguity is illustrated in Fig. 1, where both $\bar{\mathbf{y}}$ and $\mathbf{y}$ would produce the same captured point $\bar{\mathbf{z}}$. Finally, it will be useful to take into account that $\bar{\lambda}$ can be constrained to be strictly positive, as $\bar{\mathbf{y}}$ cannot be the null vector (i.e., we will assume that the document will not be located at the camera center).

The approach followed in this work for correcting the perspective distortion is based on performing the inverse operations to those involved in the capture process.

3. INVERTING THE CAPTURE PROCESS

3.1 Recovering the three-dimensional coordinates of the corners

Let $\bar{\mathbf{z}}^i$ be the image plane coordinates of the ith corner of the document, where $i = 0, \cdots, 3$, and the indices are assigned in such a way that the ith corner's neighbors are the $(i \pm 1)$mod4-th corners. Following the notation introduced in the previous section, the first two components of $\bar{\mathbf{z}}^i$ determine the position of the ith corner of the document in the captured image. Nevertheless, the value of the third component, i.e. $\bar{f}$, will be usually not known at the image processing side. Therefore, in order to invert the capture process, we will assume that the value of that parameter is f, and afterwards, as it will be shown below, the value of f will be estimated taking into account the geometrical properties of the original document. Consequently, we will denote by $\mathbf{z}^i \triangleq (\bar{z}_1^i, \bar{z}_2^i, f)$ the estimated coordinates of the document on the image plane.

According to the discussion presented above, one can compute the coordinates of every point in the three-dimensional space yielding $\mathbf{z}^i$, obtaining the points in the straight semi-line $\mathbf{y} = \lambda\mathbf{z}^i$, with $\lambda > 0$. However, one needs to estimate the particular λ_i's yielding $\mathbf{y}^i = \lambda_i\mathbf{z}^i$, using again the geometrical properties among the $\mathbf{y}^i$'s.

In order to achieve this target, we will assume, without loss of generality (as we allow an isotropical scaling), that $\lambda_0 = 1$, so $\mathbf{y}^0 = \mathbf{z}^0$. Taking into account that the opposite edges of the document are parallel, it is straightfoward to see that $\mathbf{y}^1 - \mathbf{y}^0 = \mathbf{y}^2 - \mathbf{y}^3$, so $\mathbf{z}^0 = \lambda_1\mathbf{z}^1 - \lambda_2\mathbf{z}^2 + \lambda_3\mathbf{z}^3$, or in its matrix form

$$\begin{pmatrix} z_1^1 & -z_1^2 & z_1^3 \\ z_2^1 & -z_2^2 & z_2^3 \\ f & -f & f \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \mathbf{z}^0. \tag{1}$$

This system will have a unique solution whenever the matrix above has a non-null determinant, that will be indeed the case as long as we can not write $\mathbf{z}^3$ as

$$\tau_1 \begin{pmatrix} z_1^1 \\ z_2^1 \\ f \end{pmatrix} + \tau_2 \begin{pmatrix} z_1^2 \\ z_2^2 \\ f \end{pmatrix} = \begin{pmatrix} z_1^3 \\ z_2^3 \\ f \end{pmatrix},$$

or equivalently,

$$\tau_1 \begin{pmatrix} z_1^1 \\ z_2^1 \end{pmatrix} + \tau_2 \begin{pmatrix} z_1^2 \\ z_2^2 \end{pmatrix} = \begin{pmatrix} z_1^3 \\ z_2^3 \end{pmatrix},$$

and $\tau_1 + \tau_2 = 1$, meaning that the captured image of the third vertex belongs to the straight line joining the first and second vertices of the captured images. This case happens whenever one takes the picture of the document sideways, preventing its contents from being recovered.

On the other hand, by taking into account that two adjacent edges of the original document are orthogonal, one can derive that $(\mathbf{y}^0 - \mathbf{y}^3)^T \cdot (\mathbf{y}^1 - \mathbf{y}^0) = (\mathbf{z}^0 - \lambda_3\mathbf{z}^3)^T \cdot (\lambda_1\mathbf{z}^1 - \mathbf{z}^0) = 0$, so

$$\left[\begin{pmatrix} z_1^0 \\ z_2^0 \end{pmatrix} - \lambda_3 \begin{pmatrix} z_1^3 \\ z_2^3 \end{pmatrix} \right]^T \cdot \left[\begin{pmatrix} z_1^0 \\ z_2^0 \end{pmatrix} - \lambda_1 \begin{pmatrix} z_1^1 \\ z_2^1 \end{pmatrix} \right] = -f^2 (1 - \lambda_3)(1 - \lambda_1),$$

enabling to compute the focal length, f, as the positive square root of

$$-\frac{\left[\begin{pmatrix} z_1^0 \\ z_2^0 \end{pmatrix} - \lambda_3 \begin{pmatrix} z_1^3 \\ z_2^3 \end{pmatrix} \right]^T \cdot \left[\begin{pmatrix} z_1^0 \\ z_2^0 \end{pmatrix} - \lambda_1 \begin{pmatrix} z_1^1 \\ z_2^1 \end{pmatrix} \right]}{(1 - \lambda_3)(1 - \lambda_1)}.$$

Problems would arise when computing the square root whenever the former term were negative, or its denominator were null. The first case is not possible due to the geometrical properties of the considered kind of documents, so we will focus on the second one, which will require that either $\lambda_1 = 1$, or $\lambda_3 = 1$, or both of them simultaneously, meaning that the focal length will be undetermined or infinite.

Considering $\lambda_1 = 1$, and taking into account that the product of the last row of the matrix in (1) by the vector $(\lambda_1, \lambda_2, \lambda_3)^T$ implies that $\lambda_1 - \lambda_2 + \lambda_3 = 1$, it is clear that in that case $\lambda_2 = \lambda_3$, so $\mathbf{y}^2$ and $\mathbf{y}^3$ will be located at the same distance from the image plane. Furthemore, $\mathbf{z}^1 - \mathbf{z}^0 = \lambda_2(\mathbf{z}^2 - \mathbf{z}^3)$, meaning that the edges that join the points $\mathbf{z}^0$ and $\mathbf{z}^1$ as well as $\mathbf{z}^3$ and $\mathbf{z}^2$ are also parallel in the captured image. Although the constraint on the rectangular nature of the document, summarized by (1), is verified for any value of f, the ratio between the lengths of the orthogonal edges depends on that parameter, as it is defined as

$$r \triangleq \frac{||\mathbf{z}^0 - \mathbf{z}^1||}{||\mathbf{z}^0 - \lambda_3\mathbf{z}^3||} = \frac{\sqrt{(z_1^0 - z_1^1)^2 + (z_1^0 - z_2^1)^2}}{\sqrt{(z_1^0 - \lambda_3 z_1^3)^2 + (z_1^0 - \lambda_3 z_2^3)^2 + f^2 (1 - \lambda_3)^2}},$$

showing that in this particular case the ratio between edges can not be recovered. A similar reasoning can be applied to the case $\lambda_3 = 1$.

On the other hand, if both $\lambda_1 = 1$ and $\lambda_3 = 1$, then, from the constraint provided by the last row of (1), it is straightforward to see that $\lambda_2 = 1$, so in that case the four vertices of the original document belong to the plane z_3^i, i.e. the captured document has no perspective distortion.

Summarizing, although when both λ_1 and λ_2 are equal to 1 the irrecoverability of the focal length is not relevant, as there is no perspective distortion to be corrected, this is indeed a problem when only one of them is equal to 1. Note that, although Zhang and He[11] only consider explicitly the case where both λ_1 and λ_2 are equal to 1, it can be easily proved that their approach also fails to recover the aspect ratio when only one of them is equal to 1.

3.2 Recovering the three-dimensional coordinates of an arbitrary document point

Once the three-dimensional coordinates of the four document corners have been recovered (except for an isotropical scaling), one can use that information to calculate the three-dimensional coordinates $\mathbf{y}$ of an arbitrary point on the captured document $(\bar{z_1}, \bar{z_2})$. Obviously, the derivation of $\mathbf{y}$ requires to calculate the value of λ such that $\mathbf{y} = \lambda\mathbf{z}$ lies on the plane defined by $\mathbf{y}^i$, $i = 0, \cdots, 3$. It can be seen that the value of λ we are looking for is given by[13]

$$\lambda = -\frac{\begin{vmatrix} 1 & 1 & 1 & 1 \\ y_1^0 & y_1^1 & y_1^2 & 0 \\ y_2^0 & y_2^1 & y_2^2 & 0 \\ y_3^0 & y_3^1 & y_3^2 & 0 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 & 0 \\ y_1^0 & y_1^1 & y_1^2 & z_1 \\ y_2^0 & y_2^1 & y_2^2 & z_2 \\ y_3^0 & y_3^1 & y_3^2 & z_3 \end{vmatrix}}. \tag{2}$$

One should be especially careful when the denominator of the last expression is null, an event that will happen if and only if the columns of the matrix considered in the denominator of (2) are linearly dependent. For the sake of simplicity, we will partition all the possible scenarios where this condition is verified in two subsets: 1) one can find τ_0 and τ_1 such that $\tau_0\mathbf{y}^0 + \tau_1\mathbf{y}^1 = \mathbf{y}^2$, with $\tau_0 + \tau_1 = 1$, (i.e. the first three columns of the mentioned matrix are linearly dependent) and 2) the first three columns of the matrix are linearly independent, but $\tau_0\mathbf{y}^0 + \tau_1\mathbf{y}^1 + \tau_2\mathbf{y}^2 = \mathbf{z}$ and $\tau_0 + \tau_1 + \tau_2 = 0$.

Although in the first case the numerator of (2) will be also null, which would produce an undetermined result, one can see that the considered scenario violates the geometrical assumptions about the document, as it should hold $\tau_0\mathbf{y}^0 + (1 - \tau_0)\mathbf{y}^1 = \mathbf{y}^1 + \tau_0(\mathbf{y}^0 - \mathbf{y}^1) = \mathbf{y}^2$, meaning that $\mathbf{y}^0$, $\mathbf{y}^1$ and $\mathbf{y}^2$ all lie on a straight line.

On the other hand, in the second case the numerator of (2) will be non-null, so if the four columns of the considered matrix are linearly dependent, the λ corresponding to $\mathbf{z}$ would be infinite. Nevertheless, this would happen if and only if

$$-(\tau_1 + \tau_2)\mathbf{y}^0 + \tau_1\mathbf{y}^1 + \tau_2\mathbf{y}^2 = \tau_1(\mathbf{y}^1 - \mathbf{y}^0) + \tau_2(\mathbf{y}^2 - \mathbf{y}^0) = \mathbf{z}, \tag{3}$$

i.e., $\mathbf{z}$ belongs to the plane containing the coordinate system origin (in this case, the camera center) and the points $(\mathbf{y}^1 - \mathbf{y}^0)$ and $(\mathbf{y}^2 - \mathbf{y}^0)$. Taking into account that the plane defined by the document in the three-dimensional space is given by

$$\mathbf{y} = \mathbf{y}^0 + \tau_1'(\mathbf{y}^1 - \mathbf{y}^0) + \tau_2'(\mathbf{y}^2 - \mathbf{y}^0), \tag{4}$$

as $\mathbf{y}^0$, $\mathbf{y}^1$ and $\mathbf{y}^2$ can be described through (4), it is evident that the planes defined by (3) and (4) will be parallel or coincident. By definition, for each point of the document in the three-dimensional space $\mathbf{y}$, there exists a scalar λ such that $\mathbf{y} = \lambda\mathbf{z}$ for some point of the image plane $\mathbf{z}$. Thus, if $\mathbf{z}$ verifies (3), being $\mathbf{y}$ a scaled version of $\mathbf{z}$, $\mathbf{y}$ will also verify (3). Nevertheless, since $\mathbf{y}$ belongs to the document plane in the three-dimensional space, it will simultaneously verify (4), implying that $\mathbf{y}$ belongs to both planes, which would indeed coincide. Therefore, (3) will be verified if and only if the origin of the three-dimensional coordinate system, i.e. the camera center, belongs to the plane defined by the document, meaning that the picture is taken sideways.

In the non-degenerate cases where the denominator of (2) is non-null, one can write

$$\mathbf{y} = \frac{a}{bz_1 + cz_2 + df}\mathbf{z}, \tag{5}$$

where $a \triangleq \begin{vmatrix} y_1^0 & y_1^1 & y_1^2 \\ y_2^0 & y_2^1 & y_2^2 \\ y_3^0 & y_3^1 & y_3^2 \end{vmatrix}$, $b \triangleq \begin{vmatrix} 1 & 1 & 1 \\ y_2^0 & y_2^1 & y_2^2 \\ y_3^0 & y_3^1 & y_3^2 \end{vmatrix}$, $c \triangleq -\begin{vmatrix} 1 & 1 & 1 \\ y_1^0 & y_1^1 & y_1^2 \\ y_3^0 & y_3^1 & y_3^2 \end{vmatrix}$, and $d \triangleq \begin{vmatrix} 1 & 1 & 1 \\ y_1^0 & y_1^1 & y_1^2 \\ y_2^0 & y_2^1 & y_2^2 \end{vmatrix}$.

3.3 Recovering the coordinates of the corners on the original document plane

As it was mentioned in Sect. 2, the origin of the coordinates system of the original document plane will be located at a corner of the document, that we will choose, without loss of generality, to be $\mathbf{x}^0$, where $\mathbf{x}$ denotes the estimate of $\bar{\mathbf{x}}$, and its axes will be aligned with the document edges. Therefore, in order to estimate the original document plane coordinates $\mathbf{x}$, we will move the origin of the three-dimensional coordinates system describing $\mathbf{y}$ and then rotate the new coordinates system until achieving a system such that $\mathbf{x}^0 = \mathbf{0}$, and the document edges are aligned with the X and Y axes. Mathematically, we will denote by $\mathbf{u}^i$ the translated version of $\mathbf{y}^i$, so $D_{-\mathbf{y}^0} \cdot \left(\mathbf{y}^i, 1\right)^T = \left(u_1^i, u_2^i, u_3^i, 1\right)^T$, where

$$D_{-\mathbf{y}^0} \triangleq \begin{pmatrix} 1 & 0 & 0 & -y_1^0 \\ 0 & 1 & 0 & -y_2^0 \\ 0 & 0 & 1 & -y_3^0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then, we will compute three rotation matrices, namely $R_X(\alpha)$, $R_Y(\beta)$ and $R_Z(\gamma)$, around the axes X, Y and Z, according to the angles α, β and γ, such that they translate the points $\mathbf{u}^i$ onto the XY plane, and with the edges of the document aligned with the Cartesian axes. These matrices are defined as

$$R_X(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix}, R_Y(\beta) = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}, \text{ and} R_Z(\gamma) = \begin{pmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The first rotation matrix to be used will be $R_Z(\gamma)$. We will denote the point coordinates in the new coordinate system by $\mathbf{v}$, so $\mathbf{v}^i = R_Z(\gamma)\mathbf{u}^i$. Then, $R_Z(\gamma)$ will be calculated so that the rotated version of $\mathbf{u}^1$, i.e. $\mathbf{v}^1$, verifies $v_2^1 = 0$. Therefore, γ will be given by $\gamma = \arcsin\left(\frac{u_2^1}{\sqrt{\left(u_1^1\right)^2 + \left(u_2^1\right)^2}}\right)$, which is always properly defined, except when $u_1^1 = u_2^1 = 0$;* nevertheless, in that case the original point $\mathbf{u}^1$ already lies on the plane $u_2^1 = 0$, so the sought transformation is not required, and $\gamma = 0$.

The next rotation we look for is $R_Y(\beta)$. The point coordinates in the new coordinate system will be denoted by $\mathbf{w}$, so $\mathbf{w}^i = R_Y(\beta)\mathbf{v}^i$. In this case, $R_Y(\beta)$ will be derived in order to yield $w_3^1 = 0$, so β will be computed as $\beta = -\arcsin\left(\frac{v_3^1}{\sqrt{\left(v_1^1\right)^2 + \left(v_3^1\right)^2}}\right)$, which is also properly defined, except for the case $v_1^1 = v_3^1 = 0$; as it happened when defining γ, in this degenerate case this rotation is not necessary, so $\beta = 0$.

The last rotation, $R_X(\alpha)$, will translate the rotated version of $\mathbf{w}^3$ to the XY plane. This implies that α will be computed as $\alpha = \arcsin\left(\frac{w_3^3}{\sqrt{\left(w_2^3\right)^2 + \left(w_3^3\right)^2}}\right)$, where, as it happened in the two previous cases, the angle is properly defined as long as $w_2^3 = w_3^3 = 0$ is not met; if that condition were indeed verified, then this last rotation would not be required, and $\alpha = 0$.

3.4 Recovering the original document plane coordinates of an arbitrary document point

Up to this point, we have only operated over the corners of the document. Nevertheless it is possible to generalize the obtained results to any point belonging to the possibly isotropically scaled version of the document. Indeed, by defining

$$G \triangleq \begin{pmatrix} R_X(\alpha) R_Y(\beta) R_Z(\gamma) & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \cdot D_{-\mathbf{y}_0},$$

*Throughout this paper we will assume that $\arcsin(\cdot) \in [-\pi/2, \pi/2]$, and consequently the related $\cos(\cdot)$ will be non-negative.

if $\mathbf{y}$ belongs to the three-dimensional representation of the document, due to the geometrical properties of the problem and those of the translation and rotation matrices defined above, it is straightforward to see that

$$G \cdot (\mathbf{y}, 1)^T = (x_1, x_2, 0, 1)^T . \tag{6}$$

This last equation, jointly with (2), establish a relationship between any point $\mathbf{z}$ on the captured image and the corresponding point on the original document plane $\mathbf{x}$. Nevertheless, due to the particular structure of matrix G, we show next that the computational cost of the inverse transformation can be reduced. In order to do so, we define

$$E \triangleq G^{-1} = D_{\mathbf{y}^0} \cdot \begin{pmatrix} R_Z(-\gamma) R_Y(-\beta) R_X(-\alpha) & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix},$$

where G is non-singular, as $|G| = 1$. Taking into account that the third component of the rightmost term of (6) is null, one can write

$$\lambda \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} e_{11} & e_{12} & e_{14} \\ e_{21} & e_{22} & e_{24} \\ e_{31} & e_{32} & e_{34} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix},$$

so defining

$$E_R \triangleq \begin{pmatrix} e_{11} & e_{12} & e_{14} \\ e_{21} & e_{22} & e_{24} \\ e_{31} & e_{32} & e_{34} \end{pmatrix},$$

the original plane coordinates of an arbitrary document point will verify

$$(x_1, x_2, 1)^T = \lambda E_R^{-1} (z_1, z_2, z_3)^T , \tag{7}$$

as long as E_R is non-singular. For studying the invertibility of E_R, its determinant is calculated, yielding

$$\eta \triangleq |E_R| = \left(z_1^0 \sin(\gamma) - z_2^0 \cos(\gamma)\right) \sin(\alpha) + \left[z_3^0 \cos(\beta) + \left(z_1^0 \cos(\gamma) + z_2^0 \sin(\gamma)\right) \sin(\beta)\right] \cos(\alpha) .$$

Hence, defining

$$\mathbf{s}^i \triangleq R_X(\alpha) R_Y(\beta) R_Z(\gamma) \mathbf{y}^i, \tag{8}$$

for $i = 0, \cdots, 3$, it is easy to see that $\eta = s_3^0$, so E_R will be non-singular if and only if the non-translated but rotated version of $\mathbf{y}^0$ has a non-null third component. In order to see in which scenarios this condition is verified, we will take into account that $\mathbf{x}^i = \mathbf{s}^i - R_X(\alpha) R_Y(\beta) R_Z(\gamma) \mathbf{y}^0$, where $i = 0, \cdots, 3$, so it is easy to conclude that both versions of the corners of the document ($\mathbf{x}^i$ and $\mathbf{s}^i$), will hold the same geometrical relationships between them. Specifically, the $\mathbf{s}^i$'s are the vertices of a shifted version of the recovered document, so they define a plane. Furthermore, since α, β and γ were chosen to produce $\mathbf{x}^i$'s verifiying $x_3^i = 0$, we have that $\eta = s_3^0 = s_3^1 = s_3^2 = s_3^3$. Therefore, given that the rotations defined by (8) do not modify the geometrical structure of the problem described by the $\mathbf{y}$ coordinates system, the above equation implies that E_R will be non-invertible just when the camera center and the four corners of the three-dimensional representation of the document $\mathbf{y}^i$ are coplanar, i.e. when one takes the picture of the document sideways.

Assuming that the degenerate case explained above is not verified, so $\eta \neq 0$, and taking into account that $z_3 = f$, (7) yields

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \left[\begin{pmatrix} (E_R^{-1})_{11} & (E_R^{-1})_{12} \\ (E_R^{-1})_{21} & (E_R^{-1})_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + f \begin{pmatrix} (E_R^{-1})_{13} \\ (E_R^{-1})_{23} \end{pmatrix} \right],$$

with $(E_R^{-1})_{ij}$ the (i,j) element of E_R^{-1}, so a simple relationship between $\mathbf{z}$ and $\mathbf{x}$ is obtained. The required elements of $(E_R)^{-1}$ can be computed using the following formulas

$$\begin{aligned}
(E_R^{-1})_{11} &= \frac{1}{\eta} \left[z_3^0 \cos(\alpha) \cos(\gamma) + \left(z_3^0 \sin(\beta) \sin(\gamma) - z_2^0 \cos(\beta)\right) \sin(\alpha) \right], \\
(E_R^{-1})_{12} &= \frac{1}{\eta} \left[z_1^0 \cos(\beta) \sin(\alpha) - z_3^0 \cos(\gamma) \sin(\alpha) \sin(\beta) + z_3^0 \cos(\alpha) \sin(\gamma) \right], \\
(E_R^{-1})_{13} &= \frac{1}{\eta} \left[\left(z_2^0 \cos(\gamma) - z_1^0 \sin(\gamma)\right) \sin(\alpha) \sin(\beta) - \left(z_1^0 \cos(\gamma) + z_2^0 \sin(\gamma)\right) \cos(\alpha) \right],
\end{aligned}$$

$$
\begin{aligned}
(E_R^{-1})_{21} &= \frac{1}{\eta}\left[z_2^0 \sin(\beta) + z_3^0 \cos(\beta)\sin(\gamma)\right], \\
(E_R^{-1})_{22} &= \frac{1}{\eta}\left[z_3^0 \cos(\beta)\cos(\gamma) + z_1^0 \sin(\beta)\right], \text{ and} \\
(E_R^{-1})_{23} &= \frac{1}{\eta}\left[\left(z_1^0 \sin(\gamma) - z_2^0 \cos(\gamma)\right)\cos(\beta)\right].
\end{aligned}
$$

3.5 Computational cost

Once the proposed perspective distortion correction method has been introduced, its computational cost will be discussed. For the sake of comparison we will consider the scheme by Guillou *et al.*,[9] discused in Sect. 1. This method provides exactly the same results that the algorithm proposed in this paper, although following a different approach. The number of operations required for performing the perspective distortion correction with both methods can be found in Table 1. One can see that both methods have a similar computational cost; specifically, in view of those results, our method requires the evaluation of two more square roots than the Guillou *et al.*'s approach[9] for those terms that are computed once, but the number of sums, multiplications and divisions is reduced by 45, 49 and 8, respectively, using our proposed method. Finally, it must be emphasized that the number of operations required for performing the distortion correction of each single pixel is the same for both cases.

Table 1. Required number of operations for each considered implementation (S = Number of sums/substractions, P = Number of multiplications, C = Number of divisions, R = Number of square roots).

Method	Ops. performed once				Ops. per pixel			
	S	P	C	R	S	P	C	R
Current approach	66	91	7	7	6	8	1	0
Guillou *et al.*'s approach[9]	111	140	15	5	6	8	1	0

4. EXPERIMENTAL RESULTS

As it was discussed in Sect. 1, most of existing perspective distortion correction methods are not applicable to the scenario considered in this paper. Therefore, taking into account that the target of the proposed method is to provide a functionality as similar as possible to that of a scanner, for comparison purposes we will check the results obtained by the method introduced in this work with those obtained by several scanner models. Nevertheless, a lot of nuisance factors, that is, those that are not directly related to the perspective distortion correction but are generally due to the quality of the capture device, are involved in the quality of the recovered document (e.g. non-uniform lighting, blur effect, resolution, etc.). Given that the scanners are expected to be clearly superior to mobile device cameras with respect to those factors, the feature chosen for comparing the results of the proposed scheme with the captures obtained by the considered scanners should desirably be independent of those nuisance factors. An example of such a feature, which will be used in the remainder of this work, is the the aspect ratio of the recovered document r, which is here taken to be always greater or equal than 1, i.e., $r = \max(r_0, 1/r_0)$, where

$$
r_0 \triangleq \frac{\left|\left|\mathbf{x}^1 - \mathbf{x}^0\right|\right| + \left|\left|\mathbf{x}^3 - \mathbf{x}^2\right|\right|}{||\mathbf{x}^0 - \mathbf{x}^3|| + ||\mathbf{x}^2 - \mathbf{x}^1||}.
$$

It is also worth pointing out that we are assuming that *a priori* information about the aspect ratio of the original document is not available neither for the scanners nor for the proposed scheme, providing a realistic framework.

4.1 Experimental Framework

- Document formats: Four different blank documents were considered in the developed tests, with respective sizes 216×279 mm (Letter, $r = 31/24$), 210×297 mm (DIN-A4, $r = \sqrt{2}$), 148×210 mm (DIN-A5, $r = \sqrt{2}$), and 100×100 mm ($r = 1$).

- Mobile capture device: The document photos were taken with a 5 megapixels camera, corresponding to a Nokia N82 mobile phone, being the corners of the captured document located by using a method proposed by us elsewhere.[14] The results reported in this section for the method proposed in this work were obtained using 6 letter-sized, 19 DIN-A4, 14 DIN-A5, and 14 100×100 mm document photos, captured with different camera locations and orientations.

- Scanners: The considered documents were also captured using seven different scanners: the automatic document feeding scanner Fujistu Fi-5210c, and the flat-bed scanners with manual document feeding Canon MP-550, EPSON Perfection 1200 Photo, EPSON Stylus SX-105, HP ScanJet 4470C and HP ScanJet 5100C. All of them were fed with all the considered blank documents. In order to evaluate both the mean value and the variability of the obtained results, each document was scanned 5 times, changing the position of the document for each scan.

Given that the distortion perspective correction method proposed in this work estimates the coordinates of the document corners, the derivation of the aspect ratio obtained when applying that method is straightforward. This is no longer the case when we consider scanned documents, so we will introduce here the method followed to estimate the document corners in that scenario. First of all, in order to avoid undesiderable shadows in the capture, all the documents digitized with the flat-bed scanners were scanned over a black background. Secondly, given that the scanned documents are blank, Otsu's method[15] can be used to estimate a proper threshold value for binarizing the scanners' output. Following the binarization, several equidistant points belonging to each edge of the document are obtained; this is done by manually providing the horizontal and vertical coordinates of the first and last point considered along each edge, as the corners of the scanned document are still not known. Specifically, 10 points per edge are considered for the 100×100 mm documents, and a number proportional to the edges lengths for the other formats. In order to accurately estimate the document edges, a robust regression algorithm was used.[†] The estimated document corners are just the intersection of those edges, yielding the resulting aspect ratio. Unfortunately, this method can not be applied to those documents scanned using the considered automatic feeding scanner, as in that case one can not establish a background. In order to circumvent this problem, each edge of the digitized version of the documents was estimated by manually locating two points of that edge.

4.2 Obtained results

For comparision purposes, Table 2 shows the aspect ratio mean squared error obtained by using our method, as well as by using different scanner models. As it can be seen, the mean squared error values obtained by our method are larger than those of most scanners. One reason explaining this effect could be that document photos were taken in a tougher and much less controlled scenario than that of the scanned documents; for instance, when a document is photographed with the mentioned mobile phone it might be bent (in the case of flat-bed scanners this effect is minimized due to the lid pressure). Fig. 2 shows the original capture of an A4-sized document and its rectified version.

Table 2. Mean square error of the recovered aspect ratio obtained by the different capture devices.

Capture Device	Letter	DIN-A4	DIN-A5	100×100 mm
Canon MP-550	$1.7678 \cdot 10^{-6}$	$5.4846 \cdot 10^{-5}$	$1.9146 \cdot 10^{-8}$	$1.0538 \cdot 10^{-7}$
EPSON Perfection 1200 Photo	$2.4858 \cdot 10^{-5}$	$3.2218 \cdot 10^{-4}$	$1.8975 \cdot 10^{-6}$	$2.4970 \cdot 10^{-6}$
EPSON Stylus SX-105	$4.8041 \cdot 10^{-6}$	$2.6047 \cdot 10^{-5}$	$2.9922 \cdot 10^{-7}$	$3.5861 \cdot 10^{-6}$
HP ScanJet 4470C	$9.5785 \cdot 10^{-7}$	$4.1638 \cdot 10^{-6}$	$1.2219 \cdot 10^{-6}$	$2.3761 \cdot 10^{-6}$
HP ScanJet 5100C	$1.4971 \cdot 10^{-6}$	$1.5966 \cdot 10^{-5}$	$2.2343 \cdot 10^{-9}$	$2.2726 \cdot 10^{-7}$
Fujitsu Fi-5210c	$1.4139 \cdot 10^{-9}$	$7.0777 \cdot 10^{-5}$	$1.9803 \cdot 10^{-6}$	$6.7078 \cdot 10^{-7}$
N82 + Proposed method	$\mathbf{4.8243 \cdot 10^{-5}}$	$\mathbf{1.1307 \cdot 10^{-4}}$	$\mathbf{3.5102 \cdot 10^{-4}}$	$\mathbf{1.1238 \cdot 10^{-3}}$

[†]Concretely, the algorithm based on iteratively reweighted least squares with a bisquare weighting function provided by the Matlab Statistics Toolbox was used.

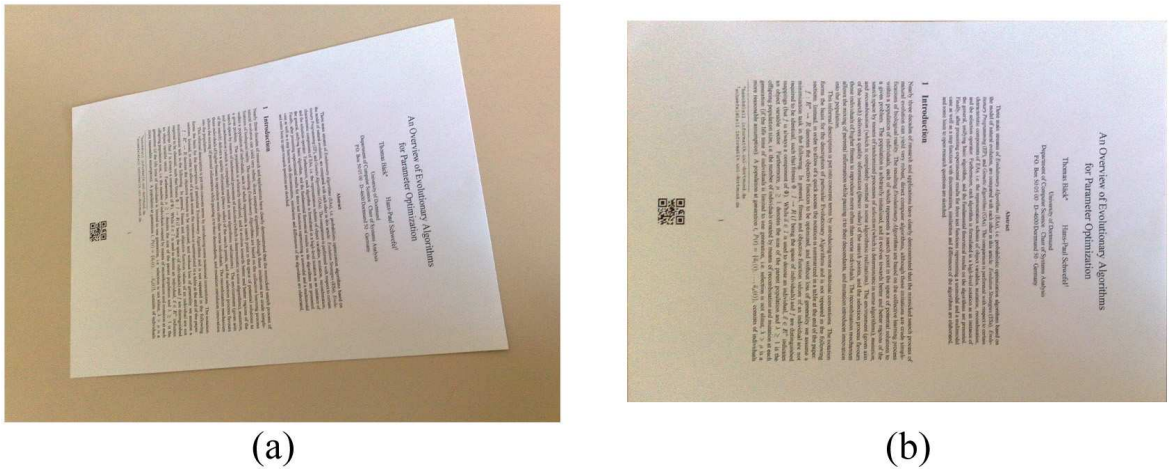

Figure 2. *(a)* Capture of a letter-sized document. *(b)* Rectified version of the document contained in the capture. The recovered aspect ratio is $\frac{1709}{1204} \approx 1.4194$, while the actual value is $\sqrt{2} \approx 1.4142$.

It must be also mentioned that the proposed scheme is not able to correct some captures. Besides the cases where the vertices estimation method is not able to find the right coordinate values, the photos where at least two edges of the captured document are parallel are not properly corrected, as it was theoretically predicted in Sect. 3.1. Indeed, even those cases where two edges of the document are nearly parallel (i.e., where the slopes of those edges are very similar) will be difficult to deal with due to numerical problems. A sample of those captures, as well as the recovered document, is shown in Fig. 3; one can see that two of the edges of the A5-sized captured document are nearly parallel, yielding an overestimated value of the focal length, and recovering a 1.3239 aspect ratio.

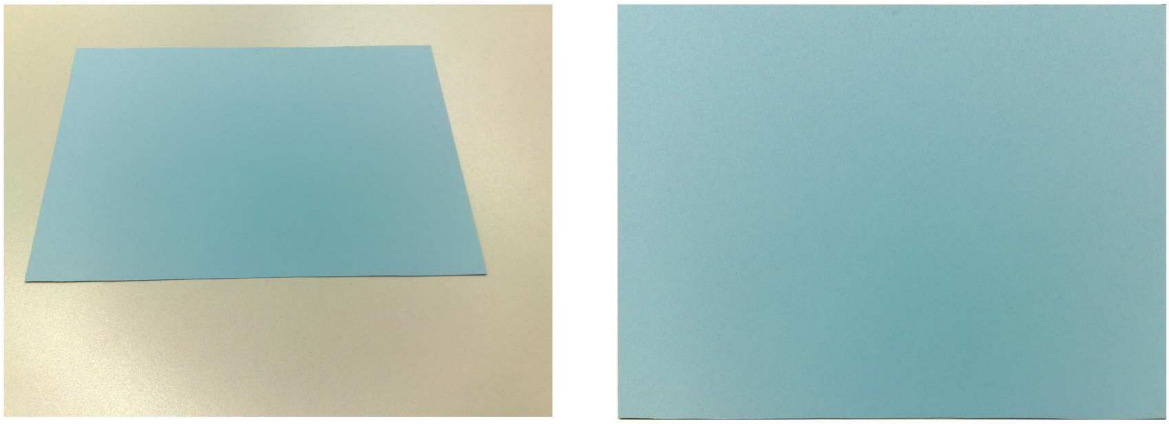

Figure 3. *(a)* Capture of an A5-sized document with two edges nearly parallel. *(b)* Rectified version of the document contained in the capture. The recovered aspect ratio is $\frac{421}{318} \approx 1.3239$, while the actual value is $\sqrt{2} \approx 1.4142$.

5. CONCLUSIONS

A perspective distortion correction method for pictures of rectangular documents based on a geometrical reasoning has been presented. Experimental results show the good performance of the proposed scheme, although scanning the document is usually better; specifically, the mean squared error of the obtained aspect ratio is lower

in the latter case. On the other hand, numerical problems could arise when dealing with documents for which at least two of their edges are (nearly) parallel in the picture.

ACKNOWLEDGMENTS

This work was partially supported by Xunta de Galicia under Projects 09TIC006E (A firma dixital impresa), 08TIC057E (SCANPHONE), 2010/85 (Consolidation of Research Units), by the Spanish Ministry of Science and Innovation under project COMONSENS (ref. CSD2008-00010) of the CONSOLIDER-INGENIO 2010 Program, and by the Iberdrola Foundation through the Prince of Asturias Endowed Chair in Information Science and Related Technologies.

REFERENCES

[1] Clark, P. and Mirmehdi, M., "Rectifying perspective views of text in 3D scenes using vanishing points," *Pattern Recognition* **36**, 2673–2686 (November 2003).

[2] Lu, S., Chen, B. M., and Ko, C. C., "Perspective rectification of document images using fuzzy set and morphological operations," *Image and Vision Computing* **23**, 541–553 (May 2005).

[3] Lu, S. and Tan, C. L., "The restoration of camera documents through image segmentation," in [*Proceedings of the seventh International Association of Pattern Recognition Workshop on Document Analysis Systems*], 484–495 (February 2006).

[4] Miao, L. and Peng, S., "Perspective rectification of document images based on morphology," in [*Proceedings of the International Conference on Computational Intelligence and Security*], **2**, 1805–1808 (November 2006).

[5] Yin, X.-C., Sun, J., Fujii, Y., Fujimoto, K., and Naoi, S., "Perspective rectification for mobile phone camera-based documents using a hybrid approach to vanishing point detection," in [*Proceedings of the Second Internartional Workshop on Camera-Based Document Analysis and Recognition*], 37–44 (September 2007).

[6] Yin, X.-C., Sun, J., Naoi, S., Fujimoto, K., Takebe, H., Fujii, Y., and Kurokawa, K., "A multi-stage strategy to perspective rectification for mobile phone camera-based document images," in [*Proceedings of the Ninth International Conference on Document Analysis and Recognition*], **2**, 574–478 (September 2007).

[7] Iwamura, M., Niwa, R., Kise, K., Uchida, S., and Omachi, S., "Rectifying perspective distortion into affine distortion using variants and invariants," in [*Proceedings of the Second International Workshop on Camera-Based Document Analysis and Recognition*], 138–145 (September 2007).

[8] Liebowitz, D. and Zisserman, A., "Metric rectification for perspective images of planes," in [*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*], 482–288 (June 1998).

[9] Guillou, E., Meneveaux, D., Maisel, E., and Bouatouch, K., "Using vanishing points for camera calibration and coarse 3D reconstruction from a single image," *The Visual Computer* **16**, 396–410 (November 2000).

[10] Kofler, C., Keysers, D., Koetsier, A., and Breuel, T. M., "Gestural interaction for an automatic document capture system," in [*Proceedings of the Second International Workshop on Camera-Based Document Analysis and Recognition*], 161–167 (September 2007).

[11] Zhang, Z. and He, L.-W., "Whiteboard scanning and image enhancement," *Digital Signal Processing* **17**, 414–432 (March 2007).

[12] Hartley, R. and Zisserman, A., [*Multiple view geometry in computer vision*], Cambridge University Press, 2nd ed. (April 2004). ISBN 0521-54051-8.

[13] Weisstein, E. W., "Line-plane intersection." From MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/Line-PlaneIntersection.html`.

[14] Rodríguez-Piñeiro, J., Comesaña-Alfaro, P., Pérez-González, F., and Malvido-García, A., "A new method for boundary estimation of document images," In preparation.

[15] Ostu, N., "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics* **9**, 62 –66 (January 1979).