Deskewing Perspectively Distorted Documents: An Approach Based on Perceptual Organization

Maurizio Pilu
Hardcopy Technology Laboratory
HP Laboratories Bristol
HPL-2001-100
May 2nd, 2001*

email: maurizio_pilu@hp.com

document capture, perceptual organization

This work deals with the recovery of illusory linear clues from perspectively skewed documents with the purpose of using them for rectification. The computational approach proposed implements the perceptual organization principles implicitly used in textual layouts. The numerous examples provided show that the method is robust and viewpoint and scale invariant.

* Internal Accession Date Only Approved for External Publication

Deskewing perspectively distorted documents: An approach based on perceptual organization

Maurizio Pilu
Hewlett-Packard Laboratories
Bristol, BS34 8QZ, UK
maurizio_pilu@hp.com

Abstract

This work deals with the recovery of illusory linear clues from perspectively skewed documents with the purpose of using them for rectification. The computational approach proposed implements the perceptual organization principles implicitly used in textual layouts. The numerous examples provided show that the method is robust and viewpoint and scale invariant.

1. Introduction

As sensor resolution increases and prices drop, cameras will one day be used to capture documents in lieu of flatbed scanners. One of the main disadvantages when capturing a document with a camera is that the non-contact image capture process causes geometric distortions dependent upon the camera orientation, in particular perspective skew, as typified in Figure 1-left.

This work addresses a fundamental problem that need to be faced when trying to passively deskew a captured document, namely the detection of linear clues that can be used as geometric primitives to determine the document plane orientation w.r.t. the camera and the rectifying homography.

Figure 1-left shows several kinds of linear clues that may arise in practice. Clue A is a vertical *illusory* [2] clue. It does not correspond to an actual linear feature but rather to a set of organized features arranged linearly. Clue B is a vertical *hard line* which is the projection of the actual document edge. Clues C are horizontal illusory lines, which have been inferred from the arrangement of characters into text lines. Similarly to clue B, clue D is a horizontal hard line. Clue E

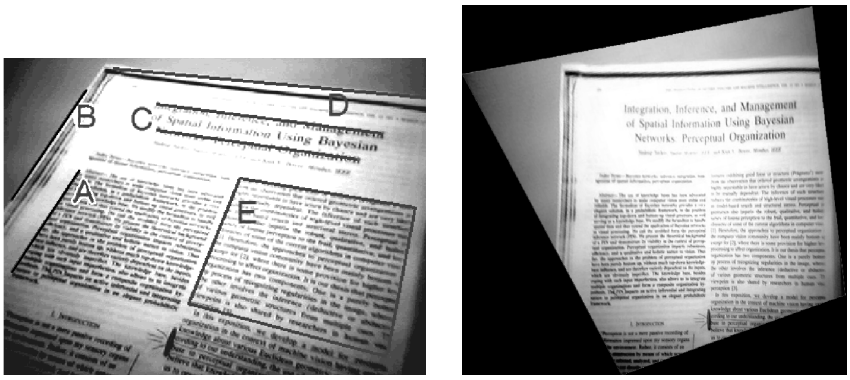

Figure 1: Left: Illustration of perspective deskew in documents and the linear clues. Right: the desired output that can be obtained with a rectifying homography using vertical and linear clues.

is a quadrilateral that can be either illusory or corresponding to an actual rectangular outline in the document (e.g. a figure box or the four document boundaries). Hard edges of the type B and D can be detected rather trivially in many ways using edge detectors (e.g. the Hough transform) and will not be addressed as such in this paper. Illusory edges such as C and A are rather difficult to find reliably in practice and most of the literature on document analysis has been only focusing on recovering groups of parallel text lines to compensate for rotation for OCR and other scanning applications.

In this paper we will be dealing with the problem of finding illusory clues of the type A and C without assumptions on the type of document, fonts or camera rotation and orientation other than the presence of some organized text.

Text had been designed long before perception studies but with the unspoken principles of perceptual saliency (see e.g. [2]) in mind. Given any organised text (even foreign and unintelligible to us) words, lines and paragraph structures pop out preattentively. The approach presented is strongly based on a computational implementation of these perceptual organization principles and will be shown to be robust, fast and general.

2. Related works

The subject of correcting perspective skew in documents has been largely neglected in the literature. Even recent works explicitly addressing camera-based document imaging such as [14] treat only rotation-induced skew.

The geometry of the rectification is well known [6] but the problem of passively and robustly detecting the geometric features needed is still open.

A substantial body of research has been dedicated to text and page segmentation in document images, but any distortion considered is again only rotation-induced. The main bottom-up methods used include many variations on projection profiles approaches, Hough-inspired techniques [12] and nearest-neighbour clustering [9][13]. Top-down methods seek to extract the high-level structure of the images, such as by using Manhattan layout analysis of the (white) background [1]. Methods that analyse connected components as we do are fairly common in document analysis. Besides the aforementioned [9][13] and others, an interesting approach that employs perceptual organization principles is [11], although assuming of parallel-lines. All the methods above assume that the text lines are still parallel in the image and could not work with perspective skew. The work in [3] is one of the extremely few works that tries to extract text from perspectively skewed documents but, it does so by extracting the document quadrilateral, which we do not assume it is visible. Works dealing with the recovery of vanishing points from images are numerous, some from edges (e.g. [4] recover them from two mutually orthogonal directions on a plane) and some others from texture and other soft clues [10]. However, both edges and regular textures are infrequent in text documents and all these methods cannot be easily extended to our situation.

3. Extraction of illusory horizontal lines

Horizontal illusory clues originates from the arrangements of characters into words and lines.

When a document is captured by a camera at an unknown angle, it is of course impossible to establish *a priori* what is horizontal and what is vertical. However, given the usual layout of western-style writing, the horizontal direction is reflected in the image in the *dominant* direction of illusory lines. Henceforth we shall indicate with *horizontal* the clues belonging to this dominant direction.

The algorithm proposed in this paper to extract the horizontal illusory lines is summarised as follows. A preprocessing stage binarizes the input image, turning it into blobs representing either single characters or (portion of) words or lines, depending upon the font size and the resolution considered. These blobs are divided into elongated (major axis longer than thrice the minor axis) or compact. A pairwise saliency measure is computed for pairs of neighbouring blobs that represent how likely they are to be part of a text line. An network is then built using the blobs and their associations. The network is then transversed to extract salient linear groups of blobs which constitute the illusory horizontal clues. Isolate elongate blobs are also considered as individual clues.

In the following sections we shall describe these stages is more detail.

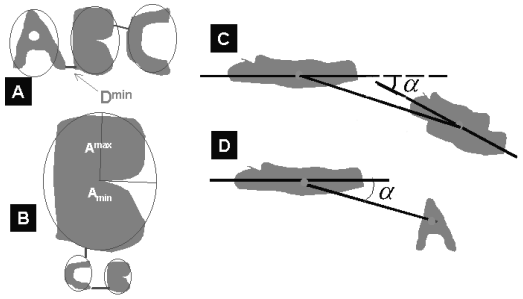

Figure 2: Illustration of the quantities used for the determination of the pairwise blob saliency.

3.1 Preprocessing of image

The input image is first down sampled to a resolution in the order of 320x200 pixels. The method is based on using a binarized version of the document, but in a typical situation, the camera is at an angle with respect to the document page and a large illumination gradient is present across the paper which may make effective thresholding difficult. Thus, we first remove the low-frequency variations (corresponding to the illumination gradient), then determine global thresholds in portions of the image and interpolate the thresholds for intermediate values. The binary image in Figure 5-A shows the result obtained (and equally well for the other examples shown).

3.2 Blobs pairwise saliency measure

The association network cast onto the image blobs is built calculating a pairwise saliency metric between a blob and its neighbours laying within a certain distance from it. Blobs can be elongated or compact and the saliency measure is tailored to their different perceptual significance.

- *Pair of compact blobs.* With reference to Figure 2-A and B we define these two saliency measures, the *relative minimum distance RMD* and the *blob dimension ratio BDR* as

$$RMD = \frac{2D_{ijMIN}}{(A_{iMIN} + A_{jMIN})}$$

and

$$BDR = \frac{A_{iMIN} + A_{iMAX}}{A_{jMIN} + A_{jMAX}}$$

where D_{ijMIN} is the minimum distance between blobs B_i and B_j, and A_{kMIN} and A_{kMAX} are the minor and major axes of blob B_k. These measures encapsulate two important saliency principles, that is the *proximity* principle and the *similarity* principle [2]: the RMD indicates how close two blobs are w.r.t. their relative dimensions, whereas the BDR will be the closer to 1 the more similar in size the two blobs are, whatever their actual dimension. Note that both these measures are scale-independent and parameter free. In order to combine these two separate saliency measures into one, we have aproximated the distribution of BDRs and RMDs across documents as two independent Gaussian distributions and define the single saliency measure as a product of the two:

$$P_{Cij} = N(BDR_{ij}, 1, 2) \cdot N(RMD_{ij}, 0, 4)$$

where $N(x, \mu_x, \sigma_x)$ is a Gaussian distribution of x with mean μ and standard deviation σ. The values of μ_{BDR}, σ_{BDR}, μ_{RMD}, σ_{RMD} have been determined experimentally from a wide range of situations and characters at different scales.

- *One or two elongated blobs.* If one blob is (or both are) elongated, the perceptual relevance of it ought to be taken into account since it (them) might have originated from one or more words that at that particular resolution could not be separated out into characters. With reference to Figure 2-C and D, a first pairwise saliency measure α_{ij} should reflect the fact that if we deal with two elongated blobs (case C) their major axis should be aligned if they belong to the same text line, and if we deal with just one elongated blob (case D) the centre of the compact one should roughly lay on the major axis of the elongated one. Their relative distance is still important too and should be considered as in the previous case. Hence, analogously with the case of compact blobs, we define the overall pairwise saliency measure as:

$$P_{Eij} = max\left(N(BDR_{ij}, 1, 2) \cdot N(\alpha ij, 0, 5^o), P_{Cij}\right)$$

The choice between the proper distribution for elongated blobs (first argument of the max operator) and the P_{Cij} as previously defined for compact blobs is dictated by the fact that if the two blobs score very well in terms of BDR and RMD, we should keep considering the pair potentially salient pending a decision at further processing stages.

Hence, with these definitions, the pairwise saliency is a probability and is thus within the $[0, 1]$ range.

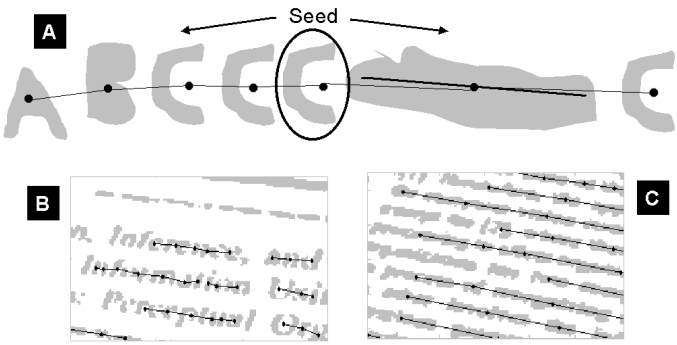

Figure 3: A: Growing a curvilinear group from a seed. B: Groups of compact blobs. C: Groups of elongated blobs.

3.3 Determination of curvilinear groups

Once we have determined the saliency measures we can cast an association network over the image where the nodes are the blob centres and the arcs are the associations between pairs and their probability (saliency) as computed in the previous section. An example of this network is shown in Figure 5-B. The problem of finding the salient linear groups of blobs can be seen as an application of the *good continuation* principle [2], which in our context translates into finding paths across the network that have both high saliency and curvilinearity. In order to do so we have followed a greedy path growing approach with random starting seeds. With reference to Figure 3-A, starting from a random seed blob, we start expanding in the direction with the highest saliency and keep growing the group until no arc that is roughly aligned with the preceding ones is found with an acceptable (> 0.6) saliency probability. When we stop, we go back to the seed and start growing in the opposite direction, thereby trying to complete the linear group. We keep repeating with different seeds, taking care of not using blobs that have been previously associated to other groups.
Figures 3-B and 3-C show linear salient groups formed with this process with compact and mixed compact/elongated blobs, respectively, reflecting regions of the 18points title and the paragraph text of the document of Figure 1-left.
Given the typical abundance of text lines in a document, a large number of linear groups are expected to be found, including occasional wrong ones at this early stage. However the robustness of algorithm is rooted in the sound perceptual organization principles used and even when the text lines are scarce and unpredictable such as the example of Figure 5-D, the algorithm manages to find enough correct groups to apply the partial deskew illustrated later. As anticipated earlier, the method does not make any assumptions about the rotation of the document as seen by the camera; in fact, the case in Figure 6-B shows a document not only perspec-

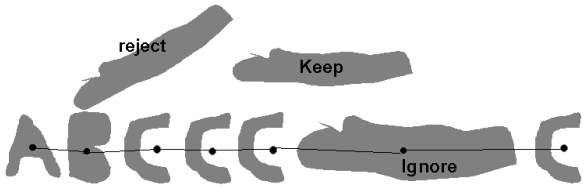

Figure 4: Perceptual rule for treating rogue elongated blobs; provided they are not inconsistent with a strong linear group or part of it, they are treated as individual linear clues.

tively skewed but also rotated by about 90^o that has been correctly rectified using the groups found.

3.4 Using rogue elongated blobs

Rogue elongated blobs are perceptually relevant in their own right. In fact, they often occur in isolation and in some cases they can even represent a whole text line. We use a simple policy to treat rogue elongated blobs, which is illustrated in Figure 4. If an elongated blob is close[1] to a longer group and its angle of incidence to the group is too high ($> 15^o$) then we treat it as non-salient; otherwise we elect it to lineaar clue as much as the groups found in the previous section.

We shall use, without distinction, both the salient groups and the salient elongated blobs as linear clues that will be used to fit the horizontal line bundle as described in the next section.

4 Partial Rectification

Once we have the pool of horizontal clues, we can perform a partial rectification (or deskew) of the document.

The first step is to fit lines to each horizontal clue, which we perform with straightforward linear regression. Figure 5-C shows an example of the fitted lines now representing the linear clues.

The second step is to find the vanishing point in the image. There are several techniques in the literature but most of them would be impractical here due to the relatively imprecise linear clues. We found that excellent results are obtained using a RANSAC approach [5] in the linear clue feature space to explicitly fit a line bundle of the form $y - v_y = m(x - v_x)$. In summary, two horizontal clues that are relatively separate are first picked at random from the set with a probability

[1] We chose less than twice the minimum distance between its minor axis length and the average thickness the blobs belonging to the group.

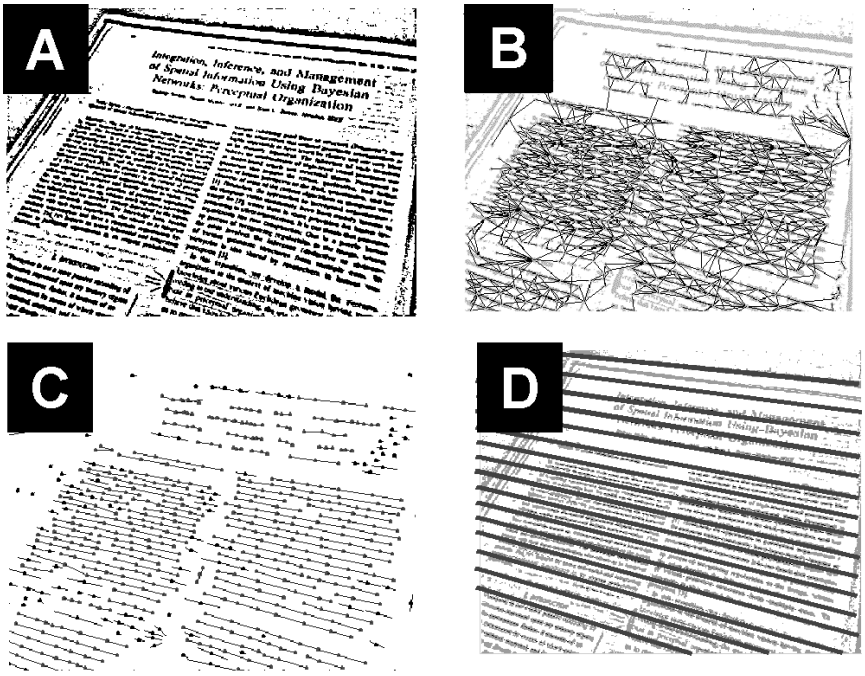

Figure 5: A: The binarized image; B: the association network; C: the curvilinear groups extracted; D: The fitted line bundle.

related to their length, and then used to fit the line bundle equation above to yield an estimate of v_x and v_y (the vanishing point); the residual is analysed to determine the amount of inliers amongst the set. We repeat this operation a number of times [5] and select the one with the largest number of inliers; finally, a final re-fitting with all the inliers is performed to produce a more accurate estimate.

In the absence of vertical clues, full rectification is clearly not possible and the best we can do is to find a homography in the image plane [7] that makes all the bundle lines horizontal. In a recent paper on projective rectification, Hartley [8] argued that amongst infinitely many (∞^6) plane homographies that would perform this transformation, good results are achieved when the homography closest to an Euclidean transformation is chosen.

Let v_x and v_y be the coordinates of the bundle centre expressed in the image reference system and let H and W be the height and width in pixels of the sensor array. First the bundle is translated by $[t_x = -W/2, t_y = -H/2]$ and then rotated by an angle $\theta = \arctan(\frac{v_y}{v_x})$ so as to make the new x coordinate of the bundle centre lay on the x axis. Next, a transform that "throws" the new bundle centre $[(v_x - t_x), 0]$ to infinity ($[-\infty, 0]$) is applied that effectively makes lines belonging to the bundle horizontal. Finally the translation is reverted to put the bundle back to its position. The overall homography comprising these transformations is then

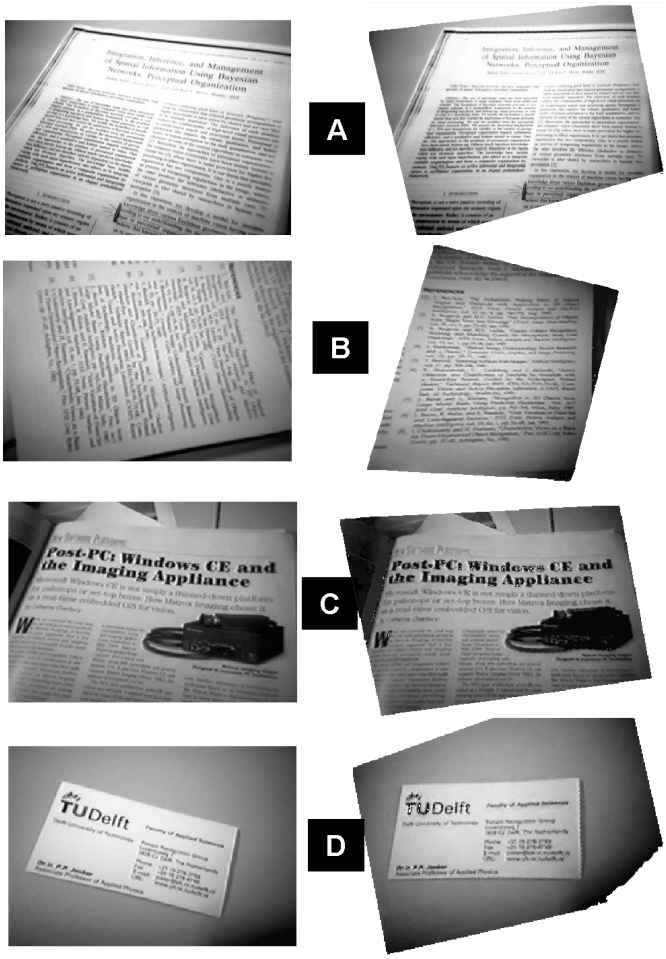

Figure 6: Four partial deskewing results using the horizontal bundle fitted to the detected illusory linear clues.

$F = T^{-1}KRT$, where

$$T = \begin{bmatrix} 1 & 0 & -t_x \\ 0 & 1 & -t_y \\ 0 & 0 & 1 \end{bmatrix} \quad R = \begin{bmatrix} cos(\theta) & sin(\theta) & 0 \\ -sin(\theta) & cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{-1}{(v_x - t_x)} & 0 & 1 \end{bmatrix}$$

are, respectively, the translation, the rotation and rectification transforms in projective coordinates.

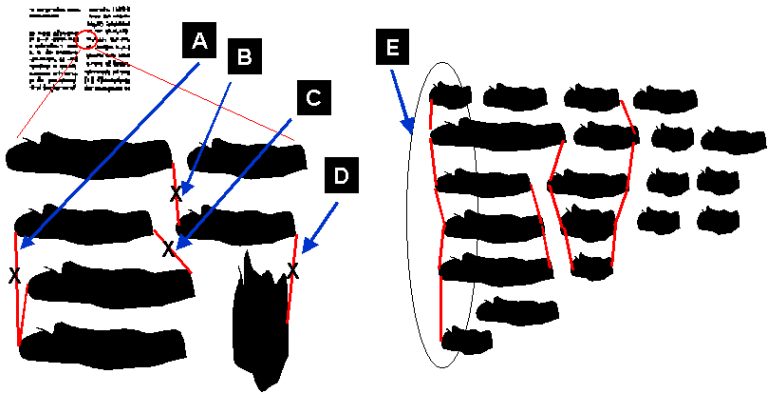

Figure 7: Illustration of the rules for pruning vertical associations.

Figure 6 shows four results of the application of this partial deskewing method using the bundles fitted to the linear clues detected as per Section 3.

This kind of rectification process makes text lines horizontal (which also implicitly corrects for rotational skew) but cannot compensate for other kinds of image distortions that do not affect the ”horizontality” of lines, such as shear and keystone. The knowledge of vertical clues is essential for removing this residual distortion.

5 Extraction of illusory vertical clues

In many situations vertical clues are scarce. Typically they may be a paragraph’s left (and less frequently right) justification, figure boundaries and the page boundary itself. We will not deal with the relatively trivial extraction of true edges (which may or may not be present) but with the more difficult extraction of illusory vertical clues arising from high-level text organization.

After partially deskewing, we made the situation easier, as we now have a coarse knowledge of the vertical direction in the document reference system. This allows us to search for vertical clues in a narrower range, thus avoiding being fooled by other clues.

The approach followed is similar to the one for horizontal clues in that we do use an association network of blobs that is searched for linear groups. However, the way the associations are made is substantially different. In fact, since the vertical associations are rather weak we have noticed that it is best not to potentially further weaken associations based on saliency measures such as those used for the horizontal clues. Rather, we have decided to only reject associations based on near-impossibility and let all the others propagate until further committal.

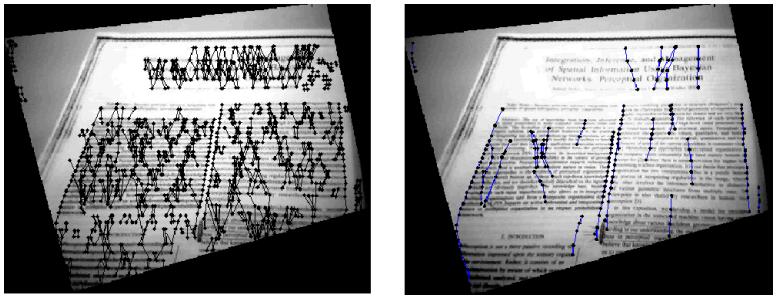

Figure 8: Left: Example of pruned association network; Right: Best illusory vertical clues with a large number of outliers.

Figure 7 summarises the four rejection rules that have been applied to the dense initial association network: (A) rejection of the longest of two overlapping associations; (B) rejection of left-end-to-right-end associations (and vice versa), as they cannot possibly originate from a justified paragraph or any other vertical clue in organized text; (C) rejection of associations at too much of an angle with respect to the vertical direction. (D) rejection of associations of blobs that have two different heights, as they are unlikely to be part of the same paragraph. The application of these pruning criteria dramatically reduce the dense association network, as shown in Figure 8-left. Noticeably, most of the associations belonging to actual illusory vertical clues are preserved, albeit amongst a sea of other meaningless ones. Finally, a greedy split and merge strategy is used to group all these associations into extended near-vertical linear groups as shown in Figure 9-right, a method very similar to that described in Section 3.3. The strongest groups are elected as vertical clues but amongst these incorrect ones always crop up. Figure 9 shows six examples of set of vertical clues detected.

6 Using the vertical clues

With the knowledge of the horizontal vanishing point, and a few correct vertical clues that can be used to determine a second vanishing point, we can perform full rectification in several, equivalent ways, e.g. by warping with a transformation mapping a virtual quadrilateral defined by the two bundles into a rectangle, using the homography calculated with the two vanishing points, or by explicitly using the plane orientation, if we knew the focal length. The result would be an undistorted document like that shown earlier in Figure 1-right.

However, since we may have in general one to four correct illusory vertical clues amongst several candidates, we may not be able to fit the bundle reliably as we did in Section 4 for the abundant horizontal clues.

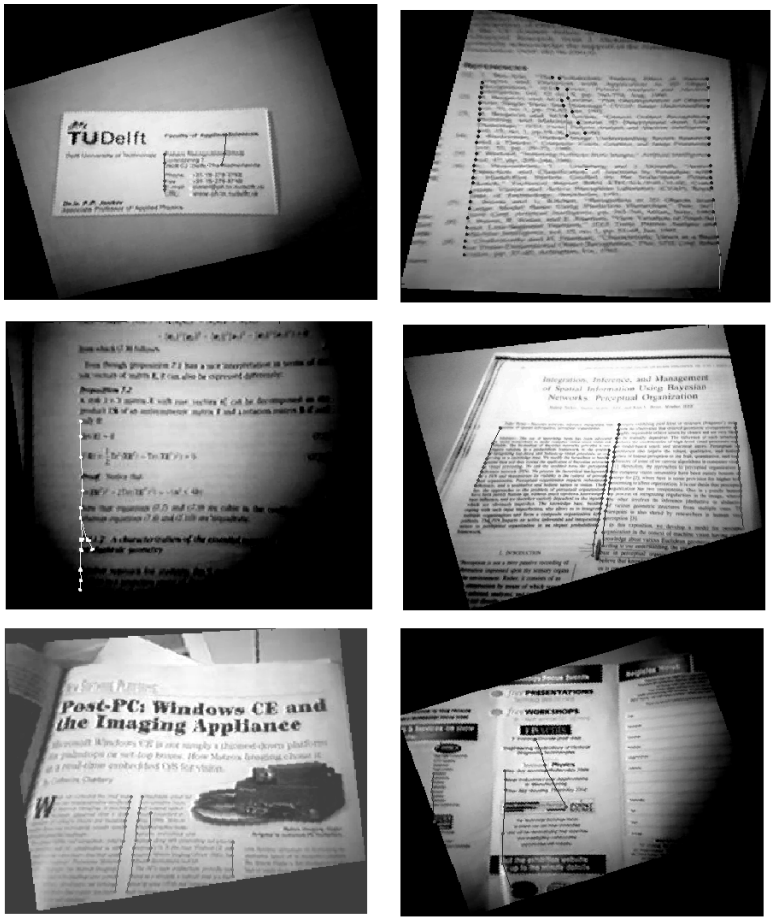

Figure 9: Six examples of vertical clues found.

There are two separate situations. If there are three or more correct vertical clues that fit a bundle, then it is likely that these are indeed the correct ones, although we have not endeavoured into establishing with what probability. If there are only two correct clues detected amongst the candidates set, it is impossible to determine whether they are correct or not. However, if the focal length of the camera is known, and given the fact that we know that the second vanishing point originates from vertical clues perpendicular to the horizontal clues, we can perform the correctness test on just *two* vertical clues [6]. Hence the knowledge of the focal lenght makes it much more likely that we can use the rare vertical clues to perform full rectification.

7. Summary and Conclusions

When capturing a document with a hand-held camera, it common to have an output image that is perspectively distorted. This paper presents a passive way of determining the illusory clues (such as text lines and paragraph margins) that can be used, along with clues such as document edges if available, to rectify the image in order produce an upright, undistorted document that can be used for storage, printing, OCR, page segmentation, etc. We have followed a novel approach completely based on a computational implementation of perceptual organization principles underlying text perception and we have done so using saliency measures and simple geometric reasoning. Experiments, illustrated in the paper, show that method is parameter free, robust and independent from scale and font size.

References

[1] H.S. Baird. Background structure in document images. *International Journal of Pattern Recognition and Artifical Intelligence*, 8(5):1013–1030, 1994.

[2] V. Bruce and P.R. Green. *Visual Perception*. 2nd edition, 1991.

[3] P. Clark and M. Mirmhedi. Location and recovery of text on oriented surfaces. SPIE Conf. on "Electronic Imaging 2000: Document Recognition and Retrieval VII", January 2000.

[4] J.M. Coughlan and A.L Yuille. Manhattan world: Compass direction from single image by Bayesian inference. In *Internation Conference on Computer Vision*, pages 941–947, 1999.

[5] M.A. Fischler and R.C. Bolles. A RANSAC-based approach to model fitting and its application to finding cylinders in range data. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 637–643, 1981.

[6] R.M. Haralick. Monocular vision using inverse perspective projection geometry: Analytic relations. In *CVPR89*, pages 370–378, 1989.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[8] R.I. Hartley. Theory and practice of projective rectification. *International Journal of Computer Vision*, 35(2):1–16, November 1999.

[9] A. Hashizume, P.S. Yeh, and A. Rosenfeld. A method of detecting the orientation of aligned components. *Pattern Recognition Letters*, 4:125–132, 1986.

[10] J.S.Kwon, H.K.Hong, and J.S. Choi. Obtaining a 3D orientation of projective textures using a morphological method. *Pattern Recognition*, (29):725–732, 1996.

[11] S. Messelodi and C.M. Modena. Automatic identification and skew estimation of text lines in real scene images. *Pattern Recognition*, (32):791–810, 1999.

[12] Y. Nakano, Y. Shima, H. Fujisawa, J. Higashino, and M. Fojinawa. An algorithm for the skew normalization of document images. In *ICPR'90*, volume 2, pages 8–13, 1990.

[13] L. O'Gorman. The document spectrum for page layout analysis. *Pattern Analysis and Machine Intelligence*, 15(11):1162–1173, Nov 1993.

[14] M.J. Taylor, A. Zappala, W.M. Newman, and C.R. Dance. Documents through cameras. *Image and Vision Computing*, 17(11):831–844, September 1999.