An Image Based Performance Evaluation Method for Page Dewarping Algorithms using SIFT Features

Syed Saqib Bukhari*, Faisal Shafait† and Thomas M. Breuel*
*Image Understanding and Pattern Recognition (IUPR)
Technical University of Kaiserslautern, Germany
†Multimedia Analysis and Data Mining (MADM)
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
bukhari@informatik.uni-kl.de, faisal.shafait@dfki.de, tmb@informatik.uni-kl.de

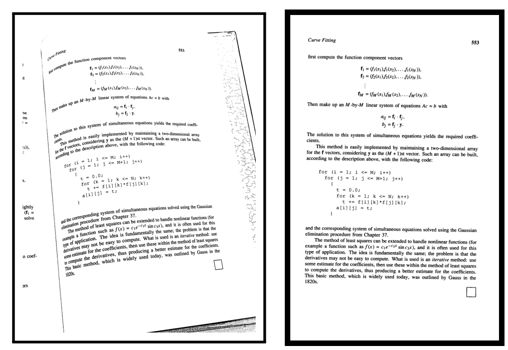
***Abstract*—Dewarping of camera-captured document images is one the important preprocessing steps before feeding them to a document analysis system. Over the last few years, many approaches have been proposed for document image dewarping. Usually optical character recognition (OCR) based and/or feature based approaches are used for the evaluation of dewarping algorithms. OCR based evaluation is a good measure for the performance of a dewarping method on text regions, but it does not measure how well the dewarping algorithm works on the non-text regions like mathematical equations, graphics, or tables. Feature based evaluation methods, on the other hand, do not have this problem, however, they have following limitations: i) a lot of manual assistance is required for ground-truth generation, and ii) evaluation metrics are not sufficient to get meaningful information about dewarping quality. In this paper, we present an image based methodology for the performance evaluation of dewarping algorithms using SIFT features. For ground-truths, our method only requires scanned images of pages which have been captured by a camera. This paper introduces a vectorial performance evaluation score which gives comprehensive information for determining the performance of different dewarping methods. We have tested our performance evaluation methodology on the participating methods of CBDAR 2007 document image dewarping contest and illustrated the correctness of our method.**

***Keywords*-Performance Evaluation, Dewarping, Camera-Captured Document Images, SIFT**

I. Introduction

The goal of page dewarping is to flatten a camera-captured document such that it becomes readable by current OCR systems. Page dewarping has triggered a lot of interest in the scientific community over the last few years and many approaches have been proposed. These dewarping approaches can be broadly divide into two main categories: i) 3-D document shape reconstruction [1], [2], [3] and ii) 2-D image processing (monocular dewarping) [4], [5], [6], [7], [8], [9].

Despite a large number of dewarping techniques, performance evaluation of page dewarping methods is still an unsolved problem. Most of the time it has been done on the basis of visual quality of dewarped images [8], [10], but it is a subjective evaluation and gives no quantitative measure. In order to objectively compare dewarping methods, OCR based [11], [9] and feature based [12] performance evaluation methods have been proposed. OCR based performance evaluation is an indirect method which can only measure the performance of a dewarping method on text regions. Nowadays commercial OCR software can handle degradations in documents to some extend, therefore, OCR based evaluation can not measure how well text elements have been dewarped with respect to their shapes. On the other hand, feature based performance evaluation do not have these problems and can measure the performance of a dewarping method for both text and non-text regions. However, existing feature based performance evaluation methods have following limitations: i) a cumbersome manual marking is required for generating ground-truth data, and ii) a single performance evaluation metric is used which may not be sufficient to compare the performance of different dewarping methods.



(a) camera-captured document (b) scanned document (ground-truth dewarped image)

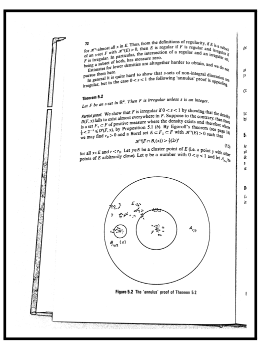
Figure 1. A sample camera-captured document and its correspoding scanned image from DFKI-I dataset. The scanned images in DFKI-I dataset are used here as ground-truth dewarped images.

In this paper, we propose an image based performance evaluation methodology for dewarping methods to overcome the limitations of the existing feature based performance evaluation methods. We use scanned images of pages, that were captured by camera, as ground-truth dewarped images. In this way, no manual efforts are required for generating ground-truth data for a publicly available dataset that contains scanned documents (like DFKI-I [11]), or a very less manual efforts are required for creating a new dataset. For measuring the performance, instead of a single performance evaluation metric, we present a vectorial score that is particularly useful in analyzing the behavior of different page dewarping algorithms. On the basis of SIFT features matching between a dewarped image and its correspoding ground-truth dewarped image, we calculate the percentage and the mean error of matching features.

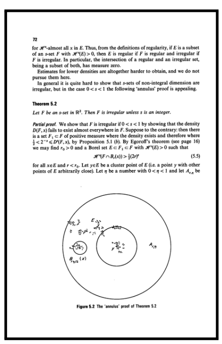
The rest of the paper is organized as follows. We describe the proposed image based performance evaluation in Sections II. Experiments and results are discussed in Section III. Section IV presents our conclusions.

II. Image based Performance Evaluation

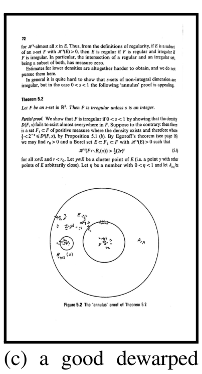
The proposed performance evaluation metrics are described here in detail along with the requirement of ground-truth dewarped images. This section is organized as follows. In Section II-A, we discuss about the ground-truth dewarped images. The performance evaluation metrics using SIFT based matches are explained in Section II-B.



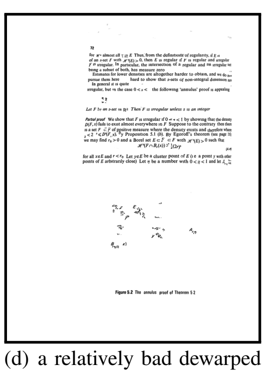
(a) camera-captured document



(b) a ground-truth dewarped image



(c) a good dewarped output



(d) a relatively bad dewarped output

Figure 2. A sample camera-captured document image and its corresponding ground-truth dewarped image and a good and a bad dewarped images.

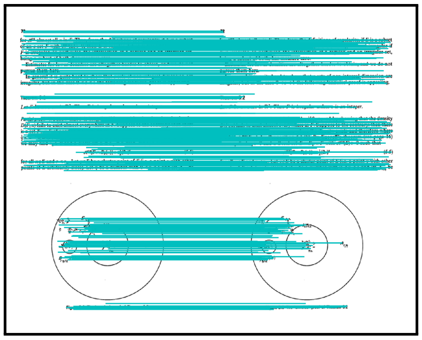
A. *Ground-Truth Dewarped Images*

The presented image based performance evaluation method requires ground-truth dewarped images. So far, DFKI-I [11] is the only publicly available dataset of camera-captured document images. We prepared this dataset to compare different page dewarping approaches in a Document Image Dewarping Contest that was held at CBDAR 2007 [11]. The following types of ground-truth were provided with the dataset: i) ground-truth ASCII text in plain text format, ii) ground-truth page segments (text-lines and zones and their types) in color coded form, iii) scanned images of pages which have been captured by a camera. A sample camera-captured document and its correspoding scanned image from the dataset are shown in Figure 1. The scanned document images in this dataset, as shown in Figure 1(b), are flat and straight. Therefore, they can be used as ground-truth dewarped images. For the purpose of performance evaluation, scanning of pages together with capturing them through camera requires very less manual effort as compared to marking images manually [12] or to generate ASCII text ground-truth [11].

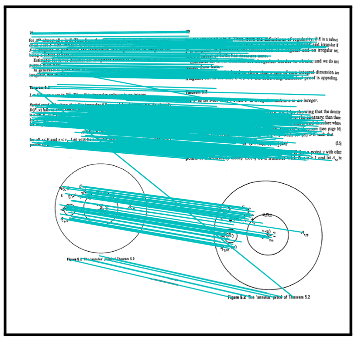
B. *Performance Evaluation Methodology*

To compare the quality of a dewarped document against a ground-truth dewarped document, image based features are calculated using SIFT [13]. For an image, SIFT estimates key features and returns their correspoding locations and descriptors. Matching between the features of two different images is done by calculating cosine inverse of the dot product of their normalized descriptors. The bad matches are removed by applying a thresholding criteria such as, a match is considered bad if the angle ratio between first and second nearest neighbors is greater than a predefined threshold. In our case, we set this threshold equal to 0.6. We have also noticed that there are some wrong SIFT based matches between two similar document images at high image resolutions, but not at low image resolutions. Therefore, we downscale document images by the factor of 4 before SIFT based comparision.

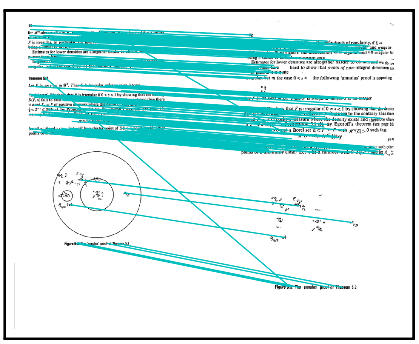
A sample camera-captured, warped document image and its corresponding ground-truth dewarped image are shown in Figure 2(a) and Figure 2(b), respectively. For the camera-captured image (Figure 2(a)), two different, a good one and a bad one, dewarped images are also shown in Figure 2(c) and Figure 2(d), respectively. Here, it can be noticed that the good dewarped image visually looks similar to the ground-truth image and contains both text and non-text elements, except slight non-linearity in text-lines and different aspect ratio. The bad dewarped image, on the other hand, missed most of the non-text elements and some of the text elements along with irregularity/non-linearity in text-lines. The SIFT based matching between: i) the ground-truth image with itself is shown in Figure 3(a), ii) the ground-truth image and the good dewarped image is shown in Figure 3(b), and



(a) feature matching of the ground-truth dewarped image with itself



(b) feature matching between the ground-truth and the good dewarped image



(c) feature matching between the ground-truth and the bad dewarped image

Figure 3. The matching between SIFT features of: a) the ground-truth image (Figure 2(b)) with itself, b) the ground-truth image and the good dewarped image (Figure 2(c)), c) the ground-truth image and the bad dewarped image (Figure 2(d)).

iii) the ground-truth image and the bad dewarped is shown in Figure 3(c). The ground-truth image matches perfectly with itself as shown in the Figure 3(a). Most of the matches in Figure 3(b) and Figure 3(c) are correct with respect to the corresponding descriptors and their locations, and some of them are only correct with respect to the corresponding descriptors, but not with the correspoding locations. In order to remove these types of wrong matches, a filtering criteria is used, according to which, all those matches that have distances greater than $T\%$ of document diagonal are removed. The value of T can be set in-between 0% to 100%, where $T = 0\%$ means that the matched descriptors should be at the perfectly same locations otherwise discarded, and $T = 100\%$ means that the locations of matched descriptors can be far apart. Both of these extreme values are not suitable for our case. The reasonable value can be set in-between 10% to 30%. It is also important to note that, the number of matches between the ground-truth image and the good dewarped image are more than the number of matches between the the ground-truth image and the bad dewarped image. Therefore, the number of matches and other related metrics can be used for the performance evaluation of page dewarping methods, which are discussed below.

Consider that we are given two dewarped images, the dewarped image I, and the ground-truth dewarped image G. Let, L_I and D_I represent the locations and normalized descriptors of SIFT features for the dewarped image I, and L_g and D_g represent SIFT features for the ground-truth dewarped image G. If the dewarped image I agrees perfectly with the ground-truth dewarped image G, there will be a perfect matching between their correspoding SIFT features as shown in Figure 3(a). If there are differences between the two dewarped images, then there will not be a perfect matching as shown in Figure 3(b) and Figure 3(c).

Here, we define two different performance measures to evaluate different aspects of the behavior of a page dewarping algorithm using SIFT based feature matching. These measures are defined as follows:

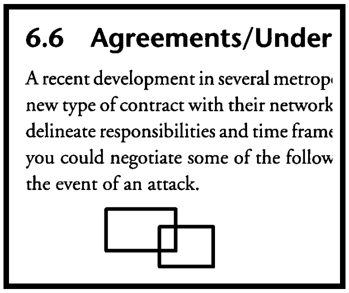
1) **Matching Percentage** M_p**:** let total number of matches between G and I is represented by N, and total number of features in G is represented by N_G. The matching percentage (M_p) is defined as:

$$M_p = \frac{N}{N_G} \tag{1}$$

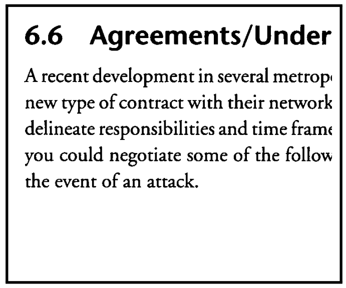
2) **Matching Error** M_e**:** for a pair of matched descriptors p, let $D_G(p)$ represents a descriptor in G, and $D_I(p)$ represent a corresponding matched descriptor in I. The mean error of all matching pairs is calculated as follows:

$$M_e = \frac{\displaystyle\sum_{p=1}^{N} \arccos(D_G(p) \cdot D_I(p))}{N} \tag{2}$$

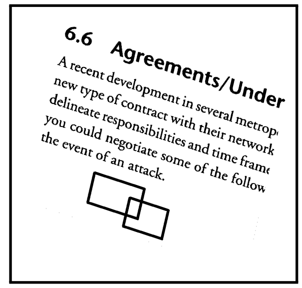
We can analyze the effectiveness and correctness of the presented metrics by comparing a ground-truth dewarped image with a good and a bad dewarped images, such an example is shown in the Figure 2. For the good dewarped image (Figure 2(c)), the values of these metrics are as follows: $M_p = 44.57\%$ and $M_e = 0.15$. Similarly, these values for the bad dewarped image (Figure 2(d)) are as follows: $M_p = 11.73\%$ and $M_e = 0.19$. As shown in the Figure 2(c), the qualities of the good and the bad dewarped



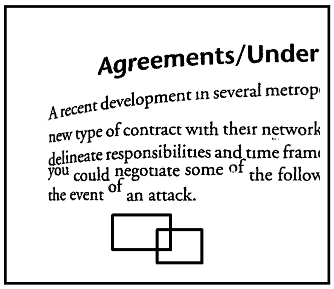
(a) ground-truth dewarped image



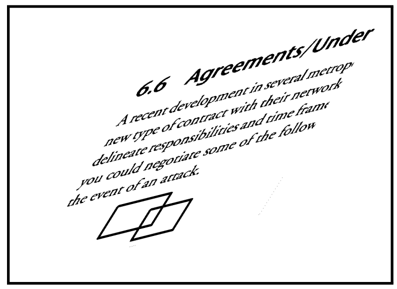
(b) dewarped image with missed non-text ($M_p = 84.34\%$ and $M_e = 0.0$)



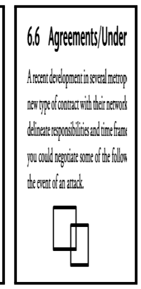
(c) dewarped image with skew ($M_p = 37.72\%$ and $M_e = 0.13$)



(d) dewarped image with warped, missed and irregular text ($M_p = 14.59\%$ and $M_e = 0.19$)



(e) dewarped image with perspective distortions ($M_p = 0\%$)



(f) dewarped image with incorrect aspect ratio ($M_p = 0\%$)

Figure 4. Behavior of the proposed performance evaluation metrics (matching percentage (M_p) and matching error (M_e) in the presence of typical errors produced by dewarping methods. A dewarped image with warped text, perspectively distorted text, and/or incorrect aspect ratio can be considered as the much more erroneous than missed non-text or global skew with respect to OCR performance.

images are consistent with their correspoding values of matching percentage (M_p) and matching error (M_e).

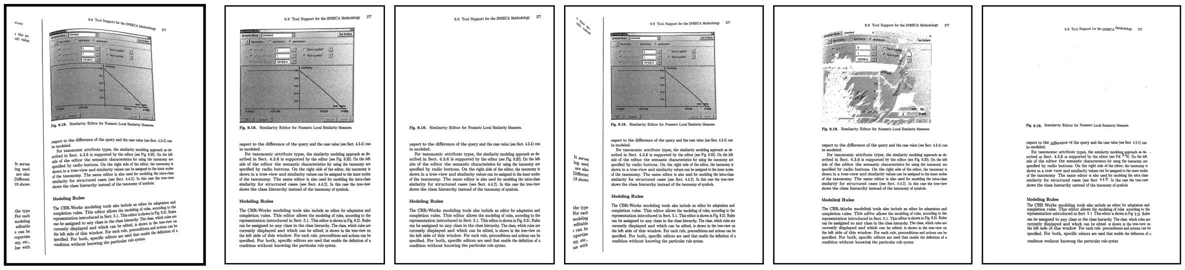
The proposed metrics are also effective in terms of indicating typical errors produced by dewarping methods such as i) missed non-text parts as shown in Figure 4(b) where $M_p = 84.34\%$ and $E_m = 0.0$, ii) global skew as shown in Figure 4(c) where $M_p = 37.72\%$ and $M_e = 0.13$, iii) warped, missed, and irregular text as shown in Figure 4(d) where $M_p = 14.59\%$ and $M_e = 0.19$, iv) perspective distortion as shown in Figure 4(e) where $M_p = 0\%$, and v) incorrect aspect ratio as shown in Figure 4(f) where $M_p = 0\%$. The main purpose of dewarping is to transform warped, non-planar documents into planar images so that traditional scanner based OCR softwares can also process them equally like scanned documents. These results are consistent with the visual (planar) quality of dewarped images as well as with respect to OCR accuracy.

In order to analyze some additional visual quality aspects of a dewarping method that do not directly influence OCR accuracy, we can estimate standard deviation of matching locations between a ground-truth image and its corresponding dewarped image. For example, the standard deviations of plain, skewed and irregular document images as shown in Figure 4 with respect to the ground-truth image are equal to 0, 8, and 4.65, respectively. It is important to note that, the skewed image (Figure 4(c)) has bigger standard deviation as compared to the irregular text (Figure 4(d)), but the skewed image may produce less number of OCR errors than the irregular text, mainly because a skew correction step is a part of standard OCR pipeline.

III. Experiment and Results

As a first step towards comparative evaluation of page dewarping techniques, a page dewarping contest using DFKI-I camera-captured documents dataset was organized along with CBDAR 2007 [11]. Three groups participated in the contest. These three method are referred as CTM [14], SKEL [15], and SEG [8]. The CTM method also used their programs to remove graphics and images from the processed pages. The results thus produced are referred to as CTM2. For the description of the participating methods please refer to [11]. We have also proposed an active contour (snake) based dewarping method in [9], referred to as SNAKE, and compared its performance with those of contest participants. For a sample camera-captured document image of DFKI-I dataset, the dewarped images of all these methods are shown in Figure 5.

These different methods have been compared with each



(a) camera-captured document (b) CTM (c) CTM2 (d) SKEL (e) SNAKE (f) SEG

Figure 5. Example results of different methods for a sample camera-captured document of DFKI-I dataset: b) CTM [14], c) CTM2 [14], d) SKEL [15], e) SNAKE [9], f) SEG [8].

Table I
COMPARATIVE OCR BASED ERROR RATE (EDIT DISTANCE) OF DIFFERENT DEWARPING METHODS ON DFKI-I DATASET.

Algorithm	**Edit Distance**
CTM2 [14]	1.758
SNAKE [9]	1.917
CTM [14]	2.113
SKEL [15]	2.162
SEG [8]	4.088

Table II
COMPARATIVE FEATURE BASED PERFORMANCE EVALUATION RESULTS OF DIFFERENT DEWARPING METHODS ON DFKI-I DATASET USING PROPOSED VECTORIAL PERFORMANCE EVALUATION METRICS (MATCHING PERCENTAGE (M_p) AND MATCHING ERROR (M_e)).

Algorithm	M_p**%**	M_e
CTM [14]	34.90%	0.13
CTM2 [14]	30.51%	0.14
SKEL [15]	25.45%	0.14
SNAKE [9]	21.52%	0.14
SEG [8]	12.44%	0.15

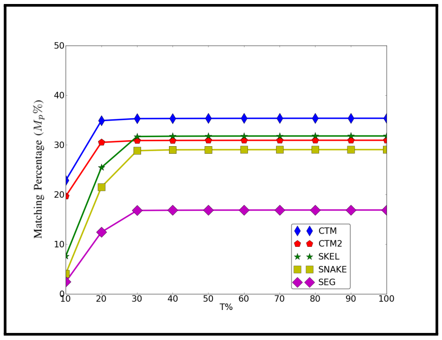
other through OCR based edit distance by using ASCII text ground-truth in [11], [9]. The OCR based performance evaluation results, that are copied from [9], are shown in Table I. The CTM2 method performs the best on DFKI-I dataset, and its results are better than CTM, i.e. after post-processing to remove graphics and images. This is because the ground-truth ASCII text contains text coming only from the textual parts of the documents, so the text that is present in graphics or images is ignored. Hence, the dewarped documents that contain text inside graphics regions get higher edit distances. On the basis of OCR based performance evaluation, CTM, SKEL and SNAKE have similar performance, and SEG has relatively inferior performance.

From the methods descriptions, we have determined that both CTM and SKEL handle non-text elements together with text elements, but SEG and SNAKE methods mainly perform dewarping for text elements and do not handle non-text elements. One of such example for DFKI-I dataset can be seen in Figure 5.

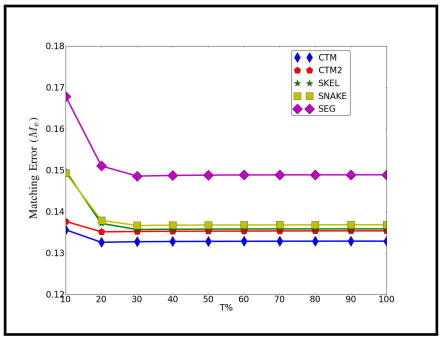
In this paper, we compare these dewarping methods using the presented performance evaluation metrics (matching percentage (M_p), matching error (M_e)) on DFKI-I dataset. The feature based performance evaluation results of the dewarping methods for different values of T (10% to 100%) are shown in Figure 6. For an optimal value of T (i.e. $T = 20\%$), feature based performance evaluation results are shown in Table II. CTM method has achieved the best matching percentage (M_p) among all other methods. The matching percentage and matching error of CTM are better than the CTM2, which is also perfectly consistent with the definition of CTM2 (i.e. removed graphics and images). CTM method has also achieved the lowest matching error (M_e) as compared to other methods. On the other hand, SEG has comparatively achieved the lowest matching percentage and highest matching error in comparison to other methods. It is very interesting to note that these feature-based performance evaluation results are also closely consistent with the OCR based results. However, feature based results give more details about the quality of dewarped images with respect to both text and non-text elements.

IV. CONCLUSION

In this paper, we have proposed an image based performance evaluation methodology for dewarping algorithms using SIFT features. Unlike OCR based performance evaluation techniques [11], [9], a feature based technique indicates how well a dewarping method performs on both text and non-text elements in warped images. Unlike previous feature based performance evaluation techniques [12], our proposed



(a) Matching Percentage (M_p%)



(b) Matching Error (M_e)

Figure 6. Comparative performance evaluation of different methods for DFKI-I dataset by using the presented feature-based performance evaluation metrics (matching percentage (M_p) and matching error (M_e)) for different values of T.

featured based technique does not require manual labeling for generating ground-truth images, and calculate vectorial performance evaluation metrics (matching percentage (M_p) and matching error (M_e)), instead of single score. We have also demonstrated that the feature based performance evaluation results are consistent with the OCR base results.

References

[1] H. Cao, X. Ding, and C. Liu, "Rectifying the bound document image captured by the camera: a model based approach," in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, Edinburgh, Scotland, 2003, pp. 71–75.

[2] M. S. Brown and W. B. Seales, "Image restoration of arbitrarily warped documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1295–1306, 2004.

[3] C. L. Tan, L. Zhang, Z. Zhang, and T. Xia, "Restoring warped document images through 3d shape modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 195–208, 2006.

[4] P. Clark and M. Mirmehdi, "Rectifying perspective views of text in 3d scenes using vanishing points," *Pattern Recognition*, vol. 36, no. 11, pp. 2673–2686, 2003.

[5] Z. Zhang and C. L. Tan, "Correcting document image warping based on regression of curved text lines," in *Proceedings 7th International Conference on Document Analysis and Recognition*, Edinburgh, Scotland, 2003, pp. 589–593.

[6] A. Ulges, C. H. Lampert, and T. M. Breuel, "Document image dewarping using robust estimation of curled text lines," in *Proceedings 8th International Conference on Document Analysis and Recognition*, Seoul, Korea, 2005, pp. 1001–1005.

[7] J. Liang, D. DeMenthon, and D. Doermann, "Flattening curved documents in images," in *Proceedings 18th International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 338–345.

[8] B. Gatos, I. Pratikakis, and K. Ntirogiannis, "Segmentation based recovery of arbitrarily warped document images," in *Proceedings 9th International Conference on Document Analysis and Recognition*, Curitiba, Barazil, 2007, pp. 989–993.

[9] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Dewarping of document images using coupled-snakes," in *Proceedings of First International Workshop on Camera-Based Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 34–41.

[10] M. S. Brown and Y. C. Tsoi, "Geometric and shading correction for images of printed materials using boundary," *Image Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1544 –1554, june 2006.

[11] F. Shafait and T. M. Breuel, "Document image dewarping contest," in *Proceedings 2nd International Workshop on Camera Based Document Analysis and Recognition*, Curitiba, Brazil, 2007, pp. 181–188.

[12] N. Stamatopoulos, B. Gatos, and I. Pratikakis, "A methodology for document image dewarping techniques performance evaluation," in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 956–960.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[14] B. Fu, M. Wu, R. Li, W. Li, and Z. Xu, "A model-based book dewarping method using text line detection," in *2nd Int. Workshop on Camera-Based Document Analysis and Recognition*, Curitiba, Brazil, Sep. 2007.

[15] A. Masalovitch and L. Mestetskiy, "Usage of continuous skeletal image representation for document images dewarping," in *2nd Int. Workshop on Camera-Based Document Analysis and Recognition*, Curitiba, Brazil, Sep. 2007.