The IUPR Dataset of Camera-Captured Document Images

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*—Major challenges in camera-base document analysis are dealing with uneven shadows, high degree of curl and perspective distortions. In CBDAR 2007, we introduced the first dataset (DFKI-I) of camera-captured document images in conjunction with a page dewarping contest. One of the main limitations of this dataset is that it contains images only from technical books with simple layouts and moderate curl/skew. Moreover, it does not contain information about camera's specifications and settings, imaging environment, and document contents. This kind of information would be more helpful for understanding the results of the experimental evaluation of camera-based document image processing (binarization, page segmentation, dewarping, etc.). In this paper, we introduce a new dataset (the IUPR dataset) of camera-captured document images. As compared to the previous dataset, the new dataset contains images from different varieties of technical and non-technical books with more challenging problems, like different types of layouts, large variety of curl, wide range of perspective distortions, and high to low resolutions. Additionally, the document images in the new dataset are provided with detailed information about thickness of books, imaging environment and camera's viewing angle and its internal settings. The new dataset will help research community to develop robust camera-captured document processing algorithms in order to solve the challenging problems in the dataset and to compare different methods on a common ground.**

***Keywords*-Dataset, Camera-Captured Document Processing, Performance Evaluation**

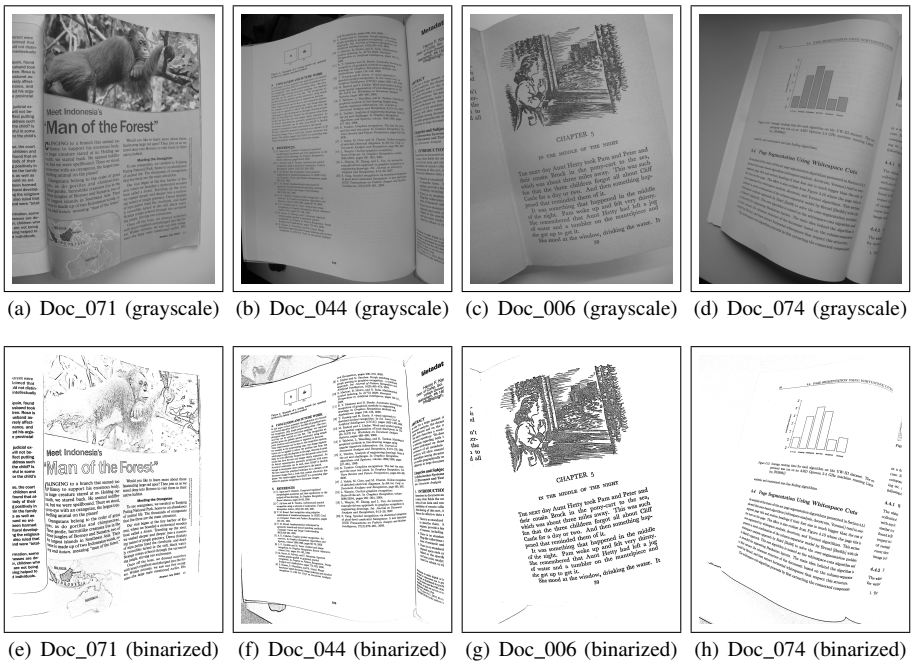
I. Introduction

Ground-truth datasets are crucial for objectively measuring the performance of algorithms in many fields of computer science. The availability of such datasets for use in research and development lays the basis for comparative evaluation of algorithms. However, collecting a real-world dataset and preparing its ground-truth is not a trivial task. Therefore, a good practice in research is to focus on developing algorithms that solve the problem at hand and use existing public datasets for evaluating the performance of the developed algorithms. In doing so, one not only saves the effort needed to create a representative dataset and its ground-truth, but also the results obtained can be directly compared to those of other algorithms on the same dataset. For instance, in the machine learning community, evaluating new classification algorithms on datasets from the UCI repository [1] has become a de facto standard.

In document analysis and recognition, collecting real-world datasets and sharing them with the community has received quite a lot of attention. As a result, several representative datasets are available for different tasks. Examples of such dataset include the MNIST dataset [2] for handwritten character recognition, the UNLV ISRI dataset [3] for optical character recognition, the UW-I/II/III datasets [4] for document layout analysis, the MARG dataset [5] for logical labeling, the UvA color documents dataset [6] for handling colored magazine pages, the IAM database [7] for off-line handwritten text-line and word segmentation, IFN/ENIT dataset [8] for Arabic handwritten word recognition, and last but not least the Google 1000 books dataset [9] for optical character recognition of old books.

While such rich datasets provide solid grounds for experimentation, all of these datasets focus on scanned documents. With the advent of digital cameras, the traditional way of capturing documents is changing from flat-bed scans to camera captures [10], [11]. Recognition of documents captured with hand-held cameras poses many additional technical challenges like perspective distortion, non-planar surfaces, uneven lighting, low resolution, and wide-angle-lens distortions [12]. These challenges have opened new directions of research like binarization and noise removal from camera-captured documents, page segmentation (zone segmentation, curled text-line extraction) and document image dewarping.

We have developed the first camera-captured document image dataset (DFKI-I) [13] for benchmark. Researchers have used this dataset for benchmarking binarization [14], [15], noise cleanup using page frame detection [16], [17], text-line extraction [18], and dewarping methods [13], [19]. All the document images in the DFKI-I dataset belong to simple technical books with single column layout and contain small skew/curl angles. Therefore, there is no variety of the images in the DFKI-I dataset. Additionally, the dataset is not provided with the details of imaging environment, camera (viewing angle, internal settings, resolution, etc.) and document contents, even though such type of information would be more helpful for understanding the experimental evaluation results of camera-based document image processing tasks.



(a) Doc_071 (grayscale) (b) Doc_044 (grayscale) (c) Doc_006 (grayscale) (d) Doc_074 (grayscale)

(e) Doc_071 (binarized) (f) Doc_044 (binarized) (g) Doc_006 (binarized) (h) Doc_074 (binarized)

Figure 1. Samples of grayscale documents and their binarized images from the IUPR dataset.

To fill these gaps, we developed a new dataset of camera-captured documents. The new dataset contains documents from a large variety of technical and non-technical books and bound pages, and the details related to imaging environment, camera settings, and document contents are also provided with each document image. Like DFKI-I dataset, we prepared ground-truth information for text-lines, text-zone, and zone-type, dewarped images (scanned documents), and ASCII text for all documents in the new dataset. We refer our new dataset as the IUPR dataset. This paper describes the IUPR dataset in detail and presents it as a foundation of comparative performance evaluation for different tasks in the camera-captured document analysis domain.

The rest of this paper is organized as follows. We present the IUPR camera-captured document images dataset in Section II. The process of generating the ground-truth is illustrated in Section III. Section IV represents our conclusion.

II. The IUPR Dataset

The dataset consists of 100 grayscale document images of pages that were captured by using a hand-held camera. The captured documents were binarized using a local adaptive thresholding technique described in [20]. Some sample grayscale documents and their binarized images in the dataset are shown in Figure 1. The details about imaging environment and camera settings that were used for capturing images, and the contents of the dataset are described here.

Imaging Environment:

Documents were captured by placing books on flat table. All documents were captured during daylight in an office-room having a normal white-light on ceiling.

Camera Setting:

A cannon PowerShot G10 camera was used for capturing document images. Images were captured by setting the camera to the "macro" mode and without any digital zoom and flash. Documents were captured with a variety of resolutions (5, 9, or 15 Mega Pixel) and different viewing angles of camera (like left, top-right, bottom-left etc.) for adding a verity of perspective distortions in the dataset. The viewing angle can be roughly estimated with respect to the document's center point. Camera settings that were used to capture the sample documents in Figure 1 are shown in Table I.

Document Content:

Documents have been selected from several different technical books, magazines, old story books, bound pages, etc. These documents belong to a large variety of layouts, some of them can be seen in Figure 1. For the sample document images as shown in Figure 1, the thickness of their correspoding books are mentioned in Table I. In general, geometric distortion in a document image depends upon book's thickness and its position (folded/unfolded).

Table I
SOME OF THE INFORMATION ABOUT CAMERA SETTING AND DOCUMENT CONTENT (THAT ARE PROVIDED WITH EACH DOCUMENT IN THE DATASET) FOR THE SAMPLE DOCUMENTS IN FIGURE 1.

Document ID	Camear Setting		Document Content	
	Viewing Angle	Mega Pixel	Book Type	book Thickness
Doc_071	Left	15	Magazine	1.5 cm
Doc_044	Right	15	Conference Proceedings	2.5 cm
Doc_006	Top-right	9	Old Story Book	2.0 cm
Doc_074	Bottom-Right	15	Bound Pages (Technical)	1.0 cm

Some statistics about the documents in the IUPR dataset are as follow. Out of 100 documents, 75 documents consist of single-column layout and 25 documents contain multi-column layout. 51 documents consist of complete page border (like Figure 1(b)) and remaining 49 documents consists of incomplete page border (like Figure 1(c)). 85 documents were captured from unfolded books (like Figure 1(a)) and remaining 15 documents were captured from folded books (like Figure 1(d)).

The following information is provided with each document image in the dataset.

- name, publisher, and thickness of book
- book type (proceedings, magazine, story, bound pages etc.)
- page number, contents detail (text, graphics, etc.) and number of columns
- folded/unfolded book
- complete/incomplete page border
- camera viewing angle
- camera resolution

III. GROUND-TRUTH

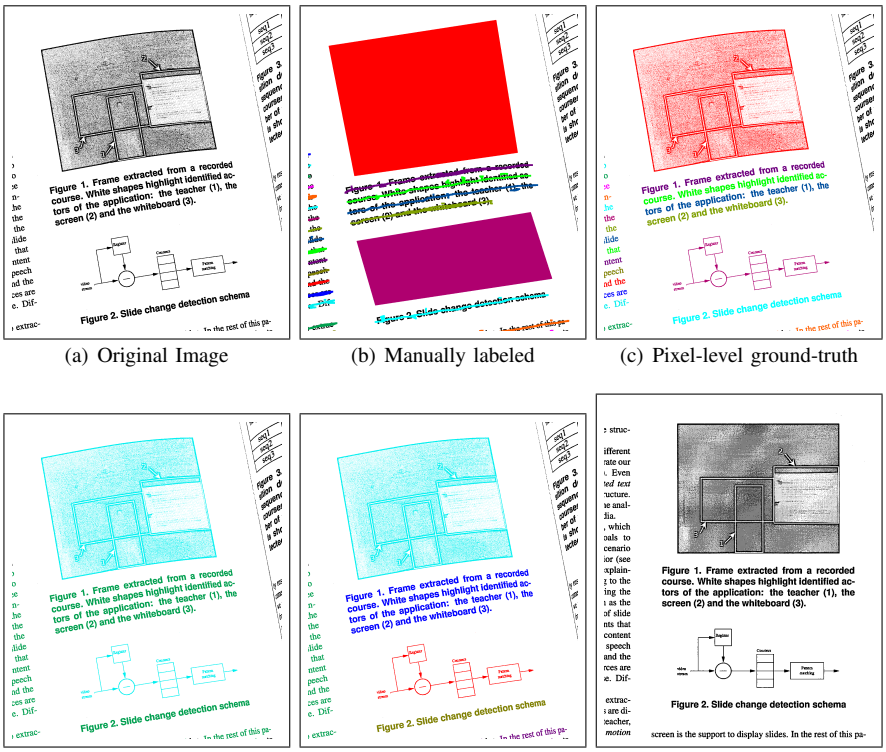
The dataset is provided with different types of ground-truth information as follows:

1) ground-truth text-lines in color coded form (Figure 2(c))
2) ground-truth text-zones in color coded form (Figure 2(e))
3) content type (half-tone/figure, equation, table, text, marginal noise) ground-truth information (Figure 2(d))
4) reading order of text-lines and text-zones
5) ground-truth ASCII text in plain text format
6) ground-truth dewarped (scanned) document images (Figure 2(f))

Generating pixel-level ground-truth can become quite cumbersome since a document image typically contains over one million foreground pixels. Therefore, we have developed semi-automatic technique [21] for preparing pixel-level ground-truth. For each text-lines/figure-captions/formulas, a line is drawn manually with a unique color, and for each table/figure/graphics, a bounding polygon is drawn with a unique color. The manual labeling for a sample image (Figure 2(a)) is shown in Figure 2(b). Manual color labeling is done in such a way that R, G, and B channel contains information about content type, zone number and text-line number in reading order, respectively, where color channel R is set to '1' for mathematical equations, '2' for tables, '3' for figure/graphics, and '4' for text-lines. The R, G, and B color channels of background and marginal noise pixels are all set equal to 255 and 0, respectively.

From manual labeling, the pixel-level ground-truth image of a document is generated as follows. First, connected components are extracted from the document image (Figure 2(a)), and then each connected component is assigned the color of the manually labeled line/polygon that touches with the connected component. The pixel-level ground-truth image is shown in Figure 2(c). In this figure, each text-line as well as non-text element can be uniquely identified. By using the information provided in color channel R and G, content type and zone level ground-truths can also be generated, respectively, which are shown in Figure 2(d) and Figure 2(e), respectively.

All documents that were captured with a camera were also scanned with a flat-bed scanner. These scanned documents are flat and straight as shown in Figure 2(f). Therefore, they can be used as ground-truth dewarped images for image based performance evaluation of dewarping methods [17]. Additionally, ASCII text ground-truth of scanned documents is intended for use as the overall performance measure of a dewarping system by using OCR on the dewarped document. A commercial OCR system was then used to generate the text ground-truth from the scanned documents. The OCR system was used in an interactive mode such that it presented



(a) Original Image (b) Manually labeled (c) Pixel-level ground-truth

(d) Content type level ground-truth (e) Zone level ground-truth (f) Dewarped (scanned) image ground-truth

Figure 2. An example image to demonstrate the process of generating different types of ground-truth that are provided with the IUPR dataset.

to the operator all characters for which the recognition confidence was not high. We also replaced all mathematical and other non-ASCII symbols with a '#' sybmol as was done in the datasets used in UNLV annual tests of OCR accuracy [22].

The dataset can be downloaded from **www.sites.google.com/a/iupr.com/bukhari/**. It is not split into training and test sets, because some algorithms need larger training sets as compared to others. It is expected that when other researchers use this dataset, they will split it into test and training sets as per requirements.

IV. Conclusion

This paper presented a new camera-captured documents dataset–the IUPR dataset. Unlike the previous DFKI-I dataset [13], the new dataset consists of images from different technical and non-technical books with a diversity of layouts as well as a large variety of perspective and/or geometric distortions. Therefore, the new dataset is much more challenging as compared to the previous DFKI-I dataset. According to the ground-truth information that is provided with the dataset, the new dataset can be used for the performance evaluation and benchmarking of camera-captured document image processing approaches, like binarization, page (text-line/zone) segmentation, zone classification, dewarping, etc. Detailed information about the imaging environment, camera settings, and document contents is also provided with each image in the dataset, which can help in analyzing the performance evaluation results. This dataset makes a good base for comparative evaluation of camera-captured document analysis algorithms.

References

[1] Http://archive.ics.uci.edu/ml/datasets.html.

[2] Http://yann.lecun.com/exdb/mnist/.

[3] Http://www.isri.unlv.edu/ISRI/OCRtk.

[4] I. Guyon, R. M. Haralick, J. J. Hull, and I. T. Phillips, "Data sets for OCR and document image understanding research," in *Handbook of character recognition and document image analysis*, H. Bunke and P. Wang, Eds. World Scientific, Singapore, 1997, pp. 779–799.

[5] G. Ford and G. R. Thoma, "Ground truth data for document image analysis," in *Symposium on Document Image Understanding and Technology*, Greenbelt, MD, USA, April 2003, pp. 199–205.

[6] L. Todoran, M. Worring, and M. Smeulders, "The UvA color document dataset," *Int. Jour. on Document Analysis and Recognition*, vol. 7, no. 4, pp. 228–240, 2005.

[7] U. Marti and H. Bunke, "The IAM-database: an English sentence database for off-line handwriting recognition," *Int. Jour. on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, 2002.

[8] M. Pechwitz, S. S. Maddouri, V. Maergner, N. Ellouze, and H. Amiri, "IFN/ENIT-database of handwritten Arabic words," in *7th Colloque Int. Francophone sur l'Ecrit et le Document*, Hammamet, Tunis, Oct. 2002.

[9] L. Vincent, "Google book search: Document understanding on a massive scale," in *9th Int. Conf. on Document Analysis and Recognition*, Curitiba, Brazil, Sep. 2007, pp. 819–823.

[10] M. J. Taylor, A. Zappala, W. M. Newman, and C. R. Dance, "Documents through cameras," in *Image and Vision Computing 17*, vol. 11, September 1999, pp. 831–844.

[11] T. Breuel, "The future of document imaging in the era of electronic documents," in *Int. Workshop on Document Analysis*, Kolkata, India, Mar. 2005.

[12] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," *Int. Jour. of Document Analysis and Recognition*, vol. 7, no. 2-3, pp. 84–104, 2005.

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

[14] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Adaptive binarization of unconstrained hand-held camera-captured document images," *Journal of Universal Computer Science (J.UCS)*, vol. 15, no. 18, pp. 3343–3363, 2009.

[15] D. M. Oliveira and R. D. Lins, "A new method for shading removal and binarization of documents acquired with portable digital cameras," in *Proceedings of Third International Workshop on Camera-Based Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 3–10.

[16] F. Shafait, J. van Beusekom, D. Keysers, and T. M. Breuel, "Document cleanup using page frame detection," *Int. Jour. on Document Analysis and Recognition*, vol. 11, no. 2, pp. 81–96, 2008.

[17] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Border noise removal of camera-captured document images using page frame detection," in *Proceedings of Fourth International Workshop on Camera-Based Document Analysis and Recognition*, Beijing, China, 2011.

[18] ——, "Performance evaluation of curled textlines segmentation algorithms on CBDAR 2007 dewarping contest dataset," in *Proceedings 17th International Conference on Image Processing*, Hong Kong, China, 2010.

[19] ——, "Dewarping of document images using coupled-snakes," in *Proceedings of Third International Workshop on Camera-Based Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 34–41.

[20] A. Ulges, C. Lampert, and T. Breuel, "Document image dewarping using robust estimation of curled text lines," in *Proc. Eighth Int. Conf. on Document Analysis and Recognition*, Aug. 2005, pp. 1001–1005.

[21] F. Shafait, D. Keysers, and T. M. Breuel, "Performance evaluation and benchmarking of six page segmentation algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, 2008.

[22] S. V. Rice, F. R. Jenkins, and T. A. Nartker, "The fourth annual test of OCR accuracy," Information Science Research Institute, University of Nevada, Las Vegas, Tech. Rep., 1995.