Automatic Ground Truth Generation of Camera Captured Documents Using Document Image Retrieval

Sheraz Ahmed*, Koichi Kise†, Masakazu Iwamura†, Marcus Liwicki*‡, and Andreas Dengel*
*German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
firstname.lastname@dfki.de
†Osaka Prefecture University, Japan
‡University of Fribourg, Switzerland
firstname.lastname@unifr.ch

***Abstract*—In this paper a novel method for automatic ground truth generation of camera captured document images is proposed. Currently, no dataset is available for camera captured documents. It is very difficult to build these datasets manually, as it is very laborious and costly. The proposed method is fully automatic, allowing building the very large scale (i.e., millions of images) labeled camera captured documents dataset, without any human intervention. Evaluation of samples generated by the proposed approach shows that 99.98% of the images are correctly labeled. Novelty of the proposed approach lies in the use of document image retrieval for automatic labeling, especially for camera captured documents, which contain different distortions specific to camera, e.g., blur, occlusion, perspective distortion, etc.**

I. Introduction

An important area in analysis of camera captured documents is recognition of text, as there are a lot of services which can be provided, if text is recognized, e.g., real time translation. Different OCR systems are already available, but they are designed specifically for scanned images. One of the reasons why existing OCRs cannot be applied to camera captured images is due to different distortions which exist in camera captured images, e.g., blur, perspective distortion, occlusion, etc. To enable these OCRs to work with camera captured documents, it is required to train them with data which contain these distortions so that they are able to handle them when encountered. The main problem in this is the lack of availability of any big dataset for camera captured document images which contains images with different distortions specific to camera captured documents.

The main problem in dataset generation is its labeling. It is not possible to manually label each word and/or character in captured images, as it is very laborious and costly. One possible solution could be to use different degradation models to build up a large scale dataset using synthetic data [1], [2]. However, researchers are still of different opinion that either degradation models are true representative of real world data or not. Hence, there is a strong need to have method for automatic labeling of large scale camera captured documents.

The problem of automatic ground truth generation for camera captured document images is not addressed in past. To automatically generate ground truth, the electronic version of camera captured image and its alignment to the captured image are needed, so that the captured image can be labeled using the ground truth information contained in the electronic version. Existing methods for ground truth generation of scanned documents [3]–[7] can not be applied to camera captured documents, as it is assumed that whole document is contained in the scanned image. Note that, camera captured documents usually contain only a part of the document, therefore it is not possible to align captured image to electronic version directly. In addition, they contain a lot of other distortions.

In this paper, we propose an approach for automatic ground truth generation of camera captured document images using a document image retrieval system. A Locally Likely Arrangement Hashing (LLAH) [8] based document retrieval system is used to retrieve the electronic version of the same document as the captured image. LLAH can retrieve the same document even if only a part of document is contained in the camera captured image. That is why LLAH is suitable for the task.

The proposed approach is fully automatic and does not require any human intervention, especially for labeling. Another highlight of the proposed method is that it is not limited to any language; this means that we can also build datasets for different languages, e.g., Japanese, Arabic, Urdu, Indic scripts, etc. In addition, it can be applied to scanned documents as is. All we need is the PDF of a document and its camera captured image.

II. Related Work

This section summarizes different approaches that are available for automatic generation of ground truth. First, different approaches for automatic ground truth generation of scanned documents are presented. Then details about degradation models for scanned and camera captured images are presented.

[5] and [6] proposed an approach for automatic generation of character ground truth from scanned documents. Documents are created, printed, photocopied, and scanned. Geometric transformation is computed between scanned and ground truth

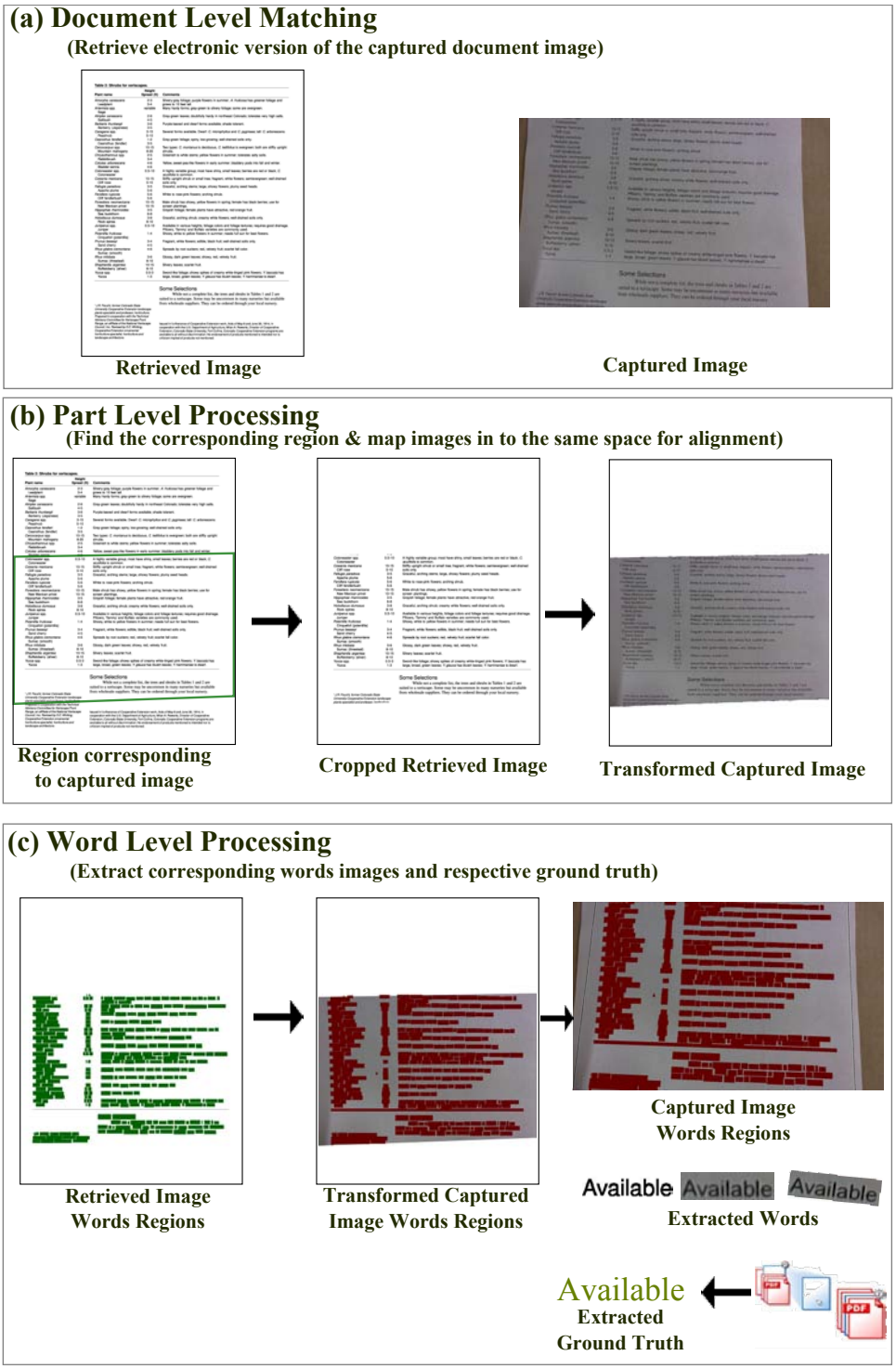

Fig. 1: Automatic Ground Truth Extraction Process

image. Finally, transformation parameters are used to extract the ground truth information for each character. [9] further improved the approach presented in [5] and [6] by using the attributed branch-and-bound algorithm for establishing the correspondence between the data points of scanned and ground truth images. After establishing correspondence, the ground truth for the scanned images is extracted by transforming the ground truth of the original image.

Similarly, [4] proposed automatic ground truth generation for OCR using robust branch and bound search (RAST) [10]. First global alignment is estimated between the scanned and ground truth image. Finally, local alignment is used to

adapt the transformation parameters by aligning clusters of nearby connected components. [3] proposed an approach for ground truth generation for newspaper documents. It is also based on synthetic data generated using an automatic layout generation system which is then printed, degenerated, and scanned. Again, RAST is used to compute the transformation to align the ground truth to the scanned image. The main focus of this approach is to create ground truth information for layout analysis which is obtained using an automatic layout generation system.

In the case of scanned documents, complete document image is available, and therefore, transformation between ground truth and scanned image can be computed using different alignment techniques [3]–[6]. However, camera captured documents are very challenging, as they usually contain a part of document along with other unnecessary objects in background. It is therefore, not possible to apply existing methods to camera captured images, as of all the proposed methods assume that complete documents are available in the scanned images.

In addition to the methods based on alignment of scanned documents using different global and local alignment techniques, there is also a possibility to use different image degradation models [11], [12]. An advantage of degradation models is that everything remains electronic, so we do not need to print and scan documents. These degradation models are applied to word or characters to generate images with different possible distortions. [7] used degradation models to synthetic data in different languages, for building datasets which can be used for training and testing of scanned documents. Recently, there are some image degradation models proposed for camera captured documents. [1] has proposed a degradation model for low-resolution camera captured character recognition. The distribution of the degradation parameters is estimated from actual images and then applied to build synthetic data. Similarly, [2] proposed a degradation model of uneven lighting which is used for generative learning. A main problem with degradation models is that they are designed to add limited distortions estimated from distorted images.

III. Methodology

The proposed approach for automatic ground truth generation is based on document image retrieval system. Figure 1 shows the complete flow of proposed approach. Document level matching (Figure 1(a), Section III-A) is performed by retrieving the electronic version of the camera captured document image using the document retrieval system. Figure 2 shows the LLAH based document retrieval system used for retrieving the same document. Along with the retrieved document, the region which corresponds to the camera captured document is also estimated by LLAH. Using this corresponding region, part level matching and transformation are performed on the retrieved and the captured image (Figure 1(b), Section III-B). Finally, the transformed parts from the captured and the retrieved images are used for word level matching and transformation to extract corresponding words in both images and their ground truth information from PDF (See Figure 1(c), Section III-C).

A. Document Level Matching

In document level matching, the electronic version of camera captured image is retrieved from the database using LLAH based document retrieval system. LLAH is used to retrieve the the same document from the large database with the efficient memory scheme. It has already shown the potential to extract similar documents from the database of 20 million images with retrieval accuracy of more than 99% [8].

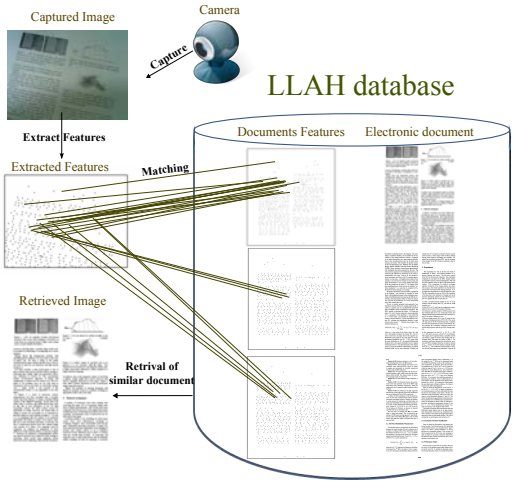

Fig. 2: LLAH Document Retrieval

Figure 2 shows the LLAH based document retrieval system. To build the database for the LLAH, document images are extracted from their corresponding PDF files. These documents include, proceedings, books, magazines, and other articles. Local invariant features based on their arrangements are extracted from each image and stored in a hash table. Hence each entry in the hash table corresponds to a document with its features. To retrieve the electronic version of the document from the database, features are extracted from the camera captured image and compared to features in the database. Electronic version of the document which has the highest matching score is returned as the retrieved document. More details about LLAH can be found in [8].

B. Part Level Processing

As camera captured documents usually contains only a part of the document, therefore part level matching is required to extract the ground truth. In the part level processing, the region of electronic document which corresponds to the camera captured image is estimated using LLAH. Using this corresponding region, the retrieved image is cropped so that only the corresponding part is used for further processing. To align these regions and to extract ground truth, it is required to first map them into the same space.

As camera captured images contain different types of distortions and transformations (Figure 1(a)), we need to find out transformation parameters which can convert the camera

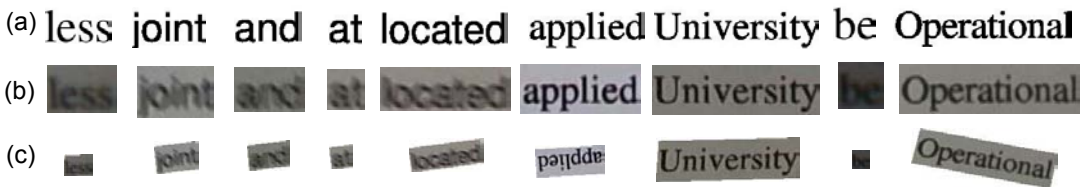

Fig. 3: Extracted words from (a) Retrieved image, (b) Normalized captured image, (c) Captured image

captured image to the retrieved image space and other way around. The transformation parameters are computed using the corresponding matched points between the query and the retrieved document image. Using these transformation parameters, perspective transformation is applied to the captured image which maps it to the space of the retrieved document image. The cropped retrieved and the transformed captured images (Figure 1(b)) are used in word level processing to extract ground truth.

C. Word Level Processing

For word level processing (Figure 1(c)), word regions need to be found. It is important to mention, it looks very trivial that if transformation parameters are applied to the word bounding boxes from PDF file, words can be cropped from the captured image. However it is not true, because even after applying transformation only some part of the cropped transformed image and the camera captured image will be perfectly aligned. Therefore if transformed bounding boxes from PDF file are used for cropping word image it would lead to false ground truth information in parts which are not perfectly aligned. Hence before cropping word it is required to first find out the region which perfectly aligns with camera captured image, so that exactly the same and complete word in cropped from captured image.

To find out word regions, Gaussian smoothing is performed on the transformed captured image and the cropped part of the retrieved image. Bounding boxes are extracted from the smoothed images, where each box corresponds to a word in each image. To find the corresponding words in both images, the distance between their centroids (d_{centroid}) and width (d_{width}) is computed. All of the boxes for which d_{centroid} and d_{width} are less than θ_c and θ_w respectively, are referred to as boxes for the same word in both images. Here, θ_c and θ_w refer to the bounding box distance thresholds for centroid and width, respectively.
The distance between centroids and width of bounding boxes is computed using the following equations.
$d_{\text{centroid}} = \sqrt{(\overline{x}_{\text{capt}} - \overline{x}_{\text{ret}})^2 + (\overline{x}_{\text{capt}} - \overline{y}_{\text{ret}})^2} < \theta_c$

$d_{\text{width}} = \sqrt{(W_{\text{capt}} - W_{\text{ret}})^2} < \theta_w$ $(\overline{x}_{\text{capt}}, \overline{y}_{\text{capt}}), W_{\text{capt}}$ and $(\overline{x}_{\text{ret}}, \overline{y}_{\text{ret}}), W_{\text{ret}}$ refer to centroids and width of bounding boxes in the transformed captured and the cropped retrieved image.

We have used $\theta_d = 5$ and $\theta_w = 5$ pixels, which means, if two boxes are at almost the same position in both images and their width is also almost the same then they correspond to the same word in both images. All of the corresponding boxes are cropped from their respective images where no Gaussian smoothing is performed. This results in two images for each word, i.e., the word image from the retrieved document image (we call it ground truth image) and word image from the transformed captured image.

The word extracted from the transformed captured image is already normalized in terms of different rotation, scale, etc., which were present in the captured image. However, an image which has different transformations and distortions is of main interest, as it can be used for training of systems insensitive to different transformation. To get this image, inverse transformation is performed on the bounding boxes, to map them back into the space of the captured image where no transformation is performed. Boxes dimensions after inverse transformation are then used to crop the corresponding words from captured image. Finally, we have three different images for a word, i.e., the word image from the ground truth/retrieved image, from the transformed captured image, and the captured image (Figure 3).

Once these images are extracted, the next step is to extract the text within these images. To extract text, we used the bounding box information of the word image from ground truth/retrieved image and extract text from the PDF for that bounding box. This extracted text is then saved as text file along with the word images.

To further extract characters from the word images, vertical projection is performed for the word image extracted from the ground truth/retrieved document image. To segment the word image, zero valleys are searched in the vertical projection profile and selected as points of segmentation. If the number of characters extracted from the PDF and the number of segments based on the vertical projection are same, then the word image extracted from the transformed captured image is cropped according to this segmentation. To extract characters from the captured image, inverse transformation is performed on bounding boxes of characters extracted from the transformed captured word. Finally, we have characters extracted from the captured image and the transformed captured image. After that for each segment the corresponding character from the word text file is extracted (Figure 4). It is possible that the captured image contains only a part of the document. Therefore, the region of interest could be any irregular polygon (see a transformed and a cropped image in Figure 1(b)). Due

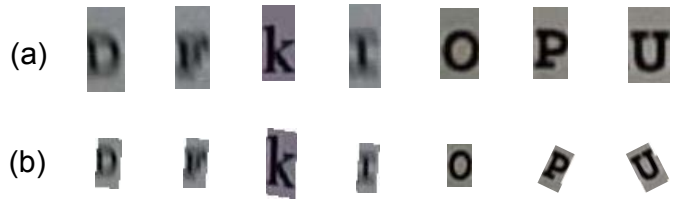

Fig. 4: Extracted characters from (a) Normalized captured image, (b) Captured image

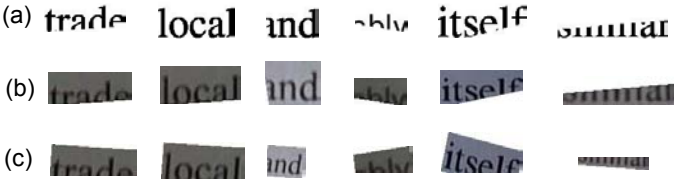

Fig. 5: Words on border from (a) Retrieved image, (b) Normalized captured image, (c) Captured image

to this, it is possible that the characters and words which occur near the border of this region are partially missing (Figure 5). These words, if included directly in the dataset, can cause problem during training, e.g., if a dot of an *i* is missing then in some fonts it looks like 1 which can increase confusion between different characters during training. To handle this problem, all the words and characters which lie near a border are marked with a flag in their names. This allows to separate these words so that they can be handled separately if included in training.

IV. Evaluation

Manual evaluation is performed to check correctness and quality of the generated ground truth. To evaluate manually, first, a 1 million words images dataset is generated from different camera captured documents. These documents include proceedings, magazines, articles, etc. Out of the generated dataset, $50,000$ samples are randomly selected for evaluation. One person has manually inspected all of these samples to find out errors. This manual check reveals that more than 99.98% of the extracted samples are correct. A word or character is referred to as correct if and only if the word in the ground truth cropped image, the transformed captured image, and the original captured image is same and the ground truth text corresponding to these images also contains only that word. While evaluating, it is also taken into account that each image should exactly contain the same information. In addition to camera captured images, the proposed method is also tested on scanned images, where it has also achieved an accuracy of more than 99%. This means that almost all of the images are correctly labeled.

V. Conclusion

In this paper a system for large scale automatic generation of ground truth of camera captured document images is presented. The proposed approach is fully automatic and does not require any human intervention for labeling. Furthermore, it is not limited to the camera captured documents and can also be applied to scanned images. Evaluation of the generated ground truth shows that our system can be successfully applied to generate very large scale dataset automatically. We are working on the development of a very large scale camera captured words dataset which can be used for evaluation as well as training of different OCRs on camera captured document images. In future, we also plan to build dataset for different languages, including Japanese, Arabic, Urdu, and other Indic scripts, as there is already a strong demand for OCR of different languages e.g., Japaneses [13], Arabic [14], Indic scipts [15], Urdu [16], etc., and each one needs a different dataset specifically built for that language.

Acknowledgment

This work is supported in part by CREST and JSPS Grant-in-Aid for Scientific Research (B)(22300062) as well as Japan Student Services Organization (JASSO).

References

[1] T. Tsuji, M. Iwamura, and K. Kise, "Generative learning for character recognition of uneven lighting," *In Proc. of KJPR*, pp. 105–106, Nov. 2008.

[2] H. Ishida, S. Yanadume, T. Takahashi, I. Ide, Y. Mekada, and H. Murase, "Recognition of low-resolution characters by a generative learning method,," *In Proc. of CBDAR*, pp. 45–51, 2005.

[3] T. Strecker, J. van Beusekom, S. Albayrak, and T. Breuel, "Automated ground truth data generation for newspaper document images," in *In Proc. of 10th ICDAR*, Jul. 2009, pp. 1275–1279.

[4] J. v. Beusekom, F. Shafait, and T. M. Breuel, "Automated ocr ground truth generation," in *In Proc. of DAS*, 2008, pp. 111–117.

[5] T. Kanungo and R. Haralick, "Automatic generation of character groundtruth for scanned documents: a closed-loop approach," in *In Proc. of the 13th ICPR.,*, vol. 3, Aug. 1996, pp. 669–675 vol.3.

[6] T. Kanungo and R. M. Haralick, "An automatic closed-loop methodology for generating character groundtruth for scanned images," *TPAMI*, vol. 21, 1998.

[7] G. Zi, "GroundTruth Generation and Document Image Degradation," University of Maryland, College Park, Tech. Rep. LAMP-TR-121,CAR-TR-1008,CS-TR-4699,UMIACS-TR-2005-08, May 2005.

[8] K. Takeda, K. Kise, and M. Iwamura, "Memory reduction for real-time document image retrieval with a 20 million pages database," *In Proc. of CBDAR*, pp. 59–64, Sep. 2011.

[9] D.-W. Kim and T. Kanungo, "Attributed point matching for automatic groundtruth generation," *IJDAR*, vol. 5, pp. 47–66, 2002.

[10] T. M. Breuel, "A practical, globally optimal algorithm for geometric matching under uncertainty," in *In Proc. of IWCIA*, 2001, pp. 1–15.

[11] H. Baird, "The state of the art of document image degradation modelling," in *Digital Document Processing*, ser. Advances in Pattern Recognition, B. Chaudhuri, Ed. Springer London, 2007, pp. 261–279.

[12] H. S. Baird, "The state of the art of document image degradation modeling," in *In Proc. of 4th DAS*, 2000, pp. 1–16.

[13] S. Budiwati, J. Haryatno, and E. Dharma, "Japanese character (kana) pattern recognition application using neural network," in *In Proc. of ICEEI*, Jul. 2011, pp. 1–6.

[14] A. Zaafouri, M. Sayadi, and F. Fnaiech, "Printed arabic character recognition using local energy and structural features," in *In Proc. of CCCA,*, Dec. 2012, pp. 1–5.

[15] P. P. Kumar, C. Bhagvati, and A. Agarwal, "On performance analysis of end-to-end ocr systems of indic scripts," in *In Proc. of DAR*, ser. DAR '12. New York, NY, USA: ACM, 2012, pp. 132–138.

[16] S. Sardar and A. Wahab, "Optical character recognition system for urdu," in *In Proc. of ICIET*, Jun. 2010, pp. 1–5.