社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Automatic Word Ground Truth Generation for Camera Captured Documents

Sheraz AHMED† Koichi KISE†† Masakazu IWAMURA†† Marcus LIWICKI††† Andreas DENGEL†

† German Research Center for Artificial Intelligence (DFKI) Trippstadter Straße 122, 67663 Kaiserslautern, Germany

†† Osaka Prefecture University Gakuenmachi 1-1,Naka-ku,Sakai-shi, 599-8531 Japan

††† University of Fribourg Boulevard de Pe'rolles 90, 1700 Fribourg, Switzerland

E-mail: †firstname.lastname@dfki.de, ††{kise,masa}@cs.osakafu-u.ac.jp, † † †firstname.lastname@unifr.ch

Abstract A database for camera captured documents is useful to train OCRs to obtain better performance. However, no dataset exists for camera captured documents because it is very laborious and costly to build these datasets manually. In this paper, a fully automatic approach allowing building the very large scale (i.e., millions of images) labeled camera captured documents dataset is proposed. The proposed approach does not require any human intervention in labeling. Evaluation of samples generated by the proposed approach shows that more than 97% 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, perspective distortion, etc.

Key words Ground truth, Locally Likely Arrangement Hashing (LLAH), Camera Captured Documents, Perspective distortion, Blur

1. Introduction

Text recognition is an important area in analysis of camera captured documents, as a lot of services can be provided, if text is recognized. Existing OCRs can not be applied to camera captured documents because different distortions 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 unavailability of any big dataset for camera captured document images.

The main problem in dataset generation is its labeling. Manual labeling of each word and/or character in captured images is very laborious and costly. One possible solution could be to use different degradation models 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. Therefore there is a strong need for automatic labeling of large scale camera captured documents.

In past the problem of automatic ground truth generation for camera captured document images is not addressed by document analysis community. The electronic version of camera captured image and its alignment to the captured image are needed, is need 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 full 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, e.g., blur, occlusion, perspective distortion, rotation, camera noise, etc.

In this paper, we propose an approach for automatic word 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 con-

tained 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 simultaneously 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.

The rest of the paper is organized as follows. A summary of existing methods for automatic ground truth genertation is presented in section 2.. Details about the proposed method to automatically generate ground truth of camera captured documents is presented in section 3.. Evaluation results are shown in section 4.. Finally, Section 5. concludes the paper.

2. Related Work

This section provides the summery of different approaches that are available for automatic generation of ground truth. First, approaches for automatic ground truth generation of scanned documents are presented. Then available 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 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.

[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. Robust Branch and Bound search (RAST) [10] 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. Similarly, [4] proposed automatic ground truth generation for OCR using RAST. 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.

In scanned documents complete document image is available, and therefore, transformation between scanned and ground truth image can be computed using different alignment techniques [3]〜[6]. However, camera captured documents 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, it is also a possible to use different image degradation models [11], [12]. An advantage of degradation models is that everything remains electronic. 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. Researchers are still of different opinion that either degradation models are true representative of real world data or not.

3. Methodology

The proposed approach for automatic ground truth generation primarily based on document image retrieval system. Figure 1 shows the complete flow of proposed approach. To perform retrieving the electronic version of the camera captured document image document level matching is performed using the document retrieval system(Figure 1(a), Section 3. 1) . Figure 2 shows the LLAH based document retrieval system which is used for retrieving the same document. In addition to the retrieved document, area which corresponds to the camera captured document is also computed by LLAH. Using this corresponding area, part level processing is performed on the retrieved and the captured image (Figure 1(b), Section 3. 2). Finally, the parts after transformation 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 3. 3).

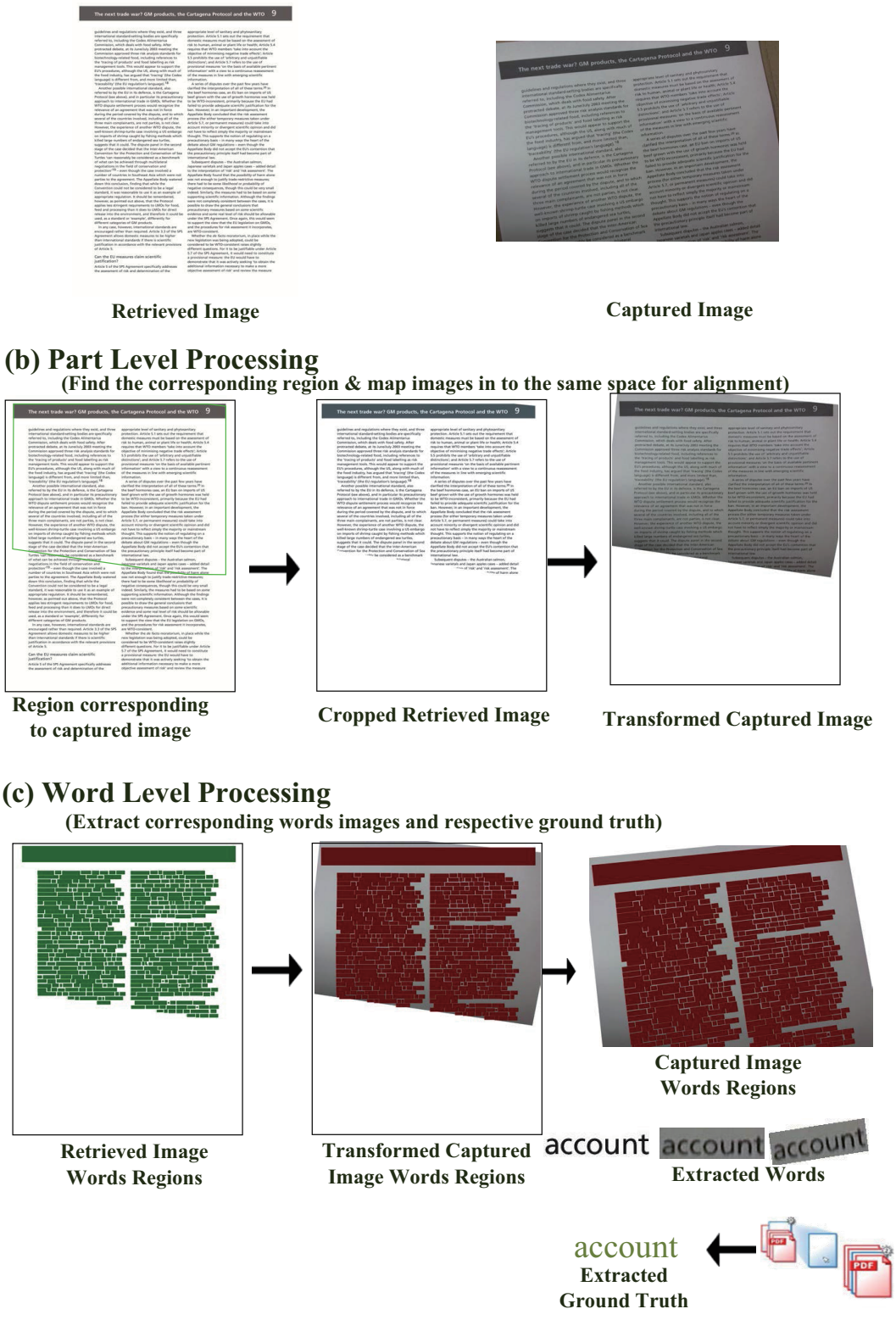

Figure 1: Automatic Ground Truth Extraction Process

3.1 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].

Figure 2 shows the LLAH based document retrieval system. Document images are extracted from their correspond-

Figure 3: Extracted Words

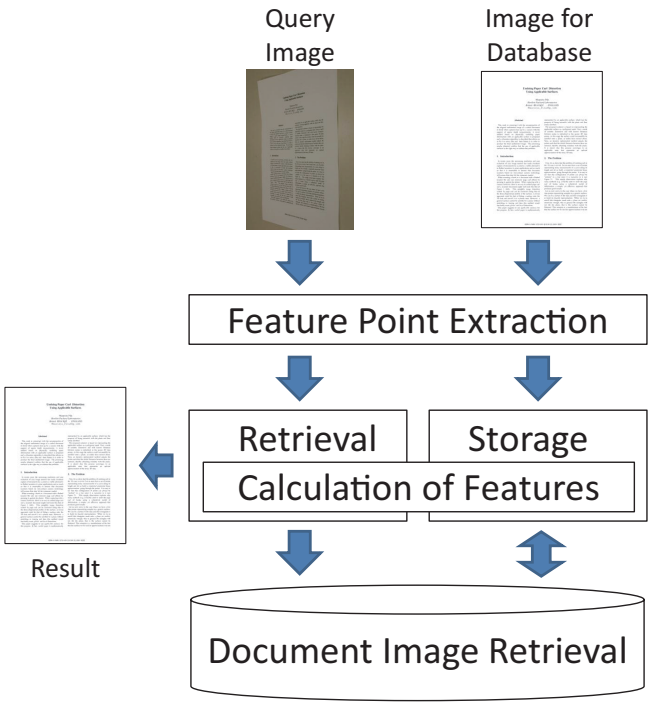

Figure 2: LLAH Document Retrieval

ing PDF files for building up the database for LLAH. These documents include, 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].

3.2 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 area of electronic document which corresponds to the camera captured image is computed 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 convert 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 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.

3.3 Word Level Processing

For word level processing (Figure 1(c)), word regions need to be found. 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 centers (d_{cent}) and width (d_w) is computed. All of the boxes for which d_{cent} and d_w is less than θ_c and θ_w respectively, are referred to as boxes for the same word in both the images. Here, θ_c and θ_w refers to the bounding box distance threshold for centers and width, respectively.

The distance between centroids and width of bounding boxes is computed using the following equations. where (x_{capt}, y_{capt}) and (x_{ret}, y_{ret}) refers to centers of bounding boxes in the transformed captured and the cropped retrieved image.

$$d_{cent} = \sqrt{(x_{capt} - x_{ret})^2 + (x_{capt} - y_{ret})^2} < \theta_c$$

$$d_w = \sqrt{(W_{capt} - W_{ret})^2} < \theta_w$$

We have used $\theta_d = 5$ and $\theta_w = 5$ pixels, which means, if two boxes are at almost the same position in both the images and their width is also almost the same then they correspond to the same word in the 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.

Figure 4: Words marked with corner

The word extracted from the transformed captured image is already normalized for different transformations which were present in the captured image. However, an image with different transformations and distortions can be used for training of systems capable of dealing with 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 text is then saved as ground truth text file along with the word images.

As mentioned before that the captured image contains only a part of the document. A region in retrived image corresponding to that part could be any irregular polygon (see a transformed and a cropped image in Figure 1(b)). Therefore it is possible that the characters and words which occur near the border of the region are partially missing (Figure 4). If these words are included directly in the dataset, they can cause problem during training, e.g., if lower part of p is missing then in some fonts it looks like D which causes confusion between different characters during training. To solve this problem, all the words and characters which occur near a border are marked with a flag in their names. This enables us to separate these words so that they can be treated separately if included in training.

4. Evaluation

To check correctness and quality of the generated ground truth, manual evaluation is performed. First, ground truth is generated for a few camera captured documents using proposed approach. These documents include magazines, proceedings, articles, etc. Words are extracted from these captured images using the proposed technique. 1000 samples are selected randomly from these extracted images. A person has manually inspected all of these samples to find out errors. This manual check reveals that 97.5% of the extracted samples are correct. A word 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 text corresponding to these images also contains only that word. In addition to camera captured images, the proposed method is also tested on scanned images, where it has also achieved an accuracy of 97.5%. This means that almost all of the images are correctly labeled.

5. Conclusion

In this paper a system for large scale automatic generation of word ground truth of camera captured document images is presented. The proposed approach is fully automatic and does not require any human intervention for labeling. In addition, it 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 which contains labeled camera captured words. 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 are planning to improve the proposed approach in terms of accuracy.

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).

Reference

[1] T. Tsuji, M. Iwamura, and K. Kise, "Generative learning for character recognition of uneven lighting," Proceedings of Third Korea-Japan Joint Workshop on Pattern Recognition (KJPR2008), 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,," Proc. 1st Intl. Workshop on Camera-Based Document Analysis and Recognition, pp.45–51, Aug. 2005.

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

[4] J.v. Beusekom, F. Shafait, and T.M. Breuel, "Automated ocr ground truth generation," Document Analysis Systems, 2008. DAS '08. The Eighth IAPR International Workshop on, pp.111–117, sept. 2008.

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

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

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

[8] K. Takeda, K. Kise, and M. Iwamura, "Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah," Document Analysis and Recognition (ICDAR), 2011 International Conference on, pp.1054–1058, sept. 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 Proc. of IWCIA, pp.1–15, 2001.

[11] HenryS. Baird, "The state of the art of document image degradation modelling," Digital Document Processing, ed. by BidyutB. Chaudhuri, pp.261–279, Advances in Pattern Recognition, Springer London, 2007.

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