AUTOMATIC CHARACTER LABELING FOR CAMERA CAPTURED DOCUMENT IMAGES

Wei Fan[†,‡], *Koichi Kise*[†] *and Masakazu Iwamura*[†]

[†]Osaka Prefecture University, Japan
[‡]Fujitsu Research & Development Center Co. Ltd., Beijing, 100027, China

ABSTRACT

Character groundtruth for camera captured documents is crucial for training and evaluating advanced OCR algorithms. Manually generating character level groundtruth is a time consuming and costly process. This paper proposes a robust groundtruth generation method based on document retrieval and image registration for camera captured documents. We use an elastic non-rigid alignment method to fit the captured document image which relaxes the flat paper assumption made by conventional solutions. The proposed method allows building very large scale labeled camera captured documents dataset, without any human intervention. We construct a large labeled dataset consisting of 1 million camera captured Chinese character images. Evaluation of samples generated by our approach showed that 99.99% of the images were correctly labeled, even with different distortions specific to cameras such as blur, specularity and perspective distortion.

Index Terms— Image registration, Document retrieval

1. INTRODUCTION

The increasing availability of high-performance mobile cameras has created a tremendous opportunity for document image recognition and analysis techniques [1]. The ubiquitous nature of camera phones makes it easy for mobile users to capture a document image without flat-bed scanners. Although OCR technologies have been around for decades, many commercial softwares are designed specifically for scanned images and give poor results on camera captured documents. Researchers in both industry and academia unanimously ascribed this performance degradation to much less constrained imaging conditions, e.g. non-uniform illumination, arbitrary viewing angles, severe paper distortion, not to mention the unpredictable sensor noise brought by cameras.

Camera based document character recognition is indeed a hard pattern classification task due to the above difficulties. However, we also get many exciting news from some equally challenging tasks. Today's advanced deep neural networks [2] have made groundbreaking improvements across a variety of applications including generic image classification, video analysis, face identification and natural scene text recognition. The key to such successes is the availability of massive amounts of training data, and powerful and efficient parallel computing architecture. We believe the bottleneck of practical camera based document recognition technologies lie on the lack of massive labeled training data. The collection of character level groundtruth, which is a set of camera captured character images and the corresponding character codes, is crucial for advancing the research activities in many aspects. Unfortunately, manual collection of accurate groundtruth for huge real data is not practical due to its prohibitively high cost and time consumption. The reCAPTCHA [3], a human-based character labeling method, transcribes text by asking web users to decipher scanned words from books that OCR failed to recognize. The biggest sources of errors in [3] are the imperfect word segmentation and alignment brought by OCR programs.

One shortcut solution avoiding real data labeling is to use different degradation models [4] to generate a large set of synthetic data. However, researchers still debated a lot whether these artificial data are really useful for training practical classifiers with good generation capability. It is difficult, if not impossible, to develop analytic representations of camera captured character degradation model. Large scale labeled real data make it possible to realize the artificial-vs-real competition to support either side of the debate. Besides, labeled real data is also useful for building generative models to create real-look data from clean data. The reverse mapping from degraded data to ideal data will find applications in document image restoration and enhancement tasks.

Most of the previous work on automatic OCR groundtruth generation is limited to scanned document images [5][6]. The degradations and distortions associated with camera-captured images, however, are very different than scanned images. Ahmed et al. [7] propose an approach for automatic groundtruth generation of camera captured document images using a document image retrieval system. To automatically generate groundtruth, the electronic version of a camera captured image and its alignment to the captured image are needed, so that the captured image can be labeled using the groundtruth information embedded in the electronic version.

One essential step of [7] is to find out the registration which can align the idea document image to the captured image. A perspective transformation is estimated using the

corresponding matched points between the query and the retrieved document image. This global model is not robust under crinkled and warped degradations since it relies on estimating a single homography which only fits a portion of the page. To avoid false groundtruth information in parts which are not perfectly aligned, bounding boxes of connected components are extracted from the smoothed query and template images, and then geometric rules are set to verify the correctness of matching. This trivial criterion will fail on images with large perspective distortion where different parts of the captured image have different levels of blurring. It is difficult to use a single Gaussian filter to extract bounding boxes of words.

The goal of this work is to find a more robust way of transforming the original page to match the captured image, improving both the coverage and accuracy of the obtained labeled data. The contributions of our work are as follows:

1) A more elastic **non-rigid alignment** method is proposed to fit the captured document image which can be either flat or with moderate warping or folding.

2) A robust local character matching is used to allow pixel accurate alignment of the captured image with the electronic document. Comparing to [7], it is more applicable to Asian scripts including Chinese, Japanese and Korean, since the connected components extracted in [7] often lead to over-segmentation or under-segmentation due to the little difference between inter-character space and intra-character space.

3) We construct a dataset of 1 million camera captured Chinese character images with groundtruth information. Our dataset can be used to train advanced powerful character classifiers and evaluate the performance of different OCR systems. It also helps understanding the camera based character degradation models.

2. GROUDTRUTH GENERATION USING DOCUMENT IMAGE RETRIEVAL

This section reviews the document retrieval based groundtruth generation methodology [5][6][7]. The approaches for OCR groundtruth generation typically use electronic documents e.g. PDF as a starting point. The electronic document is used to generate a printed version, a template image of the document and extract the character bounding boxes and the corresponding ASCII codes. Figure 1 shows the complete flow of the approach proposed in [7]. Document level matching (Figure 1(a)) is performed by retrieving the corresponding template image of the camera captured query image using a document retrieval system. Among various document retrieval systems [8][9][10][11], Locally Likely Arrangement Hashing (LLAH) [8] is adopted for its potential to extract the identical document from the database of 20 million images

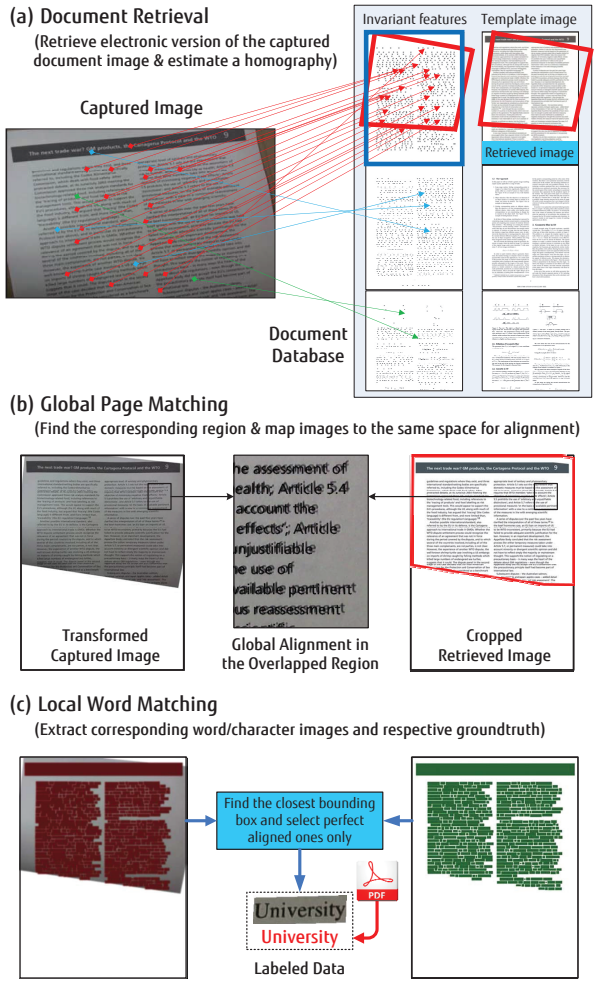

Fig. 1. Automatic groundtruth extraction process in [7] (slightly modified from the original paper).

with an accuracy of more than 99% [12]. Along with the retrieved document, the region which corresponds to the camera captured document is also estimated by LLAH. Using this corresponding region, global page transformation and cropping are performed on the query image and retrieved image respectively (Figure 1(b)). Finally, the transformed parts of both images are used for word level matching and verification to extract the perfect aligned words in both images and their groundtruth from PDF (Figure 1(c)).

LLAH is an image retrieval method targeted toward camera captured documents. A document image is represented as a set of local descriptors attached at the feature points extracted from the image. Since the descriptors consist of multiple geometric invariants, LLAH can deal with images which suffer from perspective distortion. The invariants are calculated from local arrangements of the feature points

which are centroids of word regions. These points are very stable on English documents since they have a space between words. However, discriminative descriptors are hardly extracted from Asian languages including Chinese and Japanese due to no separation between words and periodic arrangements of characters. An extension work for retrieval of documents in various languages was proposed in [13] where still relatively low accuracy was observed for Chinese documents.

Notice the high document retrieval accuracy is a necessary condition for good character labelling, but not a sufficient one. If the transformation (a homography in [7]) from the captured image to the template image is unreliable, most of the characters in the image will be neglected in the labeling process. [7] makes a strong assumption that the captured document is a flat paper. In many cases, warping and crinkling might destroy the planar assumption since the document page might not be as perfectly flat as its digital version. As a result, the global projection model with a single homography collapses since it can no longer be used to map from one point set to another. For these reasons, we attempt to find a more elastic **non-rigid alignment** that makes the entire page fit the image as well as possible.

3. THE PROPOSED METHOD

Our approach maintains a sequence of local perspective transformations which gradually fits the captured query image I_q to the template image I_t. Since it is still important to have a good starting point for the fitting, the new approach adopts the homography $\hbar_{llah}$ estimated by LLAH as the initial transformation $\hbar$. We update this estimate by propagating a set of **active feature correspondences** from the previously labeled area L to the unlabeled area so far U within the whole page. A detailed algorithm description is listed in Algorithm 1.

Step 1 first transforms I_q to a rectified version $\tilde{I_q}$ based on $\hbar$ so that $\tilde{I_q}$ lies in the same space with I_t.

Step 2-a) establishes the correspondence between a template character patch p_i and a query character patch q_i. A local template matching is performed to find this correspondence where we essentially look for the optimal patch $\tilde{q_i}$ in the transformed query image $\tilde{I_q}$. Template matching is effective to find the translation vector $\vec{d_i}$ since the perspective distortion is already corrected in $\tilde{I_q}$.

Step 2-c) creates a so-called **active set** A which consists of those matched characters with large displacement between p_i and $\tilde{q_i}$. These characters tend to locate in the boundary of the region where the homography $\hbar$ covers. It is also the frontier of the currently labeled area of the whole page.

Step 4 performs a region segmentation based on the positional and optical flow (in terms of $\vec{d_i}$) information of characters. The feature correspondences in the dominant region (the largest cluster in G) are used to estimate the new homography in Step 5.

Algorithm 1 Document character labeling framework

Input: Query image I_q; Template image I_t; Character set $\Omega = \{c_i\}$ where each character c_i is associated with its groundtruth: centroid (x_i, y_i), bounding box r_i and label l_i

Initialize: Labeled image set $\Psi = \emptyset$; Labeled set $L = \emptyset$; Unlabeled set $U = \Omega$; Active set $A = \emptyset$; Homography $\hbar = \hbar_{llah}$; A list of character clusters $G = \emptyset$; Free parameters (t_1, t_2)

Repeat until the size of L does not increase

1: Transform I_q to $\tilde{I_q}$ based on $\hbar$ and overlap $\tilde{I_q}$ with I_t

2: For each character $u_i \in U$

a) Extract from I_t a small patch p_i corresponding to r_i, perform local template matching in $\tilde{I_q}$ within a neighborhood of (x_i, y_i), and record the matched patch $\tilde{q_i}$, its corresponding patch q_i in I_q calculated by the inverse mapping $\hbar^{-1}$, the matching error e_i and translation vector $\vec{d_i} = (dx_i, dy_i)$;

b) If $e_i < t_1$, move u_i from U to L, add $(p_i, q_i, \tilde{q_i}, l_i)$ to Ψ;

c) If $|\vec{d_i}| > t_2$, add u_i to A;

3: For each $a_i \in A$, obtain geometric attributes (dx_i, dy_i, x_i, y_i);

4: Perform k-means clustering on A and push the clusters into G;

5: Re-estimate $\hbar$ by fitting a homography based on feature correspondences associated with the largest cluster in G;

6: $A = \emptyset$

Output: Labled image set Ψ

Figure 2 shows one example of the intermediate results obtained by our approach. Although the query image captures a warpped document, after 3 rounds of iteration, all the characters in the document have been successfully labeled with groundtruth information.

4. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our method, we performed some comparative experiments on a variety of document images captured by different mobile cameras. We compared our method with the results of [7] and its modified version by only replacing the geometric verification of [7] with our local template matching. Both the modified [7] and [7] relied on a single estimated homography while our method iteratively updated the homography according to the matching result of each round.

To build a promising training set, we collected 50 single page documents including magazines, newspaper, proceedings and brochures. We focused on Asian languages since to our knowledge the automatic groundtruth generation of such documents was not addressed in the past. Two groups of testing images were captured: a) D1: 50 flat document images, b) D2: 50 physically distorted document images, including paper warping, crinkling, folding and combination of them. We evaluated the labeling performance in terms

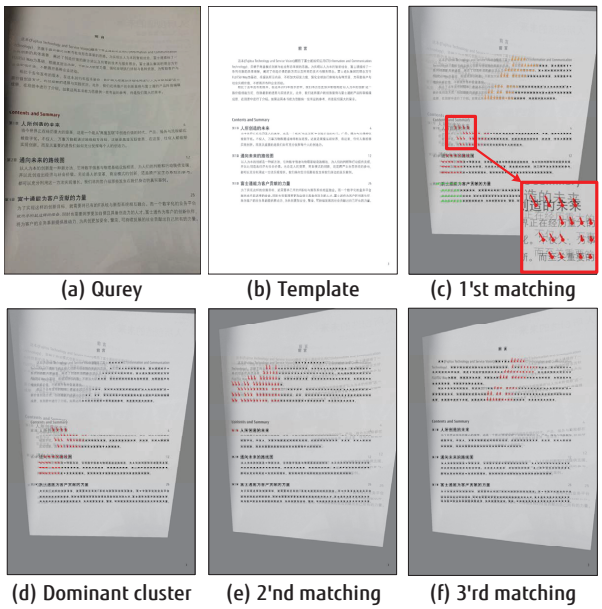

Fig. 2. A query image (a) was initially projected to the template (b) using homography $\hbar_{llah}$. Three clusters were obtained (different colored arrows) in the 1'st matching result (c), where the dominant cluster (red arrows in (d)) was used to update the homography and trigger the 2'nd matching (e). The final result (f) showed all characters were labeled although parts of them had large matching displacements (highlighted as red arrows). In (c-f) all labeled characters are either marked as black dots (which means small displacements) or colored arrows (which means large displacements). Notice that we only matched Chinese characters in this example.

of two measures: a) Recall rate $= \frac{\#labeled\_chars}{\#all\_chars\_in\_document}$, b) Precision $= \frac{\#correctly\_labeled\_chars}{\#all\_labeled\_chars}$. For all the testing images, a full document page was captured to make the calculation of recall rate meaningful.

The experimental results on two image datasets are shown in Table 1. For both datasets, our method outperforms [7] and its modified version. By replacing geometric verification with local template matching, character labelling performance is improved for both recall and precision. The superiority of our method is especially noticeable on distorted documents. Figure 3 shows some typical examples: from left to right, a warped Chinese magazine, a warped Korean paper, a crinkled Japanese paper and a folded Japanese paper.

Based on the proposed method, we collected a dataset of 1 million Chinese character images from different camera captured documents. Some representative samples are shown in Figure 4. Manual evaluation of a randomly selected 5000 samples revealed that more than 99.99% of the extracted samples were correctly labeled. Our dataset is still expanding by adding more documents, which in future can be used to train and evaluate advanced OCR systems.

Table 1. Comparative results of character labeling by our method, the modified version of [7] and [7].

	Flat paper (D1)		Distorted paper (D2)	
	Recall rate	Precision	Recall rate	Precision
Our method	0.978	1.000	0.895	0.998
[7] modified	0.827	1.000	0.668	0.998
[7]'s method	0.767	0.999	0.538	0.997

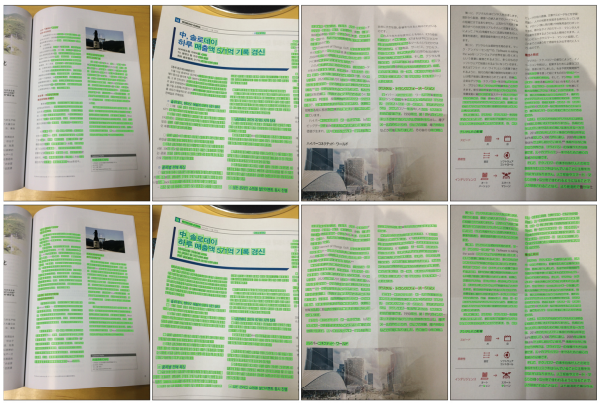

Fig. 3. Document labelling results of [7]'s (row 1) and our method (row 2). The labeled characters are highlighted by green boxes.

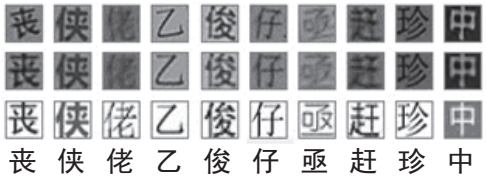

Fig. 4. Representative images of the collected dataset. From top to bottom: samples from captured images, samples from transformed captured images, samples from template images and groundtruth characters.

5. CONCLUSION

This paper presents a system for automatic groundtruth generation for camera captured document images. The approach is fully automatic and tolerant towards the typical paper distortions including severe warping, crinkling and folding. Compared to the related methods, it is more suitable for labeling documents in Asian languages. Our system can be successfully applied to generating very large scale dataset automatically, which is crucial for evaluating and training different OCR systems on camera captured documents.

6. REFERENCES

[1] Jian Liang, David Doermann, and Huiping Li, "Camera-based analysis of text and documents: a survey," *International Journal of Document Analysis and Recognition*, vol. 7, no. 2–3, pp. 84–104, 2005.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[3] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum, "reCAPTCHA: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[4] Henry S. Baird, "The state of the art of document image degradation modeling," in *Proc. of IAPR DAS'00*, pp. 1–16. 2000.

[5] Joost van Beusekom, Faisal Shafait, and Thomas M Breuel, "Automated ocr ground truth generation," in *Proc. of IAPR DAS'08*, pp. 111–117. 2008.

[6] Tapas Kanungo and Robert M Haralick, "An automatic closed-loop methodology for generating character groundtruth for scanned documents," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 21, no. 2, pp. 179–183, 1999.

[7] Sheraz Ahmed, Koichi Kise, Masakazu Iwamura, Marcus Liwicki, and Andreas Dengel, "Automatic ground truth generation of camera captured documents using document image retrieval," in *Proc. of IAPR ICDAR'13*, pp. 528–532. 2013.

[8] Tomohiro Nakai, Koichi Kise, and Masakazu Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," in *Proc. of IAPR DAS'06*, pp. 541–552. 2006.

[9] Sumantra Dutta Roy, Kavita Bhardwaj, Rhishabh Garg, and Santanu Chaudhury, "Camera-based document image matching using multi-feature probabilistic information fusion," *Pattern Recognition Letters*, vol. 58, pp. 42–50, 2015.

[10] Jorge Moraleda, "Large scalability in document image matching using text retrieval.," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 863–871, 2012.

[11] Xu Liu and David Doermann, "Mobile retriever: access to digital documents from their physical source," *International Journal of Document Analysis and Recognition*, vol. 11, no. 1, pp. 19–27, 2008.

[12] Kazutaka Takeda, Koichi Kise, and Masakazu Iwamura, "Memory reduction for real-time document image retrieval with a 20 million pages database," in *Proc. of CBDAR'11*, pp. 59–64. 2011.

[13] Tomohiro Nakai, Koichi Kise, and Masakazu Iwamura, "Real-time retrieval for images of documents in various languages using a web camera," in *Proc. of ICDAR'09*, pp. 146–150. 2009.