A MULTI-SCALE APPROACH TO EXTRACT MEANINGFUL ANNOTATIONS FROM DOCUMENT IMAGES

Yang Lei, Jian Fan, Jerry Liu

HP Labs, Palo Alto, CA 94304 USA

ABSTRACT

We propose a multi-scale approach to extract annotations from document images in a meaningful way. It compares the rectified captured image with original image, which can be obtained through image retrieval technology, at various resolutions, in order to remove the noise caused by non-uniform distortions, such as camera lens distortion and document surface curvature, while preserving the true annotations. It also provides users lots of flexibility with a voting scheme and potentially different weights at different resolution levels. In addition, we analyzed the final annotation image and found the meaningful pieces out of it, such as an image patch of a handwritten paragraph. This broadens the application of annotation extraction, and makes it easier to share the notes. It can be applied to various imaging systems, such as flatbed scanner, mobile phone camera, or fixed camera etc. Experimental results are presented and compared with previous approaches.

Index Terms— Annotation extraction, multi-scale, document imaging

1. INTRODUCTION

Extracting annotations in document images is an important research topic in document imaging. Assume the document is from a known database, image retrieval technology will find the original image. With removal of the pre-printed content, it can lead to a higher data compression rate, make content sharing, organizing, and archiving much easier, therefore has huge potential in applications such as office automation, education and training, document authentication, etc.

Annotations in print documents could be handwritten notes, or drawings added to the printed documents. Also, the imaging system used to capture them varies from flatbed scanner, fixed camera imaging system, to mobile camera. As a result the image may suffer from perspective distortion, uneven lighting, camera lens distortion, document surface curvature, etc. The major challenges for extracting annotations are duplicate content removal. A common approach is document registration.

Most previous work has been focusing on document image obtained by flatbed scanners, assuming there is no perspective distortion [1]. Also, lots of work has been on form dropout rather than general annotation extraction [2, 3]. Form dropout systems normally makes assumptions about the special structure in the document and annotations. Also, they usually only consider global registration by matching preprinted lines. However, for general documents with lots of texts, this is not enough. Some other work that addresses general annotation ex-traction calculate local displacement vectors for image blocks on a uniform grid in order to achieve better registration results [4]. But this may break the annotations into parts and result in artifacts.

There are lots of work on general image registration as well [5, 6]. They are mainly based on feature point registration. While it is good for getting an initial alignment between two images, annotation extraction requires higher registration accuracy than general image registration.

Assume there are perspective distortions, and other types of non-linear distortion are very small, such as camera lens distortion and surface curvature. Also assume the images are captured under reasonable lighting conditions. These are reasonable assumptions given the current technologies for digital cameras and normal reading/office environment. Under these assumptions, we proposes a multi-scale approach to extract annotations from print documents. Our system handles general document annotation extraction, using both feature points and image content comparison. The workflow is summarized in Fig. 1.

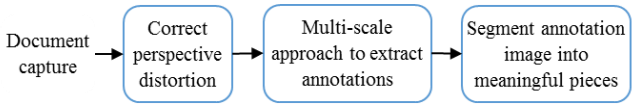

Fig. 1. Overall block diagram.

The rest of this paper is organized as follows. Section 2 describes the feature points based method to correct perspective distortions. The major part of this paper, the multi-scale approach to extract annotations is described in details in Sec. 3, followed by the process of segment annotation image into meaningful pieces in Sec. 4. Experimental results are shown in Sec. 5. Finally, the conclusions of our work can be find in Sec. 6.

2. FEATURE POINT BASED IMAGE RECTIFICATION

Feature points set provides a sparse representation of an image. A set of matching feature points from both the captured image $I_{capture}$, shown in Fig. 6(a), and original image $I_{original}$ can be used to rectify and align $I_{capture}$ with $I_{original}$ efficiently. Unlike other document rectification approaches, it does not rely on finding quadrilaterals in the captured image that should be rectangles in real world, such as document boundaries, and text lines [7, 8, 9]. Feature point based rectification tends to be more robust than quadrilateral based approach, since more than four feature points can be obtained for one document image in most cases.

Many image feature detector and descriptors have reasonable performance for document images, such as SIFT and SURF [10, 11]. LLAH (Locally Likely Arrangement Hashing) is an algorithm for document retrieval [12]. It uses word centroid and neighborhood based perspective invariant features descriptors, which can be used for image rectification as well.

The list of feature point pairs is then used to calculate the perspective transformation from the captured image to the original one. The following perspective transformation is used.

$$\vec{p}_{\text{original}} = H_{3\times3}\vec{p}_{\text{captured}}$$

where

$$\vec{p}_{\text{original}} = \begin{bmatrix} x_{\text{original}} \\ y_{\text{original}} \\ 1 \end{bmatrix}, \vec{p}_{\text{captured}} = \begin{bmatrix} x_{\text{captured}} \\ y_{\text{captured}} \\ 1 \end{bmatrix}.$$

RANSAC (RANdom SAmple Consensus) algorithm is used to remove outliers among the feature point pairs [13]. Then least square solution of the Homography $H_{3\times3}$ is calculated and used to rectify captured images. We denote the rectified image as I_{rect}.

3. MULTI-SCALE ANNOTATION EXTRACTION

Feature points set is just a sparse representation of an image, and therefore not enough for extracting annotations at high accuracy. At the meantime, the captured document image might suffer from distortions other than perspective distortion, such as camera lens distortion, non-flat document surface. The binary difference image between I_{rectB} and $I_{original}$ is shown in Fig. 6(b). As we can see, there are still lots of noise remaining, especially in regions where few feature points are detected, such as the lines on the lower half of the page. In this paper, a multi-scale approach is used to remove the non-annotation noise while preserving true annotations.

3.1. Preprocessing

First, both the rectified and original images are converted to binary images I_{rectB} and $I_{originalB}$ using adaptive threshold. Then text region of $I_{originalB}$ is dilated with an ellipsoidal structuring element as shown in Fig. 2. The dilated image $I_{originalD}$ is used as a mask to remove duplicate contents in the rectified binary image.

0	0	1	0	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
0	0	1	0	0

Fig. 2. Structuring element for morphological dilation on $I_{originalB}$.

3.2. Multi-scale annotation extraction

When we look for the differences between two images, we got more details with relatively high resolution images. However, more noises will appear compared with relatively low resolution images. The motivation of our multi-scale approach is to leverage the advantages of both cases and remove as much noise as possible, while preserving the true annotations.

In the multi-scale process, the two binary images we got from the preprocessing step, $I_{originalD}$ and I_{rectB}, are down-sampled by a factor of $\sqrt{2}$ on both x and y dimensions for next level down. This results in images with half of the area compared with the previous level. A fixed number of levels can be set or we keep down-sampling until the resolution of the lowest level is lower than some threshold. The illustration of this process is shown in Fig. 3.

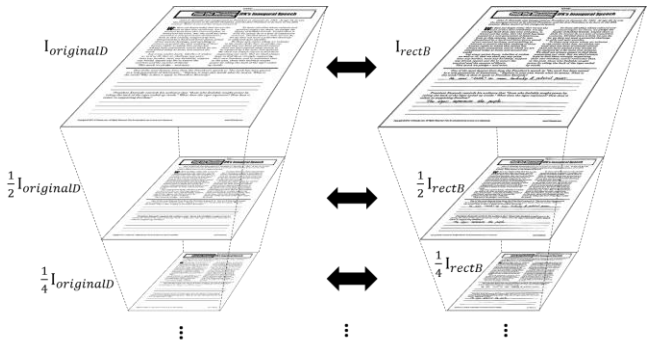

Fig. 3. Illustration of the multi-scale approach for extracting annotations.

At level i, we got two binary images of the same size, one from the rectified image and one from the dilated original image. The pixel-wise difference image between

them are calculated, and up-sampled to the highest resolution of the image set and get $I_{diff(i)}$.

When we reach the lowest level, a connected component based local adjustment is performed on the rectified binary image to compensate non-uniform distortions after the image rectification step. It costs less time to do the adjustment in the lowest level.

Each connected component in the down-sampled I_{rectB} defines a local region. The corresponding region in the down-sampled $I_{originalD}$ is its reference region. Within this region, the best translation and rotation parameters $\hat{\xi} = (\widehat{\Delta x}, \widehat{\Delta y}, \hat{\theta})$ from the down-sampled I_{rectB} to sown-sampled $I_{originalD}$ are searched within a certain range $\pm[\Delta X, \Delta Y, \Theta]$ to minimize the difference between the two binary images. The rectified image is then adjusted accordingly.

As a result, we have N difference images with the highest resolution, where N is the number of levels. We see the difference images as votes. A threshold can be set to determine if a pixel is a true annotation. For example, if more than half of the difference images think it's an annotation pixel, we label this pixel as part of the true annotation image. Also, different weights can be assigned to different levels. This gives the user lots of flexibility.

4. SEGMENT ANNOTATION IMAGE INTO MEANINGFUL PIECES

An entire annotation image has limited applications. Instead, meaningful pieces from the image, such as image patch of a whole sentence, are quite useful, since this makes it easier for people to store and share their content.

It's observed that all pixels in one connected component should belong to one piece. Also, several connected components that are close enough should be together, since people normally write in paragraphs. Based on these observations, we proposed an algorithm to understand the annotation image using connected component analysis and adaptive threshold. The flowchart is shown in Fig. 4.

First connected component analysis is applied to the whole image, and each of the component is assigned a different group ID to initialize the algorithm. Then the x and y distances between all groups are calculated. The distance is defined as the closest distance between any two pixels in the bounding box of two different groups. It applies to both x and y direction. That is, there are x distance and y distance between any two groups. For example, if the two bounding boxes overlap, then both x and y distance between the two connected components is zero.

Two adaptive thresholds on the distances along x and y axis are set for each group, based on the average width and height of all connected components within this group. If both x and y distance between two groups is smaller than the corresponding threshold, the two groups will be merged into one group. Then the adaptive thresholds and bounding box information is updated for the new group.

We apply the process described above iteratively. It converges until no change is made on the group assignment during one iteration. Intermediate results can be seen in Fig. 5. And the final results are shown in Fig. 6 (e).

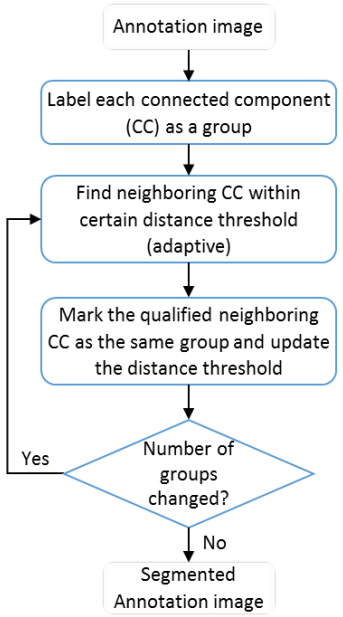

Fig. 4. Algorithm flowchart to segment annotation image into meaningful pieces.

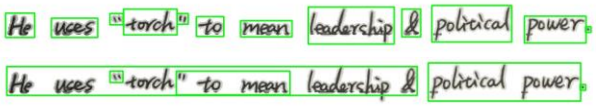

Fig. 5. Sample 1 – intermediate results of the segmentation algorithm. Upper: initial result. Lower: result after one iteration.

5. EXPERIMENTAL RESULTS

We experimented with images captured with mobile phone (as in Fig. 6(a)), and with a fixed camera imaging system (as in Fig. 7(a)). In both cases, the highest resolution for the multi-scale approach is 2200 by 1700 pixels and four levels are used. Written in C with OpenCV library, the whole algorithm runs for about 2 seconds on a laptop with Intel i5 processor and 16 GB memory.

Experimental results for mobile phone captured image and fixed camera captured image are shown in Fig. 6 and 7, respectively. They are organized in the same way. Figure 6 (b) is the difference image between I_{rectB} and $I_{original}$.

Results of the multi-scale annotation extraction method is shown in Fig. 6(c), with Fig. 6 (d) showing the segmentation result. Zoom-in version of the two meaningful pieces can be seen in Fig. 6 (e).

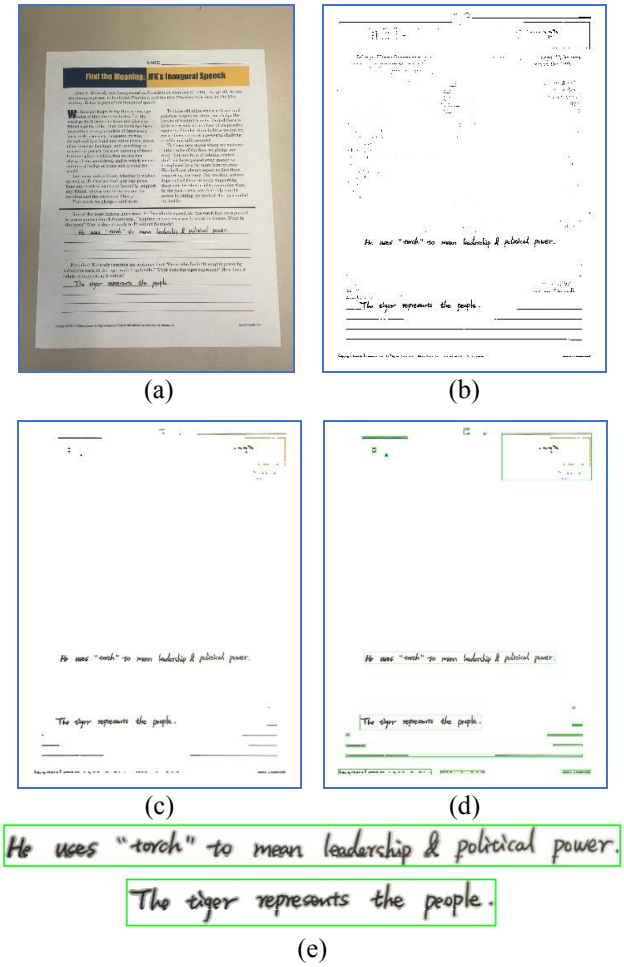

Fig. 6. Experimental results with mobile phone captured image. (a) Captured image. (b) Binary difference image between I_{rectB} and $I_{original}$. (c) Final annotation image. (d) Final annotation image with segmentation. (e) Zoom-in look at the meaningful annotation pieces.

We use the difference image, the result of Safari's approach, as the benchmark [2]. The percentages of noise removed from the difference image are shown in Table 1. The performance of our algorithm various for different imaging systems, but generally can remove more than half of the noise which remains from Safari's approach [2].

Table 1. Noise removal effectiveness.

	Fixed camera image	Mobile phone camera image
Noise removed	91.35%	52.40%

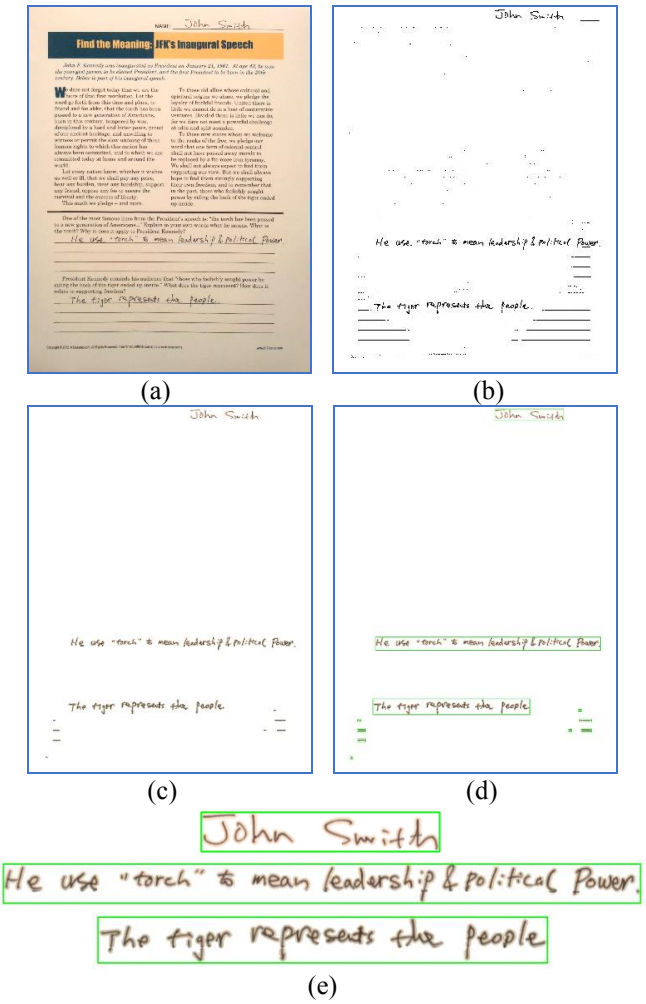

Fig. 7. Experimental results with image captured with fixed mobile camera. (a) Captured image. (b) Binary difference image between I_{rectB} and $I_{original}$. (c) Final annotation image. (d) Final annotation image with segmentation. (e) Zoom-in look at the three meaningful annotation pieces.

6. CONCLUSIONS

We proposed a multi-scale approach to extract annotations from document images in a meaningful way. The multi-scale approach compares the rectified captured image with original image at various resolutions. It provides users lots of flexibility with the voting scheme and potentially different weights for different resolutions. In addition, we analyzed the final annotation image and found meaningful pieces out of it, such as image patch of a paragraph.

From the experimental results above, we conclude that our multi-scale annotation extraction algorithm effectively removed noise caused by non-uniform image distortions, and preserved the true annotations at the meantime. Also, meaningful results are obtained from the segmentation algorithm.

7. REFERENCES

[1]. E. Dubois, and A. Pathak, “Reduction of bleed-through in scanned manuscript documents,” in Proceedings of the Image Processing, Image Quality, Image Capture Systems Conference, pp. 177-180, 2001.

[2]. R. Safari, N. Narasimhamurthi, M. Shridhar, and M. Ahmadi, "Form registration: a computer vision approach," in Proceedings of the IEEE International Conference on Document Analysis and Recognition (ICDAR), vol. 2, pp. 758-761, 1997.

[3]. J. Mao and K. Mohiuddin, “Form dropout using distance transformation,” in Proceedings of the IEEE International Conference on Image Processing (ICIP), vol. 3, pp. 328-331, 1995.

[4]. M.Ye, M. Bern, and D. Goldberg, "Document image matching and annotation lifting," in Proceedings of IEEE International Conference on Document Analysis and Recognition (ICDAR) pp.753-760, 2001,.

[5]. F. Isgro, and M. Pilu, “A fast and robust image registration method based on an early consensus paradigm,” Pattern Recognition Letters, vol. 25, no. 8, pp. 943-954, June 2004.

[6]. J. Yang, R.S. Blum, J.P. Williams, Y. Sun; C. Xu, "Non-rigid Image Registration Using Geometric Features and Local Salient Region Features," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.1, pp. 825-832, 2006.

[7]. P. Clark, and M. Mirmehdi, “Recognizing text in real scenes,” International Journal on Document Analysis and Recognition, vol. 4, no. 4, pp. 243-257, July 2002.

[8]. Y. Lu, and C. L. Tan, “A nearest-neighbor chain based approach to skew estimation in document images,” Pattern Recognition Letters, Vol. 24, Issue 14, pp. 2315-2323, October 2003.

[9]. Z. Zhang, and L. He, “Whiteboard scanning and image enhancement,” Journal of Digital Signal Processing, vol. 17, no. 2, pp. 414-432, March 2007.

[10]. D. G. Lowe, "Distinctive image features from scale-invariant key points," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[11]. H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, "SURF: Speeded Up Robust Features," Computer Vision and Image Understanding (CVIU), vol. 110, no. 3, pp. 346--359, 2008.

[12].T. Nakai, K. Kise, and M. Iwamura. “Use of Affine Invariants in Locally Likely Arrangement Hashing for Camera-Based Document Image Retrieval”, Document Analysis Systems (DAS) VII, vol. 3872, pp. 541-552, 2006.

[13]. M. Fischler, and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," Communications of the ACM, vol. 24, no. 6, pp. 381–395. June 1981.