Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Boosting OCR Accuracy Using Crowdsourcing

Shuo-Yang Wang[1], **Ming-Hung Wang**[12], and **Kuan-Ta Chen**[1]

[1]Institute of Information Science, Academia Sinica
[2]Department of Electrical Engineering, National Taiwan University

Abstract

Book digitizing is an important work in preserving ancient heritages. However, digitizing books contains a series of labor-intensive works, and one of them is to verify optical character recognition (OCR) outcomes. In this paper, we propose a crowdsourceable OCR verification method. Using our method, content holders are able to leverage the power of crowds to complete verification tasks and avoid *content leakage*. From the experiment results, our method is more efficient and reliable than the traditional method.

Introduction

Book digitizing is a major task in preserving ancient cultural heritages. Currently, book digitizing contains three steps:

1. Scan ancient books into page images.
2. Apply OCR to page images.
3. Verify the correctness of OCR outcomes manually.

Figure 1: Blurred word images in ancient books

The third step is a major bottleneck in digitizing books. Since the OCR technology does not have 100% correctness, the outcomes of OCR need to be verified manually, costing a large amount of labor works and time expenses. For example, *Siku Quanshu*, a classic Chinese book series, has 2.3 million pages in total, and each page contains about 350 words. That is, a total of 0.8 billion words in this book series need to be reviewed. This work costs approximately 800 years for a single reviewer to complete.

In order to speed up the verification step, we aim to leverage the *power of crowds* to solve this task. We propose a *crowdsoureable OCR verification method* to achieve this goal. Different from the traditional method, our method divides each page image into word images, and aggregates word images with the same recognized result in a page, as shown in Figure 2(b). Our method also *offloads* some recognition burden from reviewers to computers, and ensures the content will not be revealed when distributing review tasks to the crowds. From our experiment results, our method provides a more efficient and reliable way to verify the correctness of OCR outcomes than the traditional method.

Traditional and Crowdsourceable Methods

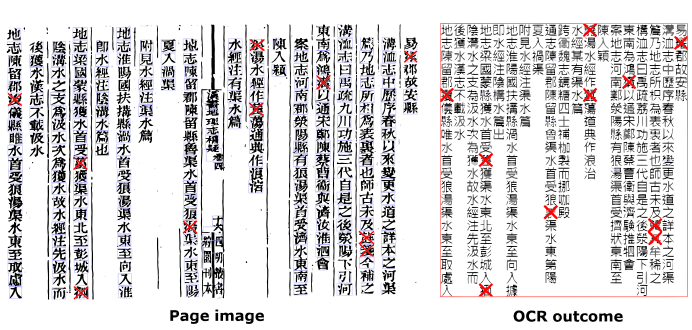
In documents printed by modern printing techniques, two same words are almost in the same shape. In contrast, ancient books are more blurred than modern documents, as shown in Figure 1. We consider that since ancient books were printed in block printing techniques, even two same words in ancient books have differences in their shapes; also, ancient books were made hundreds years ago, so it may decay with time. From our experiments, the OCR outcomes of ancient books possess higher error rates (25%) than modern printed documents (6%). To ensure the quality of digitizing books, it is important to verify OCR outcomes manually. In this section, we describe the traditional method and our crowdsourceable method for verifying OCR outcomes.

The Traditional Method

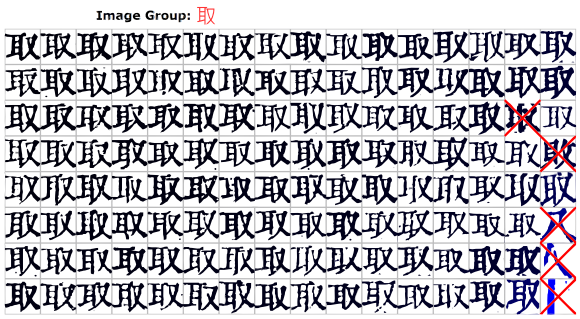
The traditional method arranges page images and their corresponding OCR results in the same page. Figure 2(a) presents a typical user-interface for reviewing OCR outcomes using the traditional method, the left side of the figure is the original page image, and the OCR outcome is on the right side. Reviewers verify the OCR outcome of the whole page word by word. In this manner, this method is not crowdsourceable, since distributing review tasks on the Internet may result in content leakage.

Our Proposal

In our method, we divide page images into word images, and each word image contains one word. We classify word images into word groups according to their OCR results. To each word group, we pre-process every word image and calculate its *dissimilar score*. The score is based on Dynamic Time Warping (DTW) algorithm, which is applied to Tamil character recognitions (Niels and Vuurpijl 2005) in previous works. In our research, we use DTW to help Chinese character recognitions. The dissimilar score is calculated according to the following steps:



(a) Traditional method



(b) Our method

Figure 2: Traditional and our crowdsourceable methods

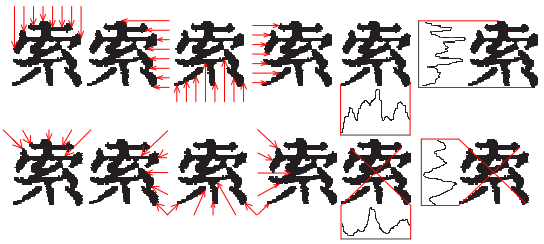


Figure 3: The 12 vectors of a word image

1. Extract the 12 vectors from a word image; each vector is calculated according to the boundary or density of the black area of a word image, as shown in Figure 3. For the k_{th} vector of image i and j, the DTW distance of the two vectors is represented as $dtw_{i,j,k}$.
2. Calculate the root mean square DTW distance from the 12 DTW distances of every two images i and j in the same image group.

$$RMS_{dtw_{i,j}} = \sqrt{\frac{1}{12}\sum_{k=1}^{12} {dtw_{i,j,k}}^2}$$

3. Add all RMS_{dtw} distances of each image i as its dissimilar score.

$$dissimilar.score_i = \sum_{j \in group} RMS_{dtw_{i,j}}$$

From the above steps, each word image has a dissimilar score relative to other word images in the same image group. We place the word images according to their dissimilar scores, as shown in Figure 2(b); words with higher dissimilar scores are placed on the right-bottom side of the page. Since word images with higher dissimilar scores are likely to be *incorrect-recognized word images*, reviewers are able to pay more attention on these word images. From the word image placement, a part of recognition burden is offloaded from reviewers to computers.

Performance Evaluation

We choose *Si Ming Cong Shu*, a classic book series written in traditional Chinese, as the material of our experiment. We recruit Internet users from Taiwan to join our experiment. A summary of our experiment is shown in Table 1.

Table 1: Comparisons between two methods

	Trad. method	Our method
# subjects	279	165
# words verified	190,415	102,530
Verification rate (images/sec)	**1.87**	**7.10**
Judgement error rate	**2.91%**	**2.17%**

From Table 1, our method has a significant improvement in image verification rates than the traditional method; also, reviewers using our method have lower judgement error rates than reviewers using the traditional method. We investigate the results as follows.

- Image verification rates: As traditional method users have to verify different kinds of word images in a page, we consider they may be more tired than users of our method, verifying the word images of the same OCR result in a page. The *weariness differences* between reviewers of the two methods may influence their verification rates.
- Judgement error rates: A major feature of our method is to place word images according to their dissimilar scores, so reviewers are able to pay more attention on the word images with higher dissimilar scores, which are likely incorrect-recognized by OCR. We consider this feature helps reviewers to decrease their judgement error rates.

Conclusion

In this paper, we propose a crowdsourceable method for verifying OCR outcomes when digitizing books. Using our method, content holders are able to distribute the verification works to the crowds and avoid content leakage. Our method also offloads a part of recognition burden from reviewers to computers. From our experiment results, our method is more efficient and reliable than the traditional method.

References

Niels, R., and Vuurpijl, L. 2005. Dynamic time warping applied to tamil character recognition. In *Proceedings of Eighth International Conference on Document Analysis and Recognition*, 730–734. IEEE.