# *Chinese document image retrieval based on recognition candidates*

*Xuhui Jia[1], Yong Xia[1], Rui Zhou[2] and Hongwei Liang[1]*
*School of Computer Science and Technology[1]*
*School of Science[2]*
*Harbin Institute of Technology*
*Harbin, P.R. China*
*xiayong@hit.edu.cn*

***Abstract*—For the sake of the low recognition rate for degraded Chinese document, the retrieval performance is not good if directly based on OCR result. In this paper, an indexing method with n-gram and recognition candidates is proposed to improve the performance of retrieval. For ease of test, this paper also presents a method to automatically generate ground-truth of imaged document, synthesized degraded document image and ground-truth of recognition candidates. Several synthesized document image collections on large-scale are built and used, and the experimental results show that the retrieval performance are improved for both collections with high or low OCR error rates.**

***Keywords-Chinese document image retrieval; indexing method with n-gram and recognition candidates; synthesized degraded document image.***

## I. INTRODUCTION

With the development of office automation, availability of working without paper hardcopy becomes increasingly popular. However, many historical materials are still saved in the form of hardcopy. As we know, it is not easy to conserve and manage paper document on giant-scale. Therefore, people usually convert them into imaged documents by photograph equipments (such as scanner or digital camera) in order to increase the validity of accessibility.

As to the document image retrieval, a number of approaches for this purpose are possible, including: (1) matching image features or constructing image feature based in pseudo-code [1, 2]; (2) optical character recognition (OCR) of document images. The first method is usually used for retrieving on-line documents or small-scale collections [3, 4]. While OCR-based retrieval is widely used because it can exploit easily existing text retrieval techniques. As existing recognition error in OCR document, especially for some badly degraded documents, so how to reduce the side effect from OCR error on retrieval has become a key issue.

Character n-grams offer a more efficient way of achieving some degree of approximate string matching. Tseng [5] applied multiple n-gram lengths for OCR-based retrieval for Chinese, finding that this improved retrieval performance over word-based indexing at OCR degradation. Sebastian [6] proposed a method using top n recognition candidates to categorize on-line handwritten documents, finding that the use of top n candidates can help counting occurrences of missing terms. In this paper, we investigate the method combination of n-gram and recognition candidates for Document image retrieval for Chinese. Since we have not discovered any similar work presently, this investigation is significant and valuable.

The remainder of this paper is arranged as follows. Section 2 presents the process of building test collection. Term-weighting methods based on OCR candidates are presented in Section 3. We show experimental results in Section 4, and finally, we draw conclusion and discuss future work in Section 5.

## II. CHINESE DOCUMENT IMAGE RETRIEVAL TEST COLLECTION

It is very difficult to collect large-scale Chinese document images with the expected degradation for retrieval, so in this Section, we present a method to automatically generate ground-truth of imaged document, synthesized degraded document image and ground-truth of recognition candidates. Four collections are built based on different recognition precision and recognition confidence evaluation methods.

### *A. System Flow for Test Collection Generation*

*1) Ground-truth generation procedure:*

*a)* For a given text document, system automatically generate PDF document after setting font, size, scan resolution and so on.

*b)* For each symbol rendered, we extract character ASCII codes, font and layout information from the metafiles. And after then generate ground-truth XML files.

*2) Degradation procedure:*

*c)* Obtaining ideal (without any noise) document images from the digital document clippings.

*d)* According to the degradation parameters that inputted, system degrade document image through corresponding degradation model (pixel drift to the document, add speckle, jitter, blur, bleed-through, rotation, lines and so on).

*3) Candidates generation procedure:*

*e)* According to corresponding ground-truth file, system obtains every character's image, and passes it to recognition engine.

*f)* The recognition engine yield as many candidates (10 in our case) as user needs and each candidate contains a recognition distance. Two approaches have been proposed to convert the distance into confidence.

978-1-4673-0024-7/10/$26.00 ©2012 IEEE

### B. *Candidate confidence*

Confidence measurement is an important issue in character recognition, which is a quantitative estimation of the potential correctness of recognition. For a given candidate, the higher the confidence, the lesser is its correlative recognition distance. In this section, two approaches are proposed to convert the distance into confidence.

Lee and Chen [7] used an empirical distance formula (EDF) to compute $p(c_k|x)$ as expressed in (1).

$$p(c_k|x) = \frac{score_k}{\sum_{i=1}^{K} score_i}, score_k = \frac{c}{d_k - d_1 + 1} \quad (1)$$

$$k = 1, 2...., K$$

The logistic regression model (LRM) was proposed Li and Ding [8] to directly convert the distance measurement of candidate $c_k$ into its confidence value. The LRM method defines $d_1, d_2, ..., d_k$ as independent variables and the correctness of $c_k$ as an independent variable (Y). If $c_k$ is the correct character, Y =1; otherwise, Y=0. The mean value of Y can be regarded as $p(c_k|x)$ which is expressed as:

$$p(c_k|x) = \left(\left(1 + \exp\left(\beta_0^k + \sum_{i=1}^{z} \beta_i^k d_i\right)\right)\right)^{-1}, 1 \le k \le K \quad (2)$$

Where $\beta_i^k$ is the regression coefficient, which can be estimated by maximum likelihood estimation [9] through the recognition results of some training samples. $z$ is the order of regression model.

### C. *Document Image Collection Generation*

An information retrieval test collection contains a set of documents, a set of topic description from which queries can be constructed, and a set of relevance judgments that identifies the relevant documents for each topic.

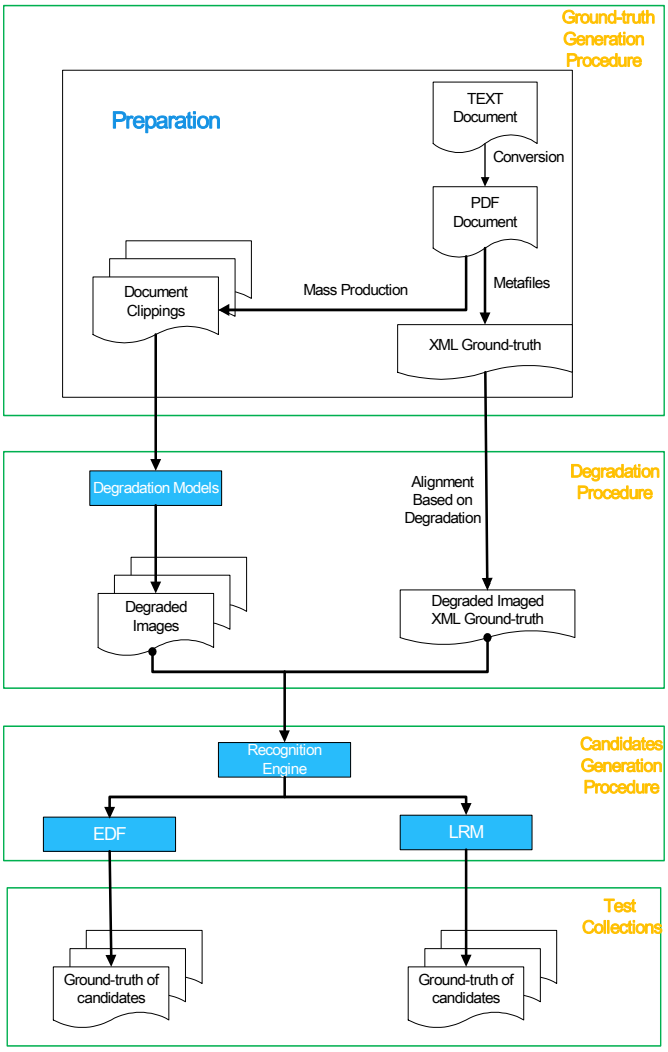


Figure1. System flow

```
<Word Value="来">
  <Font Name="SimSun" Size="23.4375" Color="Color [A=255, R=0, G=0,
  <WordCorners>
    <left-down x="836.5505" y="1504.47259471585" />
    <right-down x="859.988" y="1504.47259471585" />
    <upper-right x="859.988" y="1527.91009471585" />
    <upper-left x="836.5505" y="1527.91009471585" />
  </WordCorners>
  <CandidateList WordCount="10" Value="来火米未束未术采太耒">
    <Word Value="来" Probability="0.657659378906936" />
    <Word Value="火" Probability="0.0221176947321616" />
    <Word Value="米" Probability="0.0166089273498931" />
    <Word Value="未" Probability="0.00994307875104323" />
    <Word Value="束" Probability="0.00312365521829122" />
    <Word Value="未" Probability="0.00157949452008434" />
    <Word Value="术" Probability="0.000801768344384493" />
    <Word Value="采" Probability="0.000860031013969552" />
    <Word Value="太" Probability="0.000686287260204821" />
    <Word Value="耒" Probability="0.000704030465523489" />
  </CandidateList>
</Word>
```

Figure2. A sample ground-truth of recognition candidates

<top>
<num> 10
< title> Border Trade in Xinjiang
< description>
Xinjiang, Uigur, border trade, market,
< narrative >
A relevant document should contain information on the trading relationship between Xinjiang, China and its neighboring nations, including treaties signed by China and former Soviet Republics that are bordering China and foreign investment. If a document contains information on how China develops Xinjiang, it is not relevant.

Figure3. A sample topic of NIST in English

As to document image retrieval, there are few collections available at present. To the best of our knowledge, we found the largest-scale Chinese collection was developed by Tseng [5]. In their work, 11,108 news clippings were scanned and then were converted to text by a commercial OCR system, yielding 8,438 valid text documents. After that, they took over two month for topics set and relevance judgment (by three assessors, two of whom majored in history, with the other having majored in library science). As we see, it involved considerable cost in time and human effort. Even though, this collection (8434 documents) still can't satisfy the need for large-scale evaluation in some experiments. However, as to document full-text retrieval, TREC has built a variety of large test collections containing a set of topics, and a corresponding set of relevance judgments. All of what we need to do is just convert text into image. According to the method we presented in Sect 2.1, we can easily convert the document full-text into different levels of degraded document image by select various degradation models and degradation parameters.

Text retrieval collection is converted into image collection in this paper. A large-scale collection for Chinese document retrieval included in Track 5 and Track 6 is provided by National Institute of Standards and Technology (NIST) in the year of 2000. The collection contains 54 query topics and more than 160,000 articles from the People's Daily newspaper and the Xinhua newswire.

Large-scale collection can greatly reduce the side effect of occasional mistake. Next, Four test collections are built by our proposed method and used in the next experiments. As we can see in Table 1, these four collections in two different OCR precision and two different confidence approaches.

TABLE I. FOUR TEST COLLECTIONS

| Recognition Success Rate | Confidence Method | Named as |
|---|---|---|
| Top 1 candidate is 0.68, Top 2 candidates are 0.76. | EDF | D1 |
| | LRM | D2 |
| Top 1 candidate is 0.96, Top 2 candidates are 0.98. | EDF | D3 |
| | LRM | D4 |

## III. DOCUMENT INDEXING

In our work, n-gram indexing was applied to index OCR-based document, and also the top n recognition candidates and their associated confidence scores were considered. In this section we proposed several $\mathrm{tf}\times\mathrm{idf}$ methods for estimating the weight of terms in test collection or query topic. As to the retrieval model, we use the vector space model (VSM).

### *A. Term Weighting Schemes for Test Collection*

Top n candidates are ranked according to their confidence score. In order to evaluate the importance of candidate-term, we need to define the candidate-term frequency.

**Definition 1** Candidate-term frequency

Let $p_n(i)$ be the probability of the n-th occurrence of the candidate-term $i$, and N the occurrence of $i$ in a recognized document $d$. The frequency of the candidate-term $i$ is defined as follows:

$$\mathrm{tf}(\mathrm{i}) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{p}_{\mathrm{n}}(\mathrm{i}) \tag{3}$$

In order to reduce text-length effects, the $tf(i)$ should be normalized.

**Definition 2** Normalized candidate-term frequency

Let M be the number of indexation terms in document d, and $i$ a given candidate-term, and $\max_{\mathrm{d}} \mathrm{ctf}_{\mathrm{d,j}}$ be the most occurrence candidate-term. The normalized candidate-term frequency of $i$ are defined as follows:

$$\mathrm{ntf}(\mathrm{i}) = \frac{\mathrm{tf}(\mathrm{i})}{\sum_{\mathrm{j}=1}^{\mathrm{M}} \mathrm{tf}(\mathrm{j})} \tag{4}$$

$$\mathrm{ntf}(\mathrm{i}) = \frac{\mathrm{tf}(\mathrm{i})}{\max_{\mathrm{d}} \mathrm{tf}_{\mathrm{d,j}}} \tag{5}$$

$$\mathrm{ntf}(\mathrm{i}) = 1 + \log(1 + \mathrm{tf}(\mathrm{i})) \tag{6}$$

A normalized $tf\times idf$ score for candidate-terms in the output of an OCR system can be computed using the ordinary $idf$ and the $ntf(i)$ score.

**Definition 3** Normalized candidate- $tf\times idf$

Let N be the number of documents in a collection, and $k_i$ the number of documents in the collection containing the candidate-term $i$. The weight of candidate term $i$ in a vector is defined as follows:

$$\mathrm{tf}\times\mathrm{idf}(\mathrm{i}) = \frac{\mathrm{tf}(\mathrm{i})\times\log\frac{\mathrm{N}}{\mathrm{n_i}}}{\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{M}}\left(\mathrm{tf}(\mathrm{j})\times\log\frac{\mathrm{N}}{\mathrm{n_j}}\right)^2}} \tag{7}$$

A Chinese term like '申' might be produced by a recognition error from '中' could have a large inverse document frequency ($idf$) value, thus incorrectly affecting the weights of index terms if a mutually dependent normalization like the (7) is used. Singhal et al. [10] found that instead using a byte size normalization scheme could mitigate this source of error. For a document, the weight of candidate term $i$ using byte size normalization factor is computed as:

$$\mathrm{tf}\times\mathrm{idf}(i)=\frac{\mathrm{tf}(i)\times\log\frac{N}{n_i}}{(\mathrm{byte\_size})^{0.375}} \tag{8}$$

### B. *Term Weighting Schemes for Query Topic*

To the best of our knowledge, query topic can be treated as a document. However it is irrational to adopt the existing term- weighting methods above for query topics weighting. First, a query topic usually contains only several words, and there is less information in these words for weighting compared with normal text. Second, query topic totally different from collection document that without any noise. So, another two weighting schemes are proposed as follows:

The weight of a term i in a query q can be derived by:

$$w_{i,q}=\left(0.5+\frac{0.5\bullet tf_{i,q}}{\max_{l} tf_{i,q}}\right)\bullet\log\frac{N}{n_i} \tag{9}$$

$$w_{i,q}=\frac{3(tf_{i,q}-1)}{\sqrt{\sum_{j=1}^{t}3(tf_{j,q}-1)}} \tag{10}$$

Next, to eliminate the effect of query document (query topic in this paper) length, (7) or (8) is performed to normalize term weights obtain by (9), (10).

## IV. EXPERIMENTS

### A. *Methodology*

Existing studies have established that n-gram indexing (usually with n=2) works about as well for Chinese retrieval as word based indexing, both for ordinary text and for text produced by automatic speech recognition. As mentioned above, n-gram indexing is also known to be relatively robust in the presence of OCR errors. We therefore use n-grams in our experiments. This raises the question that how to choose the optimum value for n. Longer n-grams matching can work only in high OCR precision. But shorter n-grams might help mitigate the effect of OCR errors, and in other words shorter n-grams help improve words recall rate and improve the document retrieval effectiveness finally. We therefore tried n=1, n=2 in the next experiments. The large inventory of Chinese characters results in excessively larger indices for n>2, so we did not try larger values of n.

We use the top n recognition candidates rather than the typical output of the recognition system usually containing the top choice candidate. The use of top n candidates can help counting occurrence of missing terms, because the chance of the correct term being in the top n increase as n does. So, when recognition candidates are used, three approaches based on various term confidence computing are as follows:

*1)* Ignoring confidence. We assume that all candidate terms are of equal confidence in this approach, in other words, the influence of candidate confidence on retrieval perform is omitted. For example, the recognition result of Chinese word '恢复' as shown in Figure 4 when n-gram (n=2) and top n recognition candidates (n=2), all candidate terms confidence are set as 0.25.

*2)* Adopting direct confidence, as we see in Figure 4, These four candidate term confidences are evaluated respectively by $p_{11}\times p_{21}$ , $p_{12}\times p_{21}$ , $p_{11}\times p_{22}$ and $p_{12}\times p_{22}$ .

*3)* Adopting adjusted confidence, from the Figure 4 we can see that:

$$p_{11}\times p_{21}+p_{12}\times p_{21}+p_{11}\times p_{22}+p_{12}\times p_{22}<1,$$

Thereby let $t$ be the factor of

$$t\times(p_{11}\times p_{21}+p_{12}\times p_{21}+p_{11}\times p_{22}+p_{12}\times p_{22})=1,$$

So their confidence evaluated respectively by

$$t\times p_{11}\times p_{21}\,,\ t\times p_{12}\times p_{21}\,,\ t\times p_{11}\times p_{22}\ \text{and}\ t\times p_{12}\times p_{22}\,.$$

Previous studies have shown an interaction between the retrieval models (VSM concerned in this paper) and the term weighting schemes and normalization techniques, especially in [6, 10], they have done a series of studies to identify the effects of OCR errors on text retrieval using different weighting schemes. But can those existing term-weighting methods mentioned above still be effectively used for the retrieval of noise document with top n candidates? This is the first question we wish to address in this paper. After our experiments, we can conclude from results as following:

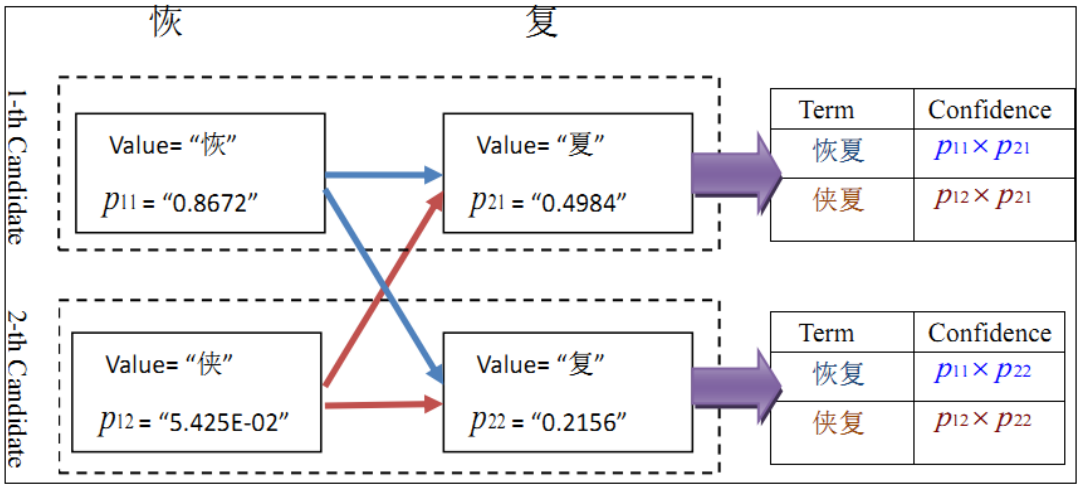


Figure 4. Recognition with top 2 candidates, and terms when n-gram(n=2)

1) In the estimation of normalized candidate-term frequency, formula (3), (4) achieving much better performance than formula (5).

2) In the estimation of weighting query topic, formula (9) achieving much better performance than formula (10).

Take all of these into consideration; we adopting the modified term-weighting method {(3), (4), (8)}, {(9), (7)} in the rest of our experiments which also has achieved the best performance in documents with top n recognition candidates. As you can see, we adopt different term-weighting methods for collection document and query topic.

### *B. Experimental Results*

To explore retrieval performance using term indexing method with n-gram and recognition candidates, four collections are used to make an overall comparison, and the results are shown in Tables 3-6. The names of indexing method are given in Table 2, and the retrieval performance is characterized by the mean (over topics) of the uninterpolated average precision.

TABLE II. THE NAMES OF INDEXING METHODS

| N-gram | Confidence Strategy | Candidates | Named as |
|---|---|---|---|
| 2-gram | Ignoring confidence | 1 | T01 |
| | | 2 | T02 |
| | Adopting direct confidence | 1 | T03 |
| | | 2 | T04 |
| | Adopting adjusted confidence | 1 | T05 |
| | | 2 | T06 |
| 1-gram | Ignoring confidence | 1 | T07 |
| | | 2 | T08 |
| | Adopting direct confidence | 1 | T09 |
| | | 2 | T10 |
| | Adopting adjusted confidence | 1 | T11 |
| | | 2 | T12 |

TABLE III. RETRIEVAL PERFORMANCE ON THE D1 COLLECTION

| Indexing Method | T01 | T02 | T03 | T04 | T05 | T06 |
|---|---|---|---|---|---|---|
| Ave.P | 0.5600 | 0.5739 | 0.5698 | 0.5872 | 0.5600 | 0.5994 |
| Indexing Method | T07 | T08 | T09 | T10 | T11 | T12 |
| Ave.P | 0.4911 | 0.4824 | 0.4927 | 0.4934 | 0.4912 | 0.5031 |

TABLE IV. RETRIEVAL PERFORMANCE ON THE D2 COLLECTION

| Indexing Method | T01 | T02 | T03 | T04 | T05 | T06 |
|---|---|---|---|---|---|---|
| Ave.P | 0.5611 | 0.5736 | 0.5654 | 0.5946 | 0.5611 | 0.6061 |
| Indexing Method | T07 | T08 | T09 | T10 | T11 | T12 |
| Ave.P | 0.4912 | 0.4818 | 0.4912 | 04967 | 0.4912 | 0.5026 |

TABLE V. RETRIEVAL PERFORMANCE ON THE D3 COLLECTION

| Indexing Method | T01 | T02 | T03 | T04 | T05 | T06 |
|---|---|---|---|---|---|---|
| Ave.P | 0.6468 | 0.6479 | 0.6468 | 0.6470 | 0.6468 | 0.6500 |
| Indexing Method | T07 | T08 | T09 | T10 | T11 | T12 |
| Ave.P | 0.5346 | 0.5237 | 0.5350 | 0.5371 | 0.5346 | 0.5600 |

TABLE VI. RETRIEVAL PERFORMANCE ON THE D4 COLLECTION

| Indexing Method | T01 | T02 | T03 | T04 | T05 | T06 |
|---|---|---|---|---|---|---|
| Ave.P | 0.6465 | 0.6480 | 0.6468 | 0.6509 | 0.6465 | 0.6512 |
| Indexing Method | T07 | T08 | T09 | T10 | T11 | T12 |
| Ave.P | 0.5346 | 0.5235 | 0.5348 | 0.5464 | 0.5346 | 0.5502 |

First of all, we analyze the retrieval performance based on following two aspects: First, different confidence estimation methods; Second, different OCR precisions. As mentioned above, confidence measurement is especially important in character recognition. With the comparison, the confidence value estimated by LRM seems to be more accurate than that designed artificially in EDF, which also has been proved in [11]. Under same condition, experiments on D2 and D4 (by LRM) achieve slight relatively higher Ave.P values than D1, D3 (by EDF). The indexing method proposed in this paper achieved better performance. And unsurprisingly, this method has a little improvement on collection with low OCR error. Because the first candidate's accuracy is relatively high, and instead using top n candidates introduce too many false occurrence of words, thus making the text noisier.

From Tables 3-6, we can make fairly clear conclusions as following:

1) n-gram length has the greatest effect on retrieval performance, and 2-gram (indexing method T01-T06) achieving a approximately 10% improvement than 1-gram (indexing method T07-T012).

2) Using the top n recognition candidates rather than the top one candidate can achieve better performance in the presence of OCR errors.

3) Adopting confidence results in better retrieval performance than ignoring confidence. Furthermore, adopting adjusted confidence produces best performance.

Besides, another one experiments is conducted to illustrate the relationship between retrieval performance and the number of candidates for one character. We select top 10 of the 54 query topics, of which associated relevant documents and another 3000 irrelevant documents from D2 and D4 (by LRM and with different recognition precision), respectively to form the test collections namely D5 and D6. As to confidence strategy, direct confidence is used

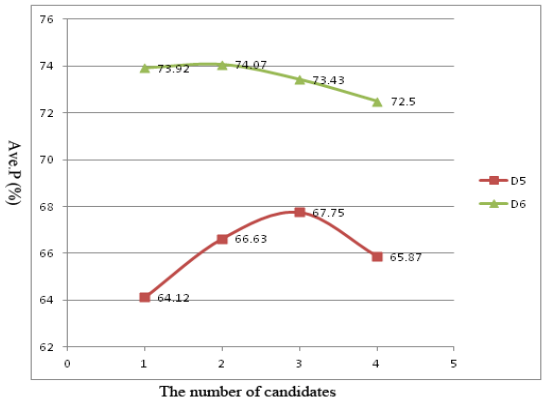


Figure 5. Peformace varying with the number of candidates

Finally, we can find out from Figure 5, retrieval performance cannot be improved all the way. Different OCR precisions need different numbers of candidates to achieve best performance.

## V. CONCLUSIONS

OCR text provides the cheapest and fastest way to make document images searchable, but optimizing retrieval performance under these conditions requires that the retrieval technique be adapted to mitigate the side effect of OCR errors. In this paper, an indexing method based on n-gram and recognition candidates is proposed. Meanwhile, a system to automatically generate ground-truth of imaged document, synthesized degraded document image and ground-truth of recognition candidate is proposed. Therefore simulated OCR collections can be achieved by this system.

By verifying the proposed indexing method with different confidence measurements on OCR text with different recognition precision, we find that this new method, especially 2-gram and 2 candidates, exhibit stable and consistent improvement over the existing indexing methods. In the future work, language models will be introduced to deal with candidates indexing for better performance.

## ACKNOWLEDGMENT

This project was supported by Shandong Excellent Young Scientist Award Fund (BS2011DX002), the Fundamental Research Funds for the Central Universities (Grant No.HIT. NSRIF.2009152),China Postdoctoral Science Foundation (20090450994) and Heilongjiang Postdoctoral Grant (LBH-Z09150).

## REFERENCE

[1] Murugappan, B.Ramachandran, P.Dhavachelvan. A survey of keyword spotting techniques for printed document images. Artificial Intelligence Review, 2010, pp.1-18.

[2] M.S.Shirdhonkar, M.B.Kokare. Document Image Retrieval: An Overview. International Journal of Computer Applications, 2010, 1(7):128-130.

[3] S. Lu, L. Li, C.L. Tan. Document image retrieval through word shape coding. IEEE Trans. PAMI, 2008, 30(11):1913-1918.

[4] S. Lu, C.L. Tan. Retrieval of Machine-printed Latin Documents through Word Shape Coding. Pattern Recognition. 2008, 41(5):1816-1826.

[5] Y.H. Tseng, D.W. Oard. Document Image Retrieval Techniques for Chinese. Proceedings of the Fourth Symposium on Document Image Understanding Technology. Columbia Maryland, April 23-25th, 2001, pp.151-158.

[6] Sebastian Pena Saldarriaga. Using top n Recognition Candidates to Categotize On-line HandWritten Documents. 10th Internatioanal Conference on Document Analysis and Recognition, 2009,(42):3374-3382.

[7] Lee YS, Chen (1996) Analysis of error count distributions for improving the post-processing performance of OCCR. Commun COLIPS 6(2):81–86.

[8] Li Y, Ding X (2002) Evaluation of character candidate confidence measure using logistic regression model (in Chinese).Pattern Recogn Artif Intell 15(2):160–166.

[9] Webb A (2002) Statistical pattern recognition. Wiley England.

[10] Amit Singhal, Gerard Salton, and Chris Buckley, “Length Normalization in Degraded Text Collections” Proceedings of Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996 pp. 149-162.

[11] Y-X Li, et al. "A Hybrid Post-processing System for Offline Handwritten Chinese Script Recognition ". Pattern Analysis and Applications, 2005, 8(3).