2009 10th International Conference on Document Analysis and Recognition

Document Image Retrieval with Local Feature Sequences

Jilin Li, Zhi-Gang Fan, Yadong Wu and Ning Le
Advanced R&D Center of SHARP Electronics (Shanghai) Co.,LTD.
jilin.li@cn.sharp-world.com

Abstract

In recent years, many document image retrieval algorithms have been proposed. However, most of the current approaches either need good quality images or depend on the page layout structure. This paper presents a fast, accurate and OCR-free image retrieval algorithm using local feature sequences which can describe the intrinsic, unique and page-layout-free characteristics of document images.

With a simple preprocessing step, the local feature sequences can be extracted without print-core detection and image registration. Then an efficient coarse-to-fine common substring matching strategy is applied to do local feature sequences matching. Beyond a single matching score, this approach can locate the matched parts word by word. It well handles the challenges including low resolution, different language, rotation and incompleteness and N-up. The encouraging experiment results on a large scale document image database show the retrieval outputs are sufficient good to be used directly as document image identification results.

1. Introduction

With the progress of office automation and digital image processing, document image retrieval techniques are largely developed in recent years. The aim of document image retrieval is try to pursuit the identical or similar document images in a database for query document image. The existing approaches can be divided into two categories, optical character recognition(OCR) based algorithms [2] and image feature based algorithms. While the state-of-art OCR technology can produce very accurate results, the sensitiveness to image resolution, heavy computation cost and language dependency limit its applications.

Different from OCR-based methods, image feature based algorithms focus on the image information instead of text information. These algorithms try to analysis the global or local layout structure for different document images and estimate the similarity among them. For example, P. Herrmann et al. proposed a method based on page layout features [3]. H. Peng et al. proposed a method based on the sizes and positions of component block list in document image [4]. D. Doermann et al. also proposed a duplicate image detection algorithm [5]. The algorithm extracts a representative text line in a document image, and then extracts a signature from this text line by character shape coding. More recently, S. Lu, L. Li and C.L. Tan proposed an algorithm using word shape coding [6].

Besides shape structure, there exist some other approaches. C. Wang et al. proposed a Chinese document image retrieval method based on the proportion of the foreground pixel area in character bounding box area [7]. H. Liu et al. proposed a method using the foreground density distribution feature and key block feature [1]. They try to combine the global and local information by using local foreground density distribution and key blocks. J.J. Hull proposed a distortion-invariant descriptor based on the lengths of consecutive words. It achieved accurate retrieval results in a fast speed with Hashing techniques [8].

The mentioned approaches achieve good performances on some document image databases. However, the local features such as character, character shape and word shape largely depend on the quality of input document images. It is hard to successfully extract these features in low resolution and distorted images. Global features like page layout is robust to image resolution and image distortion but it can not distinguish the documents with similar page layout and different content, which is a serious drawback in practice. J.J. Hull's method [8] is robust but it needs character segmentation which is still sensitive to image resolution and noise.

To overcome the challenges of low resolution, different languages, distortion, incompleteness and N-up, we propose a novel document retrieval algorithm based on the local feature sequences. Our approach can give high distinctive retrieval result without character segmentation and word shape analysis. Beyond a single matching score, our approach can locate the correspondence parts between query image and retrieved images word by word.

In Section 2, starting from the analysis of document im-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.46

age characteristics,the local feature sequences(LFS) is introduced. In Section 3, a document image retrieval method using local feature sequences is proposed. We evaluate the new method on a large scale document image database which has variations of image rotation, resolution, distortion, incompleteness and N-up. The experiment results are presented in Section 4. Finally, Section 5 gives the conclusions.

2. Local Feature Sequences

Efficient document image features are essential for document image retrieval. A sound feature should have the good properties of uniqueness, robustness and efficiency. The uniqueness provides the feature with strong discriminative ability. The robustness prevents the feature from losing reliability under image variations. The storage and computation efficiency are necessary for real-time applications.

For document image, local feature has good uniqueness but lack of reliability and global feature has good reliability but lack of uniqueness. So middle level features can balance this trade-off. We transform the whole document image into a single local feature sequence. Special local feature can be designed for different types of document image and different applications. For Latin-based document image, we define the feature as the proportion of length between adjacent words which has good properties mentioned above. We mainly concentrate on the Latin-based document image, and Asian-based document will be discussed separately.

2.1. Preprocessing

The scanned document images often contain skew and suffer from degradations such as scanned noise and low contrast. To extract the feature properly, firstly we use a mean filter with size of 3x3 to smoothen the input images. And the smoothed images are binarized by Ostu Algorithm [9]. Isolated black pixels can be removed by connected component analysis. Then we use the algorithm described in [10] to estimate the skewed angle.

2.2. Layout Analysis

Many algorithms have been proposed for document image layout analysis. In our approach, only text line segmentation is needed since the retrieval results are page-layout-free. We select the simple run length smoothing algorithm (RLSA) [11] to do layout analysis. And the run-length parameters are not so sensitive to final results. Suppose the width and height of input image is W and H. In our experiment we set run-length $(W/20, H/20)$ for horizontal and vertical respectively. The additional horizontal smoothing

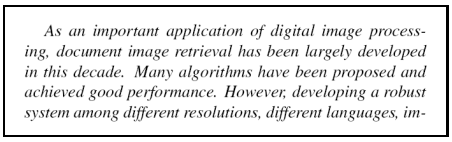


Figure 1. Original Document Image

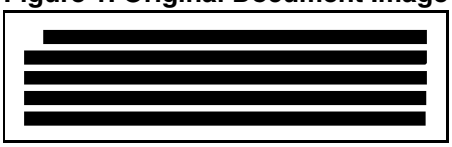


Figure 2. Extracted Text Lines

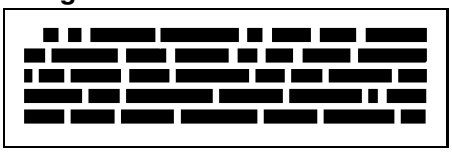


Figure 3. Segmented Words

length is set to $W/200$ pixel. With RLSA, text line patches can be extracted from original binarized image.

Then, each text line patch are segmented into several word patches. Consider the image as a matrix I with element of $\{0,1\}$. In this paper, the text pixels are 1 and background pixels is 0. Firstly I is projected onto horizontal axis by defining $P_j = \sum_{i=1}^{H} I_{ij}, (j = 1, \ldots, W)$. Then we find all zero value segments set $S = \{S_1, S_2, \ldots, S_N\}$. Each segment S_k consists of several continuous zero projection points

$$\begin{aligned} S_k &= \{P_s, P_{s+1}, \ldots, P_t\} \\ k &= 1, \ldots, N, 1 \le s \le t \le W \end{aligned} \quad (1)$$

The average length of all segments in S can be calculated by $\bar{S} = \frac{1}{N}\sum_{k=1}^{N} |S_k|$. Update P by setting the element in S_k to 1 if condition $|S_k| < \bar{S}, k = 1, \ldots, N$ holds. Then we can find all non-zero value segments $Z = \{Z_1, Z_2, \ldots, Z_M\}$. These segments indicate the position of each word in the image. A typical processing of layout analysis, text line extraction and word segmentation are illustrated in Figure 1,2,3 respectively.

2.3. Local Feature Sequence

One document image has rich hierarchical structure of paragraphs, sentences, words, characters and pixels. OCR-based and character shape based methods work on the characters level. Since these methods are sensitive to low resolution. It's natural to go upper level for more robust features. In [8], J.J. Hull proposed a document equivalence detection method based on word lengths. They estimated the count of character in each word and extracted features by concatenating the lengths of M adjacent words. Good results

and fast speed were achieved. But character segmentation needs high quality image which is a drawback in practice. We proposed a different approach which avoids character segmentation while achieving more accurate results.

2.3.1 Local Feature Representation

Different from J.J. Hull's method [8], we calculate the word length by pixels instead of character count. And the whole image is represented as a single feature sequence instead of a big descriptor set. This approach is more robust to image resolution, more accurate for retrieval, more economic on storage while keeping real-time retrieval speed.

For Latin-based document images, word length is a simple geometry feature. With certain traveling order, all words in a document can be coded as a length sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ in which x_i is the length of ith word, N is the words number. But $\mathbf{x}$ can not be applied directly for document image retrieval since word length is not scale-invariant. It is trivial to find a scale-invariant representation from $\mathbf{x}$ by defining a new sequence of proportion of word length

$$\begin{aligned} \mathbf{y} &= \{y_1, y_2, \ldots, y_{N-1}\} \\ y_i &= \frac{x_{i+1}}{x_i}, (i = 1, \ldots, N-1) \end{aligned} \tag{2}$$

The sequence $\mathbf{y}$ is resolution independent. Obviously, it captures the local structures of adjacent words.

2.3.2 Local Feature Sequence Analysis

The word length is a simple and stable feature which is not sensitive to image quality. And good uniqueness can be derived from the word length sequence $\mathbf{x}$. With the assumption of that the words in a document image is randomly selected in a fixed vocabulary, and consider x_i as a random variable, then $\mathbf{x}$ becomes a random sequence of word length. Suppose the width of each character is constant and generally Latin-word has a length from 1 to 15. Then the length proportion between two words y_i exists $(15^2 - 14) = 211$ combinations. Therefore the occurrence probability of one combination is $1/211$. And the occurrence probability for an adjacent word sequence with length of L is $(1/211)^L$. If $L = 10$, it is approximately 5.7×10^{-24}. So, the occurrence probability for two documents sharing the same feature sequence with length L is about $(5.7 \times 10^{-24})^2 \approx 3.3 \times 10^{-47}$. This indicates such feature sequence has very high uniqueness. Although our estimation is too ideal, the word usage in Latin language is not completely random and there exists special distribution for certain words, phases and sentences. So there exists the case that different word sequences may share the same length sequence. But this coincidence only happens when L is small. If L is large enough this coincidence will be vanished.

3. Document Image Retrieval using LFS

3.1. Local Feature Sequence Matching

Different from the vector-based feature representations, we use sequence-based representation for the whole document image. Document is naturally a word sequence so this representation is reasonable. The only problem remained is how to compare two feature sequences. Fortunately there exist linear time algorithms to solve this problem efficiently.

Sequence matching is a classical problem which has been deeply understood. Common subsequence matching is widely applied in bioinformatics. Inspired by its big success, we use common subsequence matching algorithm to do document image comparison. The common subsequence matching problem has continuous and discontinuous version. In this paper, we chose the continuous version.

There are two solutions to perform common substring matching. One is dynamic programming (DP). It is simple and can do approximate matching. But this approach has an $O(N^2)$ time complexity which is not feasible in large scale applications. Another solution is suffix tree. P. Weiner and E. Ukkonen proposed the linear time construction algorithms for suffix tree [12],[13] respectively. It has a time complexity of $O(N)$ and makes a powerful tool for solving string matching problems. But suffix tree can not do approximate matching. To combine the advantage of DP and suffix tree, we use a coarse-to-fine strategy. Rough matching is performed by suffix tree, the retrieved candidate images are further finely compared by DP. This strategy makes the retrieval process extremely fast while keeping high accuracy.

The feature value in the local feature sequence $\mathbf{y}$ is float number. For applying suffix tree, quantization is needed. We simply map these float numbers into an integer set $\{0, 1, \ldots, H\}$. In which the largest element H can be considered as a free parameter to control the error tolerance for matching. A smaller H will produce a coarser matching than a larger one. In our experiment, we set $H = 64$. This quantization mapping makes the feature can tolerate some words segmentation error. On the other hand, for few words, the mapping from float number to integer number will loss accuracy which is caused by hard cut on integer boundaries. Since only coarse matching results are needed and usually one document has hundreds of words, this step will not effect the final recall rate of suffix tree seriously.

3.2. Document Image Retrieval

The document image retrieval system can be implemented though the techniques discussed above. Suppose we have a document image set $\{I_1, I_2, \ldots, I_N\}$ to be registered. Then the local feature sequence set

$Y = \{Y_1, Y_2, \ldots, Y_N\}$ and the coded version $S = \{S_1, S_2, \ldots, S_N\}$ can be extracted for each document image. We can build a suffix tree noted as ST_i for each sequence S_i in linear time. Then given a query document image I^*, its local feature sequence Y^* and coded local feature sequence S^* can be calculated. To find the common subsequences between query image and registered images, we search the substring of S^* in each suffix tree ST_i which is defined as $C(S^*, ST_i) = \{c_{i1}, c_{i2}, \ldots, c_{iK}\}$. Note that different word sequences might produce same word length sequence, for removing this coincidence we just count the common subsequence with length larger than a threshold L. The initial matching score is defined as

$$F(S^*, ST_i) = \sum_{j=1,|c_{ij}|>L}^{K} |c_{ij}| \tag{3}$$

We send all S_i which satisfies $F(S^*, S_i) > 0$ to DP matcher. DP matcher will re-calculate the approximate common substring among Y^* and Y_i noted as $\tilde{C}(Y^*, Y_i) = \{\tilde{c}_{i1}, \tilde{c}_{i2}, \ldots, \tilde{c}_{iK}\}$. The approximate matching rule is defined as

$$Equal(x, y) = \begin{cases} true, & \frac{|x-y|}{|x|} \leq \epsilon \\ false, & \frac{|x-y|}{|x|} > \epsilon \end{cases} \tag{4}$$

In which ϵ is a small positive real number. We set $\epsilon = 0.1$ in our experiments. Similar to Formula (4) we can get the candidate list sorted by the matching score $F(Y^*, Y_i) = \sum_{j=1,|\tilde{c}_{ij}|>L}^{K} |\tilde{c}_{ij}|$ in which threshold L is set to the same value as coarse suffix tree search.

3.3. Asian-based Document Image

For Asian-based document images in which the word lengths are constant, other kinds of local feature should be extracted to describe the local structure. Similar to [7], we can use the proportion of foreground area in a word bounding box as the local feature. With the same logic described for Latin-based document images, we can form a sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ in which x_i is the proportion of foreground area for ith word. It can be applied directly for document image retrieval since x_i is scale-invariant.

4. Experiments and Analysis

4.1. Data Preparation

To prove our approach, we developed a document image retrieval system and evaluated it on a large scale document image database which contains different 8090 document images with variations of rotation (randomly from $-\pi/6$ to $\pi/6$), resolution (from 75 dpi to 300 dpi), color (binary-scale, gray, color), layout (combination of text, figure, table) and noise. We firstly scanned 5090 document images which are full of above variations. To test the robustness of layout change, we collected additional 3000 document images by operations include removing some text parts, exchanging the positions of whole paragraphs or sentences and merging several images together to simulate 2up or 4up scanning. In the same way, we re-scanned 5579 original document images as testing set in which 4579 images is identical to registered images. And other 1000 images have no common parts with registered images are used to test the false alarm.

Table 1. Retrieval Precision and False Alarm

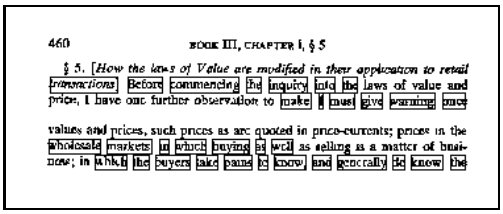
Threshold L	Top 1	Top 3	Top 5	False Alarm
5	0.9954	0.9954	0.9954	(2/1000)
7	0.9899	0.9899	0.9899	(0/1000)
9	0.9634	0.9634	0.9634	(0/1000)

Table 2. Average Retrieval Speed per Word

Threshold L	Coarse Search	Fine Search	Total
5	0.75 ms	11.99 ms	12.74 ms
7	0.72 ms	0.14 ms	0.86 ms
9	0.72 ms	0.01 ms	0.73 ms

4.2. Results and Analysis

The precision is defined as the recall rate of top K retrieved result. The false alarm is defined as the error rate of top 1 retrieved result. In this experiment, only Latin language documents are evaluated. All experiments were run on a PC with Intel® Core™ Duo 2.0GHz and 2GB memory. The retrieval accuracy is shown in Table 1. We can see our approach get high precision and extremely low false alarm. The accuracy decreased slightly with L increasing. And Top 1 result is the same as Top 5 result. This indicates the local feature sequences have very strong discriminative ability. scores is shown in Figure 6. The top line means total word count of query image. A big gap exists in related documents and unrelated documents. The matching score is beyond the concept of similarity. Identification can be determined directly based on the matching score. The strong robustness to low resolution and the invariant to page layout change can be illustrated by Figure 4 and 5. The matched parts are marked by rectangles. In the low resolution query image patch, we cropped several sentences and change the

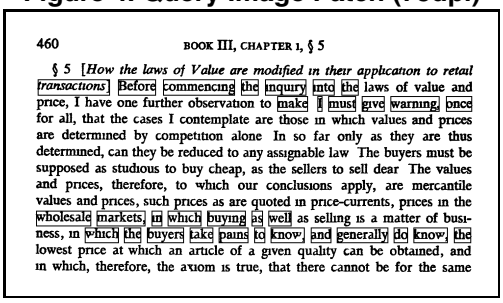


460 BOOK III, CHAPTER I, § 5

§ 5. [*How the laws of Value are modified in their application to retail transactions.*] Before commencing the inquiry into the laws of value and price, I have one further observation to make. I must give warning, once

values and prices, such prices as are quoted in price-currents; prices in the wholesale markets, in which buying as well as selling is a matter of business; in which the buyers take pains to know, and generally do know, the

Figure 4. Query Image Patch (75dpi)



460 BOOK III, CHAPTER I, § 5

§ 5. [*How the laws of Value are modified in their application to retail transactions.*] Before commencing the inquiry into the laws of value and price, I have one further observation to make. I must give warning, once for all, that the cases I contemplate are those in which values and prices are determined by competition alone. In so far only as they are thus determined, can they be reduced to any assignable law. The buyers must be supposed as studious to buy cheap, as the sellers to sell dear. The values and prices, therefore, to which our conclusions apply, are mercantile values and prices, such prices as are quoted in price-currents, prices in the wholesale markets, in which buying as well as selling is a matter of business, in which the buyers take pains to know, and generally do know, the lowest price at which an article of a given quality can be obtained, and in which, therefore, the axiom is true, that there cannot be for the same

Figure 5. Retrieved Top 1 Image (300dpi)

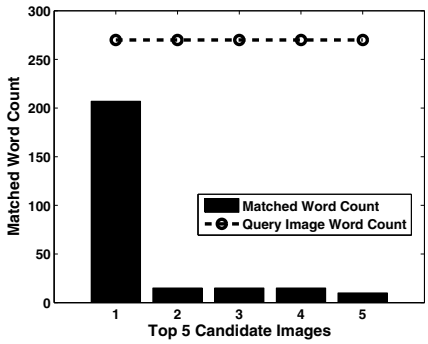


Figure 6. Typical top 5 matching scores

position of other sentences. This operation doesn't affect the Top 1 retrieval result.

Since the retrieval time is depends on the local sequence length of input document image. Table 2 shows the average retrieval speed per word. We can see coarse search by suffix tree is nearly constant with L. And fine search by DP is decreased rapidly with L increasing. So the threshold L can control the trade-off between accuracy and speed. If L is smaller than 7, coarse search will produce plenty of candidate images for DP matcher which makes retrieval slow. In our experiment, layout analysis cost 500 ms for a 300 dpi image. And suppose a typical image has 500 words, with $L = 7$, we can get good retrieval result within 1 sec from 8090 document images while keep nearly zero false alarm.

5. Conclusions

This paper proposed a novel document image retrieval algorithm based on local feature sequence and common substring matching. The word length feature sequence captures the essential local structure of Latin-based document image and has nearly perfect uniqueness among large scale database. The experiment results show our approach has good performance under the challenges of low resolution, rotation, incompleteness and N-up. The retrieval outputs are sufficient good to be used directly as document image identification results. Further study will evaluate the performance of local feature sequences on Asian-based documents.

References

[1] H. Liu, S. Feng , H. Zha , X. Liu, "Document image retrieval based on density distribution feature and key block feature", In Proc. 8th ICDAR, pages 1040-1044, 2005.

[2] G. Salton, Introduction to modern information retrieval, McGraw-Hill, 1983.

[3] P. Herrmann, G. Schlageter, Retrieval of document images using layout knowledge, In Proc. 2nd ICDAR, pages 537-540, 1993.

[4] H. Peng, F. Long, Z. Chi, W. Siu, Document image template matching based on component block list, In Pattern Recognition letters, pages 1033-1042, 2001.

[5] D. Doermann, H. Li, O. Kia, The detection of duplicates in document image databases, In Image and Vision Computing, pages 907-920, 1998.

[6] S. Lu, L. Li, C.L. Tan, Document image retrieval through word shape coding, In IEEE Trans. PAMI, pages 1913-1918, 2008.

[7] C. Wang, T. Chen, Y. Chan, R. Hwang, W. Huang, Chinese document image retrieval system based on proportion of black pixel area in a character image, In 6th ICACT, pages 25-29, 2004.

[8] J.J. Hull, Document image matching and retrieval with multiple distortion-invariant descriptor, In Document Analysis System, World Scientific, pages 379-396, 1995.

[9] N. Otsu, A threshold selection method from gray-level histogram, In IEEE Trans. SMC-9(1), pages 62-66, 1979.

[10] Chien-Hsing Chou , Shih-Yu Chu , Fu Chang, Estimation of skew angles for scanned documents based on piecewise covering by parallelograms, Pattern Recognition, Vol 40, Issue 2, pages 443-455, 2007.

[11] K.Y. Wong, R.G. Casey, F.M. Wahl, Document image analysis system,IBM Journal of Reserch and Development, 26(6), pages 647-656,1982.

[12] P. Weiner, Linear pattern matching algorithm, 14th Annual IEEE Symposium on Switching and Automata Theory, pages 1-11,1973.

[13] E. Ukkonen, On-line construction of suffix trees, Trans. Algorithmica vol.14, no.3, pages 249–260, 1995.