

A Hierarchical Algorithm for Document-images Fast Matching

Shuang Tan, Yan Jia, Hua Fan¹

Abstract—As digital libraries and document images largely use in the network, how to retrieve them quickly become one of the key issues. This paper presents a hierarchical matching algorithm to achieve fast retrieval of document images. Firstly, we can quickly find the possible matching location through approximate string matching algorithms; and then, use this location as a reference point in the target image, extract sub-block which has the same size as template image and compute the correlation coefficient between sub-block and template image. According to correlation coefficient values, we can accurately know whether the template image's information exist in the target document; finally, experimental results demonstrate the feasibility of the algorithm.

Keywords- document images; Image Retrieval; approximate string matching; correlation coefficient

1 Introduction

With the extensive use of image processing tools, such as digital cameras, fax machines, scanners and etc., we can handle document files or paper documents into digital images, which contain specific unique content features. If the main content of an image is document message, we can define this image as document image [1]. Document images mainly consist of different types of characters, electronic forms, images, and etc. In daily life, the document image can be seen everywhere. Document image retrieval, identification and other techniques have also been more extensive attention [2, 3]. As most of the document images adopt gray image coding method (such as JPEG), its data format and data stream is the same with the general form of continuous tone images, document image can utilize some conventional methods to preprocess the document images.

The rest of this paper is organized as follows. In Section II, we want to introduce some document images problem and our ideas. In section III we present some of the related work. Section IV is the impartment part of our article in this paper. In Section V, we do some experiment. Section VI concludes and discusses some future work.

1.1 Problem Statements

There are some differences between document images and normal continuous tone images. As the characteristics of the source document image information and the impact of light and equipment, digitized document images is often a homogeneous, more concentrated distribution of gray scale images. [4] proposes a calculation of correlation coefficient method in the frequency domain, the article pointed out that if non-zero elements of the template image is less than 132, the use of frequency domain correlation can achieve better results. In this paper, we implemented the algorithm by MATLAB, and found that the location of the template image cannot find in the target image, so it is not a good method to deal with document images.

The key to quickly retrieve document image is achieving rapid matching keywords. Figure 1 shows our designed document images fast matching system. This system mainly consists of two parts, the one is document images pre-processing module and the other is document images fast matching module. The designed kernel algorithm of this paper is hierarchical algorithm for fast matching document images, just as show the red dashed line in Figure 1. In this part, we can see two subparts: approximate string matching and the calculation of correlation coefficient. Under the effect of the image processing tools, object images will occur some flips and deformation, in section 2.3, we will describe clearly. There will have a certain percentage of inconsistencies between the template image and the target image's sub-image block which is corresponding to the template image block. When we locate the template image may appear in the target image, the need to further verify whether the template image is exactly the same as target image. In this step, we need to extract a sub-image from the target image, which has matched through approximate string matching, and then calculate the correlation coefficient δ between sub-images and template image. When $\delta > 0.9$ (experimental verification in this paper), we will be able to pinpoint the template image position in the target image. The algorithm accelerates the matching speed and matching accuracy.

¹ Shuang Tan(✉)
Computer Science, National University of Defense Technology, Changsha, China
Email:taylorshuang@163.com
Yan Jia
Computer Science, National University of Defense Technology, Changsha, China
Email:taylorshuang@163.com
Hua Fan
Computer Science, National University of Defense Technology, Changsha, China
fanhero@msn.com

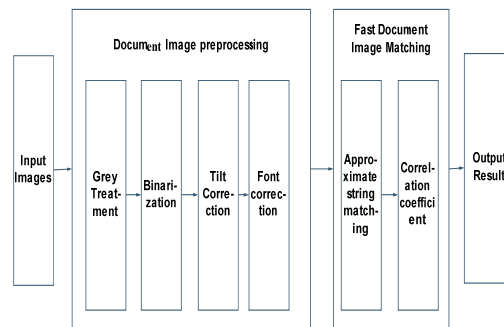


Figure 1. Schematic diagram of the document image matching

2 Related work

At present, the precise image of document retrieval technology is generally through the OCR (Optical Character Recognition, OCR) system, which converts the document image into document file, and then to retrieve, but there are still some defects in OCR practical application of the Chinese search [5]. [6] proposed a content-based document image retrieval document, which does not require OCR to support document image retrieval. [7] proposed an based on information measurement's method to classify document images, but not effective retrieve the contents of the document images. [8] designed a practical multi-body English recognition system, the system can identify up to 260 kinds of fonts, recognition rate reached 99%, but cannot achieve dual recognition in Chinese and English. Wang Hong, a researchers of Dalian University of Technology, make a systematic analysis about the recognition of printed Chinese characters and realizes the automatic document image preprocessing and recognition, but this system cannot do well when the document images have largely characters [9]. [10] proposed a character search method in the document image, this method does not require layout analysis and the treatment of OCR technology. [11] proposed an algorithm that use Hausdorff distance to determine the degree of similarity between two documents images. The algorithm will set the feature information of the target document image and the template image into two point sets, X and Y, then calculation the Hausdorff distance of two point sets. If this distance is less than a threshold, the two images are similar. [12] made a certain modifications in [11] to adapt the support of the Chinese document images, speed up the Chinese document image retrieval efficiency.

3 Document image preprocessing

Most of the document processing systems are generally handled as a gray scale image. Paper documents enter into the computer through the input device are generally a color image document images. Colored images commonly use RGB color model to storage, and its every pixel is a vector (R, G, and B). The vector size of each pixel is 8 bits, that is, values between 0 and 255, we can use weighted average method to process the color image into a gray scale image.

3.1 The binarization of Gray scale images

In order to separate the document image's words and its background, the binarization operation of gray image needs to be done. Usually we need to deal with document images is binary document images, and not gray scale images. In order to facilitate follow-up treatment, we need to convert the target image to binary black and white images. At present, there are many ripe gray scale images binarized algorithms for us to use [14, 15, and 16]. Mostly idea is that, the most critical part of gray scale document image is to determine the target document image gray threshold, and once we find the best gray scale document image threshold, can do binarization to the whole document images. When the gray value less than the gray level threshold, it will be set to 0, otherwise set to 1.

3.2 The tilt correction of document images

When scanning paper documents are not necessarily aligned with the scanning direction that may exist tilted. If the tilt angle is large, although the same content as the template and the target image, cannot match the hit. Therefore, we must do tilt correction to correct target image' angle before do next step. The document images' line is aligned, and its direction can be used as criteria for judging whether the document images are slanted. There are many ways to detect tilt, such as the Hough transform projection method [17, 18]. [19] proposed a smooth and linear based on a long tour fitting method, could find tilt angle very well. In this method, firstly, do horizontal and vertical run-length restoration; then, extract the skeleton line of document, that the center line of connected regions; last to find the sufficiently long center line and sampling, calculate the fitting line of these sampling points. According to the average slope of the line fitting, we know the titled angle of document images. Figure 1 shows the situation before and after the tilt detecting.

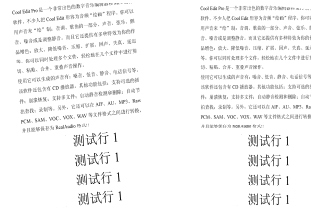


Figure 2. Document images before and after tilt detecting

3.3 Correction of the document images' font size

Document images may have different resolutions. This reason will lead document images have some differences in them, such as Font size, scale and so on. We can adjust font size before matching, so that all the target document images and the template image have the same size of font size, will have better adaptability. We can separate the same font size of the image into several regions, to find the font size within each region, and adjust the same as the template's font size. After finishing tilt correction, we can first find the text of all document areas and adjust the font size of each block to make it consistent with the font size of the template. Using projection and the recursive top-down method, we can be well divided document images into several areas [20]. In this paper, we use an improved top-down algorithm. Idea is that, we first do horizontal and vertical run-length restoration, let the separated strokes of Chinese character constitute to a connected region; and then calculate each row and column of black pixels, and if there is enough of a region of continuous blank lines or columns, where the image separately. Find out a continuous blank lines or columns in separate areas until we can't separate any. The final area is the document block.

When the image was divided into several regions, we can determine the size of the document block characters. Each connected region by calculating the length and width of the characters, extract the length and width too large or too small in the connected region, they may be noise or non-character image. Take the mean length and width of the region as a connected block length and width of Chinese characters. If a small Chinese characters than the template can insert rows or columns to increase the length and width of Chinese characters. If the character is too large, it is desirable the mean number of rows or columns instead of multiple rows or columns, and finally to the size of each block of document similar to the template. The image separate into several blocks, we can match the image block as a unit in a block to find the specified model to improve the matching accuracy.

3.4 The influence of document image's underlying code

When the conversion or transformation of image processing, the underlying code of images may cause occur flipping, which is commonly referred to as the image distortion. For a small number of flip the image, the impact is not great, but it is difficult to distinguish from distortion department. We can know that every image's data see as an array, or a matrix, pixels arranged in rows and columns. Most of the document image is black and white images after the binarization of gray scale images, and the underlying code is composed by a string of 0 and 1. Figure 3 shows that the underlying code of images may occur flipping or offset and deformation. The red figures express the correlation coefficient of two images. Figure A does not occur deformation, Figure B, the underlying code occur 8 bits flipping; Figure C occur distorted 5 degrees; Figure D occurred distorted 10 degrees, Figure E horizontal stretch of the original 105%, Figure F horizontal stretch 110%. It can be seen from the figure, although they are from the same source image, respectively, can be a very clear difference with the original. If we compare the underlying code one by one to determine whether come from the same picture, the judgment of results may not match. Therefore, we require more robust algorithm to support.

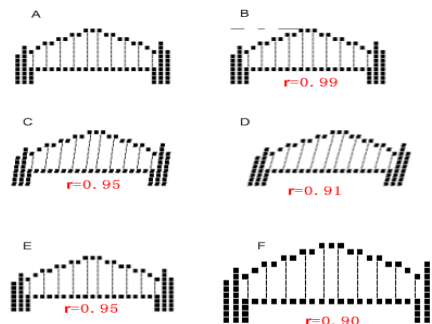


Figure 3. The impact of the underlying code (flipping, rotating, stretching), r is the correlation coefficient. A is a normal picture, B is flipped by 8 bits, C is rotated by 5 degree, D is rotated by 10 degree, E is stretched by 105%, and F is stretching by 110%

4 Hierarchical Fast Matching Algorithm

This paper presents a hierarchical algorithm to combine approximate string matching and the calculation of correlation coefficient together, experimental results show that this algorithm can obtain better results. [21] referred to a match with the sliding window algorithm that continuously sliding mode string compare string with the target model will be a larger increase in the complexity of computation is not suitable for large-scale real-time image matching. Our article designed method has high accuracy and fast speed. The algorithm is based on the following principles: we first find the pattern image in the target image that may arise about the location, and then pinpoint the exact location of the image pattern in order to avoid looking in the whole image.

<p>Algorithm A: Input: a template image's data P, object images' data T, deviation ξ Output: result, if result is equal to 1, find matched string, else no find Method Initialize: $[m, n] = \text{size of } P$, $[m_0, n_0] = \text{size of } T$; $\text{subVecB} = P(m_0/2, :)$ // ':' expresses An array of all columns of data kk is equal to 1/control value</p> <p>for each scanline T' in T from $m_0/2$ to $m - m_0/2 + 1$ do $\text{subVecA} = T(T', :)$ for each location j in scanline T' do if $\text{subVecA}[j]$ equal to $\text{subVecB}[j]$ then if kk equal to no then $\text{row} = T'$ $\text{col} = j - n_0 + 1$ return the location of matching else $kk++$; $jj++$; end if else $jj = jj - kk + 2$; $kk = 1$; end if end if end for end for end function</p>	<p>Algorithm B Input: a template image's data P, object images' data T, deviation ξ Output: result, if result is equal to 1, find matched string, else no find Method Initialize: $[m, n] = \text{size of } P$, $[m_0, n_0] = \text{size of } T$; $\text{subVecB} = P(m_0/2, :)$ // ':' expresses An array of all columns of data kk is equal to 1/control value $\text{def} = (kk - 1)/n_0$;</p> <p>for each scanline T' in T from $m_0/2$ to $m - m_0/2 + 1$ do $\text{subVecA} = T(T', :)$ for each location j in scanline T' do if $\text{subVecA}[j]$ equal to $\text{subVecB}[j]$ then if kk equal to no then $\text{row} = T'$ $\text{col} = j - n_0 + 1$ return the location of matching else $kk++$; end if else if $\text{def} < \xi$ then if kk equal to n_0 then $\text{row} = T'$ $\text{col} = j - n_0 + 1$ return the location of matching else $kk++$; $\text{def} = (kk - 1)/n_0$; end if else $d = 0$; $kk = 1$; end if end if end for end for end function</p>
--	---

Figure 4. String matching algorithm pseudo-code: (a) exactly match; (b) approximate matching

4.1 Approximate string matching algorithm

Approximate string matching algorithm is a flexible string matching algorithms, allows the target string and sub-pattern string have inconsistencies. We extract a line of data from the template image, and use it compare with target image line by line, looking for sub-Mode string may appear in the target image position. Approximate string matching algorithm is a simple change based on string matching algorithm. Through the study found that the target image in the scanning process, the underlying code may have reversal, stretch and other distortion. It will hard to match if we use normal string comparison algorithm. Approximate string matching algorithms allows the entire string flip a certain number of bits to the maximum possible for a matching sub-pattern mode string. If the sub-pattern string can find in the target image, we requires the second step that the calculation of correlation coefficient, to confirm the exact matching.

We extracted a scan line of data from the template as sub-mode string. Comparison of the location of target mode string is the first line from the target image began, and then extracts one by one and compares with sub-string pattern. If the occurrence of child pattern string matching the target string, the program will return the matching position, including the rows and columns. We only know that the location is the template image may appear in the target image; otherwise, over this matching process. Figure 3 was allowed errors and does not allow errors string matching algorithm pseudo code, ξ is matching error. We can achieve rapidly matching both English and Chinese document image. By experiment, the English document image has a low probability of occurrence of distortion, while the Chinese document image is not the same, because of its shape structure is

complicated, and the probability of occurrence may have a greater likelihood of deformation. Therefore, Chinese document images need to be more broad tolerance, Figure 4 (a) shows the accurately string matching algorithms that the probability of deformation occurs relatively small, such as English documents; Figure 4 (b) shows the approximate string matching algorithms, we adjust the size of ξ to meet the feature of document images. It is obvious that, the error ξ smaller, the efficiency of the algorithm can obtain higher efficiency.

After pretreatment, we can get the black and white document images (binary images), through analysis, encoding of the document image from the bottom string of 0 or 1, each document image has a unique serial code of 0 or 1. In the string comparison process, the middle scan line is selected as sub- model string from the template, definitely not composes of all white pixels. Recycling call Libtiff library's TIFFReadScanline function to read the target line as the target image's data string [22]. One by one with the pattern string comparison with the target string to see if it matches, the matching algorithm is to allow a certain number of bits of inconsistency, if the match go to the next step, or transferred to the next template until all templates are compared with the end.

4.2 Document image's correlation coefficient

Document image's correlation coefficient can be used to determine the consistency of the two images. [4] proposed a normal images matching method, but this is not good for document images. As including very large amount of information, it is difficult to achieve efficient matching for document images. Our algorithm can overcome this defect, through approximate string matching algorithms, we can find the template image's possible position in the target image, then extract the corresponding image block as target data block, and calculate the correlation coefficient between the target data block and template data. If the correlation coefficient less than threshold δ (0.9), we think that target image have not template image's information. $M \times N$ image $f(x, y)$ as target image, and $J \times K$ sub-picture $p(x, y)$ is template image. The correlation between f and p can be expressed as:

$$c(x, y) = \sum_{s=0}^K \sum_{t=0}^J p(s, t) f(x + s, y + t) \quad (1)$$

Where, $x = 0, 1, 2 \dots N - K$, $y = 0, 1, 2 \dots M - J$. Before matching, p is set to the origin of upper left corner of sub-image, coincides with the target image's upper left corner. $c(x, y)$ is to calculate the dot product sum of the corresponding pixel between the template image p and sub-image which is covered by p in object image f . Then, we can know correlation coefficient between the two images. when the correlation coefficient greater than a predetermined threshold δ , the sub-image and target images occur matching, otherwise, point by point move $p(x, y)$ until the match ended.

Correlation coefficient can find the position of the entire template image's location in the target image $f(x, y)$. But in terms of normal application, In order to reduce unnecessary the computational load, we only judge the target image $f(x, y)$ whether contains template image's information, and no need to find all the matching location. Therefore, we can be a sub-image p as stored by row or by column vector, the calculation process will be covered by the graphics area p as the vector stored in the same way, through this transformation, we can calculate the angle between the two vectors, and their values, when exactly the same direction of two vectors, the value is 1, the correlation coefficient is converted into a dot product between two vectors operations, such as (2) below:

$$\vec{a} \bullet \vec{b} = |\vec{a}| |\vec{b}| \cos \theta \quad (2)$$

The calculation of the corresponding correlation coefficient between the images as:

$$c(x, y) = \frac{\sum_{s=0}^K \sum_{t=0}^J p(s, t) f(x + s, y + t)}{\left[\sum_{s=0}^K \sum_{t=0}^J w^2(s, t) \bullet \sum_{s=0}^K \sum_{t=0}^J f^2(x + s, y + t) \right]^{1/2}} \quad (3)$$

In this article, we only operate the binary image, (3) can do something to simplify calculation of the equivalent, (3) simplified into equation (4). Template image's calculation is equivalent to calculate the number of black pixels, marked with the $\text{sum}_1(p)$. While fp expresses sub image that is covered by p . $\text{sum}_1(\text{fp})$ means sub-image's the number of 1. The numerator calculates the number of 1 in matrix, which is the dot product of p and fp .

$$c(x, y) = \frac{\text{sum}_1(p \bullet f_p)}{\left[\text{sum}_1(p) \bullet \text{sum}_1(f_p) \right]^{1/2}} \quad (4)$$

When the image occur a small perturbation, the target image and the template image's correlation coefficient will not be less than 0.9(the correlation coefficient's threshold is 0.9 in this paper). If the document of the image's correlation coefficient larger than 0.9, we think we find the position of the template image in target images.

Because target image is binary image, we pass through rewriting the above formula, to avoid a large number of multiplication operations, using the same add operation, improve the calculation speed.

5 Experiment

This experimental environment is running: Ubuntu10.10 operating system, CPU Core Duo dual-core frequency of 2.2GHz. Need to install open source software for spandsp3.8 terminal, Libtiff4.0 image processing library, there is no specific requirements on computer hardware.

TABLE I. PERFORMANCE COMPARISON

Image	Matching Time		
	Normal(s)	Ours' algorithm(s)	SpeedUp ratio
Image 1	54.2190	0.9380	57.8
Image 2	54.6870	2.375	22.4

Made in the template, we can take two methods. The first method, selected from the target image representative of the sub-image block as a template images, produced by this template method can filter the same image on the server; the second method, we can be interested in the text, made into a standard template, into the template library, this template method of making complex than the first method, but the application of a higher demand for its main text Image classification and retrieval. After selecting the template, the same need for a pretreatment. Doing approximate string matching, we extract the middle line from template as a model string, there are two main reasons: First, it can shield the error caused by the producers of the template; second, when the template image of the scan line is all white pixels, the template will be matched with all the images, so failure. After having the template, we will need to put the template and the corresponding information into the template library, the template information is mainly the length of the template, the width of the template, the template vector set of eigenvalues and storage path.



Figure 5. Match diagram:(a) English pre-defined templates(b) Chinese pre-defined templates(c) After completion of the English image matching(d) After completion of the Chinese image matching

This paper experiment used the document images mainly composed of two types of images: the English and Chinese images document image. As shown in Table 1, Image1 is a class of English document images; Image2 is a class of Chinese document images. Figure 5 shows the template and target images' match the diagram. Through experiments we can draw two conclusions: 1, for the same size of the two images, the English document image matching algorithm is faster than the Chinese image, this reason is that English document images have simple structure. English document images' correlation Coefficient is closed to 1, while the Chinese has a certain deformation of the image; 2, we design the hierarchical matching algorithm is significantly better than traditional methods.

6 CONCLUSION

This paper mainly studies how to solve document images fast matching's the problem. It mainly contains two issues: (1) how to quickly find the template image may appear in the location of the target image; (2) How to calculate the correlation coefficient between matched sub-images and the template image.

Pointing to document images fast matching's the problem, this paper proposes a hierarchical algorithm for images fast matching to solve it. Firstly, we use approximate string matching algorithm to quickly find the template image's position that it may appear in the target document image. Secondly, we determine the two sub-images determine whether exactly matching by computing their correlation coefficient. Finally, we validate the method by experiment, and the experiments show that the method performs very well.

ACKNOWLEDGMENT

Supported by the National Grand Fundamental Research 973 Program of China under Grant No.2011CB302601; the National High-Tech Research and Development Plan of China under Grant No.2007AA010502 and 2009AA01Z474; the National Natural Science Foundation of China under Grant 90818028;

REFERENCES

- [1] Castleman K R. Digital Image Processing[M]. Prentice_Hall International,inc.,1998:552-553.
- [2] Liu Zhongwei, Zhang liujing. A Comparative and Analysis Study of Ten color Feature-based Image Retrieval Algorithms. SIGNAL PROCESSING, 2000,16(1):79-84.
- [3] Lv Jianping, Huang Ying. Several methods of image analysis and comparison of color quantization[J]. Computer Applications, 2002,22(5)86-87
- [4] Zhang zheng, Wang Yan ping, Digital image processing and machine vision , Posts & Telecom Press, pp.430-434,2010.
- [5] Chang F , Retrieving Information from Document Images : Problem and Solution . International Journal on Document Analysis and Recognition,2001,4(1):46-55.
- [6] Tan C I, Huang W, Yu Z, et al. Imaged Document Document Retrieval without OCR[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(6):838-844.
- [7] Tong Li, Ping Xijian. Image Feature of Information Measurement and Document Image Classification. Computer Engineering. 2004.9. Vol.30 NO17.
- [8] Wu Zhenjun , Ding Xiaoqing. Implementation of Robust Multi-Font Printed English Character Recognition System.
- [9] Wang Hong. Study on Printed Chinese Character Recognition Pre-processing,2009.
- [10] Lu Y, Tan C L. Chinese Word Searching In Imaged Documents[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2004, 18(2): 229-246.
- [11] Andreev, Andrey and Kirov, Nikolay. Document Search in Document Images Based on Hausdorff Distance Measures[C]. Gabmvo, Bulgaria: Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing,2008(1): 1-6.
- [12] Han Xiaolin, A Study OR Chinese Document Images Retrieval, Computer Knowledge and Technology, Vol.5, NO.26,2009.
- [13] Zhou Changfa. Proficient in Visual C++ image processing program (second edition) [M]. Beijing: Electronic Industry Press, 2004.
- [14] GATOS B, PRATIKAKIS S J. Adaptive degraded document image binarization[J]. Pattern Recognition, 2006, (39): 317—327.
- [15] WANG Bin, LI Xiangfeng, LIU Feng, et al. Color document image binarization based on binary documenture analysis[M]. Pattern Recognition Letters, 2005, (26): 1650—1657.
- [16] Ksahoo P, Soltani S, Wong A K C, et al. A Survey of thresholding techniques[J]. Computer Graphics Vision and Image Processing, 1988, (41): 233—260.
- [17] Zheng Qinfen, Rama Chellappa. A computational vision approach to image registration[J]. IEEE Trans. Image Processing, 1993. 2(3): 311—325.
- [18] ZHOU Guan-wei, PING Xi-jian, CHENG Juan. Skew detection and correction method of document images based on improved Hough ransform. Computer Applications, 2002, 31(4): 315-317.
- [19] Wang Tongqian, Zhu Yongquan, wanghong, Document Images Skew Correction Based on Run-length Smoothing, Computer Engineering, Vol,30,NO.1, 2004.
- [20] complex Chinese newspaper layout analysis, understanding and reconstruction, Chen, Xiao-Qing Ding, Liang Jian, Tsinghua University Natural Science Edition, 2001, vol.41, No.1]
- [21] Zhu Yongsong, Gong Chengming. "The Research of Correlation Matching Based on Correlation Coefficient", signal processing, vol,19,NO.6.
- [22] LibTiff lib.download from <http://www.libtiff.org/>.