Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

# CMSOF: a structured data organization framework for scanned Chinese medicine books in digital libraries*

Jie YUAN, Bao-gang WEI†‡, Li-dong WANG, Wei-ming LU, Yue-ting ZHUANG

(*School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

†E-mail: wbg@zju.edu.cn

Received Sept. 14, 2010; Revision accepted Sept. 26, 2010; Crosschecked Sept. 14, 2010

**Abstract:** Organizing unstructured information from books into a well-defined structure is a significant challenge in digital libraries. Most digital libraries can provide only search services at the granularity of books and few libraries allow books to be accessed at the granularity of chapters, as manually constructing directory information for books is time-consuming. Extracting structured data from scanned books thus remains an urgent and important work. In this paper, we propose a novel structured data organization framework called CMSOF to organize scanned data automatically, and apply it to a Chinese medicine digital library. In the framework, image blocks and text blocks on the scanned page of books are separated based on the gray histogram projection method or a hybrid method of region growth and the Ada-Boosting classifier at first, and then the text structure is obtained from text blocks by text size and font type recognition. Finally, image blocks and structured OCRed text are correlated at the semantic level. By integrating the structured data into a Chinese medicine information system (CMIS), we can organize the Chinese medicine books well and users can access the books with flexibility, which indicates that CMSOF is an efficient framework to organize books mixed with images and text.

**Key words:** Digital library, Chinese medicine, Structured data organization, Cross media, Image separation
**doi:**10.1631/jzus.C1001007 **Document code:** A **CLC number:** TP391.4

## 1 Introduction

With the development of digital libraries, more and more books have been scanned and preserved into the libraries. Nowadays, books are more vivid than before, since they contain not only text but also images (Fig. 1). Besides, text in books often has structure, which can be used to provide more fine-grained services. However, few digital libraries use this information to enrich the services in libraries. Large digital libraries such as Google Book (Google Book Search, http://books.google.com) and CADAL (China-America Digital Academic Library, http://www.cadal.zju.edu.cn) have scanned numerous books, but they provide only metadata-based search services at the granularity of book.

To make full use of the books, there are several challenges: (1) Books are scanned in image format and the book pages are a mixture of text and images; it is difficult to separate the text block and image block because layouts and image types vary significantly in different books. (2) Text structure information, which is very important for organizing information, is discarded in almost all Chinese optical character recognition (OCR) systems. (3) The OCR errors in books make it difficult to integrate newly scanned information into the existing information system. For example, we have scanned many Chinese medicine books containing text and images which can be used to enhance the Chinese medicine information system (CMIS) built before; however, due to OCR errors, items about the same medicine in CMIS and newly scanned books did not match.

‡ Corresponding author

* Project supported by the China Academic Digital Associative Library (CADAL)


**Fig. 1 A scanned book page with mixed text and images**

There have been some studies on document structure and layout analysis in recent years (Namboodiri and Jain, 2007; Lu *et al*., 2008). Article metadata being generated automatically based on format and text features extracted from OCRed texts has also been studied (Lu *et al*., 2008). While their research object is a document, generally it has had a uniform format and images are not taken into account, so the methodology is not readily generalizable to the Chinese condition and our application.

Chinese medicine is a valuable cultural heritage of China, and we have scanned many related books in CADAL. In this paper, we propose a novel structured data organization framework called CMSOF for organizing the scanned Chinese medicine books. At first, text blocks and image blocks are separated, and then the text blocks are analyzed to obtain the structured information. Then the text and images in the same or adjacent pages are semantically correlated. Finally, the extracted information is integrated into CMIS to provide fine-grained services.

The primary contributions of the paper are summarized as follows:

1. We propose a framework called CMSOF to organize data in scanned books. CMSOF contains layout analysis with image/text separation, text structure extraction, image/text semantic relationship construction, and data integration with CMIS. It can organize structured cross-media data automatically.

2. We present two different image separating methods, the gray histogram projection method and a hybrid method of region growth and the AdaBoosting algorithm, to separate image blocks from text blocks in a scanned page.

3. We present a method to extract text structure in scanned books by text size and font style reorganization to organize text in a good structure.

4. We propose an integration sub-framework combining extracted text and image data with the CMIS established previously. The sub-framework can correct some OCR errors, and it provides administrators a very flexible operation mechanism.

## 2 Related works

There have been some studies on document layout analysis, data structure mining, and other data mining work from scanned books. le Bourgeois *et al*. (2004) introduced some general problems of digital libraries. They classified the problems into two classes: common problems and particular problems. Common problems include image details loss owing to the store format, image post-processing, metadata auto-extracting, and so on. Particular problems are some problems that occur in particular application conditions, such as digital processing of 18th century European manuscripts. Gatos *et al*. (2005) proposed a technique for automatic table detection in document images. After pre-processing of document images, they used mainly morphological operations and threshold filter to detect table lines. Namboodiri and Jain (2007) extracted document structure and analyzed layout from documents with a complex layout. Lu *et al*. (2008) proposed a supervised learning based method to generate description and structure metadata of digital books. The OCRed text was stored in DjVu XML files. This format contains not only plain OCRed text, but also the logical structure of text and the surrounding rectangle of every text word. While in Chinese OCR systems, these functions have not been provided by now. Liu *et al*. (2010) proposed a semi-supervised learning method for detecting text-lines in noisy document images. They used the seed filling algorithm for initial segmentation, then the projection profiles for estimating the vertical border of page contents, and finally a classifier for removing speckle noises embedded inside the content zones. The above methods analyze mainly document layout without taking images into account, while in

our application image blocks are as important as text blocks. Thus, it is unsuitable to use these methods to extract structure information from scanned Chinese books with mixed text and image data.

In Chinese medicine book digitalization, Shi *et al*. (2009) achieved a search engine with digital books based on manually typed catalog data; moreover, they developed a semantic recommendation framework. The framework provided users an interface, and users could select different attributes such as composition, effect, and growth area. The system could discern the book name and the corresponding chapter of the related medicines that have the same value as or similar values to the selected attributes. Zhu *et al*. (2009) used a semi-supervised text extraction method to extract information from Chinese medicine books, and then used a relationship discovering method based on Chinese medicine prescription parameterization to mine the latent relationship between different prescriptions. In their system, metadata and the relation between metadata and multimedia data were created manually.

We can see from the above that the techniques for extracting and organizing different kinds of multimedia information in Chinese medicine books are not very mature now, and that other existing Chinese medicine knowledge systems work mostly with manually inputted catalogs and other metadata. In contrast, the framework CMSOF proposed in this paper can reduce users' interference. It extracts useful information from medicine books and organizes them in a structure according to their semantic relationship. CMSOF provides users a friendly interface to learn and search for Chinese medicine knowledge, helping protect and inherit Chinese medicine.

## 3 Sketch of CMSOF

The sketch of CMSOF is as shown in Fig. 2. The scanned page images first enter the image separating module. Two methods can be used to separate image blocks: the gray histogram projection method and a method combined with region growth and the Ada-Boosting algorithm, and users can select one according to the application condition. The output of the separating module is small image blocks and the remaining text region; the text region then enters the

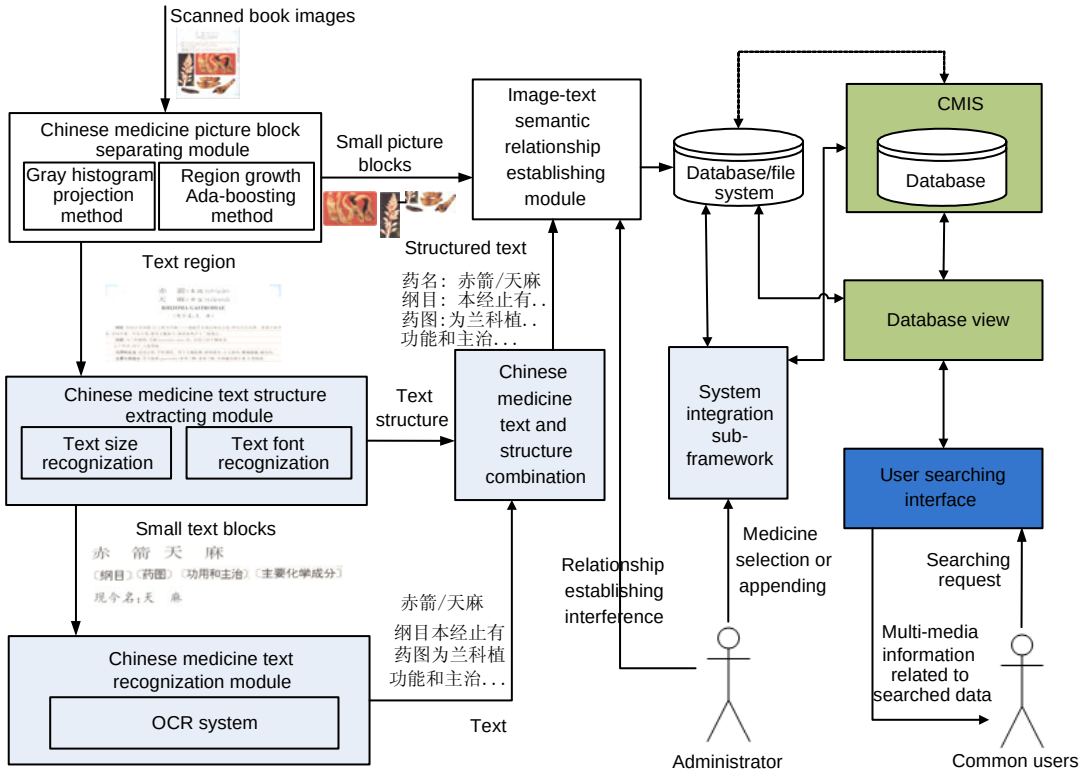



**Fig. 2 The sketch of CMSOF**

text structure extracting module. In this module, the remaining text region is further divided into some small text blocks and these small blocks are collected into several different categories according to text size and font style, and then OCR API (application programming interface) is called to recognize text in these small blocks. The recognized text is combined with logical text structure extracted above, and then the semantic relationship between these well structured text and small image blocks are established and stored in a database or file systems (generally text information and image block paths are also stored in a database while the images are stored in files). The semantic relationship establishing module usually connects small image blocks to the nearest front medicine name. Generally speaking, the medicine name and its related image blocks are on the same page or the consequent pages, so the relationship can be established automatically or with a little manual intervention. A system integrating sub-framework is used to integrate the new structured data into CMIS. If there is no existing knowledge system, this module can be ignored. Medicine names are used to connect data in CMIS and new data. Because of the limitation of the OCR technique and the poor quality of some scanned page images, the OCRed medicine names may easily be mistaken. To enhance the system robustness, medicine names not contained in CMIS are amended from two aspects of medicine attributes and image processing. On the other hand, we provide a function to permit administrators to append new medicines. When a user searches medicines, our system offers him/she multi-aspect multimedia information related to the searched medicine including text, images, videos, maps, and so on.

This paper introduces mainly the picture/image block separating module, text structure extracting module, and system integration sub-framework, because existing techniques in these domains have provided poor results in dealing with Chinese medicine books.

## 4 Picture/Image block separation

In order to arouse users' interest in Chinese medical knowledge, a number of images and other multimedia data should be provided to them. We have scanned some Chinese medicine books containing both text and pictures. In this study, we take the book *Colored Illustrations of Drugs from Bencao Gangmu* (CIDBCGM for short) as an example. For scanned books with pictures and text, the first step is to separate small picture/image blocks on every page. We propose two separation methods for use in different conditions.

### 4.1 Image separation based on gray histogram projection

For some books without a complex layout, image blocks can be separated from the text region by projecting the corresponding gray image on row and column. The number of columns on a page should be decided first and interference information such as decorating patterns should be removed. Books usually have one or two columns. To obtain the column number of a book or a page, the image's gray histogram can also be projected onto the *X* axis. If there is a blank gap of a certain width, there would be two columns on a page. To remove decorating patterns on the margin of a page in some books, gray histogram, gap detection, and the gray threshold filter can also be used. Fig. 3 shows how to use gray histogram projection to separate picture/image blocks.

Fig. 3 shows that the gray cumulative histogram corresponding to the text region looks like pulse wave, and that the pulse width is generally equal to text width. On the other hand, the histogram corresponding to the image region maintains a relatively

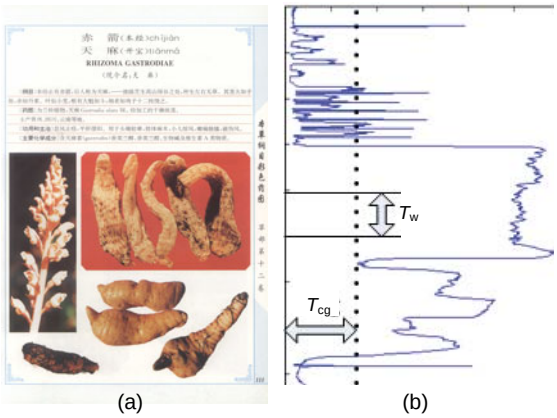



**Fig. 3 Separating image blocks on a page with simple layout based on gray histogram projection**
(a) Page sample; (b) The corresponding gray cumulative histogram. For simplification, the gray image is reversed, so white color corresponds to value 0 and black color corresponds to value 255

high value in a certain span generally larger than text width. To remove the interference of background, gray values smaller than a threshold $T_g$ are set to 0 (here white). The row span of gray histogram smaller than a width threshold $T_w$ or the gray value smaller than threshold $T_{cg}$ is also discarded to remove text areas and other unwanted regions, and the row range of picture/image blocks can be obtained. The column range of image blocks can also be obtained using the same method. Finally, the approximate range of all image blocks on a page can be identified.

Generally speaking, there are usually more than one image block in a scanned page image and the shape of image blocks are also not regular, so every image block's minimum surrounding rectangle can be obtained for simplicity When the image block is not regular, the above projection histogram method does not work well. In this case, a larger threshold $T_{cg}$ is used to obtain a small core picture/image block, and the region growth method is used to obtain the whole image block and its surrounding rectangle. There is another extreme case where the image region is divided into so many small blocks that not every small block can express an intact meaning. To solve this problem, visual features of the image block are extracted and the similarity between each other is calculated. The blocks with high similarities are merged. In this work, we use only color features.

Fig. 4 shows the results of using a simple image separation method.

(a) (b)

**Fig. 4 Image separation results on a page with one (a) and more than one (b) column using the gray histogram projection method**

Rectangles with different colors represent image blocks with dissimilar visual features

### 4.2 Image separation using region growth and the Ada-Boosting algorithm

The above separation method works poorly on some scanned book pages and newspapers that have a complex layout. To deal with more general and complex scanned data, we have developed another separation method based on region growth and the Ada-Boosting algorithm (Schapire, 1999).

The main idea of the new separation method is as follows: first all single connected blocks are found using the region growth method, and then very small blocks and blocks with a large ratio of width to length are rejected. Moreover, the connected blocks that denote text are rejected using machine learning methods. Finally, what remain are the image blocks.

The main difficulty in the above procedure is to find the genuine image blocks from the candidate image block set, which includes not only genuine image blocks, but also other disturbed blocks such as text blocks. To remove text blocks we develop a method based on the Ada-Boosting algorithm, and select SVM as the weak classifier. The algorithm is as follows.

Step 1: Collect image blocks and text blocks respectively.

Step 2: Extract wavelet energy features in every block. Create the training set

$$\mathrm{TSI}=\{(\boldsymbol{x}_1, y_1),(\boldsymbol{x}_2, y_2),\ldots,(\boldsymbol{x}_n, y_n) \mid \boldsymbol{x}_i\in\mathbb{R}^m, y_i=-1, +1\},$$

where $\mathbb{R}^m$ denotes an $m$-dimensional wavelet feature. If the block is an image block, $y_i=+1$; else $y_i=-1$. We will explain how to extract wavelet features in Section 5.

Step 3: The initial weight of every block is set to $1/N$, $w^1(i)=1/N$, $i=1, 2, \ldots, N$, where $N$ is the total number of blocks.

Step 4: For $t=1, 2, \ldots, T$, take $w^t(i)$ as the sampling probability of the $i$th sample, and use samples on TSI to obtain the train sample set $\mathrm{TS}_t$ of time $t$. Then the $t$th SVM weak classifier $\mathrm{WC}_t$ is created from the former training sample set $\mathrm{TS}_t$.

Calculate the training error of $\mathrm{WC}_t$,

$$\varepsilon_t = \sum_{i=1}^{n} w^t(i) I(y_i \neq \mathrm{WC}_t(i)), \qquad (1)$$

and set the weight of weak classifier $\mathrm{WC}_t$ as

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right). \quad (2)$$

Update every sample's weight in training set TSI:

$$w^{t+1}(i) = \frac{w^t(i)\exp(-y_t w^t(i)\mathrm{WC}_t(i))}{Z_t}, \quad (3)$$

where $Z_t$ is the normalization factor.

Step 5: Create the strong classifier by combing all above weak classifiers as follows:

$$\mathrm{SC}(x) = \mathrm{sign}\left(\sum_{t=1}^{T} w^t \mathrm{WC}_t(x)\right), \quad (4)$$

where $x$ is the wavelet feature of a to-be-classified block.

Fig. 5 is an example of comparing the region growth method without and with combining the Ada-Boosting algorithm, showing that the hybrid method has a better effect. The classification accuracy of the strong classifier reached more than 80% on the testing set containing newspapers and some books.


(a) (b)

**Fig. 5 Image separation performance of the region growth method without (a) and with (b) combining the Ada-Boosting algorithm**

### 4.3 Image separation experiment

The scanned book CIDBCGM consists of 1256 pages (corresponding to 1256 scanned images) and contains 2085 image blocks in all. We use 'precision' and 'recall' to measure the effect of the two methods we proposed above. Both methods were very effective (Table 1). Using GP, 2076 image blocks were obtained, and 2059 blocks were correct, while using the RG&B method, 1957 image blocks were obtained, and 1948 blocks were correct. Because the layout of the book is not very complex, GP is more effective than RG&B on precision. The limitation of classifier in the RG&B method results in the relatively low precision compared with GP. The difference of recall is minor.

**Table 1 Separation performance comparison**

| Method | Precision (%) | Recall (%) |
|---|---|---|
| GP | 98.75 | 99.18 |
| RG&B | 93.43 | 99.54 |

GP: gray histogram method; RG&B: hybrid method of region growth and the Ada-Boosting algorithm

## 5 Text structure analysis based on text size and font style

In general, the text sizes of titles and other important information on a page are larger than that of normal text; the structure of a book is consistent, which means that text having the same size and font style also has the same level of structure information. Thus, text structure information can be extracted based on text size and font style.

The gray projection and region growth method can also be used to find text with the same size. However, sometimes text with the same size does not share the same font style or structure (Fig. 6). Thus, text font should also be recognized. Chinese has more fonts than English, and Chinese fonts vary more greatly than English fonts. In this study, we take Song typeface, FangSong typeface, regular script, and bold typeface as examples to illustrate font recognition.

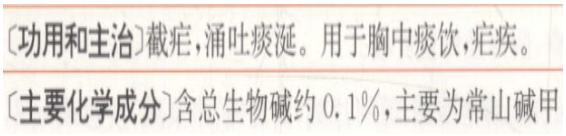



**Fig. 6 Text with the same size but different font styles**

Zhu and Tan (2001) took a scanned text image as an image that contains a special texture, then font recognition as texture identification. In their method,

text blocks are first preprocessed into blocks with the same size, layout, and density, and then 16 Gabor filters with four different directions and four kinds of frequency are used to obtain feature vectors of text blocks. Finally, all these features are sent to a classifier to obtain the font style. Their method achieved a good result, but required that the to-be-tested text block should have only one font style, so the method does not work well in text structure analysis.

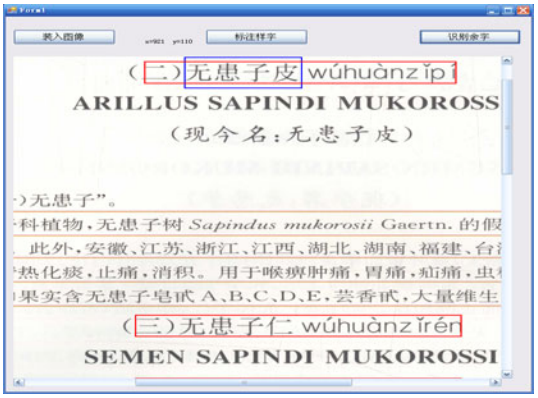
Our method can work on a single character text block. For a text block with only one Chinese character, it is first normalized to a block with size 64×64, and then two-levels Harr wavelet decomposition is imposed on it (Fig. 7). The low-frequency component varies according to special characters, so it is useless for font recognition and should be discarded. Subbands $IH_1$, $IV_1$, $ID_1$, $IH_2$, $IV_2$, and $ID_2$ are divided into small blocks with size 8×8 pixels, and then the energy ratio is extracted as the wavelet feature component:

$$f_{\text{grid}}(i,j)=\frac{\sum\left|\text{coef}_{\text{grid}}(i,j)\right|}{\sum\left|\text{coef}_{\text{level}}(i)\right|}, \tag{5}$$

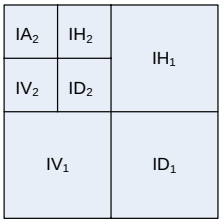
where $f_{\text{grid}}(i, j)$ denotes the wavelet feature component of the $j$th small block of the $i$th subband. $\sum\left|\text{coef}_{\text{grid}}(i,j)\right|$ denotes the corresponding sum of the coefficient absolute values, while $\sum\left|\text{coef}_{\text{level}}(i)\right|$ denotes the overall sum of the corresponding level of the $i$th subband. All the feature components are concatenated into a vector with 60 dimensions as the Chinese character image's feature, and then these feature vectors are sent to a classifier to decide the character's font style.

For a paragraph, the font styles of the first and last characters are identified. If they are of the same style, characters in the paragraph share the same style. Otherwise, font styles are identified character by character, until two continuous characters have different styles. The recognition precision is shown in Table 2.

In the example book, not all medicine name text shares the same size, so the precision is not very high. While the difference in text size between medicine aliases and other attributes is not notable, through combination with font style recognition, the precision is still not high.




(a)


(b) (c)

**Fig. 7 Text size and font style recognition**

(a) Text size recognition. Characters in the smaller rectangle are selected by users, and then the text size recognition method finds all characters with the same size on the current page. (b) Four representative Chinese fonts: Song typeface, FangSong typeface, regular script, and bold typeface. (c) Two-level wavelet decomposition

**Table 2 Text structure recognition result**

| Recognized item | Precision (%) |
|---|---|
| Medicine name | 85.7 |
| Medicine alias | 72.1 |

When the font size and font style of Chinese characters are extracted, the text structure on a page can be obtained. OCRed text with its structure information is then stored in a well-defined structure such as XML files or a database.

## 6 Integrating structured data into the Chinese medicine information system

There are 4443 traditional Chinese medicine science terminologies and their definitions, 11 161 medicines and their detail descriptions, 1338 prescriptions with their compositions and applications,

and 241 famous therapists and their specialty information in traditional Chinese medicine field in our CMIS. We can integrate the newly extracted structured text and picture information into the existing system.

The work needs two steps: (1) Correlate the separated image blocks and their corresponding structured text; (2) Relate the correlated image and text information to corresponding data in CMIS. In this work, an image block is related to its nearest front medicine name. In most cases this is correct. After correlating text and images, it is necessary to match current medicine to the corresponding medicine information in CMIS according to medicine names and establish the semantic relation between them. We present a system integration sub-framework (Fig. 8) to do this work.

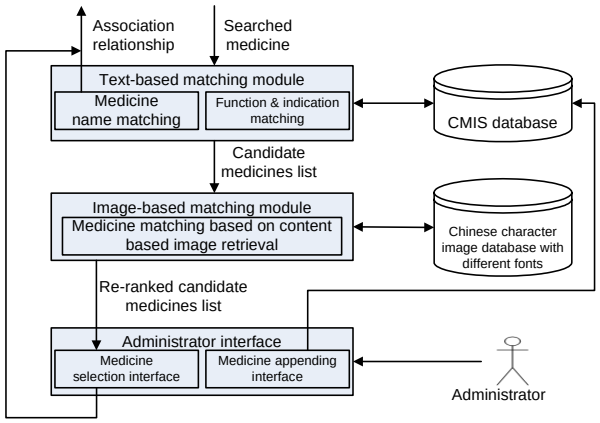



**Fig. 8 Structure of the system integration sub-framework**

The sub-framework consists of two matching modules and an interface. The text-based matching module searches medicine using string matching theory. If a medicine name in CMIS fully matches current medicine, their semantic association relationship is directly established and the control stream directly returns from this sub-framework. Otherwise, a candidate medicine list is obtained by this module, and the image-based matching module is used to re-rank the list according to character image similarity of medicine names while all these images come from a Chinese character image database with different fonts. Finally, the administrator can select a medicine from the candidate medicine name list in the interface, and he/she can also append the current medicine information into the CMIS database. If a medicine is selected, the association relationship is also established automatically.

### 6.1 Text-based matching module

For many reasons, such as the poor quality of scanned books, the poor performance of OCR software, and so on, many erroneous OCRed medicine names cannot be found in CMIS. In Fig. 9, the text block denotes medicine '茈胡' (Bupleurum), while the OCR software recognizes it as some other characters.

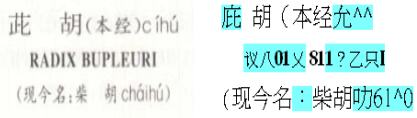



**Fig. 9 A medicine item in a medicine book and its corresponding wrong text information obtained by OCR software**

Since Chinese medicine has a long history, many medicines have different names or have had more than one name during different periods. Thus, when we search for a medicine by name, we should provide not only its current name but also its alias. Only when either the current name or the alias name is found, do we think they match. We divide the matching problem into three different cases:

1. The current name or the alias name can be found in CMIS.

2. A part of name can be found in CMIS.

3. There is no medicine in CMIS that can be found to fully or partly match the current name or the alias of the searched medicine.

In the first case, the relationship can be established directly between the current medicine and a medicine with the same name in CMIS. In the other two cases, there may be two possibilities. One is that the searched medicine is actually contained in CMIS, but the medicine name is wrong after OCR operation; the other is that the searched medicine is actually not contained in CMIS. For the second possibility, the searched medicine should be appended into CMIS, while we do not need to do so if the first possibility is true. Identifying whether the searched medicine actually is contained in CMIS or not is the issue. Two options are provided to the administrator: one is to list the most similar medicines in CMIS for the administrator to select; the other is to enable the

administrator to append new medicines manually. Below we explain mainly how to find the medicines most similar to the current medicine, using the text-based matching module.

If a part of the current medicine name can be found in CMIS, a list of medicine names partly matching the current medicine name or alias are first found in CMIS, and then these medicines are ranked by some other attributes. Besides the medicine name, common users are most interested in the medicine attributes of functions and indications. In this work, similarity between two medicines is calculated by combining the functions & indications attribute and the medicine name (including the current name and alias) attribute. We use Aminul Islam's three modified versions of the longest common subsequence (LCS) (Allison and Dix, 1986) for word semantic similarity measurement:

$$\begin{cases} v_1(r,s)=\mathrm{NLCS}(r,s)=\dfrac{\mathrm{Length}(\mathrm{LCS}(r,s))^2}{\mathrm{Length}(r)\cdot\mathrm{Length}(s)}, \\ v_2(r,s)=\mathrm{NMCLCS}_1(r,s)=\dfrac{\mathrm{Length}(\mathrm{MCLCS}_1(r,s))^2}{\mathrm{Length}(r)\cdot\mathrm{Length}(s)}, \\ v_3(r,s)=\mathrm{NMCLCS}_n(r,s)=\dfrac{\mathrm{Length}(\mathrm{MCLCS}_n(r,s))^2}{\mathrm{Length}(r)\cdot\mathrm{Length}(s)}, \end{cases} \tag{6}$$

where $\mathrm{MCLCS}_1$ denotes the longest common subsequence of strings $r$ and $s$ beginning from the first character, while $\mathrm{MCLCS}_n$ denotes the longest common subsequence beginning from any character. Length(str) returns the length of string 'str'. Note that all punctuation marks are ignored in Eq. (6). For example, if string $r$='清热消炎，凉血解毒' and $s$='清热解毒', then LCS($r$, $s$)='清热解毒', $\mathrm{MCLCS}_1$='清热', $\mathrm{MCLCS}_n$='清热' or '解毒'. $\mathrm{NLCS}(r,s)=4^2/(8\times4)=0.5$, $\mathrm{NMCLCS}_1(r,s)=2^2/(8\times4)=0.125$, $\mathrm{NMCLCS}_n(r,s)=2^2/(8\times4)=0.125$. Characters of a medicine's function & indication attribute vary in different books. For example, the description of medicine '百部' in CMIS is "润肺止咳，杀虫灭虱。1．新久咳嗽，百日咳，肺痨咳嗽。2．蛲虫、头虱及疥癣等。", while its description in CIDBCGM is "润肺下气止咳，杀虫。用于新久咳嗽，肺痨咳嗽，百日咳；外用于头虱，体虱，蛲虫病，阴痒症". If we directly calculate LCS on them, we will obtain LCS($r$, $s$)='润肺止咳杀虫新久咳嗽肺痨咳嗽头虱' while the common string '百日咳蛲虫' is lost. To avoid such a case, here we use a measurement which is independent of character order in a string. First some common words such as '用于/外/及/等' are deleted, and then the frequency histogram of every character in string $r$ or $s$ is created and another similarity measurement between $r$ and $s$ is calculated as follows:

$$v_4(r,s)=\frac{\left(\sum_{c\in C}\min(f_r(c),f_s(c))\right)^2}{\mathrm{length}(r)\cdot\mathrm{length}(s)}, \tag{7}$$

where $C$ is a character set consisting of all characters existing in string $r$ or $s$, $c$ is a character, and $f_r(c)$ is the occurrence frequency of character $c$ in string $r$ .Then the total similarity between strings $r$ and $s$ is

$$\mathrm{similarity}(r,s)=\sum_{i=1}^{4}w_i\cdot v_i(r,s). \tag{8}$$

where $w_i$ is the weight of $v_i$. In this work, $w_1$=0.6, $w_2$=0.05, $w_3$=0.2, $w_4$=0.15.

### 6.2 Image-based matching module

If every medicine name in CMIS does not fully match the current medicine, the above string matching method is first used on the medicine name and function & indication attribute to obtain 50 candidate medicines most similar to the current medicine. Then the image processing method is used to re-rank the candidate medicines and provide them to the administrator. The administrator can select one from them.

To calculate the similarity between two medicines for image processing, take the following steps:

1. Create several image sets, each of which contains image blocks of all Chinese characters or Chinese medicine characters of a Chinese font style, and regularize them to a uniform size of 64×64.

2. For the current medicine name text block in the scanned page image, obtain its font style using the method introduced in Section 5, and obtain the corresponding image set of all candidate medicines of the specific font style.

3. Take the medicine name text block as the query sample image, extract its PHOG (pyramid of histograms of orientation gradients) shape feature

(Bosch *et al*., 2007) and wavelet feature, and calculate the similarities between its feature and the features of all candidate medicine images. The $\chi^2$ distance is selected as the distance function:

$$D_j = D(s, I_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{\left(s(k)-I_j(k)\right)^2}{s(k)+I_j(k)}, \qquad (9)$$

where $s$ is the to-be-queried medicine name text block, $s(k)$ is its $k$th dimensional feature, and $I_j$ is the corresponding medicine name text block of the $j$th candidate medicine.

Gauss normalization is used to normalize the distance to a closed interval [0, 1]:

$$\mathrm{Sim}_{\mathrm{image}}(s, C_j) = 1 - \bar{D}_j, \qquad (10)$$

where $C_j$ is the $j$th candidate medicine and $\bar{D}_j$ is the normalized value of $D_j$.

4. Combine the image feature distance and text feature distance to obtain the total distance between the current scanned medicine and all candidate medicines (Eq. (11)), and sort these distances by ascending order. The ordered list then is provided to the administrator to select from.

$$\mathrm{totalSim}(s, C_j) = w\cdot\mathrm{Sim}_{\mathrm{text}}(s, C_j) + (1-w)\cdot\mathrm{Sim}_{\mathrm{image}}(s, C_j), \qquad (11)$$

where $\mathrm{Sim}_{\mathrm{text}}$ is the similarity calculated in Eq. (8) and $\mathrm{Sim}_{\mathrm{image}}$ is the same as in Eq. (10). In this study, the weight of $\mathrm{Sim}_{\mathrm{text}}$, $w$, is set to 0.5.

The correct recommending rate of the first five medicines is more than 82%, basically meeting the need of the application.

Fig. 10 displays the result after relating the Gastrodia (天麻) item in CIDBCGM to the corresponding item in CMIS. Through the matching relationship between drug names, the newly scanned image block of Gastrodia relates to the corresponding item in CMIS. Various structured multi-media information can be organized to better meet user needs and attract their attention.


**Fig. 10 Retrieval example after integration**

## 7 Conclusions

We propose a framework called CMSOF to organize information in a good structure from scanned Chinese medicine books. First, image blocks and text blocks in scanned book images are separated, and the separated small image blocks are individually stored in files. The text structure is extracted from text blocks by font size and font style reorganization. Then the stored small image blocks are related to the corresponding structured text. Finally, structured image and text information is incorporated into CMIS. CMSOF is robust to some OCR mistakes. It can also be used in other kinds of digital libraries where newly scanned books need to be integrated into existing information systems. Next we will improve the framework in the following aspects:

1. Study how to use this framework on PDF and DjVu files effectively.

2. Develop a general image/text separation algorithm and text structure mining algorithm, further reducing an individual's intervention, and improve the accuracy of these algorithms.

3. Further improve the framework efficiency.

## References

Allison, L., Dix, T.I., 1986. A bit-string longest-common-subsequence algorithm. *Inform. Process. Lett.*, **23**(5): 305-310. [doi:10.1016/0020-0190(86)90091-8]

Bosch, A., Zisserman, A., Munoz, X., 2007. Representing Shape with a Spatial Pyramid Kernel. Proc. CIVR, p.401-408. [doi:10.1145/1282280.1282340]

Gatos, B., Danatsas, D., Pratikakis, I., Perantonis, S.J., 2005. Automatic table detection in document images. *LNCS*, **3686**:609-618. [doi:10.1007/11551188_67]

le Bourgeois, F., Trinh, E., Allier, B., Eglin, V., Emptoz, H., 2004. Document Images Analysis Solutions for Digital Libraries. Proc. 1st Int. Workshop on Document Image Analysis for Libraries, p.2-24. [doi:10.1109/DIAL.2004.1263233]

Liu, Z.Y., Zhou, H.N., Yang, N., 2010. Semi-supervised learning for text-line detection. *Pattern Recogn. Lett.*, **31**(11):1260-1273. [doi:10.1016/j.patrec.2010.03.015]

Lu, X.N., Kahle, B., Wang, J.Z., Lee, C.G., 2008. A Metadata Generation System for Scanned Scientific Volumes. Joint Conf. on Digital Libraries, p.167-176. [doi:10.1145/1378889.1378918]

Namboodiri, A.M., Jain, A.K., 2007. Document Structure and Layout Analysis. *In*: Digital Document Processing, p.29-48. [doi:10.1007/978-1-84628-726-8_2]

Schapire, R.E., 1999. A Brief Introduction to Boosting. Proc. 16th Int. Joint Conf. on Artificial Intelligence, p.1401-1406.

Shi, S.M., Wei, B.G., Yang, Y., 2009. Msuggest: a Semantic Recommender Framework for Traditional Chinese Medicine Book Search Engine. Conf. on Information and Knowledge Management, p.533-542. [doi:10.1145/1645953.1646022]

Zhu, W.H., Wei, B.G., Zhuang, Y.T., Shi, S.M., Yang, Y., 2009. Content Integration in Digital Libraries. Proc. Int. Multimedia Conf. on Ambient Media Computing, p.57-64. [doi:10.1145/1631005.1631019]

Zhu, Y., Tan, T., 2001. Font recognition based on global texture analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **223**(10):1192-1200. [doi:10.1109/34.954608]