Retrieving Content Directly from Antique Book Images*

Zhifeng Chen, Yong Wang, Liang Zhang, and Baile Shi

Computer Science Department, Fudan University
220 Handan Road, Shanghai 200433, P.R. China
{zfchen,wangyyong,zhangl,bshi}@fudan.edu.cn

Abstract. The digitization and utilization of antique books are of significance. WWW lays a foundation for the dissemination of images, but more comprehensive mechanisms, such as content-based retrieval, are necessary to reduce information overload. It is well known that the content retrieval of a Chinese antique book directly from its bitmap is very hard due to intrinsic handwritten nature. In this paper, an original method for content retrieval based on visual similarity is proposed. By extracting features from book images, the method makes up a feature space and applies spatial indexing on it. A range searching strategy is then applied to retrieve all analogs to the query example. The prototypical implementation demonstrates the feasibility and supports both retrieving and browsing via popular Web browsers.

1 Introduction

Like other civilization legacy, antique books are rare and precious resources. For China, a country with long history and splendid culture, the connotation above sticks out acutely. However, raw digitized antique books are voluminous, their transmission in WWW is prone to block networks and overwhelm end-users. More comprehensive facilities, such as Optical Character Recognition (OCR) or searching mechanism are needed to reduce the bandwidth requirement.

Many projects are seeking to make use of rare and precious antique books [3,4,7]. They fall into two categories: *cataloging method* and *attached-text method*. In the cataloging method, which is widely used in library practices, some entries for a book, such as the book name, authors, publishing information, are manually produced. A user has to rely on them to retrieve information. A major conflict with the method is the limited retrieval points and unforeseen specific query purposes. On the other hand, the attached-text method constructs a text file by OCR for a specific antique book [3,7]. An index is then built on this textual file and full-text retrieval technology is applied. However, in the context of Chinese antique books, this method does not work well due to the

* This work is a sub-project of the grant from Natural Science Foundation of China for the *Research on Key Techniques in Digital Library* project (No. 69933010). One of authors is supported by the "Chun-Tsung Scholar" Foundation.

H. Lu and A. Zhou (Eds.): WAIM 2000, LNCS 1846, pp. 380–385, 2000.

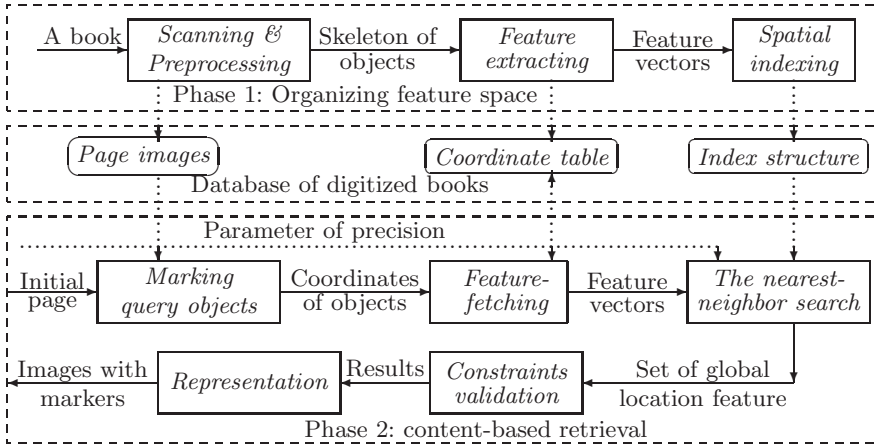

Fig. 1. Flowchart of contents retrieval by visual similarity criterion

brush-written holograph nature. Its other problems are the identity judgement between page images and the textual data, the size of lexicon and its coverage. In addition, *TONG JIA ZI* (interchangeable words or characters) might disable full-text retrieval. It is necessary to break a new path for content-based retrieval.

This paper is organized as follows: Section 2 gives the framework and its rationale. Section 3 elaborates on vital techniques we developed. Section 4 describes our prototype. Finally, we present the conclusion in Section 5.

2 Methodological Description

We believe that content-retrieval of Chinese antique books should proceed on the visual semantics of objects rather than on the linguistic semantics. This strategy enables full-text-like queries directly on page images and clears up previous problems. It can reduce the network overload by discarding irrelevant information and preserve the pleasure of reading as well.

In proposed method, we regard every virtual character or symbol, called an *object*, as an image that is a chunk of relatively clustered pixels in a page image, so the page image is composed of many such smaller images arranged in a certain order. As illustrated in Figure 1, the method consists of two phases: feature space construction and content-based retrieval. The former prepares a database for the latter. More specifically, during the *scan & preprocess* module, an original antique book is digitized into a series of page images. Further, every object is separated and is processed by thinning algorithm. Accessory functions in this module include layout alignment, noises elimination, and binarization. They are common techniques in Chinese OCR. Then, the *feature extraction* module extracts content features for every object skeleton, and feeds them into

the high-dimensional *spatial indexing* module where the object is favorably organized by its morphological feature. This process continues until all objects in the antique book are dealt with and a global index structure is established.

From now on, content-based retrieval can work repeatedly. This phase is actuated by a page image identified by a given page number or other means (e.g. cataloging method). While browsing page images, a user marks objects as he/she wishes. Marked objects and their order form the query sample. Coordinates of marked objects are supplied to the *feature-fetching* module where features of each marked object are obtained by looking up the *coordinate table* [1]. The order alone also stands for constraint for the purpose of validation. Now, *the nearest-neighbor search* module can find analogs to each object in the query sample. It produces a collection of analog sets for further process. The *constraint validation* module takes this collection and previous constraint as its input, checking combinatorial validity. The *Representation* module receives valid results from its precursor and marks proper objects on page images.

In this way, the intricate retrieval is decomposed into four tractable consecutive sub-problems: feature extraction, index structure construction, the nearest-neighbor search and constraint validation.

Because the query sample comes directly from page images that act as a stylebook, the problems of identity, character set, special symbols, interchangeability of characters and the lexicon no longer exist. Through approximate matching, we avoid the difficulty which the *attached-text method* brings about, i.e. the obstacle of concept formation.

Another advantage of this method over the attached-text method is that the decision of recall/precision balance is delayed until the retrieval moment, and can be easily controlled. In the attached-text method, the meanings of objects are stiffened at the OCR stage, i.e. a decided linguistic semantics (Chinese character) must be assigned to an object. A user has no authority to re-mapping them at later retrieval stage. In the context of antique books, the re-mapping becomes more desirable because there are various factors that disturb the mapping.

Our method is coherent with the cataloging method that is extensively used in libraries. Guided by catalogs, users can find out a specific volume of ancient books. Then, content-based information retrieval is ready to help the user find targets in the book.

To summarize, the main contribution of our work is the original approach to the content-retrieval problem in the context of Chinese antique books, and some key techniques supporting the method.

3 Key Techniques

3.1 Extracting Feature

Three basic categories of feature, say the *global location feature*, the *page feature*, and the *morphological feature*, are combined to describe an antique book. An-

[1] The coordinate table is a set of triples (Page-number, Coordinate, Feature-vector)

other feature, *book feature*, should be defined if we bound several antique books into one processing unit. In our prototype, we only elaborate upon the basic ones.

Definition 1 (global location feature, GLF). *The GLF of an object in an antique book is the sequential order number of the object within page images of the book.*

Recall that objects are correctly separated in the *scan & preprocess* module. We can identify each object according to reading habit (usually page by page and in each page from the right to the left, then from the top to the bottom). In the case of complicated layout, we can first use a recursive-curve (e.g. Hilbert's) to scan areas, then serialize objects by conventional means. Hence, a GLF uniquely identifies an object.

Definition 2 (page feature, PF). *The last GLF on each page defines the PF of an antique book.*

Definition 3 (morphological feature, MF). *The MF for an object is formed by its accumulative stroke statistic within centroid-based multi-level partitions.*

We use the MF to describe the visual semantics of an object. This is based on our observation that the configuration and componential proportion of a Chinese character remain stable [6], yet various factors, such as disparate strokes, illegible and disturbed strokes, and variety in length call for error tolerant methods of feature extraction.

Before extraction, the bitmap of each object is normalized by a MBS (Minimum Bounding Square). The object is centered within a square whose side is the maximum of width and height of its foreground. Compared with conventional MBB (Minimum Bounding Box), MBS keeps the Width/Height information.

Next, the MBS is partitioned hierarchically according to the centroid of foreground. We select centroid as the splitting point owing to above observation. It is tolerant of the change of stroke in one or two.

Finally, our method extracts stroke elements within each partition to construct the feature vector. Stroke element is defined as a segment of typical strokes, namely, Heng (two adjacent pixels horizontally), Shu (two adjacent pixels vertically), Pie(three adjacent pixels diagonally), and Na (three adjacent pixels anti-diagonally). Extracting stroke elements rather than strokes themselves facilitates the uniform processing of characters and symbols. It is also tolerant of stroke variations.

Since Heng and Shu appear more frequently than Pie and Na [6], it is adequate to reduce the spatial dimension through the multiform partition. In our prototype, partition is applied only one level to Pie and Na while two-levels to Heng and Shu. So, feature vectors of an antique book spans a 40-dimensional vector space. The feature vector f can be formulated as:

$$\begin{aligned} &f(i) = \sum_{1\le k\le i} \frac{h(k)}{p_2(k)}, \qquad f(i+16) = \sum_{1\le k\le i} \frac{s(k)}{p_2(k)},\ i = 1, 2, \ldots, 16; \\ &f(j+32) = \sum_{1\le k\le j} \frac{p(k)}{p_1(k)},\ f(j+36) = \sum_{1\le k\le j} \frac{n(k)}{p_1(k)},\ j = 1, 2, 3, 4. \end{aligned} \tag{1}$$

where, $p_1(k)$ and $p_2(k)$ are respectively the number of foreground points of one and two levels partition. While $h(k), s(k), p(k), n(k)$ are foreground points of Heng, Shu, Pie and Na stroke elements, respectively, within partition area k. The feature vector describes the layout of stroke elements within an object.

3.2 Indexing Feature Space

All MFs make up a vector space, in which each MF contributes to the location of a point. The point contains the GLF[2]. We use Euclidean distance in our prototype. Generally, points of visually similar objects are adjacent and those of visually distinct are faraway in the feature space. Analogs to a given sample can be obtained by comparing the distance between a point and the sample.

To accelerate the comparison, a spatial index structure is designed to organize feature vectors. In principle, all the spatial index structures [2] contribute to efficient indexing. However, the *dimension cruces* demand efficient algorithms. In the prototype, we choose the PK-tree for its better performance in terms of storage and query speed [5]. For further information about constructing a PK-tree, please refer to [5].

3.3 Querying Similar Objects

When the user issues a query containing several objects, the proposed method decomposes it into separated objects and looks for analogs via index structure for each of them. The nearest-neighbor search of PK-Tree [5] returns a collection of sets of GLFs.

During the process, the searching range which influences the recall/precision is decided by a parameter ϵ. In the prototype it spans 11 levels. Level 0 ($\epsilon = 0$) means strictest matching and level 10(ϵ=1) means the most relaxed query. Between the two extremes, ϵ is adjusted by step 0.1. Because the user can get feedback immediately, a satisfactory balance between recall and precision will be reached.

3.4 Validating Constrains

The constraint is the relative order of objects that the user chooses as the query sample. It is applied to the result of previous phase to filter out improper combinations. What are left is a set of GLFs of the first objects. *Constraint validation* can be formulated as:

1. First, assume that the query consists of m objects, and m lists of GLFs are obtained by *the nearest-neighbor search* module, denoted as $L_1, L_2, \ldots, L_m$;
2. Assign L_1 to **L**
3. i iterates from 2 until m, do
   3.1 For each $e \in$**L**, suppose its GLF is j. If there is no element of L_i whose GLF is $j + i - 1$, delete e from **L**;
4. Return **L**

[2] PFs are maintained separately for the sake of memory efficiency.

4 Prototypic Implementation

To verify the feasibility of proposed method, we implement a prototype in the WWW environment. It consists of the server-side services and the client-side interface. The client-side interface can run on Java-enabled browser. It supports some operations, such as Zoom-In/out, Object Marking/Retrieving, Precision control. The server-side services are made up of a three-layered organization, i.e., a web server, static HTML files, and Java servelet. The static files provide fixed information such as registration, book-selecting information. Java servelet is a daemon that handles the user requests, invokes file & database retrieval module to obtain data and produces dynamic HTML files. The third layer includes the file & database retrieval daemon and the database of antique books.

5 Conclusion

An original method of content retrieval for Chinese antique book is proposed. By vital techniques we developed, it extracts content features of a Chinese antique book and constructs spatial index structure to accelerate the searching process. A user can make the recall/precision balance at retrieval moment, and the constraint validation ensures the precision of query results. The prototype demonstrates the feasibility of the method, and shows the superiority of visual similarity criterion to conventional methods. The method can be improved further by introducing dimension reduction mechanism, such as FastMap [1].

References

1. Faloutsos,C., Lin,K.: FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets, In Proc. SIGMOD'95 (1995) 163–174 385
2. Gaede,V., Günther,O.: Multidimensional Access Methods. ACM Computing Surveys, **30**(2) (1998) 170–231 384
3. Gladney,H., Mintzer,F., Schiattarella,F.: Safeguarding Digital Library Contents and Users: Digital Images of Treasured Antiquities. D-Lib Magazine. (1997)July/August *http://www.dlib.org/dlib.html* 380
4. Thibadeau,R., Benoit,E.: Antique Books. D-Lib Magazine (1997) September *http://www.dlib.org/dlib.html* 380
5. Wang,W., Yang,J., Muntz,R.: PK-tree: a spatial index structure for high dimensional point data. In: Proc. of 5th intl. Conf. on Foundations of Data Organization (FODO'98) (1998) Kobe Japan, November 12–13, 1998. *http://dml.cs.ucla.edu/projects/oasis/PK-Tree/pk-tree.html* 384
6. Zhang,X: Chinese Character Recognizing Techniques (in Chinese). Tsinghua University Press, Beijing (1992) 383
7. Zhu, Y.: Experiences of Electronic Version of SiKu Quan Shu (Complete Library of the Four Branches of Literature. The Journal of the Library Science in China (in Chinese) 6 (1999) 82–84 380