Web-Based Chinese Calligraphy Retrieval and Learning System

Yueting Zhuang, Xiafen Zhang, Weiming Lu, and Fei Wu

The Institute of Artificial Intelligence, Zhejiang University,
Hangzhou, 310027, P.R.China
{yzhuang, cadal, wufei}@cs.zju.edu.cn

Abstract. Chinese calligraphy is a valuable civilization legacy and there are some web sites trying to help people enjoy and learn calligraphy. However, besides metadata-base searching, it is very difficult to find advanced services such as content-based retrieval or vivid writing process simulating for Chinese calligraphy. In this paper, a novel Chinese calligraphy retrieval and learning system is proposed: First, the scanned calligraphy pages were segmented into individual calligraphy characters using minimum-bounding box. Second, individual character's feature information was extracted and kept. Then, corresponding database was built to serve as a map between the feature data and the original data of individual character image. Finally, a retrieval engine was constructed and dynamic writing process was simulated to help learners get the calligraphy character they are interested in and watch how it was written.

1 Introduction

When computer and Internet become more and more popular to the general public, less and less people have chances to write with a pen, and to enjoy the beauty of writing. Calligraphy is a kind of writing, and a popular communication tool in ancient China. It is not only delight to the eye and an inspiration to the spirit, but also a creative art. Yet you don't have to be an "artist" to learn calligraphy, you can learn the skills and write them every time you want. According to thousands years of learning experience, Chinese calligraphy learning process can be divided into three main consecutive steps: reading, understanding and simulating.

In terms of web-based learning and from the view of reading and understanding, key issues in such process are: how to manage all the data to display the beauty of the different calligraphy styles of the same character to learners, and how to help learners find the context of an interested character. From the simulating point, the key issue is how to set good writing example for learners to follow since it's impossible to trace the entire history to get a video to show how a particular calligraphy character was written. Correspondingly, our system consists of a large database managing all the scanned original data and its feature data, a retrieval engine helping learners find the same calligraphy character

R.W.H. Lau et al. (Eds.): ICWL 2005, LNCS 3583, pp. 186–196, 2005.

written in different styles by different people in different dynasties, and a simulator helping learners get a vivid idea about how a calligraphy character was written.

The remainder of this paper is organized as follows. Section 2 discusses the related works. Section 3 presents the system architecture. Section 4 gives the data structure. In section 5, main functions of web-based calligraphy learning system were described in detail. In section 6, the implementation and evaluation were done. And in the final part, conclusion and future works are given.

2 Related Works

Numerous researches have been done on exploring techniques for web-based learning such as [1] and [2]. But, these techniques don't fit web-based Chinese calligraphy learning well. Some web sites have been developed trying to fit learners' needs to enjoy and learn Chinese calligraphy, such as [3] and [4]. They do provide some basic information and many useful learning materials. However, they provide no advanced dynamic services such as content-based search to find calligraphy character that interestes the learner, and they do not tell the vivid writing process of an individual character to set a good example and let learners follow.

If it is a text query, "Google" is the biggest and fastest search engine. "Google" also provides image-searching function based on the name of the image. Yet, you can't submit a text query and retrieve character images similar to it. Lots of previous content-based image retrieval works used low-level features such as colors, textures and regions. However, such features cannot represent shape properties of a character, hence irrelevant images are frequently retrieved. Recently, there has been works to handle shape features effectively such as [5] and [6]). Still, they don't work well for calligraphy character image retrieval. Our previous work (see [7]) has proposed a new approach to retrieve calligraphy characters.

3 System Architecture

Fig.1 gives out an overview of our system architecture of web-based Chinese calligraphy learning. Its infrastructure mainly includes data collection, segmentation and feature extraction, which serve for advanced web-based learning purpose.

3.1 Data Collection

The original books, mostly ancient, were scanned at 600 dpi (dots per inch) and kept in DjVu format by researchers of our China-US million book digital library project (see [8]). These digitalized resources, together with their corresponding metadata are saved and packaged. The metadata standard (Edocument Metadata,Version 2.0) we used is released by the Zhejiang University Library. It combines two kinds of metadata: DC and MARC.

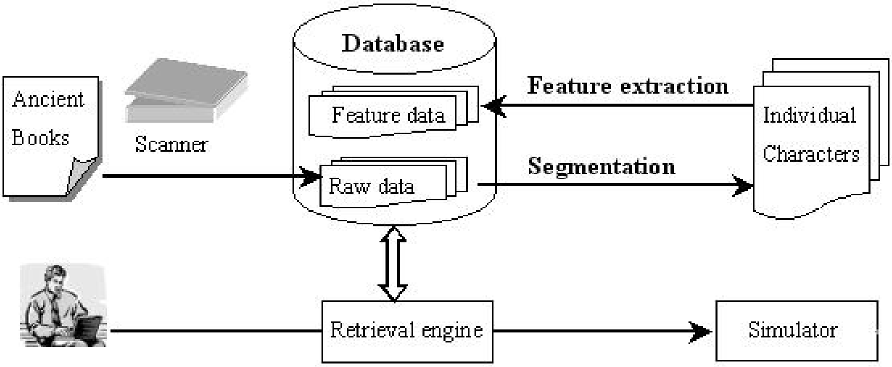


Fig. 1. Architecture of web-based Chinese calligraphy learning system

3.2 Segmentation

When digitized page images obtained, segmentation is needed in order to get feature information of individual calligraphy characters. Many researches have been done on segmentation of printed pages such as [9] and [10]. Yet no published paper has been done successfully on Chinese calligraphy page segmentation. It is mainly because calligraphy characters have more connection, and the background has more noise such as man-made seals. Our proposed segmentation approach first adjusts color luminance to get rid of red seals and smooth them to remove some noises. Then binarization was done, which is followed by projecting. After that, pages were cut into columns according to the projecting histogram, and columns continued to be cut into individual characters using minimum-bounding box as used in [10]. Fig.2 gives an example of our experiment, showing how a calligraphy page was cut into individual calligraphy characters.

Compared with [10], our segmentation approach made special constraint parameters to fit the characteristics of Chinese calligraphy. Let $x_{i,s}$ and $x_{i,e}$ denote the start and the end position of the *ith* cutting block. According to our long term segmentation experiences it subjects to the following constrains:

$$x_{i,e} - x_{i,s} \geq 5, i = 0, 1, 2 \cdot\cdot\cdot\cdot\cdot\cdot, n \tag{1}$$

$$2.5 \times \frac{1}{n} \times \sum_{i=1}^{n}(x_{i,e} - x_{i,s}) \geq 0.35 \times \frac{1}{n} \times \sum_{i=1}^{n}(x_{i,e} - x_{i,s}) \tag{2}$$

This is because according to thousands years of calligraphy writing experience, the width of individual character images in the same page tend to be similar. That is to say they have a minimum and maximum threshold for width, as described in formula 2. Let $wide_i$ and $height_i$ be the width and height of *ith* cutting block, then

$$0.6 \leq \frac{height_i}{wide_i} \leq 1.2 \tag{3}$$

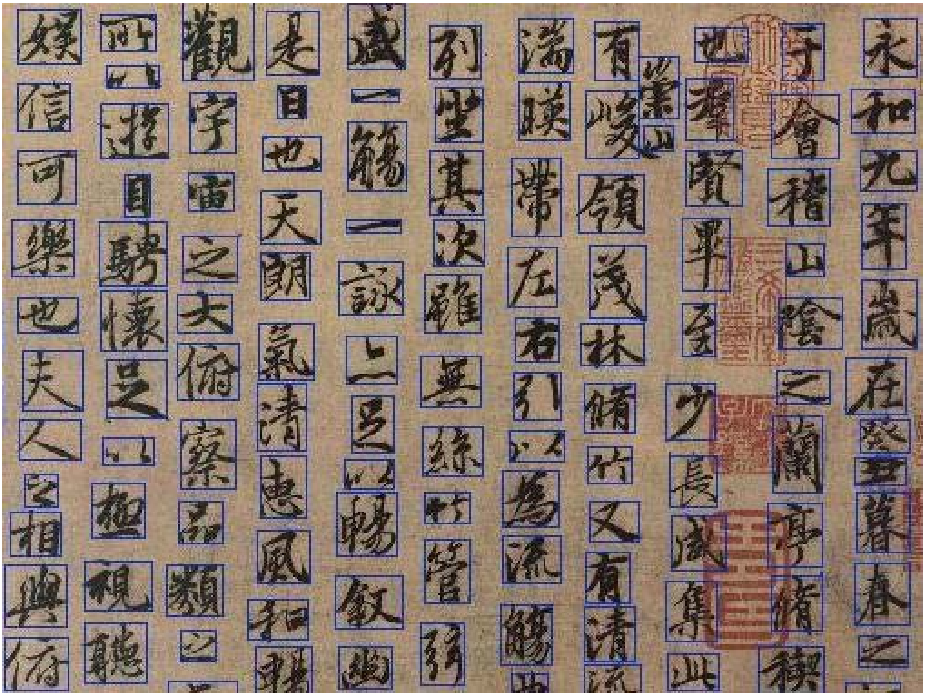


Fig. 2. An example of segmenting a page into individual characters using minimum-bounding box

Formula 3 tells the story that Chinese characters are always in square as introduced in [11].

With the above idea, most of the characters can be segmented automatically and correctly. But still, there are few man-made connections that can't be correctly segmented automatically such as in Fig.2, the fourth column from the right. In this case, we draw the minimum-bounding box to separate them. The minimum-bounding box which is in blue can be drag and drop manually.

3.3 Feature Extraction

After the segmentation was done, the next step is to extract features of individual calligraphy characters. In our approach, a calligraphy character is represented by its contour points instead of its skeleton as described in [6]. This is because skeleton representation is very sensitive to noise. As a result, it produces distorted strokes and proper shape of the character can't be detected.

According to the minimum-bounding box, we first normalize the individual character to 32×32 in pixels. Then, canny edge detector as introduced in [12] was employed to get its contour point's positions in Cartesian coordinates. Finally the values that denote the position of contour points were serialized to a string and kept in the database.

For learning purpose, a learner may want to know where an individual calligraphy character comes from and who wrote it. Therefore, the original location information of individual calligraphy characters, that is to say the location information of the minimum-bounding box should be kept too.

4 Data Structure

The scanned original image data is large and in disorder, needs further management. We build a special data structure to map the extracted feature data into the original raw data. The map consists of four tables: **book**, **works**, **character** and **author**, as shown in Fig.3. Many individual calligraphy characters compose a calligraphy works created by a calligraphist, namely an author. And many calligraphy works buildup a calligraphy book. The arrows show how these four tables are related by particular elements. In table of **character**, *co_points* is a string produced by the feature data of an individual calligraphy character.

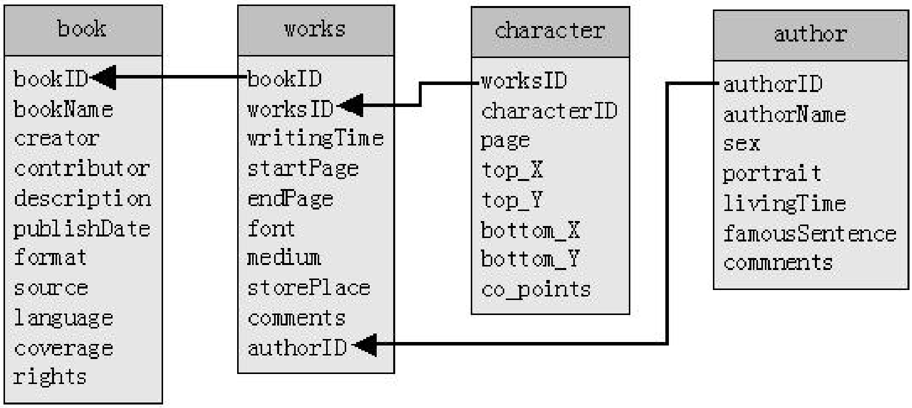


Fig. 3. Data structure for mapping feature data to raw data

With this map, it is easy to tell in which works and in which page an individual calligraphy character located by checking the table of **character** to find the *worksID*, the *page* and the *minimum-bounding box* (represented by *top_X*, *top_Y*, *bottom_X* and *bottom_Y*). It can also tells in which book this particular character can be found by searching the table of **works** using the key of *bookID*, and also who wrote this calligraphy character by checking the table of **author** using the key of *authorID*.

5 Key Calligraphy Learning Services

5.1 Learning Object Retrieval

For personalized learning purpose, different learners may be interested in different styles of the same calligraphy character. In our system, we use our new

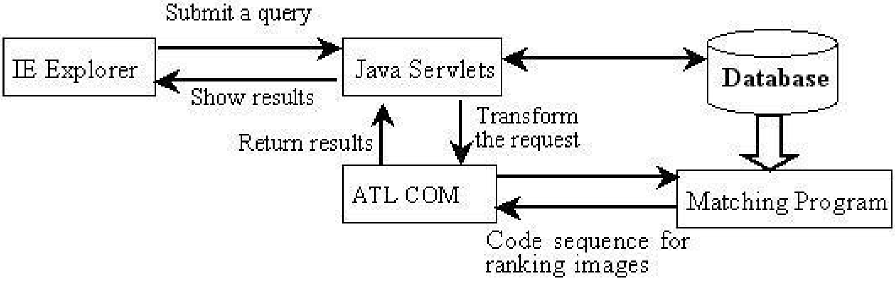


Fig. 4. Diagram flow of retrieval

content-based calligraphy character image retrieval approach (see detailed description in [7]). In that, we use inexact shape context matching. However, in terms of web-based learning its response time is beyond endurance: For a single query,the average retrieval time is about 3.6 minutes when the database consists of 336 isolated characters. In this paper, we develop a new architecture fit for web-base calligraphy character learning, shown in Fig.4.

Three ways are proposed to speed up the retrieval, one is preprocessing, another is classification, and the third is dimensionality reduction to reduce computing time. In preprocessing, shape features of every individual calligraphy character were extracted in advance, then serialized to a string and stored in the database, together with their corresponding metadata. Thus only the features of the query image need to be extracted dynamically.

For classification, complexity index was used besides the metadata. Let $f(x, y)$ be the gray value of a pixel. If a pixel belongs to the background, then let $f(x, y) = 0$, else $f(x, y) = 1$. Thus moment of the character image can be defined as:

$$m_{i,j} = \iint x^i y^j dxdy \tag{4}$$

And the root of second-order central moment in X and Y direction are defined as follows:

$$\sigma_x = \sqrt{(m_{20} - m_{10}^2/m_{00})/m_{00}} \tag{5}$$

$$\sigma_x = \sqrt{(m_{02} - m_{01}^2/m_{00})/m_{00}} \tag{6}$$

Then complexity index C can be computed as

$$C = (L_x + L_y)/\sqrt{\sigma_x^2 + \sigma_y^2} \tag{7}$$

Where L_x and L_y are the length of the longest stroke in X direction and Y direction respectively. Thus the larger C is the more complex the character is. If $|C_i - C_j| \leq 7$, then the character i and j are considered to be in the same complexity degree range, where the retrieval function works on.

Table 1. Comparison of average retrieval time of three approaches

Approach	Time
Earth Movers' Distance	16.6 minute
Projecting	5.31 minute
Shape Corresponding	1.52 minute

The number of sampled points dominates the computing time of each shape matching process. So dimensionality reduction is needed. The Number of Connected Points (NCP) is defined as the number of contour points existed in its 8-neighbourhood as introduced in [13]. If $NCP \geq 2$ and three consecutive points are in the same direction, then they are considered as parts of the same stroke. The middle point was taken out, and the reminder two points keep the structure information.

In order to measure the efficiency of these three speeding up approaches, we compare our proposed Shape Corresponding approach [7] (after implemented speeding approaches) with Projecting approach [11] and Earth Movers' Distance approach [14], as shown in Table 1. All of the tests are performed on a regular Intel(R)/256RAM personal computer. Compared with [7], the database in this paper is enlarged from 336 individual calligraphy characters to 1650 calligraphy characters, which segmented from works of about 60 calligraphists living in different dynasties. Table 1 indicates that an average single calligraphy character matching takes about 640ms (336 characters, 3.6minute) in [7], while here only 55 ms (1650 characters, 1.52 minute).

5.2 Dynamic Writing Process Simulating

After an individual calligraphy character is displayed before the learners, together with the information of where it comes from and who wrote it. The next service is to offer a visualization of how such a calligraphy character was written step by step, which may help immersion learning.

In order to simulate the dynamic writing process, stroke extraction and stroke sequence estimation should be made. We use contour segments to extract strokes of calligraphy character as introduced in [15]. Strokes in a handwritten calligraphy character are often connected together (see Fig.5). Also strokes of handwriting are not necessary corresponding to strokes of well printed character. Yet it's doesn't matter if several connected strokes are extracted as one stroke, so long as right writing sequence can be estimated. One assumption for estimating the order of the stroke sequence is based on the traditional writing rule:a calligraphy character was written from the left to the right, from the top to the bottom, and from the outside to the inside (see [11], page 14). The other assumption is that people always write a calligraphy character as fast and convenient as possible. So if strokes are connected, total distance travelled in the writing process should be minimized as introduced in [13]. Therefore when a cross corner encountered, we choose to follow the most straightforward contour segment. Because it has

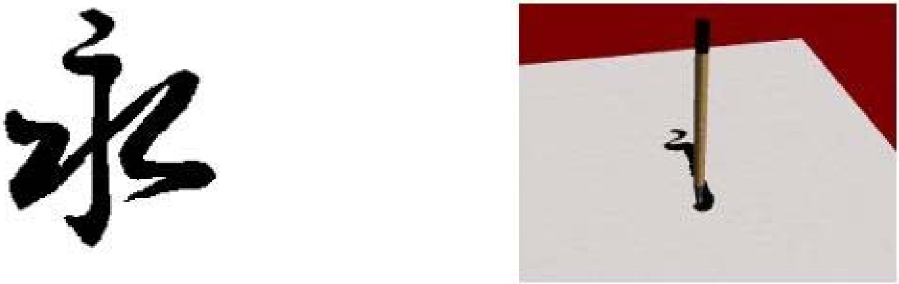


Fig. 5. A calligraphy character example and the corresponding video simulating its writing process

the biggest angle, and comply with the rule that "people write it as convenient as possible".

Based on above observation, we extract the strokes and estimated their sequence. Then develop a video to simulate how a calligraphist wrote a calligraphy character step by step, as show in Fig.5.

6 Implementation and Evaluation

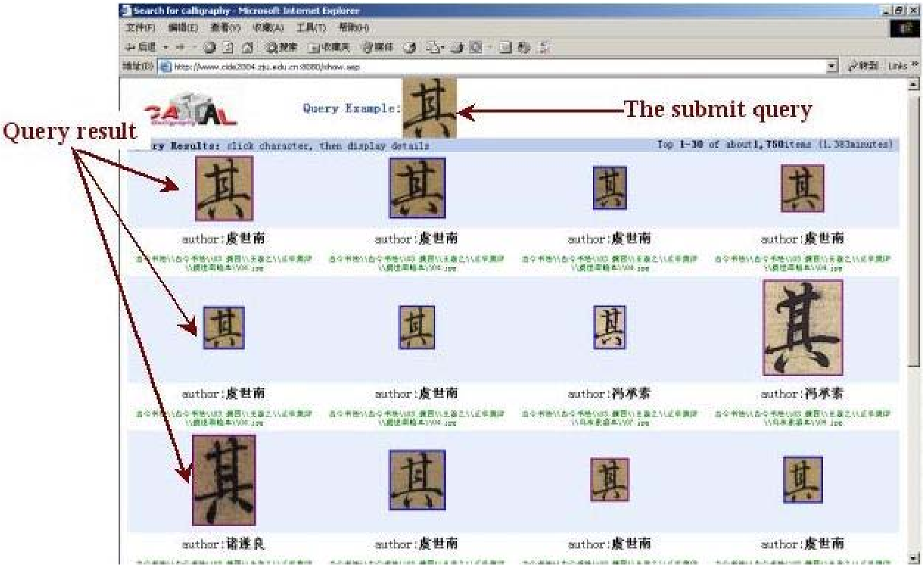
In the experiment, approaches described above are used and tested with the database consists of 1650 individual calligraphy characters. Most of these characters are segmented from 3 volumes of a book named "Chinese calligraphy Collections". Fig.6 shows a retrieval example.

If a learner is interested in one style of the character, for example the last one in the second row in Fig.6, then the learner can click this particular calligraphy character and a new web page (see Fig.7) will pop up showing its original scanned page with a minimum-bounding box marked out where it is, and also who wrote it. In Fig.7, when the name of the author is clicked, a portrait of the author accompanied by a brief resume will be shown. And if the individual character is clicked, a plug-in video (see Fig.5) will show up playing the estimated and visualized writing process.

Recall and precision are the basic measures used to quantitatively speculate the effectiveness of retrieval approach. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database, and precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Here, we use average recall and average precision. They are defined as:

$$recall_{average} = \frac{1}{C} \times \sum_{i=1}^{c} recall_i / n_i \tag{8}$$

$$presicion_{average} = \frac{1}{C} \times \sum_{i=1}^{c} precision_i / n_i \tag{9}$$


Search for calligraphy - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
http://www.cide2004.zju.edu.cn:8080/show.asp
Query result
The submit query

Query Example:

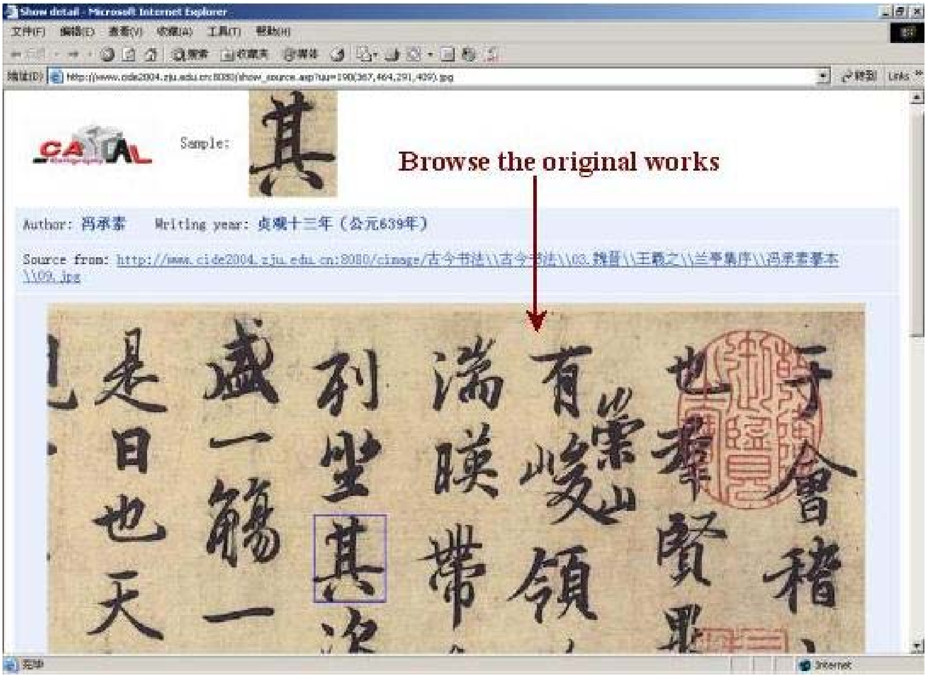
ry Results: click character, then display details   Top 1-30 of about1,750items (1.383minutes)

author:虞世南 author:虞世南 author:虞世南 author:虞世南

author:虞世南 author:虞世南 author:冯承素 author:冯承素

author:褚遂良 author:虞世南 author:虞世南 author:虞世南

Fig. 6. Screen shot of a retrieval example


Show detail - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
http://www.cide2004.zju.edu.cn:8080/show_source.asp?uu=190(367,464,291,409).jpg
Browse the original works

Sample:

Author: 冯承素   Writing year: 贞观十三年（公元639年）

Source from: http://www.cide2004.zju.edu.cn:8080/cimage/古今书法\\古今书法\\03.魏晋\\王羲之\\兰亭集序\\冯承素摹本\\09.jpg

Fig. 7. Screen shot of browsing the original works, with a minimum-bounding box mark out where the interested individual character is

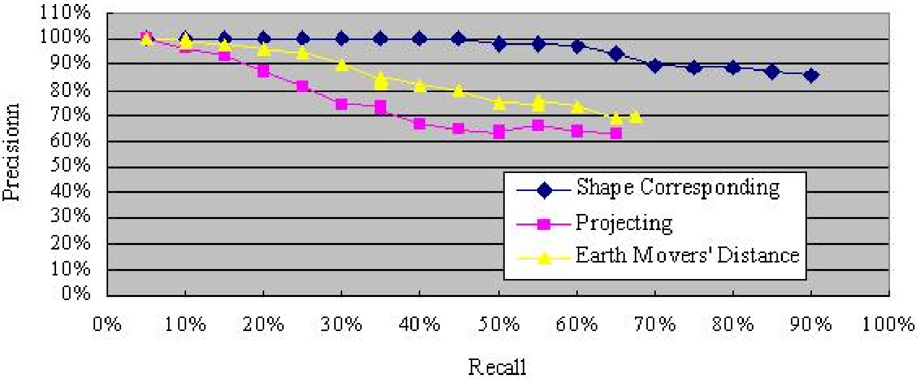


Fig. 8. Comparison of 3 approaches of their average recall and precision ratio on 20 characters with each has more than 6 different styles

Where c is the number of characters,n_i is the number of total styles of the same character i. We randomly chose 20 characters (each has more than 6 different styles) from the database, that is to say when $C = 20$ and $n_i \geq 6$, then Fig.8 can be drawn. It is obvious that the recall ratio is higher than traditional content-based image retrieval.

7 Conclusion and Future Work

We proposed a new system to help people who are interested in calligraphy to enjoy the beauty of different styles of the same Chinese character, to learn the detailed information of a particular character style (such as who wrote it, and in what environment), and also to learn how it can be written step by step. While the experiment is somewhat preliminary, it works efficiently and clearly demonstrates the applicability of our system to web-based Chinese calligraphy learning.

Our further development of this system will include continuing to speed up the retrieval service, developing more large database, and offering more convenient ways for query submitting, such as scratching a query or type in a text query by keyboard.

Acknowledgements

This research is supported by the China-US Million Book Digital Library Project (see http://www.cadal.net), and the National Natural Science Foundation of China (No. 60272031).

References

1. Yueting Zhuang, Xiang Liu, Multimedia Knowledge Exploitation for E-Learning: Some Enabling Techniques, *1st International Conference on Web-based Learning*, pp. 411-422, 2002.
2. Yueting Zhuang, Congmiao Wu, Fei Wu and Xiang Liu, Improving Web-Based Learning: Automatic Annotation of Multimedia Semantics and Cross-Media Indexing, *Third International Conference on Web-based Learning*, pp. 255-262, 2004.
3. `http://www.wenyi.com/art/shufa`
4. `http://www.shw.cn/93jxsd/jxsd.htm`
5. S. Belongie , J. Malik , J. Puzicha, Shape Matching and Object Recognition Using Shape Contexts, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, pp.509-522, No. 4, April 2002.
6. Jong-Seung Park, Visual Information Retrieval Based on Shape Similarity, *The 7th International Conference on Asian Digital Libraries*, LNCS 3334, pp. 458-461, 2004. 1336-1342.
7. Yueting Zhuang, Xiafen Zhang, Jiangqin Wu, Xiqun Lu: Retrieval of Chinese Calligraphic Character Image. 2004 *Pacific-Rim Conference on Multimedia*, LNCS 3331, pp. 17-24, 2004.
8. Jihai Zhao and Chen Huang, Technical Issues on the China-US Million Book Digital Library Project, *7th Int'l Conf. On Asian Digital Libraries*, LNCS 3334, pp. 220-226, 2004.
9. Thomas M. BreuelRepresentations and Metrics for Off-Line Handwriting Segmentation, *8th International Workshop on Frontiers in Handwriting Recognition*, pp.428 - 433,2002.
10. R. Manmatha, Chengfeng Han,E. M. Riseman,W. B. Croft, Indexing handwriting using word matching, *Proceedings of the 1st ACM international conference on Digital libraries*, pp.151 - 159,1996.
11. Wu You-Shou and Ding Xiao-Qing, Chinese character recognition: the principles and the implementations. Beijing: Advanced Education Press, 1992.
12. `http://homepages.inf.ed.ac.uk/rbf/HIPR2/canny.htm`
13. Lau K.K., Yuen P.C., Tang Y.Y, Stroke extraction and stroke sequence estimation on signatures, *16th Int'l Conf. On Pattern Recognition*, vol. 3, pp: 119 - 122,2002.
14. S. Cohen and L. Guibas. The Earth Mover's Distance under Transformation Sets. In *Proceedings of 7th IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 173-187.
15. Chungnan Lee, Bohom Wu, Wen-Chen Huang, Integration of multiple levels of contour information for Chinese-character stroke extraction, *4th Int'l Conf on Document Analysis and Recognition*, vol.2, pp.584 - 587,1997.