2012 International Conference on Medical Physics and Biomedical Engineering

On Tangut Historical Documents Recognition*

Changqing Liu

School of Mathematics and Computer Science; Xixia Research Institute, Ningxia University, Yinchuan, China

Abstract

As the Tangut studies have made progress, a considerable number of Tangut historical documents’ copies have been published. It is of great importance to carry out digitalization and domestication of these copies. The paper firstly makes an initial processing of images by global threshold, then dissect the photocopies by scanning. Finally adopts the recognition approach of principal component analysis. The experiment shows that a better recognition can be achieved by calculation without extra time.

1. Introduction

Over past 10 years, Tangut Studies has attracted the academic attention abroad and at home. Great progress has been made with a number of publications of Tangut historical documents. However, current Tangut studies rely on the traditional approach of documentation thus time-consuming and inconvenient. Thanks to the growth of computer technology, Tangut studies are believed to be benefited from the ever-advancing digital technology. Recognition of historical characters is the key issue in the digitalization and text of Tangut historical documents.

At present, no comprehensive processing system for Tangut characters and Tangut historical documents has been established online. Scholars are looking forward to a translation platform of Tangut languages, Chinese, and English. The online platform will be of great help for Tangut research. On the one hand, some Tangut scholars have set out the work, but inexperienced with computer knowledge, the project is slow and ill-efficient. On the other hand, computer scholars are capable of programming needed software, but have difficulty to join with Tangut studies. Experts in two fields come to realize that it is necessary to work across disciplines.
2. Preprocessing of Tangut Historical Documents

The first task for digitalization of Tangut historical documents is the layout analysis. Character dissecting is the basic work. Because the character dissection is based on the results of binaryization on text area, so it is crucial to realize binaryization[1]. The common approach is to separate the character background from the characters by locating threshold. Currently, optimal segmentant threshold can solved by global threshold of image, it can be a single threshold or double threshold [2].

Step1: gray the image. Convert the color picture into black and white one. Firstly turn it into gray. There are many graying methods available; one simple approach is listed as follows:

\[ \text{Gray} = 0.299 \times R + 0.587 \times G + 0.114 \times B \]  

Step2: set a threshold, image binaryzation. Processing the 256 brightness level by appropriate threshold, those lower than the value is set as 0, higher representing pixel of font as 255.

There are lots of noise dots after turning into black and white, which makes it difficult for further recognition. First of all, the images are expanded and erosion. Then put it in a suitable model. Setting the value as 0 if the black color pixel is lower than a certain value [3].

The testing documents are the complete photocopies, characters are to be separated to make recognition. Through observation, most characters in historical documents follow regular rules in writing. Therefore, images are separated by column and rows. Specifically, the process is illustrated as follows:

Step1: separating the characters into different areas and scanning from left to right. Started from the first separating line, scan from left to right. When the black color pixel is less than a certain value, the point will be the end of character column. The documents are readily preserved and named and catalogued with specific name. Likewise, all the character columns are created. Figure1 is the result of separation.

Step2: separating the single image. After reading the preservation information, scan the image from top to bottom. The first black point is the start of scanning, and when the black color pixel is less than a certain value, the current position will be the end. Preserve the single character and name it according to its preset number.

Figure 1. Tangut character

A Tangut historical document contains a large sum of characters, so it is time-consuming to make manual location of Tangut characters. For those documents with clear image, auto-separation by computer is a better option.

Procedures of auto location
1. Graying Tangut historical documents
2. Turning into black and white
3. Deciding its character border by pixels around the black pixel.
According to Figure 2, the auto-located Tangut characters are shown in blue boxes. Most of the characters are identified and with specific location. Further separation can be based on the location.

3. PCA recognition of Tangut Characters

Principal Component Analysis, PCA, also named as K-L, is orthogonal transformation based on target second-order statistics[4]. A new component is generated through transformation, but it is orth or uncorrelated. The transformation decrease data redundancy, data are processed in a low-dimensional feature space, and save most of the raw data information [5].

There are a large number of feature variables based on multi-sample analysis. Great differences existed in those variables. Important features will be reduced without proper judgment. However, it is impossible to extract every feature. PCA collects information of a group of feature variables by using reduction dimensions and apply to a number of irrelevant principle components [6]. If a feature variable is similar to sample, it is impossible to distinguish samples. PCA is to pick up the representative variables to achieve recognition.

PCA is generally applied to human face recognition [7]. A lot of research has proved that it can reach a great accuracy in pattern recognition [8]. In the process of experiment, 6,000 Tangut characters models have been created based on Tangut fonts and extract by PCA. It has been proved effective in Tangut characters recognition.

1. Reading image matrix

1,000 Tangut characters samples are chosen with a feature of 60*62. A matrix (1,000*3,720) can be established. The character images are shown as Figure 3:

![Figure 3. Tangut character sample](image_url)

After reading 1,000 Tangut characters images, the computer gets the result by program:

![Figure 4. Average character sample image](image_url)

1,000 Tangut characters are put together and get a comprehensive sample data with all the information of these 1,000 characters.

2. Calculation of decomposition coefficient

```
cov mat = x'*x;                   % covariance matrix
[V,D] = eig(cov mat);            % eigenvalues and eigenvector
V = x*V*(abs(D))-0.5;           % normalization
KLCoef = x'*V;                   % decomposition coefficient
```

It is necessary to reduce dimension because of the huge sample data due to the large amount of sample data.

3 Calculation of similar characters

Step1: Calculation of decomposition coefficient of base vector:
```
filename1 = sprintf('c:\\images\\no1.bmp');
image_data1 = imread(filename1);
```
x(:,1) = image_data(:);
KLCoef1=x'*V;

Step2: The distance between the testing image and the sample matrix. The shorter the distance, the closer the feature of images. The program operated is illustrated as follows:
for i=1:1:40
dist_comp(i)=sqrt(dot(KLCoef1(1,:)-KLCoef(i,:),KLCoef1(1,:)-KLCoef(i,:)));
dist_comp(i)
end;

4. Conclusion

Digitalization of Tangut historical documents is the trend of Tangut studies, which is based on the recognition of Tangut characters. PCA is an effective approach to carry out character recognition. The research has found that it applies better to printed historical documents with clear image, but hard to manual or broken documents, for clarity and completion are two essential elements in recognition. The future research will be focus on manual Tangut characters recognition and the recovery techniques for incomplete documents.

*Supported by national Natural Science Foundation of China (No: 60803104)

References


<table>
<thead>
<tr>
<th>TABLE I. RECOGNITION RESULT BASED ON PCA</th>
<th>Testing character</th>
<th>The closest 6 characters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.11</td>
<td>40.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II. RECOGNITION RESULT BASED ON PCA</th>
<th>Testing</th>
<th>The closest 6 characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>character</td>
<td>37.41</td>
<td>38.49</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>🗄️</td>
<td>🗄️</td>
<td>🗄️</td>
</tr>
</tbody>
</table>

**TABLE III. RECOGNITION RESULT BASED ON PCA**

<table>
<thead>
<tr>
<th>Testing character</th>
<th>40.30</th>
<th>41.57</th>
<th>41.68</th>
<th>41.94</th>
<th>41.96</th>
<th>42.36</th>
</tr>
</thead>
<tbody>
<tr>
<td>🗄️</td>
<td>🗄️</td>
<td>🗄️</td>
<td>🗄️</td>
<td>🗄️</td>
<td>🗄️</td>
<td>🗄️</td>
</tr>
</tbody>
</table>