A Method for Removing Inflectional Suffixes in Word Spotting of Mongolian Kanjur

Hongxi Wei, Guanglai Gao
School of Computer Science
Inner Mongolia University
Hohhot, China
{cswhx, csggl}@imu.edu.cn

Yulai Bao
Library
Inner Mongolia University
Hohhot, China
lbbyl@imu.edu.cn

***Abstract*—According to characteristics of Mongolian word-formation, a method for removing inflectional suffixes from word images of the Mongolian Kanjur is proposed in this paper. By removing inflectional suffixes, the amount of clusters equivalent to indexing terms might be reduced in word spotting. For the above purpose, we need to determine whether or not one word image contains inflectional suffix. If the word image contains inflectional suffix, the inflectional suffix would be segmented from the word image. The proposed method is as follows: first, many parts are segmented from the bottom of the word image according to the cutting positions of the inflectional suffixes. Then, the segmented parts are represented by a number of profile features and classified by multi-BP neural networks. Finally, the outputs of BP are confirmed by template matching using DTW. Experimental results on our data set prove the feasibility of the proposed method.**

Keywords-Mongolian Kanjur; word soptting; inflectional suffix; BP neural network; template matching

I. INTRODUCTION

Historical documents are precious cultural heritages of the human beings. At present, many countries are digitizing their native historical documents in order to protect them as long as possible and enable public access to them more convenient and fast such as via Internet. In Inner Mongolia University, a project for protecting Mongolian Kanjur is in process. The Mongolian Kanjur is the most famous Mongolian book around the world. It is a Mongolian encyclopedia including history, medicine, astronomy, literature and so on. The Mongolian Kanjur, which is preserved in Library of Inner Mongolia University, was made by woodblock printing in 1720 (Qing Dynasty). The printing process is as follows: Mongolian words were engraved in woodblock and then printed on paper by cinnabar. It contains 108 volumes in total and about 45,000 pages with twenty million words more or less.

Although public can browse such digital Mongolian Kanjur need not to travel to the library, it is difficult to retrieve them without indexing. Traditionally, there are two ways to create indexing. The first one is manual annotation, which is a very expensive and tedious task for a large collection of document images. The second one is an automatic approach. It utilizes OCR (Optical Character Recognition) technology to convert image into text. However, the words in the Mongolian Kanjur are equivalent to a kind of off-line handwritten Mongolian and are degraded due to the passage of time. Moreover, it is difficult to segment the words into the corresponding characters. And there is no available OCR software for this kind of off-line handwritten Mongolian. Therefore, OCR technology can not be easily applied to the Mongolian Kanjur.

When OCR is poor or hard, word spotting technology is an effective alternative especially for historical handwritten documents. Word spotting was originally proposed for speech processing and was firstly introduced by Manmatha et al. [1] for indexing George Washington's manuscripts. The idea of word spotting is as follows. It treats a collection of document images as a collection of word images and uses image matching for calculating pairwise distances between word images. According to the distances, word images can be clustered and each cluster can be considered as an indexing term. Ideally, each cluster only contains all instances of the same word. In [2], Rath and Manmatha adopted lots of profile features including projection profile, word upper profile, word lower profile and background/ink transitions for representing each word image. Several image matching algorithms were compared with each other using the above features by Rath and Manmatha [3, 4]. They concluded that DTW (Dynamic Time Warping) was the best one. Moreover, they have detailedly studied each step of the word spotting technology in [3].

As well as historical handwritten English documents, word spotting technology has been used to historical handwritten or printed documents in other languages.

Gatos et al. [5] proposed a segmentation-free approach to keyword spotting in historical typewritten Greek documents. In their work, synthetic keyword images would be created according to user typed queries. Then, the synthetic keyword images were matched to word images of collection using features based on zones and projections. User feedback technology was also added to the retrieval procedure to improve performance.

Ataer et al. [6] used SIFT operator for detecting and representing salient points (such as connection points, dots or high curvature points) in historical printed and handwritten Ottoman documents. Each Ottoman word image was represented by a set of visual terms obtained by vector quantization of the feature vectors. The pairwise similarities of words were calculated by the symmetric KL-divergence

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.27

on the visual terms' distributions of words. Their method can also capture similarities of the semantically similar words. But there was only qualitative analysis of one word in their work.

Terasawa et al. [7] realized word spotting on historical Japanese and Chinese manuscripts by an Eigen-space method. First, document images were segmented into text lines after preprocessing. Then, these text lines were transformed into a sequence of slits along the writing direction. Each slit with N pixels was regarded as a N-dimensional vector and each slit was mapped into a D-dimensional vector (D is much smaller than N) by PCA (Principal Component Analysis). Thus, document image can be represented by the eigenvectors of a certain amount of slits. DTW was also selected as image matching algorithm to achieve higher performance.

Bilane et al. [8] focused on word spotting for ancient Syriac manuscripts written in Serto calligraphy. First, document images were segmented into text lines after preprocessing as well as [7]. Then, they used a fixed size sliding window to pass on each text line at a step of one pixel and analyzed the content of the window at each step in order to retain the window or not. Each retained sliding window was divided into several sub-windows in equal size and directional roses of 8 directions were extracted in each sub-window. Then, each retained sliding window was represented by a feature vector using directional roses with equal length. And Euclidean distance between two feature vectors was calculated to represent similarity.

In aforementioned references, the handling objects are word images in [3], [5] and [6], because the words can be achieved relatively easy from the corresponding document images. But in [7] and [8], it is quite hard to extract words from document images opposite to [3, 5, 6]. So, the researching objects in [7] and [8] are text line images. In the Mongolian Kanjur images, words can be extracted relatively easily. Therefore, the objects in our study are the Mongolian word images.

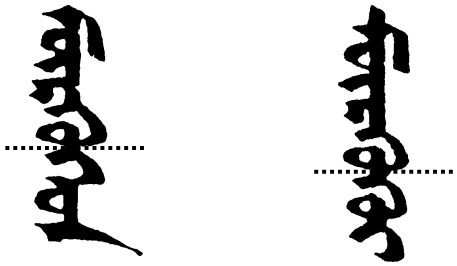
In this paper, we mainly concentrated on the method for removing the inflectional suffixes from word images of the Mongolian Kanjur. By removing the inflectional suffixes, the number of clusters in word spotting might be reduced so that the recall level would be improved.

The rest of the paper is organized as follows: our motivation is given in Section II. The proposed method is explained in Section III, along with the details of each step. Experimental results of the proposed method are shown in Section IV. Section V provides the conclusions and future work.

II. MOTIVATION

Mongolian is an agglutinative language. Its word formation and inflection is built through connecting different suffixes to the roots or stems. These suffixes are classified two categories ordinarily. One is word-formation suffix that can produce variations of part-of-speech or meaning. The other one is word-inflection suffix that often causes variations of person or tense. Generally, inflectional suffixes appear at the end of the words. Thus, there are lots of words with the same part-of-speech and meaning, but they include different inflectional suffixes at the end of the words. For example, there are two different inflectional suffixes below the dotted lines in Fig. 1. Word A is pronounced "murguged" in Latin and its meaning is "hit" in the past tense. Word B is pronounced "murguhui" and its meaning is "to hit". Word A and B have the same parts after removing the inflectional suffix respectively. Therefore, in order to reduce the amount of clusters in word spotting, the inflectional suffixes should be removed from word images before clustering.

In this paper, only the word-inflection suffixes are considered and the word-formation suffixes will not be processed. To the best of our knowledge there is no literature about removing inflectional suffixes from word images.



Word A Word B

Figure 1. Tow word images of the Mongolian Kanjur with different inflectional suffixes.

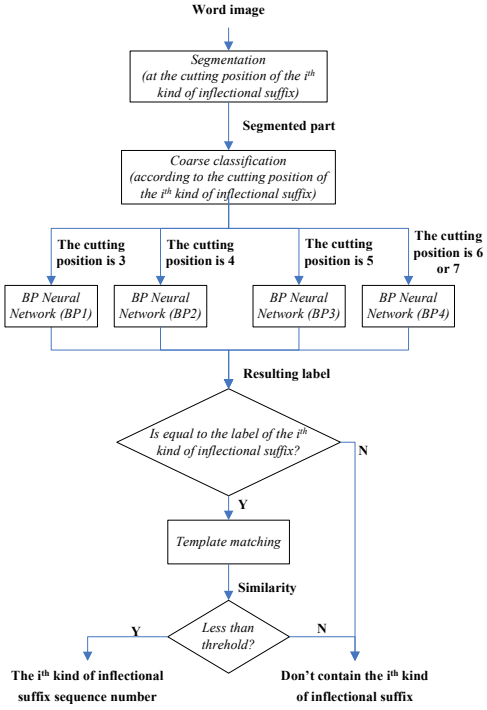


Figure 2. A flowchart for determining whether a word image contains the i^{th} kind of inflectional suffix.

III. PROPOSED METHOD

In order to accomplish our motivation, we need to determine whether or not one word image contains inflectional suffix. If one word image does not contain any inflectional suffix, the word image would not be changed. Or else, the part of the inflectional suffix in word image should be removed and the rest part is reserved. Thus, the problem of removing inflectional suffixes from word images is converted to the problem of determining whether or not one word image contains inflectional suffix. Our solution to this problem is as follows: for each time, a certain part from the bottom of a word image is segmented and classified by a BP neural network; then, the result of the BP is confirmed by template matching so as to determine whether the part is this kind of inflectional suffix (such an example is shown in Fig. 2); each kind of inflectional suffix should be processed in the above same way. Occasionally, a word image may be considered as containing several inflectional suffixes by the above way. Under the circumstance, the final result is the one with the minimum similarity. The proposed method is detailed in the following subsections.

A. Determining Cutting Positions in Word Images for Each Kind of Inflectional Suffix

In our study, we find that the left sides of word images vary more abundantly than the right sides. That is, the left profile curves (Left Profile Curve abbr. LPC) of word images appear much more rises and falls than the right profile curves (Right Profile Curve abbr. RPC). LPCs have more power than RPCs on discriminating different word images. One example is presented in Fig. 3. The two different words have the same RPCs in (e) and (f) of Fig. 3, but LPCs are different in (c) and (d) of Fig. 3.

It is the fact that each kind of inflectional suffix contains a fixed amount of rises and falls on their LPCs. Therefore, for each kind of inflectional suffix, the number of rises and falls on LPC can be used as cutting positions in word images. In Fig. 4, the word  contains the inflectional suffix  and the cutting position of the inflectional suffix is the bottom third valley points (see Table I). Thus, we can extract the red dotted line in Fig. 4 (c), which is from the cutting position to the end of the word.

In order to locate the rises and falls on LPC more accurately, some preprocessing tasks should be done. First, LPC of word image is smoothed by a one-dimensional Gaussian filter (standard deviation is 5). And then, all peak points and valley points on the LPC are extracted. Neighboring peak points and valley points with small difference value (below 0.01) need to be removed. Specially, if the difference value between the last peak point and valley point is below 0.1, the last peak point and valley point will be removed too (see Fig. 4 (c) and (d)).

Here, 15 frequently-used kinds of inflectional suffixes are selected. They appeared from 17 times to more than 200 times in our data set. These inflectional suffixes and their cutting positions are displayed in Table I. Each cutting position of Table I represents the bottom j^{th} (e.g. 3 represents the bottom third) valley point on the LPC of one word image.

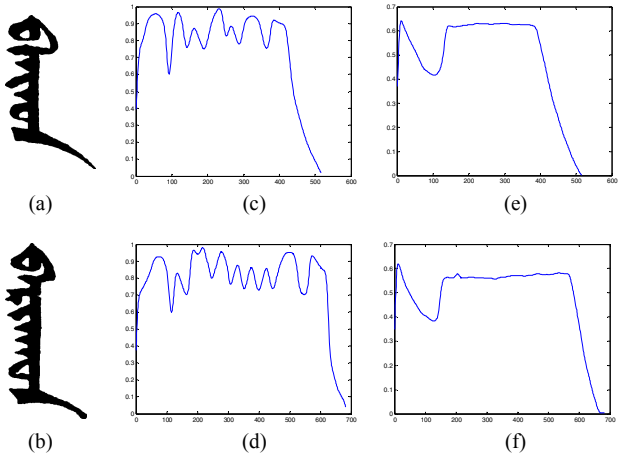


Figure 3. Two different words in **(a)** and **(b)**; **(c)** The left profile curve of (a); **(d)** The left profile curve of (b); **(e)** The right profile curve of (a); **(f)** The right profile curve of (b).

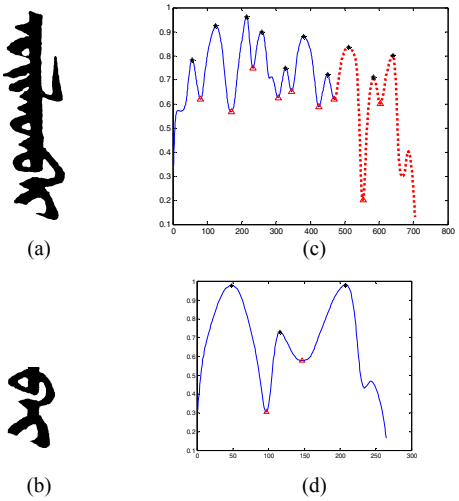


Figure 4. **(a)** Mongolian word 'uiledbei' means 'manufacture' in English; **(b)** Inflectional suffix 'bei' represents perfect tense; **(c)** The left profile curve of (a); **(d)** The left profile curve of (b).

TABLE I. 15 INFLECTIONAL SUFFIXES AND CUTTING POISTIONS

Inflectional suffix								
Cutting Position	*3*	*5*	*5*	*4*	*6*	*5*	*3*	*3*
Inflectional suffix								
Cutting Position	*4*	*3*	*3*	*6*	*4*	*6*	*7*	

TABLE II. THE NUMBER OF INPUT NEURONS IN EACH BP

BP identifier	*BP1*	*BP2*	*BP3*	*BP4*
Normalized scale (Width*Height)	*200*250*	*200*350*	*300*350*	*300*450*
Feature dimension	*1350*	*1650*	*1950*	*2250*
Input neurons	*1350*	*1650*	*1950*	*2250*

B. Classification for the Inflectional Suffixes

All the inflectional suffixes in Table I are coarsely divided into four categories by theirs cutting positions in word images. The cutting position of the first category is 3; the second category is 4; the third category is 5 and the last category is 6 or 7. In the same way, segmented parts from word images using different cutting positions would be also divided into corresponding categories.

BP neural network is used to finely classify in each category. We utilize four BP neural networks for the above four categories inflectional suffixes respectively. The BP neutral networks used in this study are fully connected and have four layers: an input layer, two hidden layers and an output layer. The number of the first hidden layer's neurons is 200; the number of the second hidden layer's neurons is 25 and the number of output layer's neurons is 1.

For each BP neutral network, the number of input layer's neurons equals to the dimension of input feature vector. Each input for BP should be normalized in pre-defined size. The normalized scales are presented in the second row of Table II. Here, three features including left profile, right profile and horizontal projection are extracted from per image row. Another three features including upper profile, lower profile and vertical projection are extracted from per image column. Thus, there are six features for representing each input. The fourth row of Table II lists the number of input neurons in the four BP neural networks separately.

C. Template Matching for Confirming Results

In this step, we propose a method with discriminative information to select the corresponding template set for each kind of inflectional suffix. The proposed method is described as follows. Given a collection containing M (M is 15) kinds of inflectional suffixes images, the subcollection in the collection of each kind of inflectional suffix images is denoted as S_i (i=1, 2, ..., M). The template set is denoted as T_i (i=1, 2, ..., M) and its size is K (K is 5 in this paper). For each kind of inflectional suffix, do the same following steps:

(1) Do for j=1, 2, …, $|S_i|$
   - Calculate the DTW distances between j^{th} inflectional suffix and the other ($|S_i|$-1) inflectional suffixes.
   - Compute average DTW distance of the j^{th} inflectional suffix.
(2) Choose the one (denoted as ***centroid***) that has the smallest average DTW distance to others and put it into T_i.
(3) Sort the DTW distances of the ***centroid*** with others in descending order.
(4) Choose the first (K-1) suffixes from the sorting result and put them into the T_i.

Here, the DTW distance between inflectional suffixes images is calculated as well as [4]. But, four profile features were extracted from per image column only for calculating in [4]. In our study, the same four profile features are extracted not only from per column but also from per row.

If the output of BP is the i^{th} (i=1, 2, ..., 15) inflectional suffix, each template of the template set T_i should be matched with the segmented part from the word image using the above DTW algorithm and the smallest DTW distance is selected. If the smallest DTW distance is smaller than a pre-defined threshold value, the segmented part is the i^{th} inflectional suffix. The threshold values for each kind of inflectional suffix are defined:

$$\boldsymbol{threshold = \alpha \cdot min\_average\_distance} \quad (1)$$

where ***min_average_distance*** is the average DTW distance of the ***centroid*** and $\boldsymbol{\alpha}$ is a coefficient ($\boldsymbol{\alpha = 1.2}$ in this paper).

IV. EXPRIMENTAL RESULTS

A. Data set

We selected 50 pages (one page contains 200 words more or less) from the digital Mongolian Kanjur and converted them into binarization images using our previous method [9]. Then, these binarization images were segmented into word images by layout analysis based on connected components. Finally, 5500 word images with good quality were selected to form our experimental data set. And each word image was annotated using the corresponding glyph codes.

By analyzing the annotations, the number of the vocabulary in our data set is 1235 and the number of the words containing 15 kinds of inflectional suffixes is 1371. If the 15 kinds of inflectional suffixes are removed, the number of the vocabulary will reduce to 1122. That is, the amount of indexing terms can be reduced about 9%.

B. Experiment I

In this experiment, we examined the accuracy for achieving the inflectional suffixes from word images according to the cutting positions. Firstly, we selected all word images which contain any kind of inflectional suffix by analyzing theirs annotations. And then, for each selected word image, we segmented it at the corresponding cutting position and achieved the inflectional suffix image from the cutting position to the end of the word. Each achieved inflectional suffix image need to be checked up. The detail results are given in Table III. Its accuracy is 97% in average. That is, if the word image contains a certain kind of inflectional suffix, we can achieve the correct inflectional suffix from the cutting position to the end of the word image with 97% accuracy.

C. Experiment II

In this experiment, 4000 word images of our data set were forming the training set. There are 1157 word images contain inflectional suffix with 26 segmentation errors. So, 1131 inflectional suffix images were extracted from the cutting positions to the end of the word images and used to train the four BP neural networks and select template sets. They were normalized in pre-defined size before training the four BP neural networks. But, they were not normalized in template selection.

The remaining 1500 word images were used for testing the performance of our proposed method. In the 1500 word

images, 214 words contain inflectional suffixes with 8 segmentation errors.

The precision and recall are used to evaluate the performance of our proposed method. Let ***Ground Truth Data*** (***GTD***) is the number of each kind of inflectional suffix in testing set; ***Returned*** is the number of achieved by our proposed method; ***Correction*** is the number of correctly achieved by our proposed method. The precision (***Pr***) and recall (***Re***) are defined as follows:

$$\mathbf{Pr} = \frac{\boldsymbol{Correction}}{\mathbf{Re}\,\boldsymbol{turned}} \tag{2}$$

$$\mathbf{Re} = \frac{\boldsymbol{Correction}}{\boldsymbol{GTD}} \tag{3}$$

The experimental results are shown in Table IV.

TABLE III. ACCURACY FOR ACHIEVING INFLECTIONAL SUFFIXES USING CUTTING POSITIONS

Inflectional suffix	Ground Truth Data	Achieved Correction	Accuracy (%)
1	115	114	99.13
2	114	114	100.00
3	124	122	98.39
4	146	145	99.32
5	58	57	98.28
6	82	80	97.56
7	217	206	94.93
8	63	63	100.00
9	266	254	95.49
10	41	40	97.56
11	21	19	90.48
12	38	37	97.37
13	17	17	100.00
14	33	33	100.00
15	36	36	100.00
Total	**1371**	**1337**	**97.52**

TABLE IV. EXPERIMENTAL RESULTS OF THE PROPOSED METHOD

Inflectional suffix	Returned	Correction	GTD	Pr (%)	Re (%)
1	6	5	6	83.33	83.33
2	6	4	4	66.67	100
3	38	32	52	84.21	61.54
4	59	56	92	94.92	60.87
5	7	2	2	28.57	100
6	1	0	0	——	——
7	13	12	16	92.31	75.00
8	7	5	5	71.43	100
9	14	10	10	71.43	100
10	3	1	1	33.33	100
11	10	8	10	80.00	80.00
12	2	0	0	——	——
13	3	3	3	100	100
14	1	1	1	100	100
15	3	3	4	100	75.00
Total	**173**	**142**	**206**	**82.08**	**68.93**

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a method for removing inflectional suffixes from word images of the Mongolian Kanjur. On our experimental data set, the precision is about 82% with 69% recall level and the F-measure is about 75%, which proves the feasibility of our method. The proposed method provides an approach to solving the same problem in other agglutinative languages.

We will test the performance for reducing clusters on a larger data set. This is our next work. Meanwhile, we can count the frequent errors in removing inflectional suffixes and take them as a sort of garbling information. The garbling information can be used for query expansion. That is, if query images contain the garbling information, the query images would be segmented in the wrong way. Thus, the word with the same segmentation error in indexing would be returned by this way and the recall level could be improved. So, gathering garbling information is also our future work.

ACKNOWLEDGMENT

This paper is supported by the Natural Science Foundation of China (NSFC) and the project numbers are 60865003 and 70863008.

REFERENCES

[1] R. Manmatha, C. Han, E. M. Riseman and W. B. Croft, "Indexing handwriting using word matching," In Proceedings of 1st ACM International Conference on Digital Libraries, Bethesda, Mar. 1996, pp. 151–159.

[2] T. M. Rath and R. Manmatha, "Features for word spotting in historical manuscripts", In Proceedings of 7th International Conference on Document Analysis and Recognition, Edinburgh, Aug. 2003, vol. 1, pp. 218–222.

[3] T. M. Rath and R. Manmatha, "Word spotting for historical documents", Int. J. of Document Analysis and Recognition, vol. 9, 2007, pp. 139–152.

[4] T. M. Rath and R. Manmatha, "Word image matching using dynamic time warping", In Proceedings of 28th International Conference on Computer Vision and Pattern Recognition, Madison, Jun. 2003, vol. 2, pp. 521–527.

[5] B. Gatos, T. Konidaris, K. Ntzios, I. Pratikakis and S. J. Perantonis, "A segmentation-free approach for keyword search in historical typewritten documents", In Proceedings of 8th International Conference on Document Analysis and Recognition, Seoul, Aug. 2005, vol. 1, pp. 54–58.

[6] E. Ataer and P. Duygulu, "Matching Ottoman words: an image retrieval approach to historical document indexing", In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, Jul. 2007, pp. 341–347.

[7] K. Terasawa, T. Nagasaki and T. Kawashima, "Eigenspace method for text retrieval in historical document images", In Proceedings of 8th International Conference on Document Analysis and Recognition, Seoul, Aug. 2005, vol. 1, pp. 437–441.

[8] P. Bilane, S. Bres, K. Challita and H. Emptoz, "Indexation of Syriac manuscripts using directional features", In Proceedings of 16th International Conference on Image Processing, Cairo, Nov. 2009, pp. 1841–1844.

[9] Hongxi Wei, Guanglai Gao, Yulai Bao and Yali Wang, "An efficient binarization method for ancient Mongolian document images", In Proceedings of 3rd International Conference on Advanced Computer Theory and Engineering, Chengdu, Aug. 2010, vol. 2, pp. 43–46.