# Eigenspace Method for Text Retrieval in Historical Document Images

Kengo Terasawa,  Takeshi Nagasaki,  and  Toshio Kawashima
School of Systems Information Science,  Future University-Hakodate,
116–2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041–8655, Japan
{g3103004, nagasaki, kawashima}@fun.ac.jp

## Abstract

*A new method for text retrieval that does not need segmentation is described. Segmenting the images in historical documents into individual characters is difficult. Therefore, the conventional OCR method, which uses segmentation, does not work well. Our method instead divides the text image into a sequence of small slits. The image region that corresponds to the query image region is retrieved by solving the matching problem of these sequences. Applying the eigenspace method to the slit images enables us to solve the matching problem efficiently. Moreover, using dynamic time warping (DTW) further improves the results. Our method has higher accuracy than the simple template matching method, and it has far higher efficiency in computational cost.*

## 1. Introduction

Many libraries or regional government offices plan to create publicly available digital archives of historical documents or cultural assets. These facilities have an enormous amount of historical documents, including highly valuable ones and ones that are rarely referenced. Most of them will be stored in an image format. Therefore, efficient indexing or retrieval techniques for historical document images are indispensable.

One idea is to make text format transcription by using the optical character recognition (OCR) method. However, traditional OCR techniques cannot be easily applied to historical documents for many reasons. One is that historical documents are mostly handwritten and sometimes are significantly degraded due to the passage of time. In addition, for cursive style writing, character segmentation becomes very difficult because many of the characters are written continuously without clear spacing. Unlike Western languages, segmentation into words is still difficult in some Eastern languages such as Japanese. Furthermore, Japanese historical documents are mainly written by brushes, and they make

the result of thinning process unstable. All of these limit the application of traditional OCR to historical document images. Of course, the dependency on language is also a problem in particular if the usage of words or characters of that era differs from today's usage.

In this paper, we describe text retrieval without character segmentation or recognition. When given one text image, our method retrieves other images of similar appearance from the whole document image. Despite an inability to automatically recognize, our method will help librarians make an index of documents and will help readers to find the information they are looking for. Also, it may help researchers to read unrecognizable words by retrieving the same word in another context.

### 1.1. Related Work

The idea of text retrieval without recognition is seen in the work of Manmatha et al. [1], which they called "word spotting". Rath and Manmatha proposed a set of features suitable for word image matching [2], and they applied a dynamic time warping method to match the images [3]. Their matching target was a set of segmented words because their works were intended for English manuscripts, where word segmentation is possible. A study of word spotting for Eastern languages, where word segmentation is difficult, was done by Yue Lu and Chew Lim Tan [4]. They developed a method for searching for words in Chinese newspapers, but it was only applied to machine printed fonts.

Our method is intended for handwritten manuscripts of Japanese or some other Eastern languages. Our image matching method was inspired by the eigenspace method. The eigenspace method is famous for its application called "eigenface" [5, 6], which is widely used in the area of face recognition. In this method, face images are compared in reduced dimensional space. We extended the eigenspace method to compare the sequences of images.
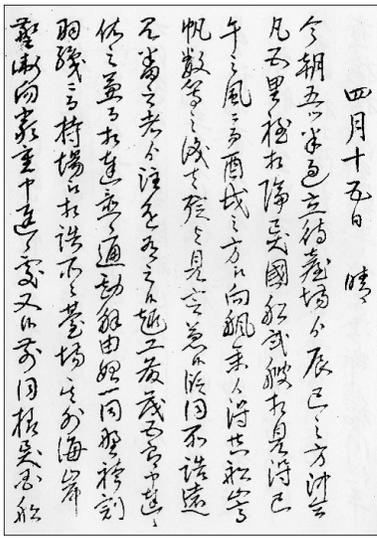
**Figure 1. One Japanese Historical Document Image: "Akoku Raishiki" written in the mid-19th century.**



**Figure 2. Images divided into slits with duplications**

## 1.2. Outline of our method

Our method does not need to segment images into single characters or words. Instead, it segments images into a sequence of small slits. A low dimensional descriptor is generated for each slit image by use of the eigenspace method. Text retrieval is executed by matching the sequences of these descriptors.

## 2. Algorithm description

### 2.1. Preprocessing

The materials for our study were digital images of historical documents, which were obtained by scanning. One example is presented in Fig. 1.

Some preprocessing steps needed to be performed at the beginning. The first step was to remove background noise from the image. For this step, we used a quite simple thresholding method: if a pixel value was higher than a certain threshold value, it was regarded as background, and the value was set to white(255). We did not convert it to black(0) even if the pixel value was lower than the threshold because our brush written materials sometimes show gray level variation in their strokes. The materials may have somewhat useful information.

The next step was to segment the document image into text lines. The segmentation into lines was far easier than segmentation into words or characters. We also used a quite
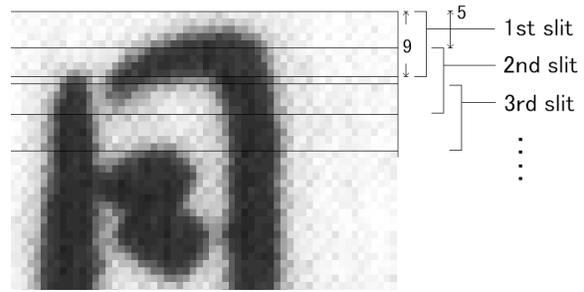
simple method for this step. A projection of stroke pixels along the line axis was calculated, and the lower peak was regarded as a separating border.

In the third step, we realigned the horizontal position of the text image to remove the horizontal perturbation of text lines, as is often the case with handwritten documents. The amount of horizontal drift was estimated from the center of the mass in a relatively long window that slides along the text line.

Finally, we applied the smoothing of a Gaussian filter so that it was robust to noise. The deviation parameter was set to $\sigma = 2$.

### 2.2. Transformation into Slit Sequences

Preprocessed images were then transformed into a sequence of slits. "Slits" mean narrow rectangular windows that scan images along the line axis. The width of the slits (length of the window along the line axis) should be sufficiently narrow relative to the size of single character. In this study, since the characters were each about 60 pixels large, we set the slit width as 9 pixels. To avoid dependency on the origins of slit cutting, the sliding step of the slit window was set to 5 so that the slit sequences had 4 pixels overlapping (see Fig. 2).

### 2.3. Eigenspace Projection

Generating low dimensional descriptors by means of PCA is widely used especially in face recognition (called "eigenfaces"). The images are reshaped to $n$-dimensional column vectors $x_i$, $i = 1, \ldots, m$, where $n$ is the number of pixels in the image, and $m$ is the number of images. A mean image vector $c = (1/m) \sum x_i$ is subtracted from each image vector, and the following matrix

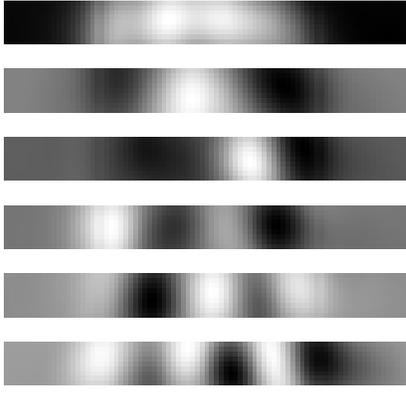$$A = \begin{pmatrix} x_1 - c & x_2 - c & \cdots & x_m - c \end{pmatrix} \qquad (1)$$

**Figure 3. Examples of eigenslits. From top to bottom, the first to sixth eigenslits.**

is created. The covariance matrix is

$$C = AA^T. \tag{2}$$

After choosing the first $d$ eigenvectors of $C$ sorted by the eigenvalue, we obtain a set of principal vectors: $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_d$. These vectors formed the basis of eigenspace.

Each one of the mean-subtracted images was then projected to these basis vectors, and the generated $m$ $d$-dimensional vector became a good descriptor of the original image. These low dimensional vectors allowed us to solve matching problems more easily.

In fact, this formulation as an $n \times n$ eigenvalue problem involves high computational cost. Fortunately, we can solve the problem with an easier $m \times m$ eigenvalue problem, where $m$ is far smaller than $n$ in general. That is, if $V$ is the eigenvector matrix of $AA^T$, and $U$ is the eigenvector matrix of $A^T A$, they satisfy the equation $AU = VD$, where $D$ is the $n \times m$ matrix with eigenvalues on the diagonal. Based on this, we can derive $V$ from $U$.

However, in our case, the number of images, meaning the number of slits, increased as the target document increased in length. Fortunately, for dealing with document image, the produced principal vectors varied only slightly based on the number of slits. We observed that 50 to 100 is a sufficient number of slits for computing eigenvectors to represent the document image, and even if more slits are computed in, the effect is limited. Fifty to 100 slits are equivalent to four to eight characters. In the following experiment, we set the number of slits at 200, including the safety factor.

Figure 3 is an example of the resulting eigenvectors.

## 2.4. Matching by Slit Feature Descriptors

After the image document was transformed into vector sequences, the remaining problem was matching the vector sequences.

Let the descriptors of slit sequences be $\{\boldsymbol{y}(t)\}$, where $t$ is the slit ID. If the query image is represented as $\{\boldsymbol{y}(t)|t_0 \leq t \leq t_0 + \tau\}$, the matching cost between the query image and the part of database image beginning from $t_0'$ is defined as

$$D(t_0, t_0') = \sum_{0 \leq t \leq \tau} |\boldsymbol{y}(t_0 + t) - \boldsymbol{y}(t_0' + t)|. \tag{3}$$

In this equation, $|\boldsymbol{y}(t_0 + t) - \boldsymbol{y}(t_0' + t)|$ represents the distance between two descriptor vectors. Although several ways can be used to define this distance, we employed the simplest $L1$-norm, called the Manhattan distance, i.e.,

$$|\boldsymbol{y}(t_0+t) - \boldsymbol{y}(t_0'+t)| = \sum_i |y_i(t_0+t) - y_i(t_0'+t)|, \tag{4}$$

where $y_i$ represents the $i$-th element of vector $\boldsymbol{y}$. This distance is so easy to compute that it involves low computational cost. We observed that the results changed only slightly when other distance measure such as an L2-norm or an Euclid distance were used.

The distance $D(t_0, t_0')$ was calculated with various $t_0'$, and the $t_0'$ that gives the minimum $D(t_0, t_0')$ was output. The retrieved image was the image represented as $\{\boldsymbol{y}(t)|t_0' \leq t \leq t_0' + \tau\}$.

## 2.5. Dynamic Time Warping

To make the matching algorithm more robust, we applied dynamic time warping (DTW). DTW is a widely used method in the area of speech recognition. If two time series are given, DTW considers every conceivable time correspondence including non-linear time coordinate transformation, and it outputs the path with minimum matching cost. For our document image retrieval, regarding a slit ID as a time coordinate, sequences of slits can be regarded as a time series.

The time normalized distance between two vector sequences $A = \{\boldsymbol{a}(i)\}$ and $B = \{\boldsymbol{b}(j)\}$ is defined as follows:

$$D(A, B) = \min \left[ \frac{\sum_k |\boldsymbol{a}(i_k) - \boldsymbol{b}(j_k)|}{k} \right], \tag{5}$$

where $i_1$ and $j_1$ are the first subscript of $A$ and $B$, $i_k$ and $j_k$ are the last subscript of $A$ and $B$, $(i_1, j_1), \ldots, (i_k, j_k)$ represents the matching path, $k$ is the length of the path, and the minimum is searched in all possible paths. In most situations, the matching path is restricted to a certain range. In our case, the restriction was defined as:

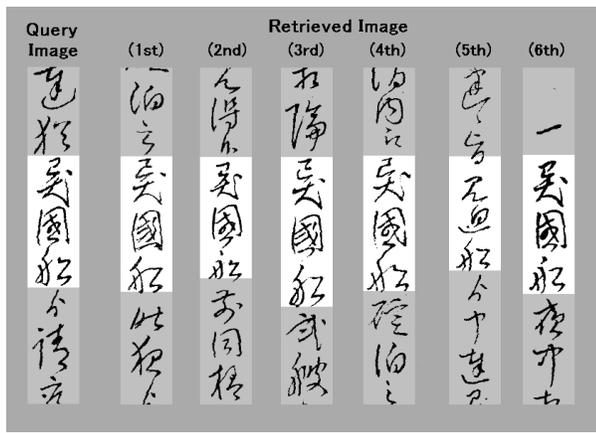$$(1/\alpha) \cdot i_k \leq j_k \leq \alpha \cdot i_k, \tag{6}$$

**Figure 4. Result for "Akoku Raishiki". The leftmost column is the query image, the following are retrieved images ranked first to sixth.**



(Manuscript A)          (Manuscript B)

**Figure 5. Tested images: "Ranteijo" written by Ogishi. The same words are written in manuscripts A and B.**

where $\alpha$ is the stretching allowance ratio. In the following experiment, $\alpha$ was set to 1.2, which shows the best performance.

# 3. Experimental Results

## 3.1. Results for "Akoku Raishiki"

Our method was first tested on images of "Akoku Raishiki (The diary of Matsumae Kageyu)" (Fig. 1). It is a historiography written by a Japanese government worker in the mid 19th century. The tested images consisted of 22 pages, 179 lines, and 2771 characters. The resolution per single character was about $60 \times 60$ pixels. Our method was tested with a keyword "Ikokusen(foreign ship)", which appears several times in this manuscript. One of the images corresponding to this keyword was used as query image, and if the images of the same word were retrieved, the results seemed to be correct.

The results are shown in Fig. 4. The leftmost column is the query image, and the following are retrieved images sorted by the matching cost. The white-back region in the figure is actually the retrieved regions, and the gray-back region is displayed only to show the context. Five of the six retrieved images were successfully retrieved. The fifth image was not successfully retrieved; it seems to have been caused by the fact that the final character of the word corresponds to the query image.
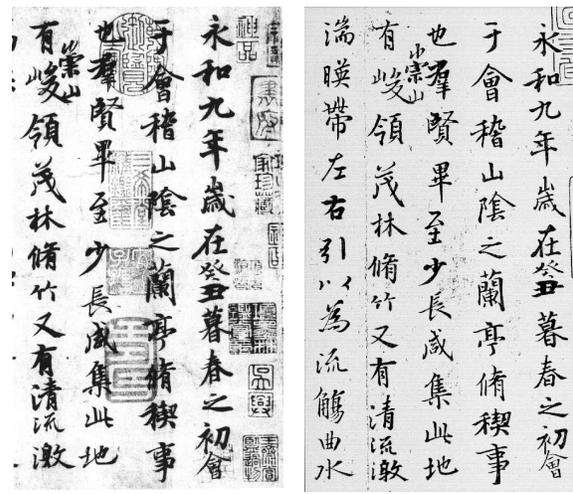
## 3.2. Quantitative Evaluation for "Ranteijo"

We conducted another experiment to evaluate our method quantitatively. The materials used were two manuscripts of "Ranteijo (Lan Ting Xu)"(Fig. 5). "Ranteijo" was written by a famous Chinese calligrapher, "Ogshi (Wang, Xizhi)". There exists several manuscripts written by the same author, and the materials are two of them.

Manuscripts A and B contain exactly the same words, so a quantitative evaluation was possible. That is, when a part of manuscript A was used as a query image, if the same part of manuscript B was retrieved, it was assumed to be a correct retrieval.

Both manuscripts A and B consist of 28 lines and 321 characters. The resolution per single character was about $60 \times 60$ pixels. Note that this image contains many stamps, which were not removed by preprocessing, therefore these images were assumed to be quite degraded. The basis vector for projection was composed from the first 200 slits of manuscript A, and all images including manuscript B were projected on the same basis. The query images were arbitrary slit sequence of 26 length (it is equivalent to about two characters).

For the sake of comparison, we used a simple template matching method for the retrieval, i.e., the distance of the query image and the target image were defined as the accumulated pixel value differences.

We also tested Rath and Manmatha's method [3] as a benchmark. Both a with-DTW case and a without-DTW
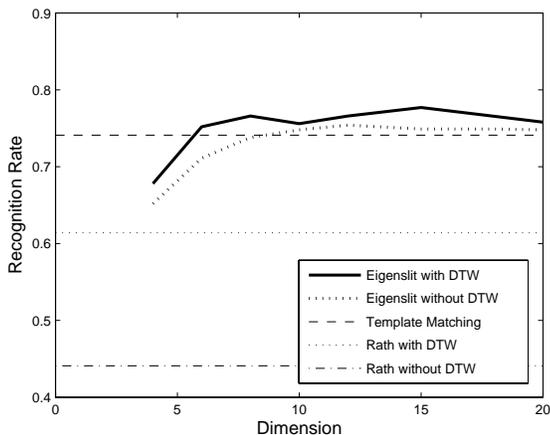
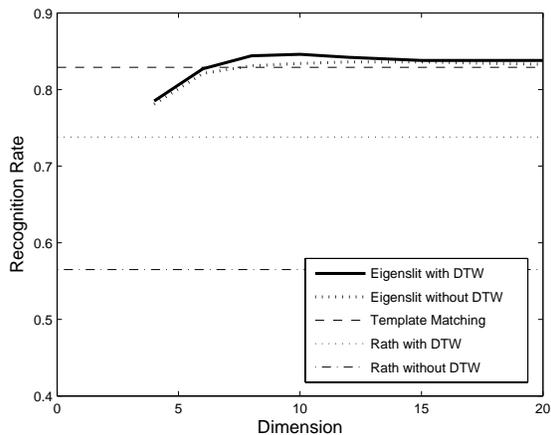**Figure 6. Percentage of Correct Answer Retrieved in the First Rank**



**Figure 7. Percentage of Correct Answer Retrieved within Third Rank**

case were tested. Let us point out that their method was intended for English manuscripts and needs additional preprocessing. Therefore, the results do not exactly describe their availability. It was just used for the sake of comparison.

All the results are shown in Fig. 6 and 7. Figure 6 represents the percentage of correct answers retrieved in the first rank, and Fig. 7 represents the percentage of correct answers retrieved within third rank. In both figures, both the case with DTW and the case without it are represented. The horizontal axis represents the dimensions of the descriptors.

Our method exceeded the simple template matching method regarding accuracy when the dimensionality was over five. Furthermore, our method had far less computational cost than the simple template matching method. The with-DTW method required more computational cost than the without-DTW method, but the accuracy was improved.

## 4. Conclusions and Future Work

We presented a new method for text retrieval in historical document images. The method describes document images as a sequence of small slits. After this translation, the eigenspace method is used to enable image matching at low computational cost. By translating images into such low dimensional representation, improvement based on dynamic time warping is possible. Our experimental results show that the method achieves higher accuracy than simple template matching, although the computational cost is far lower.

Future work for our approach includes improving the preprocessing for further enhanced performance. Our

method for dividing images into slits should be investigated further. A more sophisticated design for distance definition of sequences or single slits is also needed. Our method is conceptually independent of languages. Therefore, we will also apply it to other languages.

## References

[1] R. Manmatha, Chengfeng Han and E.M. Riseman, "Word Spotting: A New Approach to Indexing Handwriting," Proc.of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 631–637, 1996.

[2] T.M. Rath and R. Manmatha, "Features for Word Spotting in Historical Manuscripts," Proc.of International Conference on Document Analysis and Recognition, pp. 218–222, 2003.

[3] T.M. Rath and R. Manmatha, "Word image matching using dynamic time warping," Proc.of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 521–527, 2003.

[4] Yue Lu and Chew Lim Tan, "Word spotting in Chinese document images without layout analysis," Proc.of IEEE International Conference on Pattern Recognition, pp. 30057–30060, 2002.

[5] M.A. Turk and A.P. Pentland, "Eigenfaces for recognition," Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71–86, 1991.

[6] M.A. Turk and A.P. Pentland, "Face recognition using eigenfaces," Proc.of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 586–591, 1991.