

## *An Approach to Script Identification in Multi-Language Text Image*

Mingji PIAO

Intelligent Information Processing Lab.  
Dept. of Computer Science & Technology  
Yanji, China  
piaomingji123@hotmail.com

Rongyi CUI<sup>†</sup>

Intelligent Information Processing Lab.  
Dept. of Computer Science & Technology  
Yanji, China  
cui rongyi@ybu.edu.cn

**Abstract**—A character level script identification method to identify Korean, Chinese and English scripts using PCA is proposed in this paper. First, the space of eigenvectors was constructed by using PCA, and the segmented character was reconstructed by projecting the character into the space. Second, relative entropy between original and reconstructed image is computed for vertical and horizontal histogram. Finally, the written language was identified according to Euclidean distance and relative entropy between original and reconstructed image. The experiment results show that proposed method achieved 99.78% high accuracy for correct segmentation which effectively solved the script identification problem for multi-language text image contains Korean, Chinese and English.

**Keywords**—script identification; principal component analysis; relative entropy; Euclidean distance; character segmentation

### I. INTRODUCTION

In human society linguistic nature is something about defining population, that is to say each language defines a population, and character, which is viewed as visual manifestation, is one of significant bases for population identity recognition. Script identification plays an important role in international polylingual information service and indexing etc. area, and is meaningful to extend existent OCR system and develop multilingual OCR system as well[1].

Texture characteristic varies from different types of character and thus which can be used as characteristic for script identification[2,3], and the method for extracting texture characteristic may be divided into two broad categories—structured-based and visual-contour-based methods[4]. In previous work, several methods of automatic script identification have been developed so far, such as, steerable pyramid[2,5], multi-wavelet transform[7], Gabor filters[11,13] etc. feature extracting methods combine with SVM[7,8,9,11], decision tree[10], K-NN[11,12], neural network[7,14] etc. classifiers. In general, there are two problems in given methods:

- Most of the methods are combination of feature and classifier, but training classifier is time-consuming, and slight deviation in coefficients may cause a big influence on results.

- The methods identify script at the level of page-wise, paragraph, text-block, textline-wise and therefore limit the flexibility to adapt to different forms of script.

To overcome above problems, we proposed a PCA (principal component analysis)-based method at character level for multi-script which includes Korean, Chinese, English characters. Based on the structure characteristics of the three kinds of character mentioned above, firstly, the space of eigenvectors was constructed by using PCA for each set of characters, and furthermore, the reconstructed character was obtained by segmenting and followed by mapping which onto the space, and finally, script identification was accomplished according to the Euclidean distance and relative entropy between original and reconstructed character.

Section 2 presents pre-processing for script identification. Section 3 describes the proposed method in details. Section 4 is devoted to experimental results and error analysis, and Section 5 concludes this paper.

### II. PRE-PROCESSING OF SCRIPT IDENTIFICATION

#### A. Pre-processing and flowchart of script identification

There are skew and noise etc. phenomena during scan and thus it is necessary to employ skew correcting and noise removing before script identification. Egozi et al. [14] achieved high performance in correcting skew using combination of statistic multi-model and EM algorithm. BI Xiao-jun et al. [15] proposed higher-order cumulant-based denoising method which is ideal for removal of additive Gaussian noise in text images.

Fig.1 shows the procedure of script identification and the steps are as follows:

- The space of eigenvectors was constructed for each kind of character set, and pre-processing and character segmentation were done for the text image.
- Reconstructed character is obtained by mapping a character onto the English space, and the horizontal and vertical histogram of reconstructed image and original image were computed.
- Computing the Euclidean distance and relative entropy between reconstructed image and original image. If the Euclidean distance and relative entropy satisfy the restricted condition then the character is

<sup>†</sup> Corresponding author: e-mail: cuirongyi@ybu.edu.cn

identified as English or re-mapping the original image onto the Korean and Chinese space in sequence.

- The rest can be done in the same manner, if the image does not satisfy the restricted condition then it will be rejected to identify.

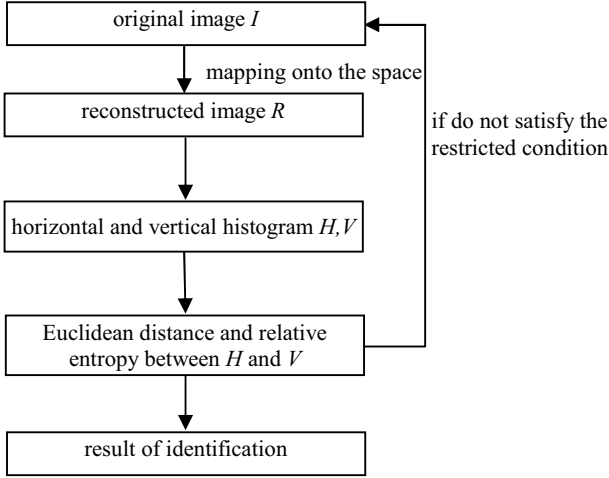


Figure 1. Procedure of script identification

### B. Character segmentation

The proposed method is at character level, and therefore character segmentation step is necessary that is greatly related to identification accuracy. Most prevailing segmentation method is valley-position-based method, but different character structures make it impossible to segment correctly only using valley information. To improve the segmentation accuracy, according to the structure characteristic of three kinds of character (i.e. Korean, Chinese and English), we use width and centroid of character and valley position of vertical histogram as segmentation criteria.

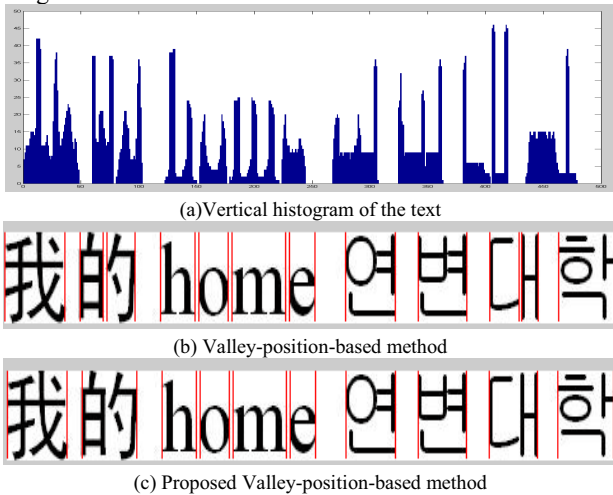


Figure 2. Two different segmentation methods

Fig.2 (a) shows vertical histogram of the text image, and the result segmented by valley position-based method as shown in (b), in which Chinese character ‘的’ and Korean character ‘ ’ were separated into two parts, but the proposed method segmented correctly as shown in (c).

## III. PCA-BASED SCRIPT IDENTIFICATION

### A. Constructing the space of eigenvectors

Each set of characters has their own structure characteristics and thus in one set of characters they have some common features, based on this idea we can use less data to describe the set of characters. Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [17]. By using only a few largest principal components we can attain the lower-dimensional data that best explain the structure characteristics of the specific characters. Let  $N$  by  $M$  image  $I(x, y)$  becomes a vector  $\mathbf{P}_i \in R^k$  ( $k=N \times M$ ), and then the average image is given by

$$\mathbf{avg} = \frac{1}{n} \sum_{i=1}^n \mathbf{P}_i \quad (1)$$

and covariance matrix of the image is

$$\mathbf{C} = \frac{1}{n-1} \sum_{i=1}^n \Phi_i \Phi_i^T \quad (2)$$

where  $\Phi_i = \mathbf{P}_i - \mathbf{avg} \in R^k$  ( $k = N \times M$ ), and  $n$  indicates the total number of characters. The space of eigenvectors can be constructed by using eigenvectors which corresponding to the largest associated eigenvalues.

Korean characters can be divided into 12 kinds of structures according to the statistic characteristic [18, 19], based on this idea we can obtain 5, 2 and 1 spaces of eigenvectors for Korean, English and Chinese separately. The eigenvector of capital letters, small letters, Korean, Chinese are as shown in Fig.3, in which the eigenvectors are presented in two-dimensional image. It may be noted from Fig.3 that English letters are mainly concentrated in the center, less dense and the distribution area are small, and Korean letters are more regularly arranged compare to Chinese, and the components of Chinese characters are relatively dense and complex.



Figure 3. Eigenvectors of three kinds of characters

### B. Character reconstruction and script identification

Reconstructed image is attained by projecting the original image onto the space of eigenvectors using (3).

$$w = (\mathbf{P} - \mathbf{avg})^T \mathbf{V} \quad (3-a)$$

$$\hat{\mathbf{P}} = w\mathbf{V} + \mathbf{avg} \quad (3-b)$$

where  $\mathbf{V} \in R^k$  ( $k=N \times M$ ) is the eigenvectors of covariance matrix  $C$  mentioned in (2).

There is structure correlation in each set of characters, so when the character is projected onto the corresponding feature space the reconstructed image resembles the original image in appearance, and while projected onto the other feature space the reconstructed image appears quite different from the original image. Fig. 4 illustrates two cases mentioned above.

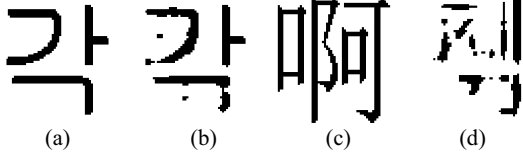


Figure 4. Original image and reconstructed image

As seen in Fig.4, Korean character (a) does not change radically when projected into the corresponding space, and the reconstructed image is as shown in (b), while the Chinese character (d) which is reconstructed in Korean space appears so different from original image (c) that it is impossible to identify whether it is Chinese character.

### C. Script identification algorithm

Euclidean distance  $|\mathbf{p} - \hat{\mathbf{p}}|$  between original image and reconstructed image can be adopted to describe the difference between the two images, and, in addition, the relative entropy of horizontal and vertical histograms between two images is also important to measure the difference. After the histogram normalization, relative entropy can be computed by

$$E_{rel}(\mathbf{P}, \hat{\mathbf{P}}) = \sum_{i=1}^n Q_i \log_2 \frac{Q_i}{\hat{Q}_i} \quad (4)$$

where  $Q_i$  and  $\hat{Q}_i$  denote the  $i$ -th value in horizontal or vertical histogram of original image and reconstructed image respectively, and  $n$  means the width or height of the character. The smaller value of relative entropy indicates the more similar in appearance of histograms.

The script algorithm involves following steps:

Step 1: Compute the average images according to (1) for Korean, Chinese and English character.

Step 2: Compute the covariance matrix of the three types of characters according to (2), and find the eigenvectors of the matrix; establish 2, 5, 1 feature spaces for English, Korean, Chinese respectively using eigenvectors corresponding to the top 30, 80, 150 largest eigenvalues as orthogonal vectors of the feature space.

Step 3: By (3), reconstruct image that projected into the English space.

Step 4: Compute Euclidean distance between the two images and relative entropy by (4); if the restricted

condition is not satisfied then project the original image into the Korean space or identified as English character.

Step 5: After the same procedure for Korean and Chinese space in sequence, if the restricted condition is not satisfied in all cases then the character is rejected to be identified.

In above steps the restricted conditions for English and Korean denote that Euclidean distance between two images and horizontal and vertical relative entropy are less than the given thresholds  $D, EH, EV$  respectively, and for Chinese it means horizontal and vertical relative entropy are less than the given threshold  $E$ . Fig.5 illustrates the script identification result of multi-script image which includes Korean, Chinese and English, where the mark circle, cross and plus indicate the identified result of Korean, Chinese and English character respectively.

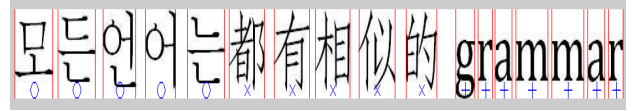


Figure 5. Script identification result

### D. Experimental result

Experiments are performed on the dataset that include two parts, to make sure total characters of Korean, Chinese and English can be tested; the first part is obtained by computer system, and the other part is consists of scanned images. It is founded that top 30, 80, 150 largest eigenvalues corresponding to English, Korean, Chinese account for 75.89%, 67.80%, 84.45% of the sum of eigenvalues respectively, and the thresholds  $D, EH, EV$  and  $E$  mentioned in algorithm take the value of 250, 0.1, 0.2 and 0.15 respectively.

It is noted from TABLE I that the proposed method achieved high performance and is flexible, since it is at character level so there is no limitation for script image. Where IER, SER and RER stand for the rate of misidentification, and rejection, and IER means without mis-segmentation (Error No.=Character No. $\times$ (IER+SER)). There are certain reasons bring out the misidentification:

- Structure simplicity. For example, when Chinese '一' is projected to Korean space it also get the similar image.
- Overlap between two characters in special font style. For example, English words 'face'.

TABLE I SCRIPT IDENTIFICATION RESULTS

Script Type	Character Num.	Error Num.	IER (%)	SER (%)	RER (%)
Korean	8000	19	0.1	0.13	0
Chinese	15000	45	0.2	0.1	0.28
English	1000	14	0.4	1	0

## IV. CONCLUSIONS

It is essential to know the type of script used in writing a text before an appropriate character recognition method and

document analysis algorithm can be chosen. In view of this, we proposed PCA-based method in combination with Euclidean distance and relative entropy, which achieved high accuracy. The previous work focused on script identification at page-wise, paragraph, text-block and textline-wise, and combination of features and classifier. The proposed method addressed the problem of time-consuming caused by training classifier and the limitation to adapt to different forms of script.

The future work includes that generalize the proposed method, in order to adapt to the other characters and to improve the performance of identification by adopting character-recognition-based post-processing.

#### ACKNOWLEDGMENT

Our work was supported by Sci. & Tech. Department of Jilin Prov. of China under Grant No. 20050703-1, Educational Foundation of Yanbian University, Key Subjects Foundation of Yanbian University.

#### REFERENCES

- [1] A. Lawrence Spitz, Determination of the Script and Language Content of Document Image, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 1997, No.3, Vol.19, pp. 235-245.
- [2] Gu Lijuan, Shao Mingshan, Hao Yubao, LANGUAGE IDENTIFICATION METHOD BASED ON SUB-BAND ENERGY FEATURES OF STEERABLE PYRAMID, Computer Application and Software(Chinese edition with English abstract), 2011, No.3, Vol.28, pp. 91-94.
- [3] Hidayet Takci, Tunga Gungor, A high performance centroid-based classification approach for language identification, Pattern Recognition Letters, 2012, No.33, pp. 2077-2084.
- [4] D. Ghosh, T. Dube, A. P. Shivaprasad: Script Recognition - A Review, IEEE Transaction on Pattern Analysis and Machine Intelligence, 2010, No.12, Vol.32, pp. 2142-2161.
- [5] GU Lijuan, PING Xijian, CHENG Juan, HAO Yubao, A Robust Rotation-invariant Script Identification Method of Document Images, Journal of Image and Graphics(Chinese edition with English abstract), 2010, No.6, Vol.5, pp. 879-886.
- [6] GUO Hai, ZHAO Jing-ying, WEI Zong-wei, A Method of Chinese Minority Script Identification Using Wavelet Packet Decomposition and RBFN, COMPUTER ENGINEERING & SCIENCE(Chinese edition with English abstract), 2010, No.8, Vol.32, pp. 78-80.
- [7] GU Li-juan, LIU Cai-bin, WU Yong, HAO Yu-bao, Script identification of document images based on multi-wavelet transform, Electronic Design Engineering(Chinese edition with English abstract), 2011, No.15, Vol.19, pp. 152-155.
- [8] GUO Long, PING Xi-jian, ZHOU Lin, TONG Li, Identification of Scripts in Document Images Using Basic Image Features, JOURNAL OF APPLIED SCIENCES-Electronics and Information Engineering(Chinese edition with English abstract), 2011, No.1, Vol.29, pp. 56-60.
- [9] Script Identification-A Han & Roman Script Perspective. International Conference on Pattern Recognition. 2010, pp. 2708-2711.
- [10] Bilal Bataineh, Siti Norul Huda Sheikh Abdullah, Khairuddin Omar. A novel statistical feature extraction method for textual image: Optical font recognition. Expert Systems with Applications. 2012, No.39, pp. 5470-5477.
- [11] Peeta Basa Pati, A. G. Ramakrishnan. Word level multi-script identification. Pattern Recognition Letters, 2008, No.29, pp. 1218-1229.
- [12] P. S. Hiremath, S. Shivashankar. Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image. Pattern Recognition Letters. 2008, No.29, pp. 1182-1189.
- [13] W. M. Pan, C. Y. Suen, T. D. Bui. Script Identification Using Steerable Gabor Filters. Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition. 2005.
- [14] Amir Egozi, Its'hak Dinstein. Statistical mixture model for documents skew angle estimation. Pattern Recognition Letters. 2011, No.32, pp. 1912-1921.
- [15] BI Xiao-jun, ZHAO Wen, Text image denoising based on higher-order cumulant, Applied Science and Technology(Chinese edition with English abstract), 2007, No.10, Vol.34, pp. 1-4.
- [16] Amjad Rehman, Tanzila Saba. Performance analysis of character segmentation approach for cursive script recognition on benchmark database. Digital Signal Processing. 2011, No.21, pp. 486-490.
- [17] Matthew Turk, Alex Pentland. Eigenfaces for Recognition. Journal of Cognitive Neuroscience. 1991, No.1, Vol.3, pp. 71-72.
- [18] Mingji PIAO, Sejin Kim, Rongyi CUI. Structure Based Modern Korean Character Set Partitioning and Pre-Classification Method of Korean Character Recognition. International Conference on Computer Science and Information Processing. Xi'an. 2012, pp. 660-663.
- [19] CUI Rongyi, KIM Sejin, Research on Information Structure of Korean Characters, JOURNAL OF CHINESE INFORMATION PROCESSING(Chinese edition with English abstract), 2011, No.5, Vol.25, pp. 114-119.

# An Approach to Script Identification in Multi-language Text Image

[Full Text](#)  
[Sign-In or Purchase](#)**Need Full-Text?**

Request a free trial to IEEE Xplore for your organization.

[FREE TRIAL](#)

2

Author(s)

Piao, Mingji ; Cui, Rongyi

[Abstract](#)[Authors](#)[References](#)[Cited By](#)[Keywords](#)[Metrics](#)[Similar](#)

A character level script identification method to identify Korean, Chinese and English scripts using PCA is proposed in this paper. First, the space of eigenvectors was constructed by using PCA, and the segmented character was reconstructed by projecting the character into the space. Second, relative entropy between original and reconstructed image is computed for vertical and horizontal histogram. Finally, the written language was identified according to Euclidean distance and relative entropy between original and reconstructed image. The experiment results show that proposed method achieved 99.78% high accuracy for correct segmentation which effectively solved the script identification problem for multi-language text image contains Korean, Chinese and English.

**Published in:**

Intelligent Networks and Intelligent Systems (ICINIS), 2013 6th International Conference on

**Date of Conference:**

1-3 Nov. 2013

**Page(s):**

248 - 251

**Print ISBN:**

978-1-4799-2808-8

**Conference Location :**

Shenyang, China

**DOI:**

10.1109/ICINIS.2013.70

**Publisher:**

IEEE

Tweet

0

[Share](#)

FREE

Multiphysics  
Simulation  
Online  
Magazine[READ NOW](#)

COMSOL

[Sign In](#) | [Create Account](#)**IEEE Account**[Change Username/Password](#)[Update Address](#)**Purchase Details**[Payment Options](#)[Order History](#)[Access Purchased Documents](#)**Profile Information**[Communications Preferences](#)[Profession and Education](#)[Technical Interests](#)**Need Help?**[US & Canada: +1 800 678 4333](#)[Worldwide: +1 732 981 0060](#)[Contact & Support](#)



Record output:

NOTE: Your selected records (to a maximum of 500) will be kept until your session ends.



However, to delete them after this task:

- Return to the Search results page and click Delete Selected Records, or
- Go to the Selected records page and click Remove All, or
- Click the End session link at the top of the page

1.

**Accession number:** 20143718147583

**Title:** An approach to script identification in multi-language text image

**Authors:** Piao, Mingji<sup>1</sup> ; Cui, Rongyi<sup>1</sup> 

**Author affiliation:** <sup>1</sup> Intelligent Information Processing Lab., Dept. of Computer Science and Technology, Yanji, China

**Corresponding author:** Cui, R. (cuirongyi@ybu.edu.cn)

**Source title:** Proceedings - 2013 6th International Conference on Intelligent Networks and Intelligent Systems, ICINIS 2013

**Abbreviated source title:** Proc. - Int. Conf. Intelligent Networks Intelligent Syst., ICINIS

**Monograph title:** Proceedings - 2013 6th International Conference on Intelligent Networks and Intelligent Systems, ICINIS 2013

**Issue date:** 2013

**Publication year:** 2013

**Pages:** 248-251

**Article number:** 6754719

**Language:** English

**ISBN-13:** 9781479928088

**Document type:** Conference article (CA)

**Conference name:** 2013 6th International Conference on Intelligent Networks and Intelligent Systems, ICINIS 2013

**Conference date:** November 1, 2013 - November 3, 2013

**Conference location:** Shenyang, China

**Conference code:** 107267

**Sponsor:** et al.; Institute of Electrical and Electronics Engineers (IEEE); Intelligent Networks and Systems Society (INASS); Japanese Neural Network Society; Shenyang Institute of Engineering; WSEAS Japan Chapter on Intelligence and Informatics

**Publisher:** IEEE Computer Society

**Abstract:** A character level script identification method to identify Korean, Chinese and English scripts using PCA is proposed in this paper. First, the space of eigenvectors was constructed by using PCA, and the segmented character was reconstructed by projecting the character into the space. Second, relative entropy between original and reconstructed image is computed for vertical and horizontal histogram. Finally, the written language was identified according to Euclidean distance and relative entropy between original and reconstructed image. The experiment results show that proposed method achieved 99.78% high accuracy for correct segmentation which effectively solved the script identification problem for multi-language text image contains Korean, Chinese and English. © 2013 IEEE.

**Number of references:** 19

**Main heading:** Image reconstruction

**Controlled terms:** Entropy - Image segmentation - Intelligent networks - Intelligent systems - Principal component analysis

**Uncontrolled terms:** Character level - Character segmentation - Euclidean distance - Reconstructed image - Relative entropy - Script identification - Script identificationt - Text images

**Classification code:** 641.1 Thermodynamics - 723.2 Data Processing and Image Processing - 723.4 Artificial Intelligence - 741 Light, Optics and Optical Devices - 741.1 Light/Optics - 922.2 Mathematical Statistics

**DOI:** 10.1109/ICINIS.2013.70

**Database:** Compendex

Compilation and indexing terms, © 2014 Elsevier Inc.

---

Copyright © 2014 [Elsevier B.V.](#) All rights reserved.