Visual Cues for Automatic Identification of Languages

Ching Y. Suen †
† CENPARMI, Concordia University
Montreal, Quebec H3H 1M8, Canada
e-mail: suen@cenparmi.concordia.ca

Masahiko Hamanaka †‡
‡ C&C Media Res. Labs., NEC Corporation
Miyamae-ku, Kawasaki 216-8555, Japan
e-mail: hama@ccm.cl.nec.co.jp

Abstract

Computers with the capability of identifying languages printed in documents have many applications including document classification, language understanding, information compression and retrieval. Each language exhibits certain types of visual appearances. This paper briefly describes the problem of identifying languages using their visual properties. It then focuses on the identification of three major oriental languages based on the extraction of such visual cues as complexity, density, lines, curvatures and circles. A new feature based on contour density has been proposed. Encouraging results will be presented.

1 Introduction

Hundreds of languages exist today [1, 2]. It would have been much easier for humans to communicate with each other if every person knows all these languages, or if there is only one language in this world. Also, very few human beings know many languages other than their mother tongue. Indeed it takes a long time to acquire a second language. Then, with advances in computers and computation, we may ask whether it is possible for us to teach the computer to understand different languages and have it translate an unknown language into a known one to us. This concept seems quite simple, but it will take an immense effort to do so. Actually, machine translation is still far from perfect and much human intervention is required to translate one language to another. Hence it is difficult to understand a document if it is not printed in the language we are familiar with. However, one can still take up the challenge of trying to teach the computer to understand different languages step by step, e.g. by first digitizing the document, identifying the language of the print, and then trying to read and understand it. And if a language can be identified by the computer, several interesting operations are possible, e.g. the classification of the document, and pursuit of reading and understanding the print for information compression, classification, and retrieval, plus many other applications in document processing.

During the past few years, several studies have already been made on the identification of both European and Asian languages, and very interesting results have been generated [3, 4, 5, 6]. They have used visual cues such as letter and word shapes [3, 5], horizontal projection profiles [4, 6], upward and downward concavities [4], optical density [3, 4], and distribution of connected components and their sizes [6].

This paper attempts to introduce the subject of language identification by computer, and describe how visual cues such as density and curvature features can be used effectively to identify languages printed on documents. Due to the limitation of space, this paper will be confined to the study of oriental languages only, viz. Chinese, Japanese, and Korean. These languages are actually related, and they are the most widely used in the Pacific Rim [7]. A brief introduction about them will be given in the beginning.

2 Characteristics of Oriental Languages

Many oriental languages (scripts) have very different characteristics from occidental ones. Especially Chinese script is unique, and there exist many scripts influenced by Chinese in nearby countries [2]. In this paper, the authors pay attention to Chinese,

Japanese and Korean because of their popularity and the vast number of documents written in these three languages.

In China, although several languages are actually used, Chinese is the national language, and is written mainly in Chinese characters, which are ideograms. It is believed that there are about 55,000 Chinese characters, while only about 3,000 – 4,000 are used daily [7]. Some of them are very complex; the average number of strokes of the traditional characters is approximately 10. Nowadays, simplified Chinese characters are used in China, while the traditional form is still used in Taiwan and Hong Kong.

In Japan, Japanese is the official language and is written in Kanji and two sets of Japanese Kanas (Hiragana and Katakana). Kanji, which means Hanzi in Chinese, came from China in the ancient times, and many characters still take the same shape as Chinese ones, while there also exist Japanese original Kanji characters. Kana characters are pure syllabaries, obtained by simplifying some Kanji characters; their average number of strokes is 2 or 3. Hiragana is a cursive script which is used together with Kanji, while Katakana is an angular script which is used for loanwords and emphasis. Each set contains approximately 80 characters. It is said that in usual Japanese texts about 50% is occupied by Hiragana and about 30 – 40 % is by Kanji.

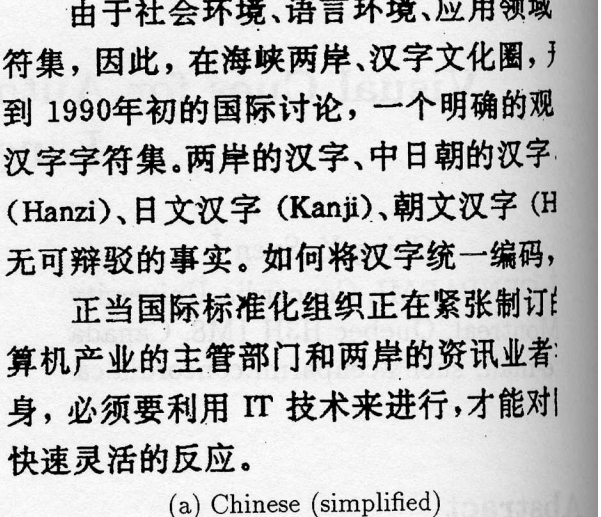
In Korea, Korean is the official language and is written in Hangul (Korean script). In South Korea, traditional Chinese characters can also be used with Hangul. Hangul is the unique script where two or three syllabic elements are combined; basically there are 14 consonant and 10 vowel elements. Therefore there exist thousands of Hangul characters.

Chinese, Japanese and Korean languages can be written either in vertical columns which shift from right to left or in horizontal rows. Of course, sometimes alphanumerics are also used in their texts.

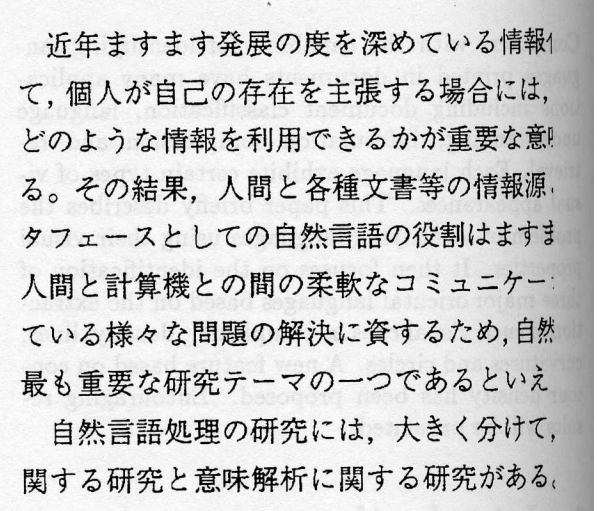
3 Language Identification of Oriental Documents

3.1 Visual Cues

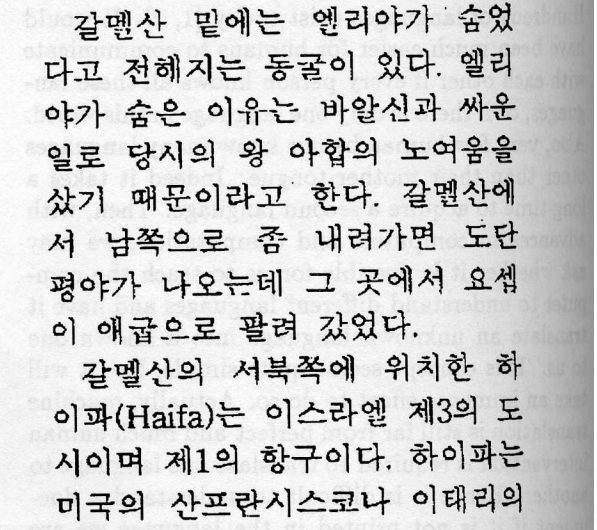
Figure 1 shows examples of document images of the three oriental languages, which exhibit some differences among them. At first, Chinese characters including Japanese Kanji have various complexities: their stroke numbers range from 1 to about 30.


由于社会环境、语言环境、应用领域
符集，因此，在海峡两岸、汉字文化圈，
到 1990年初的国际讨论，一个明确的观
汉字字符集。两岸的汉字、中日朝的汉字
(Hanzi)、日文汉字 (Kanji)、朝文汉字 (H
无可辩驳的事实。如何将汉字统一编码，
正当国际标准化组织正在紧张制订
算机产业的主管部门和两岸的资讯业者
身，必须要利用 IT 技术来进行，才能对
快速灵活的反应。

(a) Chinese (simplified)


近年ますます発展の度を深めている情報
て，個人が自己の存在を主張する場合には，
どのような情報を利用できるかが重要な意
る。その結果，人間と各種文書等の情報源
タフェースとしての自然言語の役割はます
人間と計算機との間の柔軟なコミュニケー
ている様々な問題の解決に資するため，自然
最も重要な研究テーマの一つであるといえ
自然言語処理の研究には，大きく分けて，
関する研究と意味解析に関する研究がある

(b) Japanese


갈멜산 밑에는 엘리야가 숨었
다고 전해지는 동굴이 있다. 엘리
야가 숨은 이유는 바알신과 싸운
일로 당시의 왕 아합의 노여움을
샀기 때문이라고 한다. 갈멜산에
서 남쪽으로 좀 내려가면 도단
평야가 나오는데 그 곳에서 요셉
이 애굽으로 팔려 갔었다.
갈멜산의 서북쪽에 위치한 하
이파(Haifa)는 이스라엘 제3의 도
시이며 제1의 항구이다. 하이파는
미국의 산프란시스코나 이태리의

(c) Korean

Figure 1: Example of document images

However, generally traditional Chinese characters are slightly more complex than Kanji, while simplified Chinese characters are a little simpler than Kanji. Korean characters (Hangul) are not so simple, but on the average less complex than simplified Chinese characters. Japanese Kana characters are simpler than Hangul.

Next, the oriental characters contain many straight lines and curves, especially Hangul characters which contain many simple lines and circles. Japanese Kana, especially Hiragana characters include many intermediate curves, while Chinese characters also have a few.

3.2 Identification Methods

In the past several years, some identification methods including identification among oriental languages have been proposed. They can be classified into two main approaches [6]: searching for specific tokens [8] and using statistical information [3, 4]. The former approach can be used for specific classification, because it does not need many character images for identification, while there is a problem that it needs a large database to extract enough specific tokens and large memory space to store them; for example about 500 KB for 4,000 templates. On the other hand, the latter tends to be used for gross classification, because it needs many character images to get enough statistics, while it has the merit of not needing much computational time and space.

In this paper, the authors use the statistical approach, and investigate its effectiveness. Based on visual cues of the oriental languages, complexity (density) and curvature features are expected to be effective for identification.

3.3 Density Feature

Spitz [3] proposed a method using optical density to classify oriental languages (scripts). In this method, a document image is first segmented into character cells, within each such cell the number of 'on' pixels is counted as an optical density. Spitz's optical density $D1_i$ of the i-th cell is defined as follows:

$$D1_i = \frac{B_i}{H_{L(i)} W_i} \tag{1}$$

where B_i is the number of 'on' pixels in the i-th cell, W_i is the width of the i-th cell, and $H_{L(i)}$ is the height of the line to which the i-th cell belongs. He showed that histograms of optical density when applied to Chinese, Japanese and Korean documents show characteristically different distributions: Chinese has only one significant mode while Japanese and Korean have two modes; in addition the low density mode is greater than the high density in Japanese while the low density mode is smaller than the high density in Korean. And he showed that these three languages can be classified by applying linear discriminant analysis (LDA) to the multivariate data generated from the specific areas of distributions.

However, Spitz's optical density is influenced by stroke width. In addition, Spitz's method assumes that the inter-character spacing is so large as a character splitting process is not needed. Lee *et al.* [4] also pointed out these problems. They modified Spitz's method slightly; at first a document image is segmented into square cells after all white gaps (inter-character spaces) are eliminated, then within each square cell, a density is calculated as the fraction of its area that is black, multiplied by the average number of runs per cell computed over the page. They showed that Chinese and Japanese documents can be classified using the mean and variance of the density. However, there are problems that Lee's density is still influenced by stroke width, and it uses the average number of runs that cannot be calculated in each cell. In addition, Lee's segmentation method makes densities insensitive, because adjoining complex and simple characters are mixed into a cell. Density features depend on character segmentation, therefore the characters should be segmented almost successfully, because complexities in oriental languages change considerably from character to character.

Optical densities are influenced by stroke width, while those extracted from thinned images are insensitive to stroke width. However, since thinning process usually needs much time, the authors propose contour density of the original image, which is calculated as the fraction of its area that is on contour, and can be used instead of optical density of a thinned image. The contour density $D2_i$ of the i-th cell is defined below:

$$D2_i = \frac{C_i}{H_{L(i)} W_i} \tag{2}$$

where C_i is the number of contour pixels in the i-th cell.

3.4 Curvature Feature

In the continuous case, curvature can be defined exactly. However, in the case of digital images, it is difficult to calculate curvatures exactly because successive angles can only differ by a multiple of 45 degrees. Many curvature extraction methods for

digital images have been proposed [9], based on the deviative of tangent orientation, the norm of the second deviative of the path, or the inverse of the radius of the osculating circle.

Rosenfeld and Kak [10] defined the k-curvature based on the deviative of tangent orientation. For the digital arc $S = p_0, ..., p_{n-1}$, the left and right k-slopes of S at p_i are defined as the directions from p_{i-k} to p_i and p_i to p_{i+k} respectively, where $k \geq 1$. And the k-curvature of S at p_i is defined as the difference between the left and right k-slopes. The k-curvature (and k-slope) can take on angular values; a curvature on a straight line is 0. In this method, how to choose the factor k is difficult. The factor k controls the locality of the angle; if k is too small, the angle is influenced by digital quantization, while if k is too large, the angle is far from the real curvature.

By using the k-curvature, straight lines and curves can be detected. In addition, to detect circles, the authors propose to use only inner contours for curvature extraction. If an inner contour is a circle, only constant curvatures are extracted.

In addition, curvature has a merit that character segmentation is not required, while it is needed to extract density. However, curvature is dependent on character sizes, therefore characters have to be normalized to the same height.

4 Experimental Results

4.1 Database

In the experiments, 6 Chinese, 6 Japanese and 6 Korean document images were used (Fig. 1). They were scanned from books, newspapers and magazines. From an image, two areas which include about 200 characters respectively were cut out, therefore for each language about 2,400 character images were used. The data include some alphanumerics, but the proportion is very small.

4.2 Preprocessing

Document images have been binarized. The preprocessing includes text line segmentation, deskewing and normalization. At first, text lines and their skew angle are estimated. After noises are removed by a median filter [11], text lines are estimated using the constrained run length algorithm (CRLA) [12]. The constrained parameters in CRLA, which depend on the sizes of characters and spaces, are estimated from the average height of connecting blocks. Small lines containing fewer than 50 pixels are neglected. A skew angle for each line is estimated based on the least moment of inertia [11] respectively. An average skew angle for lines is calculated as the weighted mean of the estimated skew angles of lines in an area. Next, text lines are segmented from the original images based on the information of the text line estimation. The segmented lines are rotated based on the estimated average skew angle for lines. Finally, the deskewed lines are normalized as differences between the upper and lower lines become a constant height of 64 pixels, where they are estimated from the histograms of upper and lower bounds of connecting blocks (Fig. 2).

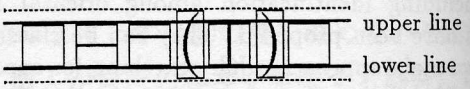


Figure 2: Normalization of the height of a line of characters

4.3 Character Segmentation

When density features are used, text lines are segmented into character cells. In this paper, merge and split techniques are used for character segmentation. At first, connecting blocks are extracted. Then small blocks whose widths are smaller than $a_m H$ are merged to adjacent blocks respectively, where H is the height of the line and a_m is a parameter (0.5 in this paper) that controls merging. The adjacent block of a merged block is selected when its width after the merge is smaller than the other's. After all widths of blocks became larger than $a_m H$, large blocks whose widths are larger than $a_s H$ split into several blocks with widths smaller than $a_s H$, where a_s is a parameter (1.0 in this paper) that controls the maximum width. A splitting point is determined based on the projection histogram, where its projection value is the smallest in the range of $a_m H$ and $a_s H$. After all widths of blocks became smaller than $a_s H$, finally small blocks whose widths are smaller than $a_n H$ are neglected, where a_n is a parameter (0.1 in this paper) that controls the minimum width. The left blocks are character cells for feature extraction.

For the experiment data, all text lines were segmented successfully, and most (about 90%) characters were segmented correctly (Fig. 3).

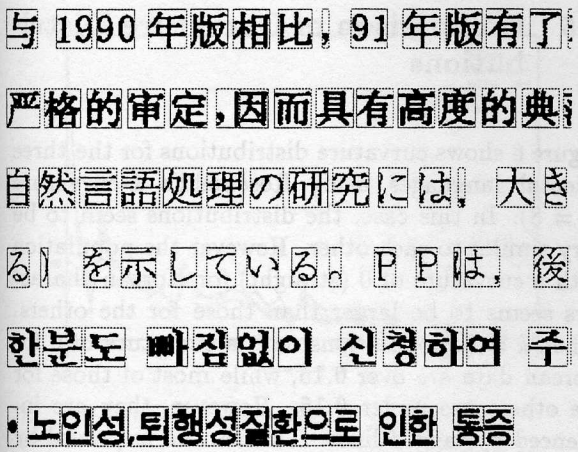
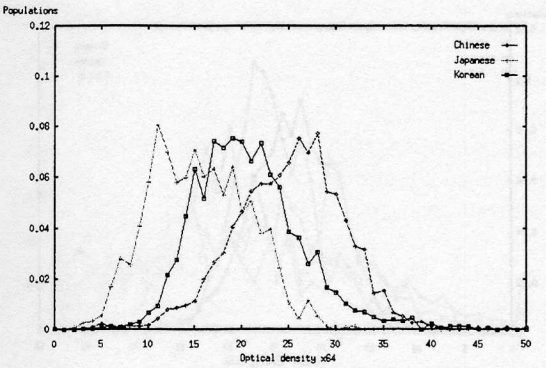


Figure 3: Examples of character segmentation

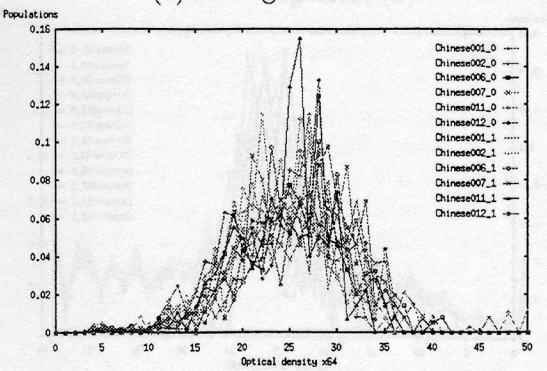
4.4 Comparison of Density Distributions

Figure 4 shows density distributions for the three oriental languages based on Spitz's optical density. The three languages seem to be identifiable using the mean values. However, this result is different from expected. For example, Japanese images do not have large densities like Chinese in Fig. 4, while Japanese actually includes complex characters. This is caused by stroke width; in this database Japanese images have narrow strokes, while Chinese images have wider strokes. Since Korean images have several ranges of stroke width, their density distributions are not stable (Fig. 4(d)). Especially, the distribution of data "Korean007_0" is largely different from those of the others, because this image has wide stroke width. As a result, since the optical densities are influenced by stroke width, their distributions are not stable in a language for various document images.

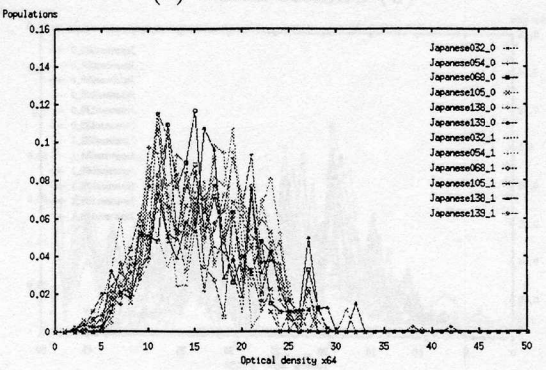
Figure 5 shows density distributions for the three oriental languages based on the contour density, and they seem to be reasonable. For example, the Japanese distribution has two main modes for Kana and Kanji; Kana characters are simpler than Hangul and simplified Chinese, and some Kanji characters are more complex than simplified Chinese. In addition, the density distributions for Korean images are stable (Fig. 5(d)). As a result, since the contour density is insensitive to stroke width, their distributions are stable in a language, independent of images. It is considered that the oriental languages can be identified in some cases by contour density; especially Japanese seems to be identified, however it is difficult to identify them solely by density features (complexities) like contour density.



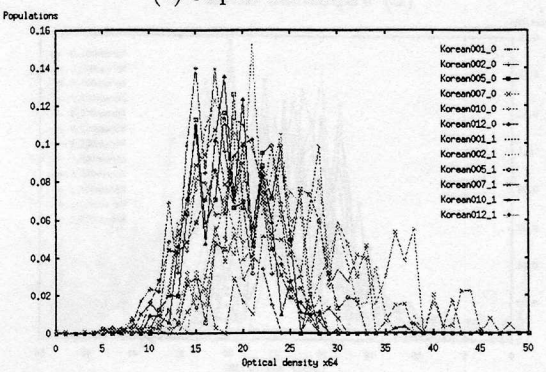
(a) Average



(b) Chinese data

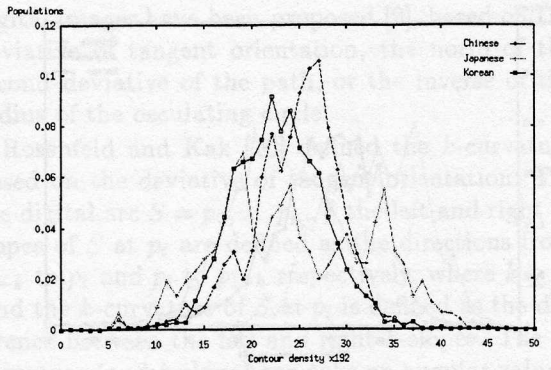


(c) Japanese data

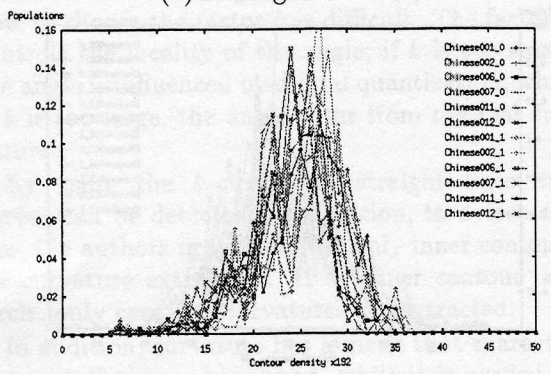


(d) Korean data

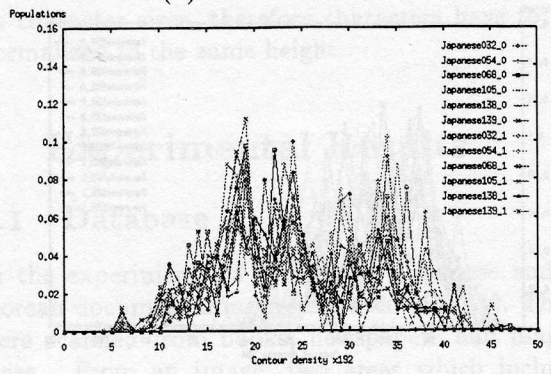
Figure 4: Distributions of optical density $D1$



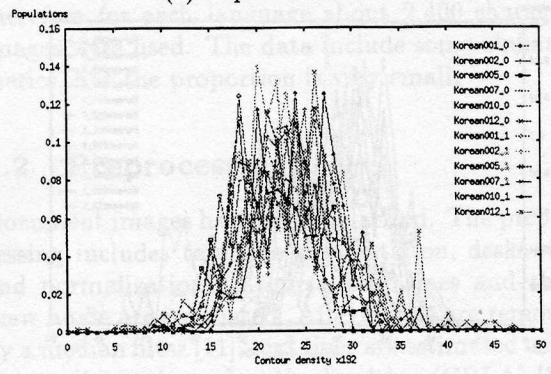
(a) Average



(b) Chinese data



(c) Japanese data



(d) Korean data

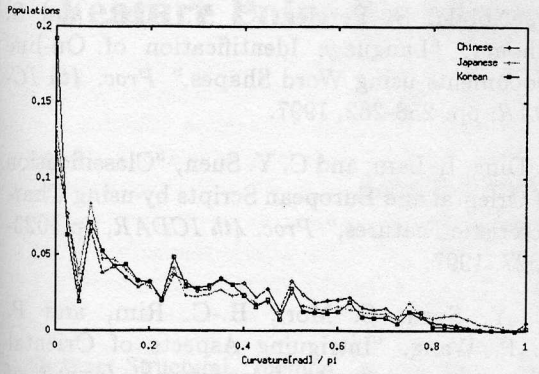
Figure 5: Distributions of contour density $D2$

4.5 Comparison of Curvature Distributions

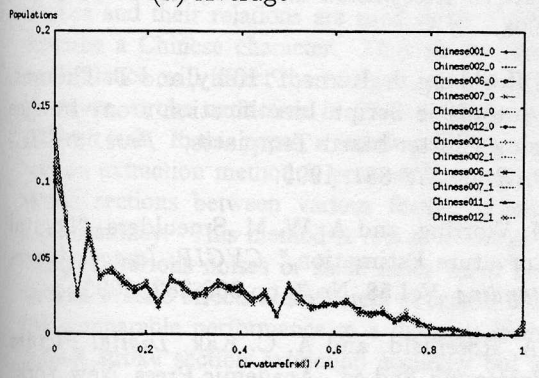
Figure 6 shows curvature distributions for the three oriental languages using Rosenfeld's k-curvature ($k = 8$). In this case, the distributions seem to be very similar to each other. However the population with a curvature of 0 (straight) for Korean characters seems to be larger than those for the others. When $k = 8$, populations with a curvature of 0 for Korean data are over 0.15, while most of those for the others are under 0.15. However, they are influenced by jagged lines. This can be improved by smoothing lines or by extracting curvatures more globally.

Next, Figure 7 shows the results when curvatures are calculated at those points on the inner contours only. In this case, the distribution of Korean language is different from the others, because circles that can be seen frequently in Korean characters are emphasized by using only inner contours. When $k = 8$, most distributions for Korean data have peaks whose populations are over 0.06 in the range of curvature between 0.3π and 0.4π (Fig. 7(d)). However, some distributions for Korean data do not have clear peaks in this range. For example, the populations of data "Korean007_0" are low in this range, because this image has wide strokes so that inner circles become smaller. This can be improved by thinning. However, many characters are needed to extract enough statistics, therefore these distributions may be a little unstable. As a result, Korean is considered to be identified by detecting both straight lines and circles.

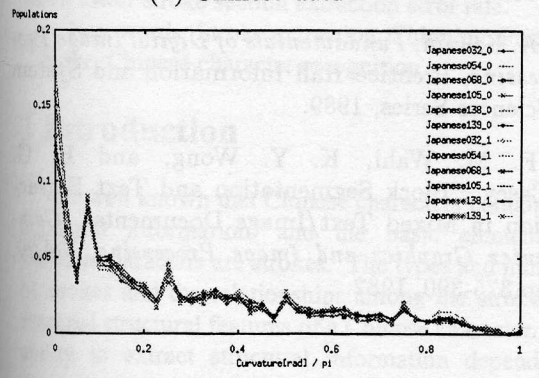
It seems to be difficult to classify Chinese and Japanese by curvature feature alone. In Fig. 6, small differences can be seen around a curvature of 0.1π (18 degree); Japanese data have a little larger populations than Chinese ones. However this is not considered to be enough to classify them. Since Chinese characters including Japanese Kanji are complex, many intermediate angles are detected around corners and jagged lines. Therefore, if characters consist of only straight lines, non-zero curvatures may exist. To remove or decrease these pseudo-curvatures, some methods can be considered. For example, pseudo-curvatures around corners can be neglected by detecting corners, and pseudo-curvatures from jagged lines can be reduced by extracting curvatures more globally.



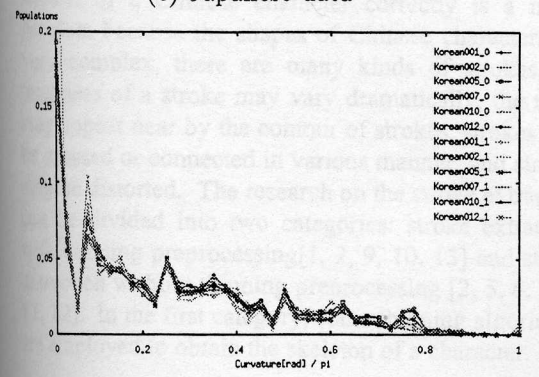
(a) Average



(b) Chinese data

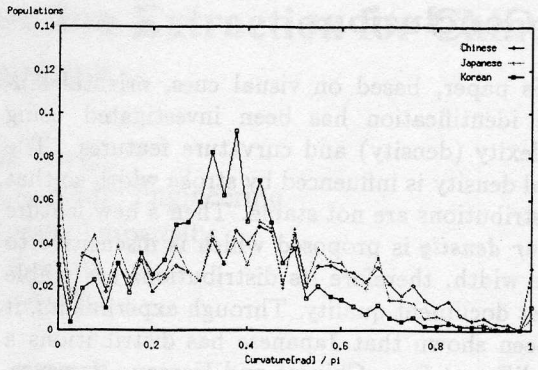


(c) Japanese data

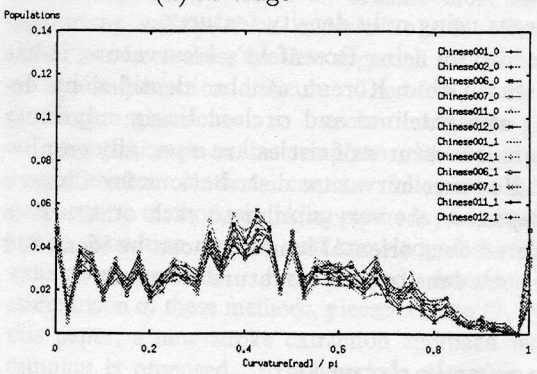


(d) Korean data

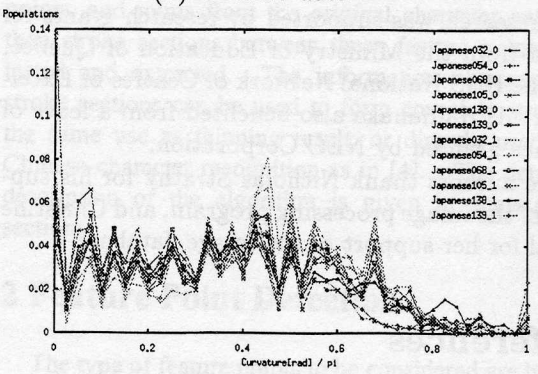
Figure 6: Distributions of curvatures ($k = 8$)



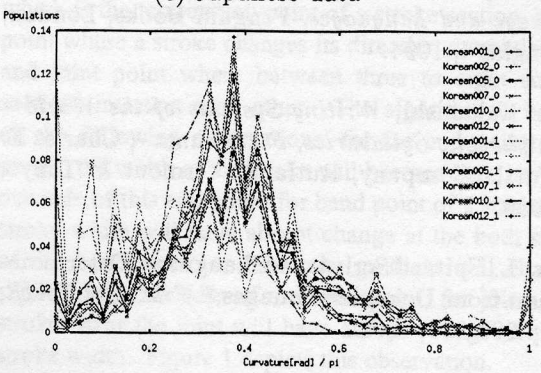
(a) Average



(b) Chinese data



(c) Japanese data



(d) Korean data

Figure 7: Distributions of curvatures on inner contours ($k = 8$)

5 Conclusions

In this paper, based on visual cues, oriental language identification has been investigated using complexity (density) and curvature features. The optical density is influenced by stroke width so that its distributions are not stable. Then a new feature *contour density* is proposed which is insensitive to stroke width, therefore its distributions are stable against document quality. Through experiments, it has been shown that Japanese has distributions a little different from Chinese and Korean. However, it seems to be difficult to identify all three oriental languages using only density features.

In addition, using Rosenfeld's k-curvature, it has been shown that Korean can be identified by detecting straight lines and circles. Using only inner contours, curvatures of circles are especially emphasized. However curvature distributions for Chinese and Japanese are very similar to each other. It is considered that oriental languages can be identified using both density and curvature features.

Acknowledgment

This research was supported by research grants received from the Ministry of Education of Quebec, and the IRIS National Network of Centres of Excellence. M. Hamanaka also benefited from a leave of absence granted by NEC Corporation.

The authors thank Nicholas Strathy for his support of the image processing program, and Christine Nadal for her support of the image database.

References

[1] D. Crystal, *An Encyclopedic Dictionary of Language and Languages*, Penguin Books, London, England, 1994.

[2] A. Nakanishi, *Writing Systems of the World – Alphabets, Syllabaries, Pictograms* –, Charles E. Tuttle Company, Rutland, Vermont & Tokyo, 1980.

[3] A. L. Spitz, "Script and Language Determination from Document Images," *Proc. 3rd DAIR*, pp.229-235, 1994.

[4] D.-S. Lee, C. R. Nohl, and H. S. Baird, "Language Identification in Complex, Unoriented, and Degraded Document Images," *Proc. DAS*, pp.76-98, 1996.

[5] N. Nobile, S. Bergler, C. Y. Suen, and S. Khoury, "Language Identification of On-line Documents using Word Shapes," *Proc. 4th ICDAR*, pp. 258-262, 1997.

[6] J. Ding, L. Lam, and C. Y. Suen, "Classification of Oriental and European Scripts by using Characteristic Features," *Proc. 4th ICDAR*, pp.1023-1027, 1997.

[7] C. Y. Suen, S. Mori, H.-C. Rim, and P. S. P. Wang, "Intriguing Aspects of Oriental Languages," in press, *International Journal of Pattern Recognition and Artificial Intelligence*, 1998.

[8] J. Hochberg, L. Kerns, P. Kelly, and T. Thomas, "Automatic Script Identification from Images using Cluster-based Templates," *Proc. 3rd ICDAR*, pp.378-381, 1995.

[9] M. Worring, and A. W. M. Smeulders, "Digital Curvature Estimation," *CVGIP: Image Understanding*, Vol.58, No.3, pp.366-382, 1993.

[10] A. Rosenfeld, and A. C. Kak, *Digital Picture Processing*, 2nd ed., Academic Press, New York, 1982.

[11] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall Information and System Sciences Series, 1989.

[12] F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents," *Computer Graphics and Image Processing*, Vol.20, pp.375-390, 1982.