Simultaneous Mixed Vertical and Horizontal Handwritten Japanese Character Line Detection

Tomotaka Kimura[1(✉)], Chinthaka Premachandra[2], and Hiroharu Kawanaka[3]

[1] Department of Electrical Engineering, Tokyo University of Science,
6-3-1 Niijuku, Katsushika-ku, Tokyo 125-8585, Japan
kimura@ee.kagu.tus.ac.jp

[2] Department of Electronic Engineering, Shibaura Institute of Technology,
3-7-5, Toyosu, Koto-ku, Tokyo 135-8548, Japan
chintaka@shibaura-it.ac.jp

[3] Graduate School of Engineering, Mie University,
1577 Kurimamachiya-cho, Tsu, Mie 514-8507, Japan

Abstract. Teachers consume considerable time and their energy in process of marking examination sheets. To reduce this burden on teachers, we have been developing an automatic marking system. To mark examination sheets automatically, handwritten character lines are extracted, and then the characters on those lines are recognized. In this paper, we discuss how character lines are extracted from Japanese handwritten examination sheets without ruled lines. Japanese characters can be written vertically and horizontally, so examination sheets written in Japanese are consisted of mixed vertical and horizontal (MVH) character lines. Conventional character line extraction methods cannot deal with MVH lines, because they have been developed to consider only horizontal character lines. This paper focuses on the simultaneous detection of MVH character lines. The result of experiments using appropriate examination sheet images shows that our method can detect MVH character lines effectively.

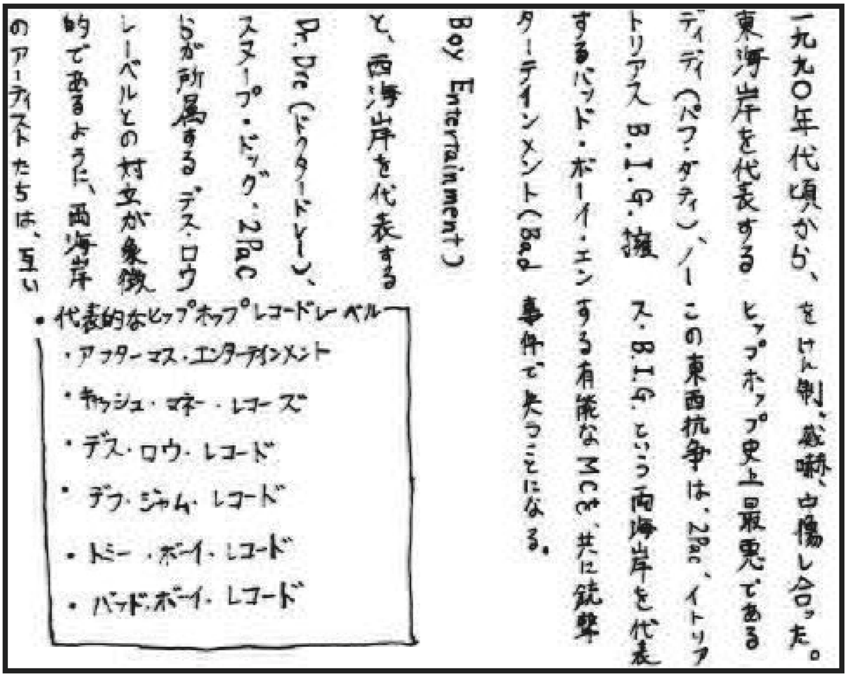
1 Introduction

In academic institutions, various paper-based examinations are conducted to evaluate the academic performance of students. In general, academic institutes in Japan provide two types of answer sheets: *marking sheets* and *writing sheets*. The former is a special mark sheet, and can thus be marked automatically by automated marking systems. These marking sheets are used in most university entrance examinations in Japan. In such examinations, students place a mark next to their chosen answer on the sheet, and the sheet is then automatically evaluated by a computer. In contrast, writing sheets require handwritten answers. Although the latest character recognition software can recognize most printed characters with high accuracy, the accuracy with handwritten characters is not especially high. In particular, characters written in Japanese are very complicated, and some characters are very similar in shape, making them difficult to distinguish.

© Springer International Publishing AG 2016
L.J. Chmielewski et al. (Eds.): ICCVG 2016, LNCS 9972, pp. 564–572, 2016.
DOI: 10.1007/978-3-319-46418-3_50

We have been developing an automatic marking system for Japanese handwritten examination sheets. An automatic marking system that can accurately process these handwritten examination sheets is a desirable applications, because teachers expend considerable time and their energy for marking handwritten examination sheets. The creation of an automatic marking system for such complicated handwritten examination sheets will reduce the burden on teachers in academic institutes.

To recognize the characters within handwritten documents such as examination sheets, character line extraction is often used. Figure 1 illustrates a typical example of a Japanese handwritten examination sheet. Japanese characters can be written vertically and horizontally, so the sheet consists of mixed vertical and horizontal (MVH) character lines. In Fig. 1, the lower left section is written horizontally, whereas the other parts are written vertically. Therefore, to extract character lines from the image, horizontal and vertical character lines must be detected.



一九九〇年代頃から、　をけん制、威嚇、中傷し合った。
東海岸を代表する　ヒップホップ史上最悪である
ディディ（パフ・ダディ）、ノー　この東西抗争は、2Pac、イトリア
トリアス B.I.G.擁　ス・B.I.G.という西海岸を代表
するバッド・ボーイ・エン　する有能なMCを、共に銃撃
ターテインメント（Bad　事件で失うことになる。
Boy Entertainment）
と、西海岸を代表する
Dr.Dre（ドクタードレー）、
スヌープ・ドッグ、2Pac
らが所属するデス・ロウ
レーベルとの対立が象徴
的であるように、西海岸
のアーティストたちは、互い

・代表的なヒップホップレコードレーベル
・アフターマス・エンターテインメント
・キャッシュ・マネー・レコーズ
・デス・ロウ・レコード
・デフ・ジャム・レコード
・トミー・ボーイ・レコード
・バッド・ボーイ・レコード

Fig. 1. An example of a Japanese handwritten sheet, where the lower left of the image is written horizontally and other parts are written vertically.

This paper focuses on the problems encountered by MVH handwritten character line extraction. Several character line extraction methods have been considered in the literature [1–9]. These methods are generally time-consuming and deal with only horizontal character lines. Hence, they cannot simultaneously extract MVH handwritten lines. Previously, we have considered the horizontal handwritten character line extraction problem of reducing the computation time [6]. In this paper, we tackle the MVH handwritten character line extraction problem. Our method for MVH character lines results in a high extraction rate. In addition, our method reduces the computation time compared with existing horizontal character line extraction methods.

The reminder of this paper is organized as follows. Section 2 introduces and discusses previous approaches to handwritten character line extraction. Section 3 describes the details of our method for the simultaneous extraction of MVH character lines. Section 4 presents and discusses the experimental results obtained in this study. Finally, Sect. 5 concludes the paper.

2 Related Studies on Character Line Detection

Several methods for character line extraction have been reported in the literature. Most of these methods consider both printed and handwritten character line detection, though some studies have only targeted handwritten character lines. Moreover, these existing studies focus on character line extraction from horizontally written documents. To the best of our knowledge, no study has yet targeted the problem of MVH character line detection.

The methods proposed by Adachi et al. [1] and Tsuruoka et al. [7] use a thinning approach to detect character lines. This technique thins all characters before using their gravity points to detect the character lines. Unfortunately, experiments suggest that these methods require more than 40 s to process a single image. Hirabayashi et al. [3] proposed a method that detects character lines via the Hough Transform. Their approach detects the gravity points of characters are detected, and then uses Hough Transform to identify the character lines. This method can also be used to simultaneously detect MVH character lines. However, it is very time-consuming because of the voting-based processing involved in the Hough Transform. In addition, this method cannot be used to detect handwritten curved character lines, because the classical Hough Transform cannot detect randomly curved lines. Examination sheets may contain lots of curved character lines, depending on the individual writing style.

Louloudisa et al. [5] have proposed a multi-step method for the character line extraction problem. The first step conducts image binarization and enhancement is conducted, connected component extraction, partitioning of the connected component domain into three spatial sub-domains, and average character height estimation. In the second step, a block-based Hough Transform is used to detects potential text lines. A third step then corrects possible splitting, detects text lines that the previous step did not reveal, and separates vertically connected characters and assigns them to text lines. This method is also very time-consuming, because connected component detection and the Hough Transform are computationally expensive.

Chaudhuri et al. [2] proposed a method that detects character lines by following the gap between two lines. This method is interesting, though it is less accurate when the gap between two lines is very small. In addition, it is difficult to employ this method for MVH character line detection problem since the gaps between and vertical and horizontal character lines become confused.

Yin and Liu [8,9] proposed a character line detection method based on minimum spanning tree (MST) clustering with new distance measures. First, the connected components of the document image are grouped into a tree by the MST clustering with a new distance measure. The edges of the tree are then dynamically cut to form text lines using a new objective function to determine the number of clusters. However, this method also requires time-consuming connected component analysis, which is especially problematic when processing large document images.

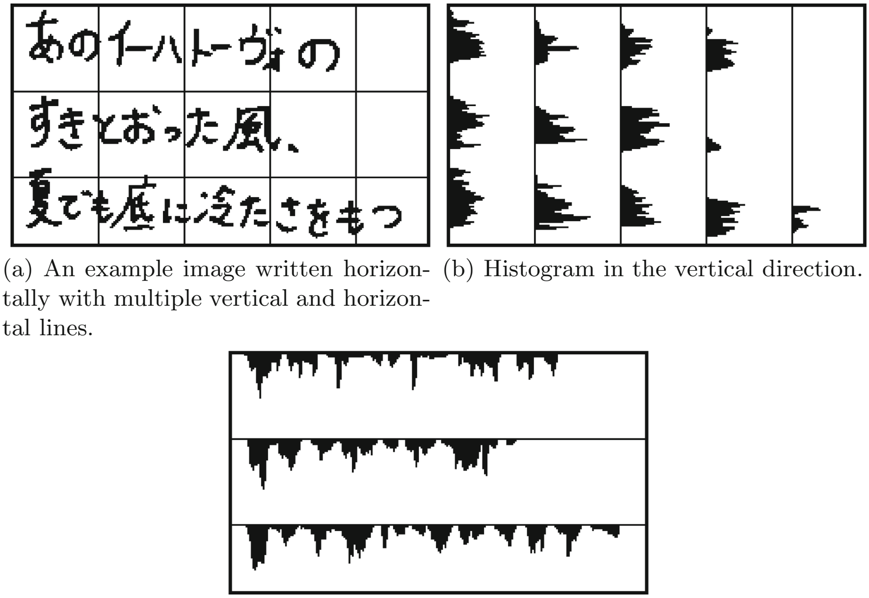
Khayya et al. [4] developed a handwritten text line detection method by applying an adaptive mask to morphological dilation. This method first identifies the characteristics of the document and its connected components to set the parameters and thresholds of the algorithm. The final smearing of the document is then determined by the dynamic mask. A recursive function plays an important role in the method as it breaks up blobs according to the attraction and repulsion of the text within those blobs. This is another interesting concept, though the connected component analysis process requires very time-consuming.

More recently, Premachandra et al. [6] proposed a method that reduces the character line detection time by analyzing block-based histograms, which is a relatively simple process. Experiments on a number of examination sheets demonstrated that the detection time of their method is lower than that of the method in [9].

3 Proposed MVH Line Detection Method

The proposed method draws multiple vertical and horizontal lines, such as depicted in Fig. 2(a), and then examines histograms of black pixels in both directions. To distinguish between vertical and horizontal writing, the proposed method uses the difference between the deviations of each histogram. When creating a histogram in the vertical direction for an document written horizontally, the line spacing between each sentence causes a number of peaks to appear (see Fig. 2(b)). In contrast, when creating a histogram in the horizontal direction, a series of connected peaks appear because the characters have been written continually (see Fig. 2(c)). These observations indicate that the histogram formed horizontally more closely represents a uniform distribution than the histogram formed vertically. Therefore, the calculation of the Kullback–Leibler distance between the uniform distribution and the histogram for the horizontal orientation is less than that between the uniform distribution and the histogram for the vertical orientation. Using this observation, we can determine whether the document is written horizontally or vertically.

The procedure of the proposed method is detailed as follows, where we assume that figures in examination sheets can be removed using the technique described in [6].



(a) An example image written horizontally with multiple vertical and horizontal lines.

(b) Histogram in the vertical direction.

(c) Histogram in the horizontal direction.

Fig. 2. Concept of the proposed method.

Step 1: A $N \times M$ pixel image is divided into small regions. By setting each small region to have width w, we draw $\lfloor N/w \rfloor$ straight lines in the horizontal direction and $\lfloor M/w \rfloor$ straight lines in the vertical direction. Therefore, the image is divided into $\lceil N/w \rceil \times \lceil M/w \rceil$ regions. Further, $A_{n,m}$ represents the nth region in the horizontal direction and the mth region in the vertical direction.

Step 2: The number of black pixels is counted for each region $A_{n,m}$, and the total number of black pixels is denoted as $S_{n,m}$.

Step 3: For each region $A_{n,m}$, we set $v_{n,m}(i)$ for $i \in \{0, 1, \ldots, w-1\}$, as follows:

$$v_{n,m}(i) = \sum_{j=0}^{w-1} \frac{f(nw+i, mw+j)}{S_{n,m}},$$

where $f(k,l)$ is a function such that when the pixel at (k,l) is black, $f(k,l) =$ 1; otherwise, $f(k,l) = 0$.

For each region $A_{n,m}$, we also set $h_{n,m}(j)$ for $j \in \{0, 1, \ldots, w-1\}$ as follows:

$$h_{n,m}(j) = \sum_{i=0}^{w-1} \frac{f(nw+i, mw+j)}{S_{n,m}}.$$

Step 4: We set $u(i) = 1/w$ for $i \in \{0, 1, \ldots w - 1\}$. For each region $A_{n,m}$, we calculate the Kullback-Leibler distance for $u(i)$ and $v_{n,m}(i)$, and denote its value as $V_{n,m}$.

$$V_{n,m} = \sum_{i=0}^{w-1} v_{n,m}(i) \log \frac{v_{n,m}(i)}{u(i)}.$$

We also calculate the Kullback-Leibler distance for $u(i)$ and $h_{n,m}(i)$, and denote its value as $H_{n,m}$.

$$H_{n,m} = \sum_{i=0}^{w-1} h_{n,m}(i) \log \frac{h_{n,m}(i)}{u(i)}.$$

Step 5: $V_{n,m}$ is compared with $H_{n,m}$. For $V_{n,m} < H_{n,m}$, region $A_{n,m}$ is considered to contain vertical writing. For $V_{n,m} > H_{n,m}$, region $A_{n,m}$ is considered to contain horizontal writing. For $V_{n,m} = H_{n,m}$, we cannot distinguish whether the region $A_{n,m}$ contains vertical or horizontal writing. In the following, we refer to regions for $V_{n,m} < H_{n,m}$ and $V_{n,m} > H_{n,m}$ as *vertical-writing regions* and *horizontal-writing regions*, respectively. Figure 3(b) shows the estimation result of Fig. 1, where black and gray regions indicate horizontal-writing and vertical-writing regions, respectively.

Step 6: The peak of each region is determined and all peaks are plotted on the borders of regions, as illustrated in Fig. 3(c). For vertical-writing and horizontal-writing regions, peaks are plotted on the horizontal and vertical borders, respectively.

Step 7: We extract straight horizontal or vertical lines with a technique similar to that described in [6]. If there are more than two consecutive regions with horizontal writing, we draw a straight horizontal line between these regions. More specifically, we draw a straight line from the first peak to the last peak in the consecutive regions (see Fig. 3(d)). If there are more than two consecutive regions with vertical writing, we draw a straight vertical line between these regions.

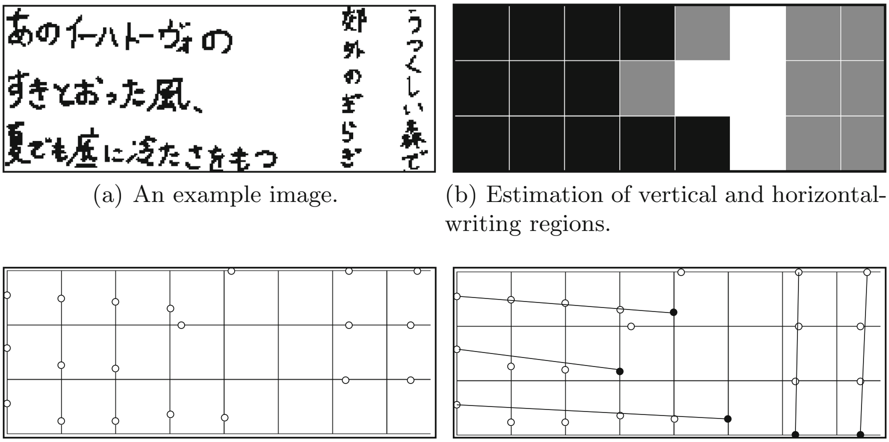
4 Experiments

4.1 Experimental Environment

All experiments were conducted using a computer with the following configuration:

OS: Mac OS 10.10, CPU: Core i5 3.5 GHz
RAM: 8.00 GB, Programming language: C++

The images used in the experiments had the following specifications:



(a) An example image.

(b) Estimation of vertical and horizontal-writing regions.

(c) Peaks are plotted on the borders of regions.

(d) Line extraction.

Fig. 3. An example of applying the proposed method.

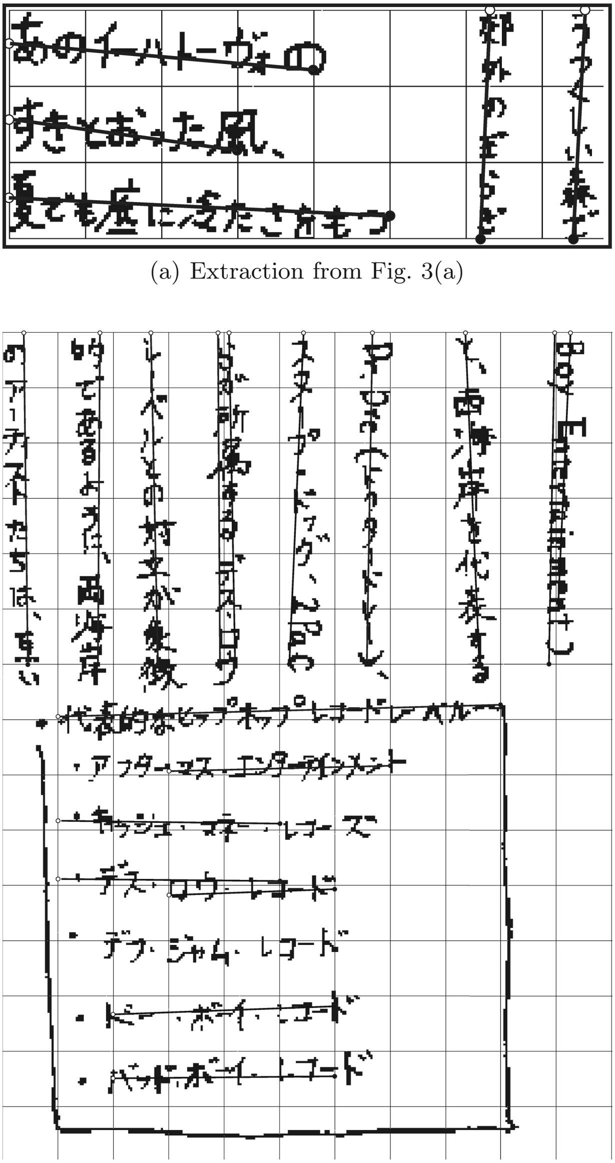
Handwritten examination sheets: 30
Image Size: 1646 × 2079
Number of character lines: 660

Experiments were conducted to verify the performance of the proposed character line extraction method and its required processing time. Further, we compared the performance of our method with that of a HT (Hough Transform)-based method.

4.2 Experimental Results

Figure 4 shows the character line extraction results obtained using our method. All lines in Fig. 4(a) are extracted correctly, whereas in Fig. 4(b) the third horizontal line from the bottom cannot be extracted because the line spans multiple blocks. This result indicates that the setting of the block-width w is important to extract lines with the proposed method. We leave the appropriate setting of the block-width w for future works.

Table 1 shows the character line extraction rate of the proposed method and the HT-based method. The average processing time of our method is 3.2 s. The results indicate that our proposed method gives better character line extraction and reduced processing time compared with the existing method.



(a) Extraction from Fig. 3(a)

(b) Extraction from Fig. 1

Fig. 4. Character line extraction results.

Table 1. Comparison of extracted character lines.

Method	Character line extraction rate [%]	False positive rate [%]	Average processing time [s]
Proposal	95.6	1.2	3.2
HT-based method	95.2	23.5	12.8

5 Conclusion

In this paper, we have considered the MVH character line extraction problems and described a method for simultaneous MVH detection. Through experiments with handwritten Japanese examination sheets, we have demonstrated the effectiveness of our proposed method.

References

1. Adachi, Y., Yoshikawa, T., Tsuruoka, S.: Character string segmentation using thinning algorithm from handwritten document image (in Japanese), Technical report of IEICE (The Institute of Electronics Information and Communication Engineers), PRMU98-208, pp. 121–126 (1999)
2. Chaudhuri, B.B., Bera, S.: Handwritten text line identification in Indian scripts. In: Proceedings of 10th International Conference on Document Analysis and Recognition, pp. 636–640 (2009)
3. Hirabayashi, K., Tsuruoka, S., Kawanaka, H., Takase, H., Ozaki, T.: Character line segmentation from blackboard image using hough transform. In: Proceedings of Mie Section of the Society of Instrument and Control Engineers (SICE-Mie), pp. B11-1–B11-4 (2008)
4. Khayyat, M., Lam, L., Suen, C.Y., Yin, F., Liu, C.L.: Arabic handwritten text line extraction by applying an adaptive mask to morphological dilation. In: 10th IAPR International Workshop on Document Analysis Systems, pp. 100–104 (2012)
5. Louloudisa, G., Gatosb, B., Pratikakisb, I., Halatsisa, C.: Text line detection in handwritten documents. Pattern Recogn. **41**, 3758–3772 (2008)
6. Premachandra, C., Goto, K., Tsuruoka, S., Kawanaka, H., Takase, H.: Speedy character line detection algorithm using image block-based histogram analysis. In: Image Analysis and Recognition, pp. 481–488 (2015)
7. Tsuruoka, S., Kimura, F., Yoshimura, M., Yokoi, S., Miyake, Y.: Thinning algorithms for digital pictures and their application to hand-printed character recognition. IEICE Trans. Inf. Syst. **J66–D(5)**, 525–532 (1983)
8. Yin, F., Liu, C.L.: Handwritten text extraction based on minimum spanning tree clustering. In: Proceedings of International Conference on Wavelet Analysis and Pattern Recognition, pp. 1123–1128 (2007)
9. Yin, F., Liu, C.L.: A variational bayes method for handwritten text line segmentation. In: Proceedings of 10th International Conference on Document Analysis and Recognition, pp. 436–440 (2009)