Text-line Extraction and Character Recognition of Japanese Newspaper Headlines with Graphical Designs

Minako Sawaki and Norihiro Hagita
minako@apollo3.brl.ntt.jp hagita@apollo3.brl.ntt.jp
NTT Basic Research Laboratories
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-01, Japan

Abstract

The conventional OCR fails to recognize most characters in Japanese newspaper headlines with graphical designs because of the difficulty of removing the designs. This paper proposes a method that recognizes such characters without removing the designs. First, text-line regions are extracted from a local distribution of the combination of black and white runs observed in a rectangular window while the window is shifted pixel-by-pixel in the direction of the text-line. Characters in the extracted text-line region are then recognized by displacement matching. Adaptive thresholding against the degree of degradation suppresses spurious candidates yielded by displacement matching even with graphical designs. Experimental results for fifty Japanese newspaper headlines show that the method achieves a recognition rate of 97.7%, much higher than a conventional method (17.0%).

1. Introduction

The optical character reader (OCR) plays an important role as a convenient input device in constructing a document image database, namely a digital library system. Since headlines in newspapers or magazines usually include keywords for queries, character recognition of the headlines is essential. Japanese newspaper headlines often use graphical designs to inform readers of hot articles. However, conventional OCR software hardly recognize such characters.

To tackle this problem, several methods for removing the graphical designs before recognition have been studied [1-4]. However, these methods often extract incomplete regions of character parts because the stroke configurations of complicated Kanji characters can be quite similar to textured backgrounds.

This paper, therefore, proposes a method that recognizes characters with graphical designs without removing them. The method consists of two stages: text-line extraction and character recognition. Three different techniques are introduced. The first technique is related to the matching process in character recognition. We have proposed a similarity measure, called complementary similarity measure [5], which is robust against graphical designs. This similarity measure has high recognition accuracy for degraded characters under the assumption that character heights or widths are correctly detected [5]. To apply the similarity measure to the recognition of headlines, the number of text-line regions and character heights/widths are to be extracted before recognition. Assuming headlines with graphical designs, conventional methods using projection profiles of black or white pixels cannot readily extract them. Therefore, as the second technique, we adopt a projection value [6], which takes high values for character part and low ones for backgrounds over a variety of graphical designs. The number of text-line regions and the averaged character heights/widths are extracted from a local distribution of the projection values. For the extracted text-line regions, previous methods using specific features for cut positions [7-9] rarely extract individual characters from text-lines with graphical designs. Therefore, we utilize displacement matching [10] to recognize individual characters in the extracted text-line regions. In displacement matching, a crucial problem is to suppress spurious character categories and determine correct character positions. As the third technique, this paper proposes a degradation-estimation method for adaptive thresholding to determine correct categories and their position. We will show that high recognition reliability is achieved by using the complementary similarity measure [5] and the adaptive thresholding.

We introduce the complementary similarity measure in Section 2 and a projection value for extracting text-line regions in Section 3. Section 4 describes the recognition of individual characters in the text-line region by displacement matching and adaptive thresholding. Section 5 presents experimental results.

2. Complementary similarity measure

We take a brief look at the complementary similarity

1015-4651/96 $5.00 © 1996 IEEE
Proceedings of ICPR '96

measure [5]. Each input image is binarized and then normalized in size to n (= $N \times N$) pixels. That is, a normalized binary image is used as a feature vector. Now, let $\boldsymbol{F} = (f_1, f_2, ..., f_i, ..., f_n)$(where $f_i = 0$ or 1) be an input image and $\boldsymbol{T} = (t_1, t_2, ..., t_i, ..., t_n)$(where $t_i = 0$ or 1) be a binary reference pattern. The complementary similarity measure S_c of $\boldsymbol{F}$ to $\boldsymbol{T}$ is defined as

$$S_c(\boldsymbol{F}, \boldsymbol{T}) = \frac{a \cdot e - b \cdot c}{\sqrt{T \cdot (n-T)}} \tag{1}$$

where

$$a = \sum_{i=1}^{n} f_i \cdot t_i , \qquad b = \sum_{i=1}^{n} (1-f_i) \cdot t_i , \tag{2}$$

$$c = \sum_{i=1}^{n} f_i \cdot (1-t_i) , \qquad e = \sum_{i=1}^{n} (1-f_i) \cdot (1-t_i) ,$$

$$a + e + b + c = n , \quad T = \| \boldsymbol{T} \| . \tag{3}$$

The complementary similarity measure S_c has the following characteristic against the reverse contrast pattern $\boldsymbol{F}^c$ of $\boldsymbol{F}$.

$$S_c(\boldsymbol{F}, \boldsymbol{T}) = - S_c(\boldsymbol{F}^c, \boldsymbol{T}) . \tag{4}$$

In this paper, we use the measure $|S_c|$ (absolute value of S_c) as a discriminant function in character recognition from the property of Equation(4), since newspaper headlines have different character colors. The dimension of the feature vector is determined as n = 1,024 (= $N \times N$, N = 32) pixels, which is large enough to express even complicated Kanji characters. Figure 1 shows recognition rates for the complementary similarity measure under the assumption that character heights or widths are correctly detected [5]. The reference patterns were made of black-plain characters. The measure achieves over 98% recognition accuracy with printed Kanji characters with graphical designs, when the gothic style character database ETL-2 consisting of 571 categories was used.

	Black-plain character	Textured character	Character with textured backgrounds	Outline character	Reverse contrast character
Examples					
Recog. rates	99.68	99.79	99.72	98.74	99.82

Fig.1 Recognition rates for characters with graphical designs [5]. (Test data was synthetically made of ETL-2)

3. Text-line region extraction based on complementary similarity measure

A projection value [6] for extracting text-line regions is introduced. We assume that headline images are extracted from a newspaper image before text-line region extraction using conventional methods such as [11]. Japanese newspaper headlines fall into five types in terms of character parts and backgrounds. Table 1 shows the occurrence rates of these types for two of the main newspapers in Japan. Type V is seldom used, because it is not legible.

The projection profile of black or white pixels [11] is not applicable for headlines with graphical designs (Types II - IV) but is applicable to Type I (no graphical design). Therefore, we developed an alternative projection value for Types II - IV as well as for Type I. This value focuses on the complementary relationships between characters and backgrounds in terms of black and white runs.

We will explain the algorithm for a headline with horizontal text-lines. For a headline with vertical text-lines, the scanning direction is vertical.

Let $\boldsymbol{G}$ ($N_x \times N_y$ pixels) be the input headline image. Also, let $\boldsymbol{G}_w$ ($g_w(u,y)$; $W \times N_y$ pixels; u = 1, 2, ..., W; y = 1, 2, ..., N_y) be a rectangular window that can include at least one character. Since textured backgrounds can change gradually along the horizontal or vertical axis, the text-line regions are locally estimated using the projection profile in local rectangle window $\boldsymbol{G}_w$, which is shifted pixel-by-pixel in the direction of the text-line.

The projection value $p(y)$ which enhances the difference between the character parts and backgrounds for Types I - IV is defined by:

$$p(y) = \frac{a_p \cdot e_p - b_p \cdot c_p}{\sqrt{r_T(r-r_T)\, r_X(r-r_X)}} \tag{5}$$

where

Table 1 Types of headline images in Japanese newspapers. (1 week data)

Type	Characters	Backgrounds	Occurrence(%)
I	Black-plain White-plain	White-plain Black-plain	52.3
II	Black-plain White-plain	Textured Textured	37.4
III	Textured Textured	White-plain Black-plain	7.7
IV	Outline	White-plain Black-plain Textured	2.2
V	Textured	Textured	0.4

$$a_p=\Sigma_{u=1}^{W-1} g_w(u,y)\cdot g_w(u+1,y)\ ,$$
$$b_p=\Sigma_{u=1}^{W-1}(1-g_w(u,y))\cdot g_w(u+1,y)\ ,$$
$$c_p=\Sigma_{u=1}^{W-1} g_w(u,y)\cdot(1-g_w(u+1,y))\ ,$$
$$e_p=\Sigma_{u=1}^{W-1}(1-g_w(u,y))\cdot(1-g_w(u+1,y))\ , \qquad (6)$$

$$r_T=a_p+b_p\ ,\ r_X=a_p+c_p\ , \qquad (7)$$
$$a_p+b_p+c_p+e_p=r. \qquad (8)$$

This projection profile makes use of a complementary relation of four parameters (a_p, e_p, b_p, c_p). They correspond to the four possible changes (black-to-black, white-to-white, white-to-black, and black-to-white) of black and white pixels during scanning along the direction of the text-line.

$p(y)$ lies in the range $-1 \le p(y) \le 1$. $a_p \cdot e_p$ corresponds to the product of the total of the black run-length and the total of the white run-length, and $b_p \cdot c_p$ corresponds to the square of line complexity in each scanning line. In general, $a_p \cdot e_p$ takes higher values in character parts and lower ones in black-plain or white-plain backgrounds. Also, $b_p \cdot c_p$ takes higher values for the textured backgrounds and lower ones for plain characters. As a result, $p(y)$ takes higher values in character parts than in backgrounds for Types I - IV. $p(y)$ is also invariant for character color. $p(y)$ is the same expression as the four-fold point correlation.

In order to avoid notches, the projection axis is divided into N sections and projection values $p(y)$ are averaged in each section. When the headline consists of two text-lines, the vertical projection profile generates two groups with high values. Its range h corresponds to the character height at location G_w. Figure 2 shows an example of a headline image G and a rectangular window G_w. Figure 3 shows the local distribution of $p(y)$ for four types of headline images at $x = 1$. Fig. 3 shows that two text-lines and their character heights h can be estimated for each headline image of Types I to IV. In this paper, when the averaged projection value exceeds 30% of the maximum of the projection profile, the section is determined to be a text-line region candidate. This threshold was defined empirically. In practice, the text-line regions and character heights fluctuate somewhat over G_w due to different character heights at each location and graphical designs. Therefore, the text-line region and its character height are averaged over all G_w.

4. Character recognition by displacement matching

4.1 Displacement matching

Since it is difficult to select cut positions of individual characters even with $p(y)$, displacement matching is applied to the extracted text-line region for character

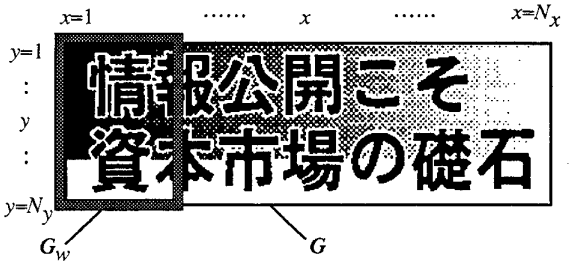

Fig.2 Observation window for text-line region extraction.

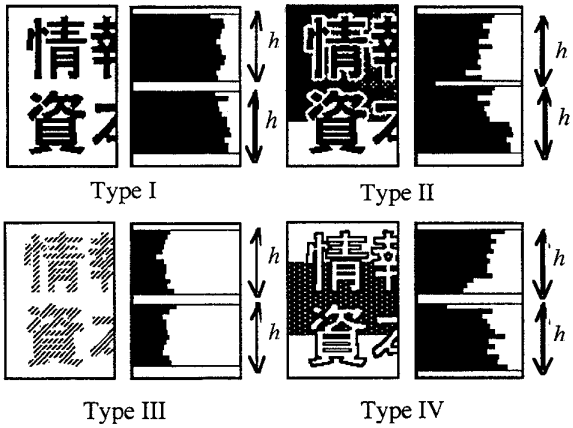

Fig. 3 Local distribution of $p(y)$ in a rectangular window for various graphical designs.

recognition. The extracted text-line region with height h is normalized to yield normalized text-line region L with height N (=32). A $N \times N$-pixel square window F is selected for matching since binary reference patterns consist of 32 ×32 pixels. F is shifted pixel-by-pixel along the direction of the text-line and is compared to the reference patterns using the absolute value of complementary similarity measure $|S_c|$ at each position.

In general, displacement matching is faced with the problem of extra candidates, that is, false categories may be selected at the location where no correct character category is located [10]. However, $|S_c|$ and adaptive thresholding may suppress these spurious candidates. The complementary similarity measure is sensitive to position translation and takes high values when the input character is matched with the reference pattern of its character category while taking low values for the other categories. These properties hold even for characters with graphical designs. Therefore, whenever F is located at a correct character position, the similarity between the window and a reference pattern of the correct category is maximal; local peaks may be observed in the distribution of maximum similarity value at the correct character position. Figure 4 shows an example of the distribution

of maximum $|S_c|$ and the maximum conventional similarity value S at each position along the horizontal axis. Figure 4 (b) and (d) show that $|S_c|$ has dominant peaks around the left-most side of $\boldsymbol{F}$ of the exactly correct category, while the conventional similarity measure $S(\boldsymbol{F},\boldsymbol{T})=a/\sqrt{F\cdot T}$, $F=\|\boldsymbol{F}\|$ has few peaks.

4.2 Adaptive thresholding

When the maximum $|S_c|$ at each position exceeds the threshold, the recognized category and its position are determined. To recognize characters with high precision, a relevant threshold should be determined using the degree of degradation in $\boldsymbol{F}$ and the reference pattern of category with the maximum $|S_c|$. When $\boldsymbol{F}$ is located at the correct character position during displacement matching, $|S_c|$ of $\boldsymbol{F}$ to the reference pattern of the correct category decreases according to the degree of degradation while $|S_c|$ of $\boldsymbol{F}$ to the reference patterns of the other categories decrease more rapidly. Therefore, by assuming that the square window $\boldsymbol{F}$ at each position contains the reference pattern having the maximal similarity value, the degree of degradation for $\boldsymbol{F}$ and the adaptive threshold against the degree can be estimated.

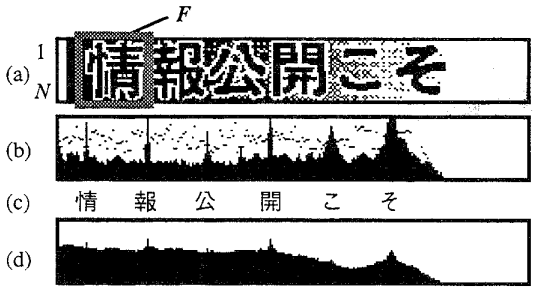

Fig.4 Character recognition by displacement matching.
(a) normalized text-line region L
(b) maximum $|S_c|$ (solid bars) and estimated thresholds $Th(T_i, Z_i)$ (dot plots)
(c) recognized categories
(d) maximum S

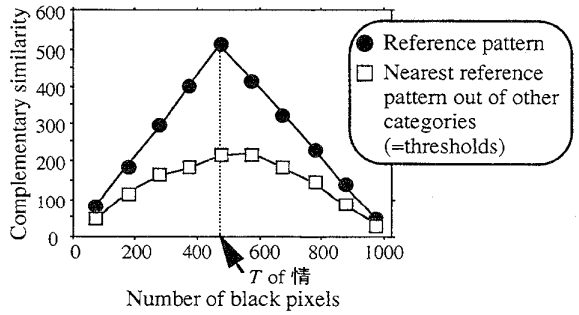

Fig.5 Relationship between complementary similarity and number of black pixels. (Example for reference pattern 情)

The thresholds for different degradation degrees are determined in advance by a learning process for each reference pattern. Let $\boldsymbol{T}_i$ be a reference pattern of the ith category without noise, $\boldsymbol{Z}_i$ and Z_i be a noisy image of $\boldsymbol{T}_i$, and the number of black pixels in $\boldsymbol{Z}_i$, respectively. Also, let $Th(\boldsymbol{T}_i, Z_i)$ be a threshold value for $\boldsymbol{T}_i$. A noisy image $\boldsymbol{Z}_i$ is synthetically formed from $\boldsymbol{T}_i$ and random dot image [5]. The similarity $|S_c(\boldsymbol{Z}_i, \boldsymbol{T}_j)|$ $(j \neq i)$ are calculated for different numbers of black pixels Z_i. Figure 5 shows an example of the threshold-to-degradation table, that is the relationship between the maximum $|S_c(\boldsymbol{Z}_i, \boldsymbol{T}_j)|$ for $i =$ 情 and Z_i. For comparison, Fig. 5 also shows $|S_c(\boldsymbol{Z}_i, \boldsymbol{T}_i)|$ for $i =$ 情. In order to eliminate false candidates from recognized categories, the maximum $|S_c(\boldsymbol{Z}_i, \boldsymbol{T}_j)|$ is determined as the threshold $Th(\boldsymbol{T}_i, Z_i)$ for the reference pattern $\boldsymbol{T}_i$. For $|S_c|$, we select $max[Th(\boldsymbol{T}_i, Z_i), Th(\boldsymbol{T}_i, n-Z_i)]$ as the threshold.

In [6], we estimated the threshold using $P=\sum_{y=1}^{N} p(y)$ of $\boldsymbol{F}$ instead of the number of black pixels. In this case, the relationship between P and the degradation level α and the relationship between α and thresholds were stored in each table. However, the estimation of degradation level from P was not always accurate enough. The threshold-to-degradation table introduced in this paper is an improvement on the previous method.

5. Experimental results

5.1 Headline data

We used 50 headlines for Types II-IV in Japanese newspapers (25 horizontal text- and 25 vertical text-lines) including 529 characters as test data. They were gathered by using three binarization thresholds, level 1 (low), level 2 (fine), and level 3 (high). The number of text-lines in one headline was either one or two. The character font in the headlines was gothic. Reference patterns without graphical designs were extracted manually from 121 headline images at level 2. As the aspect ratio of characters differs with the direction of text-lines, the reference patterns were stored in either a horizontal text-line dictionary or a vertical text-line dictionary according to the direction of the text-line of the headline. The number of reference patterns was 913 (500 categories) for the horizontal text-line dictionary and 988 (525 categories) for the vertical text-line dictionary.

5.2 Character recognition

Character recognition was conducted using the test data. The horizontal (vertical) text-line dictionary was used

when the width of the input headline image was larger (smaller) than its height.

The recognition results for 50 test headline images are shown in Table 2. Table 2 shows that the recognition rate for level 2 is 97.7% and the recognition rates for levels 1 and 3 are 81.1% and 82.4%, respectively. The number of extra candidates are 43 (level 1), 23 (level 2) and 47 (level 3), respectively. Fig.6 shows several examples of correctly recognized headlines using the proposed method.

For comparison, a conventional method was applied to the test data with level 2 scanning. Graphical designs were removed using the method in [2] and the resulting images were recognized with commercial OCR software. Figure 7 shows an example of unsuccessful results of the graphical design removal by [2]. The recognition rate was 17.0%. This shows that the proposed method achieves much higher recognition rates than the conventional method. Table 2 also shows recognition results achieved with previous adaptive thresholding [6] based on the summation of $p(y)$. The adaptive thresholding in this paper yields higher recognition rates than that in [6].

Recognition errors in our method falls into two main sources: unsuccessful extraction of text-line regions and adaptive thresholding. Error on adaptive thresholding occurred when the maximum similarity of the correct category was less than the estimated threshold. The complementary similarity measure achieves over 98% recognition accuracy against both learning and test samples when the character height is known [5]. The accuracy of headline image recognition will be improved by eliminating these two sources of recognition error.

Table 2 Recognition results

Threshold level for binarization	level 1	level 2	level 3
Recognition rate	81.1% (73.7%)	97.7 (91.1)	82.4 (74.5)
Recognition rate of conventional method		17.0	

() : using previous thresholding [12]

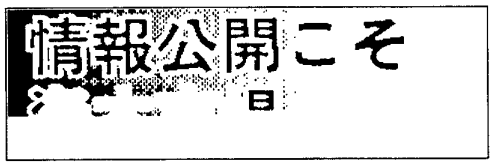

Fig.7 Result of graphical design removal by a conventional method [2].

6. Conclusion

We have proposed a method for text-line extraction and character recognition of Japanese newspaper headlines with graphical designs. The projection value based on the complementary similarity measure successfully estimates the number of text-lines in the headlines and their heights/widths. Characters in the extracted text-lines are then recognized with the complementary similarity measure by displacement matching. Spurious candidates in displacement matching are suppressed by adaptive thresholding. Experimental results for 50 newspaper headlines show that this method achieves high recognition rates of over 97%, which is much higher than the 17% of a conventional method. Improving text-line extraction and adaptive thresholding are future tasks.

7. Acknowledgments

We would like to acknowledge the encouragement and support of Dr. Ken'ichiro Ishii and stimulating discussions with Dr. Kazumi Odaka.

8. References

[1] H. Sakou, H. Matsushima and M. Ejiri, "Texture Discrimination Using Self-Organized Multiresolution Filtering", *IEICE D-II*, Vol. J73-D-II, No.4, pp.562-573, 1990 (in Japanese).

[2] M. Okamoto and H. Hayashi, "Character Extraction from Headlines with Background Patterns by Using Shrinking/ Expanding Methods", *IEICE Technical Report PRU90-151*, p.47-54, 1991 (in Japanese).

[3] S. Liang and M. Ahmadi, "A Morphological Approach to Text String Extraction from Regular Periodic Overlapping Text/Background Images", *CVGIP*, Vol.56, No.5, Sep, pp.402-413, 1994.

[4] H. Ozawa and T. Nakagawa, "A character image enhancement method from characters with various background images," *Proc. of Second ICDAR*, Tsukuba, Japan, October, pp.58-61, 1993.

[5] M. Sawaki and N. Hagita, "Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning", *IEICE Trans. Inf. & Syst.*, Vol. E79-D, No.5, 1996.

[6] M. Sawaki and N. Hagita, "Character Recognition of Japanese Newspaper Headlines with Graphical Designs", *Proc. of SPIE*, San Jose, U.S.A, Jan., Vol. 2660, pp.175 - 183, 1996.

[7] S. Kahan and T. Pavlidis, "On the Recognition of Printed Characters of any Font and Size", *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-9, 274-287, March, 1987.

[8] R. Fenrich, "Segmentation of automatically located handwritten words", *In Proc. 2nd International Workshop on Frontiers in Handwriting Recognition*, pp. 33-44, Chateau de Bonas, France, 1991.

[9] S. Tsujimoto and H. Asada, "Major Components of a Complete Text Reading System," *Proc. of the IEEE*, Vol. 80, No.7, pp.1133-1149, 1992.

[10] V. A. Kovalevsky, "Image Pattern Recognition", Springer, Berlin, 1980.

[11] T. Akiyama and N. Hagita, "Automated entry system for printed documents", *Pattern Recognition*, Vol.23, No.11, pp.1141-1154, 1990.

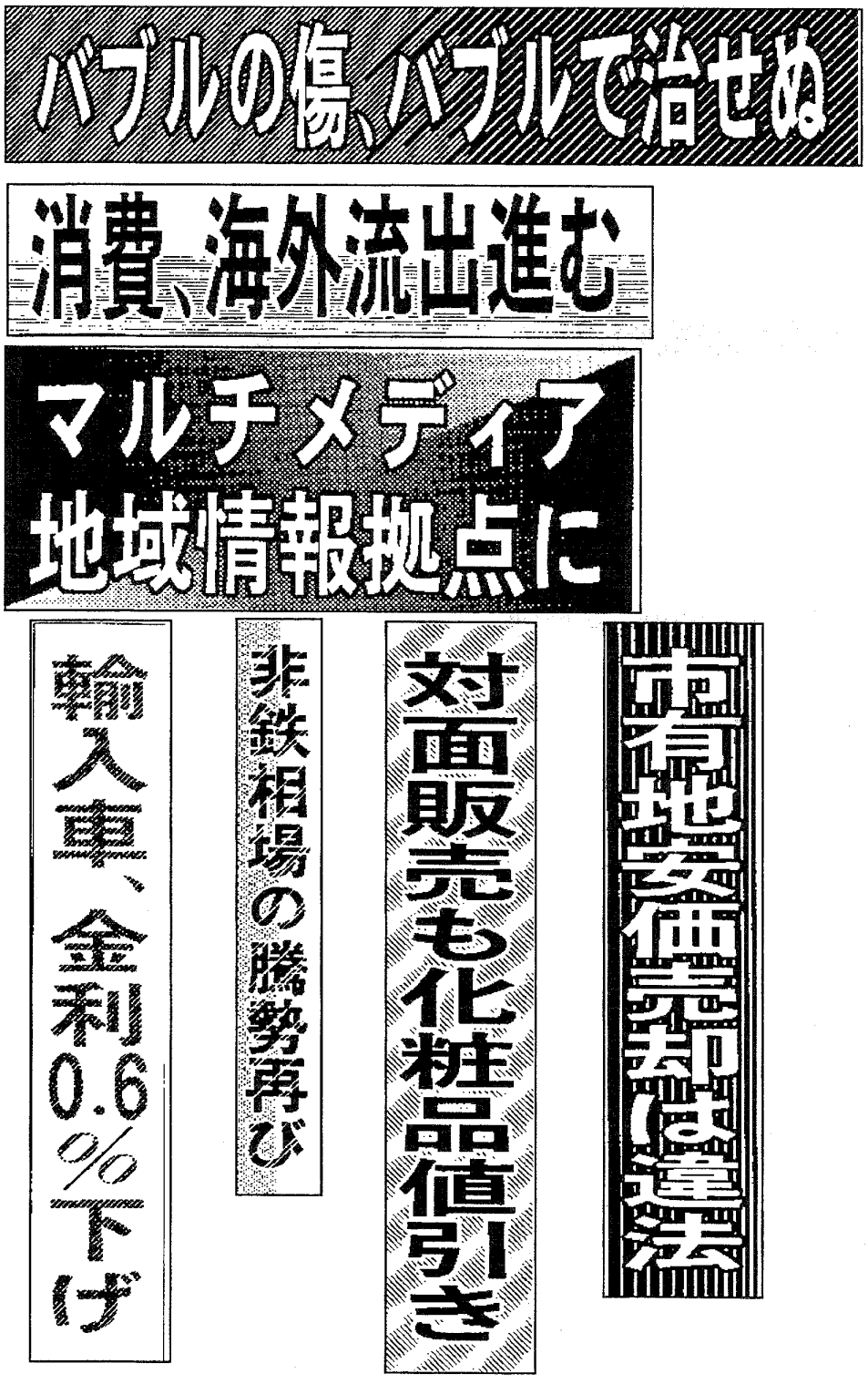

Fig.6 Examples of correctly recognized headlines using the proposed method.