

## **Separating drawings, formula and text from free handwriting**

Keisuke MOCHIDA<sup>1</sup>, Masaki NAKAGAWA<sup>2</sup>

*Department of Computer, Information and Communication Sciences,  
Tokyo University of Agriculture & Technology, JAPAN.  
<sup>1</sup>mcdksk@hands.ei.tuat.ac.jp, <sup>2</sup>nakagawa@cc.tuat.ac.jp*

**Abstract.** This paper describes a method for separating on-line handwritten patterns into characters, figures, and formulas. Today, Tablet PCs and electronic whiteboards provide much larger writing area for pen interfaces unlike PDAs, so that users can easily input text, write mathematical formulas and draw figures on the screen. The fact that people can write these objects by a single pen (marker) without switching the device, mode, software or whatever else and without any writing restriction such as grids or boxes is one of the most important benefits of the pen interfaces but it requires difficult task of segmenting these objects. We apply a probabilistic model for this problem and employ stroke features, stroke crossings and stroke densities. Moreover, we partially apply the approach of segmentation by recognition. Although the current recognizer for formulas is not a true recognizer, we have achieved about 81% correct segmentation for all the strokes in our newly prepared database of the mixed patterns from these objects.

### **1. Introduction**

The method of writing characters into character writing boxes one by one has been practiced on PDAs as a practical way to input Japanese text without using a keyboard. This UI dispenses with the problem to segment handwriting into each character and people have been used to writing characters into boxes already in manuscript papers so that the writing characters into boxes has been a good compromise for machines and people in Japan.

As for tablet PCs and electronic whiteboards that have much larger writing surfaces, however, people tend to write text freely. Furthermore, they may also input figures, tables and formulas and hope that their handwriting can be automatically stored, searched, recognized and so on. For these needs, a method to separate these objects and to process them is essential. The opposite approach would be to require people to switch the device, mode, software or whatever else for every object, but this definitely kills the benefit of pen-interfaces. Therefore, we attempt to solve the problem of separating the mixed handwritten patterns into correct objects. After the users have written something without being bothered by operations, they would be helpful to correct mis-segmentations if not so many.

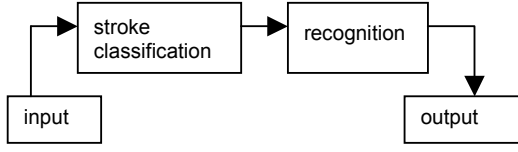
For printed document recognition, the problem has been studied for more than two decades (Lloyd, 1988). Recently, methods of separating text into several different kinds of languages have been also studied (Pal, 1999 & 2001). However this problem is rather new for handwriting recognition. There are not so clear differences between spatial frequencies among handwritten text, tables, figures and formulas, compared with printed documents. For off-line handwriting recognition, several papers have been published since there has been the demand for handwritten paper forms. For on-line handwritten recognition, however, it was not strongly recognized until recently. Nevertheless, some researchers have also started to attack this difficult problem. For western characters, stroke curvature and length are shown to be effective to distinguish text and figures (Jain, 2001). However, few reports were made to distinguish text and formulas.

We have been working on handwriting recognition and our methods have been employed in commercial on-line character recognizers both within writing boxes and without boxes (Fukushima, 2000). This paper describes a further extension to process handwriting mixed of text, figures and formulas. Tables are left for the moment but we think that they could be included without intrinsic extensions.

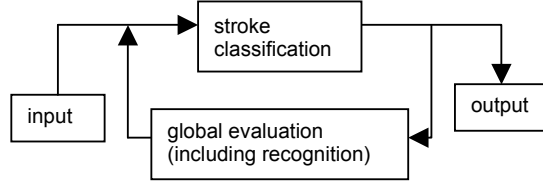
This paper describes pattern separation by stroke classification following a probabilistic model, considers several useful features for this task, and partially applies the approach of "segmentation by recognition". This paper also presents a preliminary evaluation using a database of mixed patterns.

### **2. Basic scheme of segmentation**

It is difficult to judge whether each handwritten stroke (a trace of the pen tip from pen-down to pen-up) is a part of text, formula and figure without considering its neighboring strokes. Our method is to make a soft decision for stroke classification based on each stroke's feature (local evaluation) before applying global evaluation that is consistency evaluation of each stroke with neighboring strokes or their recognition. The global evaluation modifies the initial result in local evaluation. Although the recognition is generally applied after segmentation (Figure 1; a



**Figure 1:** Linear model of segmentation and recognition.



**Figure 2:** "Segmentation by recognition" model.

linear model), our method applies "segmentation by recognition", model based on the global evaluation as shown in Figure 2. We follow this scheme and consider the local and global evaluation in more details in Chapter 4.

### 3. Probabilistic model

We apply a probabilistic model for separating mixed handwritten patterns into each class. Each stroke is evaluated and the similarity scores that should be approximations of probabilities for being member of each object type are estimated. Finally, the similarity scores obtained by the all evaluation methods are averaged by object type and every stroke is classified into an object type whose average score is highest.

Now, we describe how to calculate the similarity scores from a histogram. Let us consider a histogram of the stroke size as an example as shown in Figure 3 where we define the stroke size as the diagonal length of the bounding box of a stroke (minimum size rectangle that covers it) as shown in Figure 4. Consider the case that the stroke size is 30.

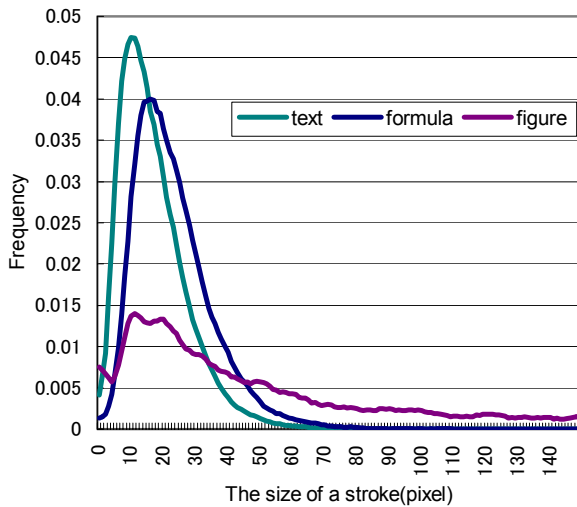
The similarity scores  $S_t$ (text),  $S_{fo}$ (formula), and  $S_{fi}$ (figure) are calculated from the frequencies  $F_t$ (text),  $F_{fo}$ (formula), and  $F_{fi}$ (figure) as shown in Figure 5 according to eq. (1).

$$S_t = \frac{F_t}{F_t + F_{fo} + F_{fi}}, \quad S_{fo} = \frac{F_{fo}}{F_t + F_{fo} + F_{fi}}, \quad S_{fi} = \frac{F_{fi}}{F_t + F_{fo} + F_{fi}} \quad (1)$$

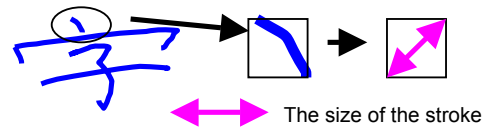
## 4. Stroke classification using local and global evaluation

### 4.1 Soft decision by local evaluation

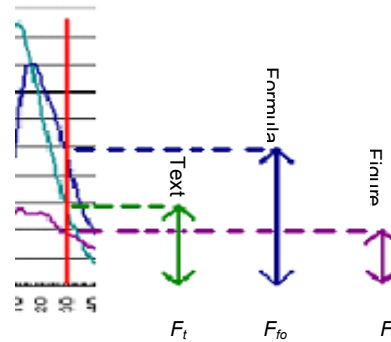
First, the size of each stroke is evaluated as a local evaluation. The histogram of the stroke size has already been shown in Figure 3. Although there are some other definitions as a stroke size such as a stroke length, a longer side's length of the bounding box etc., we employ the diagonal length of the bounding box. There are strong correlations between these values (the correlation coefficients are more than 0.9). While strokes for both text and formula are mostly distributed in small stroke sizes, those for figures are widely distributed in their sizes. Therefore, this evaluation is effective in distinguishing typical large sized figure strokes from others.



**Figure 3:** A histogram of the stroke size.



**Figure 4:** Size of a stroke.



**Figure 5:** Frequencies when the stroke size is 30.

## 4.2 Global evaluation

When two or more strokes cross each other, they are likely a part of the same object. This is useful to classify figure strokes. Assume that a stroke  $A$  crosses another stroke  $B$ , the similarity score of  $A$  as figure in local evaluation, i.e.,  $A_{fi}$  is less than that of  $B$  i.e.,  $B_{fi}$  and  $B_{fi}$  is higher than the threshold (it is set 0.5 at this time), then every similarity score of  $A$  is updated by those of  $B$ , with the effect of updating the score of  $A$  as figure to the same as  $B$ . This method is realized from the statistical result that the combination of figure strokes has the higher probability of crossing than that of a figure and a non-figure (Machii, 1993).

Other global evaluations mainly distinguish text and formula, and they are performed per line to the strokes whose similarity score of figure is not highest among its three scores. We use a length of off-stroke (a distance from a last point of previous stroke to a first point of the stroke) in order to detect the line. Then, the stroke density, the formula recognition score and the text recognition score are evaluated. The similarity scores calculated per line are applied to all strokes included in the line. Since Japanese text hardly consists of the characters for formula, these evaluations seem to be effective in Japanese text.

The stroke density is defined as a ratio of the number of strokes contained in a line to the size of the line. The size of a line is measured by the diagonal length of the bounding box for the line as shown in Figure 6. The evaluation is performed by the histogram method using the histogram shown in Figure 7.

Formula recognition score is used to evaluate a line as a formula. At this moment, we employ a very simple method to recognize a limited class of formula (elementary functions). It produces recognition scores from 0 to 1023 for a line. A histogram of formula recognition score is shown in Figure 8 so that we can obtain the similarity scores by applying the histogram method.

We also apply text recognition to a line of strokes as well as formula recognition and obtain the ratio to evaluate a line of strokes as text or formula as shown in Figure 9. The similarity scores are calculated according to eq. (2). Since text recognition produces the most probable segmentation and recognition of characters in a line and output recognition scores to every character, it takes the average of all the recognition scores and assigns the average to the recognition score of the line as text as shown in Figure 10. Evaluation by this method is more direct application of the principle of “segmentation by recognition”.

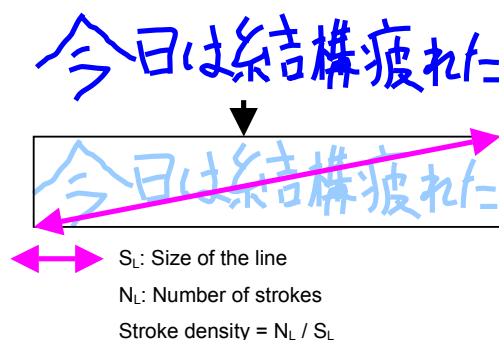


Figure 6: Stroke density within a line.

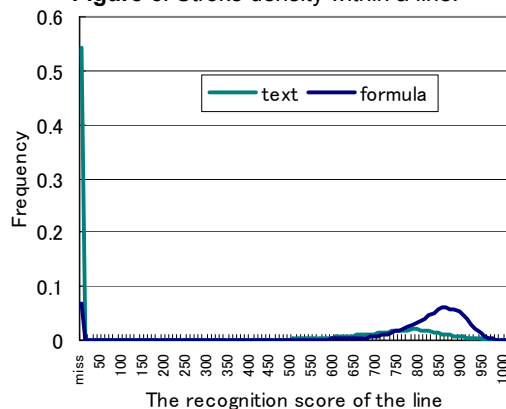


Figure 8: A histogram of formula recognition score.

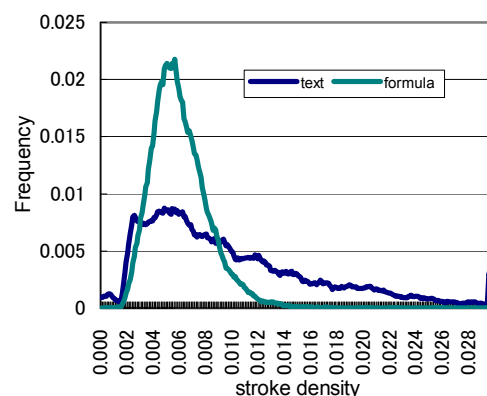


Figure 7: A histogram of the line stroke density.

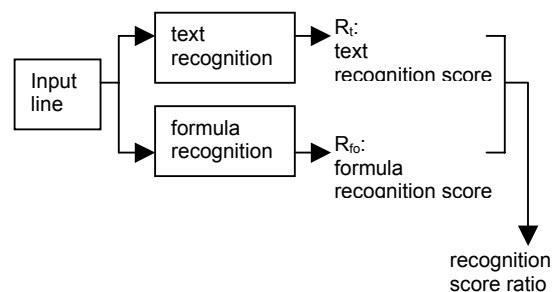


Figure 9: Model of the recognition score ratio.

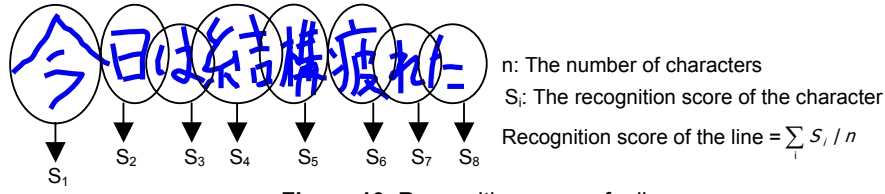


Figure 10: Recognition score of a line.

$$S_t = \frac{R_t}{R_t + R_{fo}}, \quad S_{fo} = \frac{R_{fo}}{R_t + R_{fo}} \quad (2)$$

## 5. Experimental Results

We experimented the method described above. We used 51 people's data stored in our database "HANDS-Kondate\_t\_bf-2001-11" for training and 16 people's data for evaluation. The database stores on-line handwriting patterns of mixed objects written without any writing box or grid and without any writing constraints. The number of total strokes tested for evaluation is 39,081 (25,941 text strokes, 8,221 formula strokes, and 4,919 figure strokes).

Separation accuracy attained was about 81% (text 87.5%, formula 78.6% and figure 77.6%). On the other hand, about 79% (text 84.3%, formula 77.9% and figure 76.1%) was attained for the same test data when formula and text recognition were not employed. Therefore, the increase of 2% is due to the scheme of "segmentation by recognition".

Most of separation mistakes are the cases where text strokes are mis-classified into formula. About 27% of all the strokes classified into formula were text strokes. Otherwise, it is about 33% when formula and text recognition were not employed. This result means usefulness of "segmentation by recognition" for classifying text and formulas.

Classification using a histogram of stroke sizes is too deterministic when text or formulas are written in a large size or figures are written by many short strokes. Classification using a histogram of stroke density also caused errors for a text line of "hiragana" or "katakana" characters (phonetic characters of simple shapes) to be mis-classified as formulas because their stroke density is sparse and similar to that of formulas. Therefore, histogram method must be empowered and improved.

## 6. Future views

Although the employment of text and formula recognition was proved effective to improve the classification of text, formula and figures, their recognizers, especially the formula recognizer must be improved. Moreover, some methods to evaluate the likelihood of figures seem to be useful, but the domain of figures is not well-defined (unlike characters, we don't know what can be written) so that we will need other types of engines rather than recognizers. At the moment, we are not considering the cases when formulas appear in a line of text. In order to cope with these cases and also to improve the classification, global evaluation including recognition must be empowered and improved.

This research is being supported by the Strategic Information and Communications R&D Promotion Scheme under the Ministry of Public Management, Home Affairs, Posts and Telecommunications.

## References

- Lloyd, A. F., & Rangachar, K. (1988). A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, Vol. 10, No. 6 (pp. 910-918).
- Jain, A. K., & Nambodiri, A. M. (2001). Structure in On-line Documents. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*. (pp. 844-848). IEEE Computer Society.
- Pal, U., & Chaudhuri, B. B. (2001). Automatic Identification of English, Chinese, Arabic, Devnagari and Bangla Script Line. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, (pp. 790-794).
- Pal, U., & Chaudhuri, B. B. (1999). Script line separation from Indian multi-script document. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, (pp. 406-409). IEEE Computer Society.
- Machii, K., Fukushima, H., & Nakagawa, M. (1993). On-line Text/Drawings Segmentation of Handwritten Patterns. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, (pp. 710-713). Tsukuba Science City: IEEE Computer Society.
- Fukushima, T., & Nakagawa, M. (2000). On-line Writing-box-free Recognition of Handwritten Japanese Text Considering Character Size Variations. In *Proceedings of the 15th International Conference on Pattern Recognition*, Vol. 2, (pp.359-363). Barcelona: IEEE Computer Society.