MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

A Stroke Index For Document Image Analysis Based On The MCR Expression Method

AbdelMalek B.C. ZIDOURI, Supoj CHINVEERAPHAN, and Makoto SATO
Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta-cho, Midori-ku, Yokohama, JAPAN 227
Tel.(045) 922-1111 ext.2050 E-mail: malek@pi.titech.ac.jp

Abstract

In this paper we introduce a new feature called stroke index for document image analysis. It is based on the minimum covering run expression method (MCR). This stroke index S is a function of the number of horizontal and vertical runs in the original image and of number of runs by the MCR expression. As document images may present a variety of patterns such as graph, text, picture or dithered image, it is necessary for image understanding to classify these different patterns into categories beforehand. This index gives an insight on the possibility of stroke extraction from document images and classifies different patterns in a compound image.

1 *Introduction*

One of the current trends in document image processing such as image compression, database management, image analysis and understanding is to consider these tasks through an unified approach. The MCR expression method was developed in this scope. Using information from the MCR expression for classifying various types of regions according to the possibility of strokes extraction, we introduce a new feature called stroke index for document image analysis. As document images may present a variety of patterns such as graph, text, picture or dithered image, it is necessary to classify these different patterns into categories for image analysis and understanding. The proposed index aims at the classification of these different image patterns.

2 *The MCR Expression*

The MCR expression method has been developed to express binary document images by a minimum number of runs both in the horizontal and vertical directions rather than being expressed by either the horizontal or the vertical run representation. Fig. 1 illustrates this for a simple binary image: it needs 14, 12, and 10 runs to represent it by horizontal runs, vertical runs, and MCR expression method respectively.

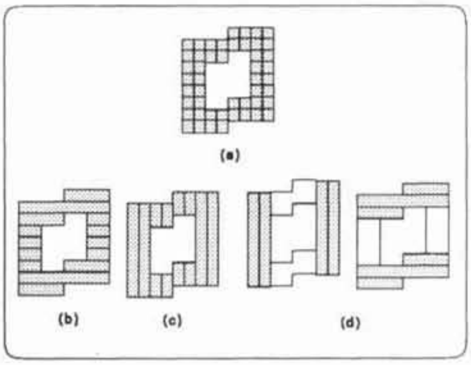


Figure 1: An illustration of (a) a simple binary image (b) its horizontal run representation (c) its vertical run representation and (d) its MCR expression .

The MCR expression is based on maximum matching in a bipartite graph. A finite and undirected graph G = (V,E) is said to be bipartite if its vertex set V can be partitioned into disjoint subsets X and Y, called partite sets, such that every edge in E joins a vertex in X with a vertex in Y. It has been shown that horizontal and vertical runs of binary image can be thought of as partite sets of a bipartite graph. From this correspondence between the binary image and the bipartite graph, where runs correspond to partite sets and edges of the graph correspond to pixels in the image, finding the MCR expression amounts to solving the problem of maximum matching or minimum covering in bipartite graph .

The maximum matching is a subset M of edge set E which has the largest number of edges such that no two edges are adjacent. The image is scanned line by line then representative horizontal and vertical runs are registered. Generally, information in document images such as characters or lines is composed of horizontal and vertical strokes. These can be represented by a minimum number of covering runs by the MCR expression method.

3 *Stroke Index*

3.1 *Definition*

Based on the MCR expression method we propose a stroke index which is a function of the number of horizontal and vertical runs in the original image and of number of runs by the MCR expression. The vertical stroke index S_V and the horizontal stroke index S_H, are given by:

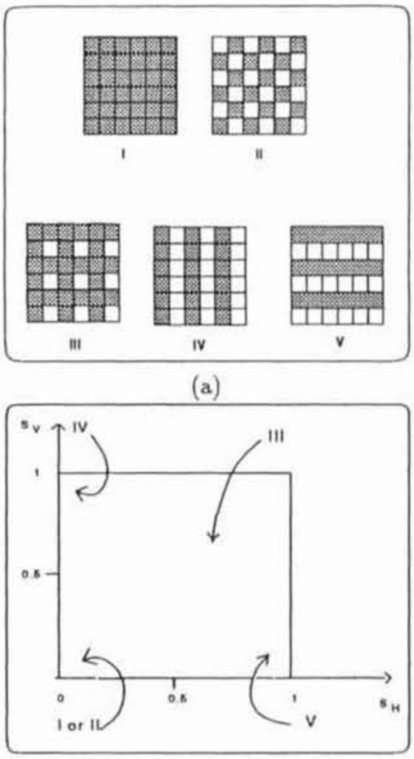
$$S_V = (n_r - n^*)/n_r,$$

$$S_H = (n_c - n^*)/n_c,$$

where
n_r is the number of horizontal runs, n_c is the number of vertical runs, n^* is the number of runs by the MCR expression. This is adopted because in case of an image with a large number of horizontal strokes, the number of horizontal runs (and covering runs by MCR) is small compared to the large number of vertical runs necessary to cover the image. Therefore the value of S_H tends to one, and that of S_V tends to zero, and viceversa for a vertical strokes pattern.

3.2 *Basic Properties*

To illustrate mapping of different document image patterns into the index plane (S_H, S_V) we consider five typical patterns as shown in Fig.2. and show the different parameter values and values of stroke index. These are summarized in Table1.
For a dithered or black image (I or II), as there are no strokes in these two, they will map into the region at the origin (0,0) on the index plane, a vertical strokes image (IV) will map into region near (0,1), a horizontal strokes image (V) will map into region near (1,0), and a type (III) composed of both horizontal and vertical strokes maps as shown in Fig. 2.



(a)

(b)

Figure 2: Typical Image Patterns and their mapping on the index plane

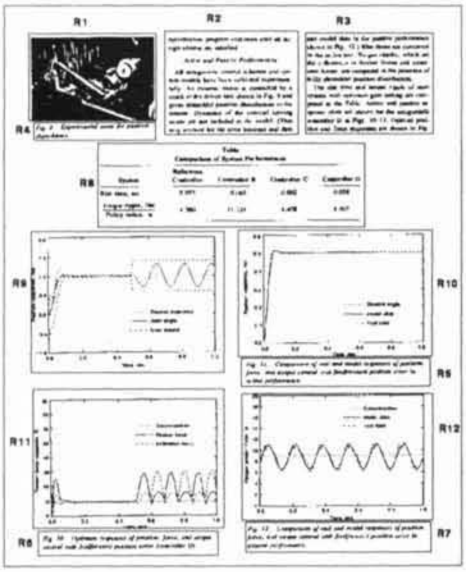
	I	II	III	IV	V
n_r	6	18	12	18	3
n_c	6	18	12	3	18
n^*	6	18	6	3	3
S_H	0.0	0.0	0.5	0.0	0.8
S_V	0.0	0.0	0.5	0.8	0.0

Table 1.

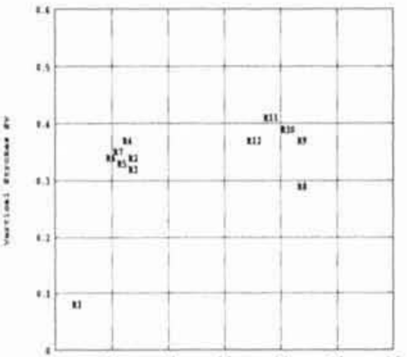
4 *Experimental Results*

As document images are in general a mixture of the above typical patterns, the mapping of any pattern on the index plane (S_H, S_V) should be according to the above result. To show this we take an example of document as shown in Fig.3(a). We first segment the image into 12 blocks, R1 through R12. R1 is a picture, R2 up to R7 is text blocks, R8 is a table with only horizontal lines and text, and R9 through R12 are blocks of graph type. We then computed the values of the stroke index for each block and the result is mapped on the index plane as shown in Fig.3 (b). We see that blocks such as R2 up to R7 mapped on the same region representing a similar pattern (which is text in this case).

Blocks R8 through R12 mapped in another region representing another similar pattern (which is of graphic type) with higher coordinates of stroke index. We can also notice that block R8 has a high horizontal stroke value and a vertical value less than the average for text. This is because it is a table and has lines only in the horizontal direction and its text has larger font.



(a)



(b)

Figure 3: (a) An Example of Test Document and its Regions R1: picture, R2-R7: text, R8-R12: graph, (b) Mapping of regions on index plane

5 *Application: Block Classification*

One of applications of this stroke index would be the classification of different patterns on a document image. We have scanned some document images containing graphs, a picture and different type of text, see Fig.5 . The compound documents have been divided into different blocks to compute the stroke index values for each block. Mapping values of the stroke index for different blocks of the documents on the index plane classifies the various types of patterns into distinct regions as shown in Fig.4.

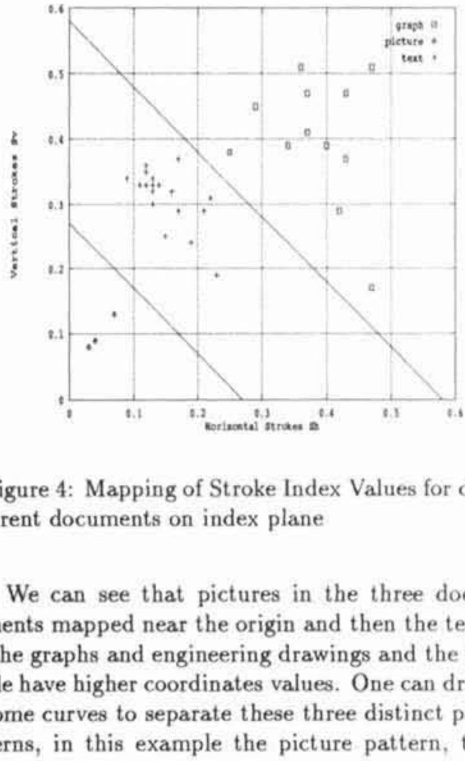


Figure 4: Mapping of Stroke Index Values for different documents on index plane

We can see that pictures in the three documents mapped near the origin and then the text. The graphs and engineering drawings and the table have higher coordinates values. One can draw some curves to separate these three distinct patterns, in this example the picture pattern, the graphs and drawings pattern, and text could be separated into regions. The result of this classification is given in Fig.5. We will note that the graph type pattern is the most scattered on a wide area as expected because of the different shapes that it has. For exemple we can recognize the value of the graph of document “c” which is composed mainly of horizontal discontinued lines as being the point with the lowest vertical stroke component in the graph area. We can also tell from the value of S_H and S_V whether the strokes in the image are mainly horizontal or vertical strokes.

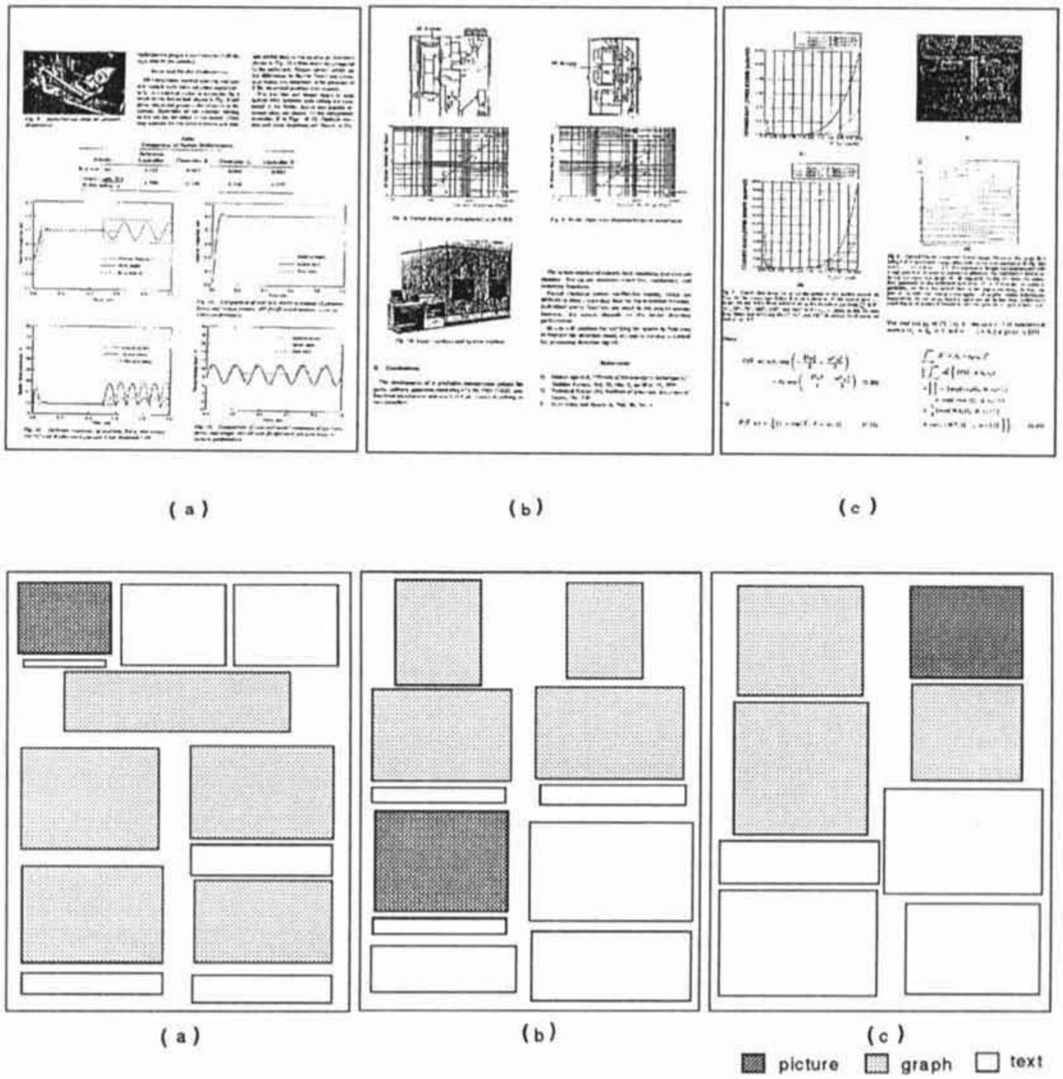


Figure 5: Block Classification Result for Test Documents a, b and c

6 *Conclusion*

As document images are composed of various types of regions; such as text, graph, or picture it is necessary to classify these patterns for image processing tasks. The capability of the strokes to be used as a basic part in representation of characters and lines makes it important to know the possibility of strokes extraction from document images. In this paper we have proposed a new feature called stroke index for document image analysis. This has been applied to typical binary image patterns as well as compound document images scanned locally and the results proved to be interesting. We have shown that this new feature could be readily used for classification of various patterns in a document image. We can tell from the value of this stroke index whether the possibility of strokes extraction from a document image is high or not. The experimental results are included.

References:

1. K. Douniwa S. Chinveeraphan M. Sato *Minimum Covering Run Expression of Document Images Based on Matching of Bipartite Graph.* First Korea-Japan Joint Conf. On Computer Vision. Oct. 10-11, 1991.

2. A.N. Jain: *Fundamentals of Digital Image Processing* Prentice-Hall, Inc. (1989).