TRADEMARK IDENTIFICATION THROUGH TEXT RECOGNITION

GHAZALI SULONG
ZAIDAH IBRAHIM
Faculty Of Computer Science & Information System
University Technology Malaysia
Skudai, 80990 Johor Bahru

ABSTRACT

Trademark identification is of great interest in the document domain due to two reasons. Given a document that consists of a trademark, the trademark needs to be conclude whether present or not in the database and given a known trademark, we need to index into a database of documents and extract all information related to it. This paper proposes a new method to recognise the text on mixed-mode trademarks that consists of various font style and sizes, focusing on uppercase Roman alphabets. A stroke-based feature extraction method represented by chain codes is used. A modified rule-based classifier recognises the characters with satisfactory results.

KEYWORDS : Chain Codes, Strokes, Character Recognition, Trademarks

ABSTRAK

Terdapat dua sebab mengapa topik mengenal pasti logo diminati di bidang dokumen. Pertama, dengan adanya sebuah dokumen yang mengandungi logo, logo tersebut perlu dikenal pasti samada terkandung di dalam pangkalan data atau tidak. Kedua, dengan adanya logo yang telah dikenal pasti, kita perlu mencari di dalam pangkalan data dokumen dan mencapai semua maklumat yang berkaitan dengan logo tersebut. Kertas kerja ini membincangkan satu kaedah baru untuk mengecam aksara yang beranika fon dan saiz yang terdapat pada logo, khusus untuk aksara berhuruf besar. Kaedah berdasarkan pada strok

digunakan sebagai cara untuk pengambilan ciri-ciri menerusi penggunaan kod rantai. Teknik pengaturan yang telah diubah suai telah digunakan sebagai kaedah untuk mengecam aksara dan keputusan ujikaji amatlah memberangsangkan.

1.0 INTRODUCTION

Generally, trademarks are characterised as mixed text and graphic symbols and they are associated with a given group or organisation. They represent a class of images that may be useful for businessmen, graphic designers or mass communication operators. In the document domain, trademark identification or recognition tasks are of great interest due to two reasons. Firstly, given a document that consists of a trademark, the trademark needs to be concluded whether it is present or not in the database. It is possible that new trademarks may closely resemble existing ones. Secondly, given a known trademark, we need to index into a database and extract all information related to the trademark or documents that contain that trademark. This is an important issue as we may face with the problem of enormous difficulties in material retrieving and automatic archiving. Text recognition can be one of the ways for recognising the trademarks and used as index to databases.

There are many commercial devices capable of accurately reading clean machine printed documents. The first commercial devices were developed in the late 1960s. They were specialised machines of very high cost designed to handle large amounts of documents. Matrix matching techniques have been the preferred approach since they are easier to implement. The unknown image needs to be compared with a set of templates representing known characters. However, it is not well adapted to size variations and font styles.

[Shlien 1988] developed a feature-based recognition algorithm that uses a decision tree classifier with 197 binary features which flag the presence or absence of linear, convex, concave and hole features. [Muelenaere et. al. 1992] developed a character recognition technique using topological features where the character is described in terms of bars and

holes, and the approximate positions of the bars and holes using 21 feature vectors. [Shih et. Al. 1992] developed a rule-based classifier which is based on stroke detection.

This paper proposed a character recognition technique that is based on a combination methods developed by [Muelenaere et. al. 1992] and [Shih et. al. 1992]. This proposed technique reduces the number of feature vectors used to only 9.

2.0 FEATURE EXTRACTION

Trademarks consists of a combination of text and graphic symbols. And usually there is only one graphic symbol and one word of text involved, and no interflow between them. The work presented here is designed specifically for black (represented by 1's) and white (represented by '0') trademarks that are within closed borders. The trademark is digitised by a scanner and thresholded into binary image. Then, the binary image is smeared horizontally and vertically to close up any broken borders. The borders are deleted by changing the white pixels to black. Since the characters are nicely separated, character segmentation is based on the occurrence of white columns or columns of 0's. More discussions on character segmentation can be found in [Lu, Yi 1995].

One of the most important aspects of any character recognition system is feature extraction. Its purpose is to extract features from the character that would simplify the recognition process. Basically, a character image consists of several lines or curve segments (also called as strokes) that are drawn and adjoined at specific position which differentiates a character from other characters [Shirvaikar et. al. 1988].

The feature extraction method proposed here is divided into two levels of analysis where in the first level, a global analysis is performed and if ambiguity occurs, the second level is invoked where a magnified analysis is performed. In global analysis, the strokes are

analysed generally for its straightness where a slight curve or slant is ignored. But, in magnify analysis, the details of the strokes are analysed where a slight curve or slant is detected in determining the results.

In [Muelenaere et. Al. 1992], the binary image is scanned in 4 directions which is, left to right, right to left, top to bottom and bottom to top to obtain the character projections, outlines and stroke densities. Calculations are performed to produce the number of bars and holes that present in the image. The image in this research is also scanned in 4 directions to obtain 4 strokes, referred as Left, Right, Top and Bottom, represented as chain codes [Wilf 1981].

There are 8 possible directions between a point and its neighbour. Figure 1(a) shows the 8 direction of chain codes and figure 1(b) shows an example of an image and its chain codes

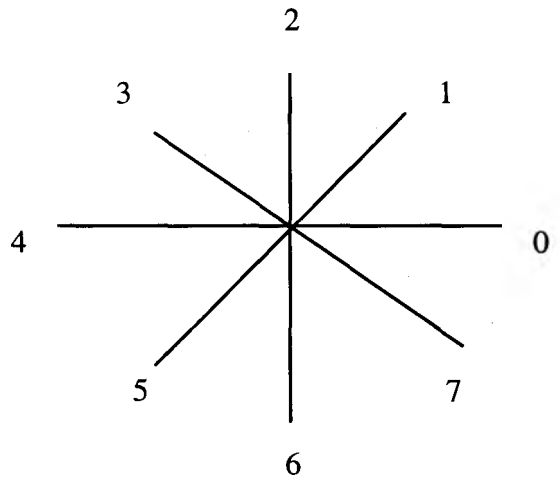

Fig. 1(a) 8 direction of chain codes

Left chain codes
6565665656656566565665656566565665656656565655654544
Right chain codes
6766767676676767667676767667676766767676767770706
Top chain codes
00
Bottom chain codes
000000000000134432222212122121100000000000000000007767667676666554457000000000000000000

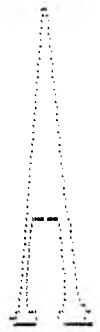

Fig. 1(b) A sample of character A and its four strokes represented as chain codes.

2.1 STRAIGHT LINE DETECTION

Various algorithms have been developed to determine the straightness of chain codes [Freeman 74, We 82, Hung 85]. Yuan et. al. [Yuan et. al. 1995] have developed an algorithm to detect straight lines based on the quantization scheme. In this scheme, a passing area is constructed around pixels so that a line formed by these pixels must pass through this area if it is straight. The passing area is modified as a new point is accepted until a point is rejected as a pixel on the straight line.

However, before the straight line detection process can be applied, the chain codes needs to be cleaned or smoothed to delete any noises that may occur [Wilf 1981]. Smoothing can be time consuming. Thus, to overcome this problem, the straight line detection algorithm can be modified to cater this noise where a small inclination to the next direction other than specified by Yuan would also be considered as part of the same chain code of a straight line. This can be performed by using certain rules where the codes of the same

straight line are grouped together. For instance, a chain codes that consists of values 1, 0 or 2 belong to the same group, values 5, 6 or 7 belong to another group, and so on. By applying this rule, the smoothing process can be discarded.

2.2 RULE-BASED CLASSIFIER

[Shih et. Al. 1992] developed a rule-based classifier where the character image is compared with a character recogniser. Each character recogniser consists of a group of characters that have similar topological structures. Once a rule is matched, the input character is recognised. Otherwise, the next character recogniser will be tested.

This paper proposes a modified rule-based classifier where the differences in characters among the different fonts are detected and deleted. For instance, the bottom of the left stroke of font Times New Roman and Courier consists of a curve to the left. Thus, the curve is detected and then ignored before straight line detection is performed.

The next step is to check whether the input image satisfies the common properties of the group where the characters are divided into two groups based on the hole feature as follows:

Group 1 - A, B, P, R, Q, D, O

Group 2 - C, E, F, G, H, I, J, K, L, M, N, S, T, U, V, W, X, Y, Z

The last step is to match the input image against each recognizer in the group. For global analysis, each stroke is tested for the number of straight lines. A line with small curvature or a perpendicular line is also considered as a straight line. For Group 1, ambiguity occurs for characters D and O, and to differentiate between them, the left stroke is analysed further under the magnify analysis level. Its top left stroke is tested for the occurrence of a curve as character O has that feature but not character D.

For Group 2, two cases of ambiguity occurs and they are further divided into two subgroups, that are, Z and T, and U, V and Y. Under the magnify analysis level, the left chain codes for characters Z and T are looked at. Character T has a vertical line but not character Z. The depth of the top stroke of character U and V is always deeper than the depth of the top stroke of character Y. And to differentiate between characters U and V, a vertical line is detected on the top stroke.

3.0 EXPERIMENTAL RESULTS

Trademarks that consists of machine printed character sets of the Arial, Times New Roman and Courier fonts served as test sets for the experiments. 20 different trademarks have been tested with 122 characters from the three different fonts. The recognition rate obtained was 96%. Errors occur in recognising a few samples of characters G and S. This is because the difference occur in the middle of the stroke and the algorithm could not detect it. See fig. 3 for some samples of trademarks where the text have been successfully recognised. Character G in this sample is recognisable by the algorithm.

4.0 CONCLUSION

The proposed algorithm for character recognition described in this paper has produce satisfactory results. This technique able to cater only 9 feature vectors, that are, horizontal and vertical lines on all four strokes, plus the hole feature. Reducing the number of features greatly reduces the memory size needed by the software. Nonetheless, more research is being done to make this algorithm more flexible and improve the recognition rate.

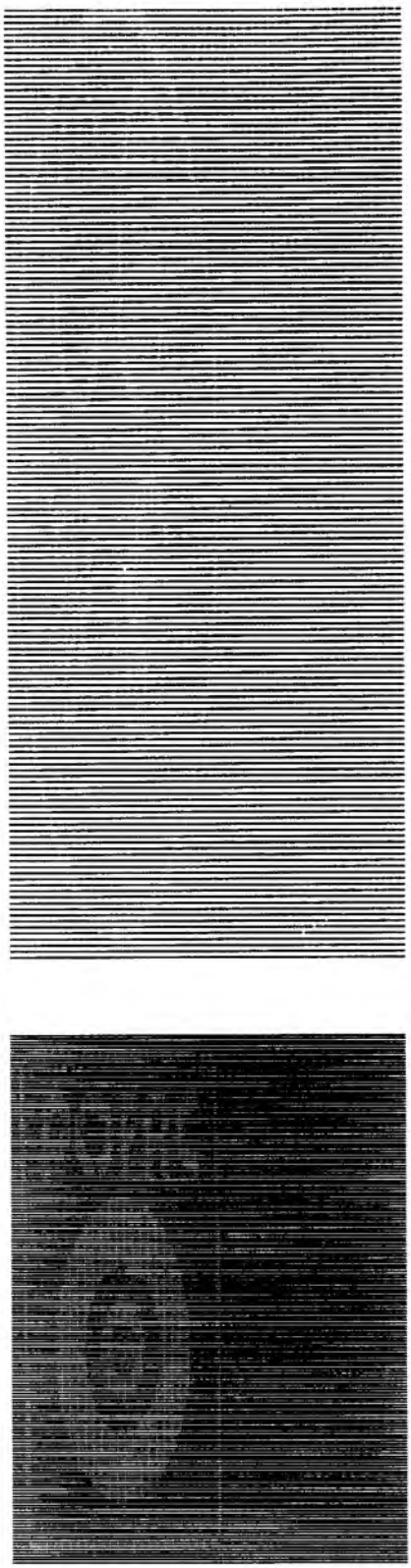

Fig. 3 Some samples of trademarks whose text have been correctly recognised.

5.0 REFERENCES

1. Bai, G. "Multifont Chinese Character Recognition Using Side-Stroke-End-Feature", Proc. IEEE, 1993, pp. 794-797.
2. Hung, S.H.Y. "On The straightness Of Digital Arcs", IEEE Trans. PAMI, vol. PAMI-7, No. 2, March 1985, pp. 203-215.
3. Freeman, Herbert. "Computer Processing Of Line Drawing Images", Computer Surveys, Vol. 6, No. 1, March 1974, pp. 57-97.
4. Muelenaere, P. De; Dauw, M.; and Legat, J.D. "Omnifont Recognition Of Text Using Topological Recognition Techniques", Proc. IEEE, 1992, pp. 410-413.
5. Shih, F.; Chen, S.; Hung, D.; and Ng. P. "A Document Segmentation, Classification And Recognition system", Proc. IEEE, 1992, pp. 258-267.
6. Shlien, S. "Multifont Character Recognition For TypeSet Documents", Int'l. Journal Of Pattern Recognition & Artificial Intelligence, Vol. 2, No. 1, March 1988, pp. 603-620.
7. Shirvaikar, M.V. and Musavi, M.T. "A Stroke Feature Distribution Scheme For The Recognition Of Alphanumeric Characters", Proc. IEEE, 1988, pp. 192-195.
8. We, Le-De, "On The Chain Code Of A Line", IEEE Trans. PAMI, vol. PAMI-4, No. 3, May 1982, pp. 347-353.
9. Wilf, J.M. "Chain Codes", Robotics Age, Vol. 3, No. 2, Mar/Apr 1981, pp. 12-19.
10. Yuan, J. and Suen, C. "An Optimal O(n) Algorithm For Identifying Line Segments From A Sequence Of Chain Codes", Pattern Recognition, Vol. 28, No. 5, 1995, pp. 635-646.