CLASSIFICATION OF TABLE FORM DOCUMENTS USING HIGH ORDER CORRELATION METHOD

Ren-Jean Liou, Brian Shu and Mu-Song Chen

Department of Electrical Engineering
Da-Yeh University
Changhua, Taiwan 515, ROC
e-mail: renjean@mail.dyu.edu.tw

Abstract

The recognition of table form documents is useful in office automation and file management. This paper presents a new approach for automatic document classification using high order correlation (HOC) method. HOC was originally used to recursively compute the cross-correlations between consecutive data in order to extract moving target tracks in three-dimensional (3-D) space. The most similar application in 2-D space is curve detection. A table form document consists of lines, characters and sometime graphs. It would be very convenient to use HOC to perform segmentation and classification on this type of images. The results contribute to many applications such as document identification and optical character recognition (OCR). It was shown that HOC could be implemented using a neural-network type of structure. This will greatly improve the efficiency of computation. The effectiveness of this approach will be demonstrated in the simulation results.

Keywords: High order correlations, Pattern recognition, Document analysis, Office automation, Neural networks.

1. Introduction

As the development of new image processing and pattern recognition techniques, the issues office automation and machine intelligence have been continuously improved in the aspects of cost, speed and quantity. Table form documents are enormously used in today's offices and factories. The required processing time and storage space are increased considerably. An automated processing system would be very helpful in preventing over flood of such tasks.

The research on classification of table-form documents [10]-[17] has become popular in recent years. The first step to this problem would be identification and classification. The content of the document can then be described. Nevertheless, there is still a lot of room for improvement in accuracy and speed. Most document processing systems perform image enhancement first. A preprocessing is then applied to separated texts, graphs and tables. At this stage for segmentation, various techniques are utilized for edge and line detection. The results are then compared to the database for document identification and classification. There are two major difficulties in performing such tasks. The first is that the content and types of documents are very complicated and diverse. The second is that most of the time the actual inputs are deviated from its ideal case. For example, variation in position and angle often occur due to operation. Human and machine errors also create interference for processing.

Based on human instinct, the number of lines, length and geometric arrangement are important features to investigate a document. However, many conventional approaches do not take full advantages of these features. The goal of this paper is to use a new approach named high order correlation (HOC) method for identification and segmentation of table form documents. The correlation property of lines and curves in a document will be fully utilized. More accurate and efficient results can thus be achieved.

HOC [1]-[2] was originally proposed for detect-

ing point target tracks in 3-D space. Owing to the extremely low signal-to-noise ratio (SNR) and lack of target and clutter *a priori* information, the problem is very complicated. Many conventional methods [3]-[6] were not able to overcome all the difficulties. HOC has successfully solved the problems. Its nature of detecting correlations can be easily modified into 2-D form for other applications [7]. The problem of detecting lines and curves in a document has great similarity to track detection. We will show in this paper the potential of using HOC in table form document recognition and classification.

This paper is organized as follows. Section 2 introduces HOC in 3-D and derives its formula in 2-D. The processes of HOC for table form document classification will be demonstrated in Section 3. Simulation results are presented in Section 4. Finally, Section 5 is conclusion.

2. The HOC Method

The original HOC is a 3-D image processing algorithm. The images are constructed by stacking series of 2-D images. The coordinates are defined as (x, y, t_n) where (x, y) represents the space variables and t_n is time. A moving target in the scene forms a track in this 3-D image. However, the targets are under highly cluttered background and are disturbed by serious sensor noise. Hence they are very difficult to be identified. Since there is no presumed knowledge pertaining to the intensities of targets and noise, the intensity does not carry much useful information in differentiating target and clutter. Therefore, the input image can be converted into binary form so that less assumption is made on this problem.

The high order correlation method continuously exploits the spatial and temporal dependencies of the points on a track. This is similar to line detection in a 2-D image. The computation is performed recursively on the spatio-temporal cross-correlations between images of consecutive scans. Since the shape of a real moving target track possesses some regularity, it enables the algorithm to discriminate real tracks from noisy ones with random nature. If the assumed maximum target velocity is v from on scan to the next, we shall check the correlation in all directions in order to find the dependency. The formula for finding this dependency was derived in [2], which gives

$$\begin{aligned} Y^{(k)}(x,y,t_n) &= Y^{(k-1)}(x,y,t_n) \\ &\quad g[\sum_{i_1=-|v|}^{|v|}\sum_{j_1=-|v|}^{|v|} Y^{(k-1)}(x+i_1, y+j_1, t_{n+1}) \\ &\quad \sum_{i_2=-|v|}^{|v|}\sum_{j_2=-|v|}^{|v|} Y^{(k-1)}(x+i_1+i_2, y+j_1+j_2, t_{n+2})] \end{aligned} \tag{1}$$

where $Y^{(k)}(x, y, t_n)$ indicates the processing results of position (x, y, t_n) at the kth order, k is the order of recursion. When k=0, $Y^{(0)}$ is set to the original image of sensor returns. Therefore, the operation can be continued recursively until k reaches a preset order. Also in equation (1), g is a hard limiter threshold function that is defined as

$$g(t) = \begin{cases} 1, & t > 0 \\ 0, & t \le 0 \end{cases} \tag{2}$$

The results are thus clearly a binary image as well. If $Y^{(k)}(x, y, t_n)$=1, it indicates that correlations exit from scan t_n to t_{2k+n}, which represents that there is a track starting at location (x, y, t_n) with length $2k$+1. The track can be in arbitrary shape as long as consistent correlation is given. In [2], a neural-network architecture is also proposed for real-time implementation.

In equation (1), we have applied velocity constraint by limiting the range of (i_1, j_1) and (i_2, j_2). However, heavy clutter noise may produce tracks with jagged shape or abrupt speed change that will also satisfy equation (1). These phenomena usually do not happen to a real target track. Therefore, we shall apply another constraint to remedy this problem. This can be achieved by refining the values of (i_2, j_2) so that the moving direction and speed changes are within a limited range [1].

It is convenient to modify the HOC for 2-D processing. The variable t_n is first removed and the movement is applied to either x or y, which respectively corresponds to performing the proc-

esses vertically or horizontally. For detecting the correlations vertically, we have

$$Y^{(k)}(x,y) = g[Y^{(k-1)}(x,y) \sum_{i_1=-v}^{v} Y^{(k-1)}(x+i_1, y+1) \sum_{i_2=-v}^{v} Y^{(k-1)}(x+i_1+i_2, y+2)] \quad (3)$$

where i_1 and i_2 are used to determine the shifting angle of the curve in the vertical direction. If the maximum shift v is equal to 1, this equation can be used for detection of curves with absolute slops more than 1. Since v is set from –1 to 1, a jagged line, which has opposite signs for i_1 and i_2, will also satisfy equation (3). To eliminate this situation, we can simply apply the condition such that $|i_2\text{-}i_1| \leq 1$.

The use of x, on the other hand, as the reference will calculate correlation in the horizontal direction. This can be expressed as

$$Y^{(k)}(x,y) = g[Y^{(k-1)}(x,y) \sum_{j_1=-v}^{v} Y^{(k-1)}(x+1, y+j_1) \sum_{j_2=-v}^{v} Y^{(k-1)}(x+2, y+j_1+j_2)] \quad (4)$$

where j_1 and j_2 are used to determine the angle of the curve in the horizontal direction. This equation can be used for detection of curves whose slops are between –1 and 1. If only straight lines are to be detected, we can set $i_1=i_2$ in equation (3) and $j_1=j_2$ and in equation (4). This will generate the same results as the traditional Hough transform but provide much simpler computation and implementation.

3. Document Processing Systems

The processing procedures proposed in this paper are shown in Fig. 1. The original table form document is input from a scanner using 300dpi grayscale resolution. The preprocessing step includes converting images into binary form and adjusting resolution. This can be done using some simple software. Thinning process [16]-[17] is then performed so that all lines are slimmed to one pixel in width. Not only the fullness of lines is preserved after this step, but also the computational complexity can be reduced. Higher efficiency and accuracy can then be achieved.

After the lines were slimmed, line detection and description is then performed using HOC that was described in equations (3) and (4). The results of HOC can provide information on the origin and length of each line. If enough order k is computed, the texts and graphs can be removed owing to their short length in nature. Therefore, document classification can be achieved by counting the number of lines and their corresponding location and length.

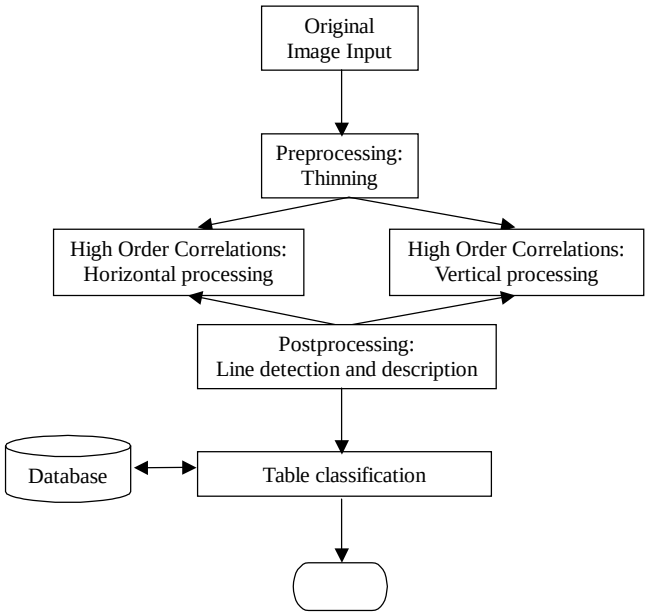


Fig. 1 The flow chart of the proposed document processing system.

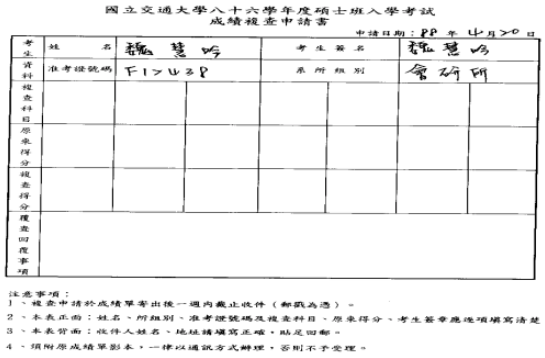
In regular scanning of documents, it is difficult to ensure that the paper is aligned. An alignment process [8]-[11] normally has to be applied first. This process is not needed in our approach. Though the length of each line in an image may be affected by angle of view and change of resolution, same number of lines and length can be obtained after a simple normalization process. This information can be obtained easily from the results of HOC. Accurate classification can still be obtained. This is an important feature called rotation invariance. Even the image input is rotated or shifted by some angle, the variation can be adapted by HOC. The rotated image can be

corrected by simple coordinate transformation.

4. Simulation Results

By following the above steps, we have successfully tested numerous samples and obtained excellent results. Fig. 2 is an original scanned image. It is easy to select a threshold to convert the image into binary form. The image after thinning process is shown in Fig. 3. We can see that the texts and table lines are still mixed together. They can be easily separated using HOC. Fig. 4 and Fig. 5 shows the vertical and horizontal processing results of HOC, respectively. HOC only preserved lines that are long enough. Texts are thus removed. When order k is used for processing, the length of line has to be longer than $2k+1$ in order to satisfy HOC criterion. Fig. 6 is the composite image of Fig. 4 and Fig. 5. We can clearly see the table lines of the original input document.

Table 1 lists the results of vertical processing in numbers. It clearly indicated the origins and lengths of all vertical lines. Using the longest line as base, Table 2 lists the normalized length of each line of Table 1 in order. This table has great importance for classification and the feature of rotation invariance. Similarly, the results of horizontal processing are listed in Tables 3 and 4.


國立交通大學八十六學年度碩士班入學考試
成績複查申請書
申請日期：PP 年 ㄐㄧㄡˋ 月 ㄕˊ 日

考生資料	姓名	魏慧吟	考生簽名	魏慧吟
	准考證號碼	F1ㄕㄨㄥP	系所組別	會研所
複查科目				
原來得分				
複查得分				
覆查回覆事項				

注意事項：
1、複查申請於成績單寄出後一週內截止收件（郵戳為憑）。
2、本表正面：姓名、所組別、准考證號碼及複查科目、原來得分、考生簽章應逐項填寫清楚。
3、本表背面：收件人姓名、地址請填寫正確，貼足回郵。
4、須附原成績單影本，一律以通訊方式辦理，否則不予受理。

Fig. 2 The original scanned image


國立交通大學八十六學年度碩士班入學考試
成績複查申請書
申請日期：PP 年 ㄐㄧㄡˋ 月 ㄕˊ 日

考生資料	姓名	魏慧吟	考生簽名	魏慧吟
	准考證號碼	F1ㄕㄨㄥP	系所組別	會研所
複查科目				
原來得分				
複查得分				
覆查回覆事項				

注意事項：
1、複查申請於成績單寄出後一週內截止收件（郵戳為憑）。
2、本表正面：姓名、所組別、准考證號碼及複查科目、原來得分、考生簽章應逐項填寫清楚。
3、本表背面：收件人姓名、地址請填寫正確，貼足回郵。
4、須附原成績單影本，一律以通訊方式辦理，否則不予受理。

Fig. 3 Preprocessing of Fig. 2.

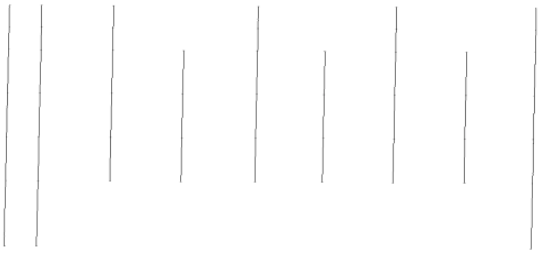


Fig. 4 Vertical processing of HOC

Fig. 5 Horizontal processing of HOC.

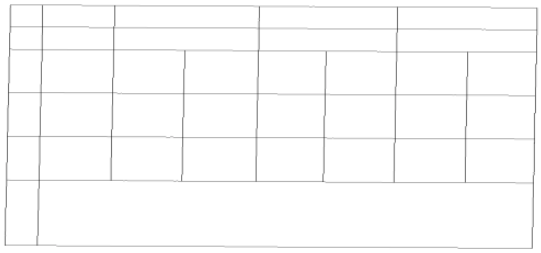


Fig. 6 Composite image of Fig. 4 and Fig. 5.

Table 1 The origins and lengths of vertical lines in pixels.

Line number	*X* coordinate	*Y* coordinate	Length
1	173	116	164
2	302	117	163
3	431	118	164
4	241	62	218
5	109	60	219
6	367	62	219
7	15	60	299
8	43	60	299
9	494	64	299

Table 2 Normalizing and sorting the results of Table 1.

0.545151	0.548495	0.548495	0.729097	0.732441
0.732441	1.000000	1.000000	1.000000	

Table 3 The origins and lengths of horizontal lines in pixels.

Line number	*X* coordinate	*Y* coordinate	Length
1	9	358	482
2	10	277	482
3	11	223	482
4	12	169	482
5	13	87	482
6	13	115	482
7	14	61	482

Table 4 Normalizing and sorting the results of Table 3.

1.000000	1.000000	1.000000	1.000000	1.000000
1.000000	1.000000			

In order to test the property of rotation invariance, Fig. 7 presents the same input document with variation in angle. The composite image of vertical and horizontal processing using HOC is shown in Fig. 8. The result was not affected by rotation. The upper part of Table 5 lists the normalized lengths of all lines after vertical processing. We can see no much variation in compare to Table 2. Their differences are listed in the lower part of Table 5.

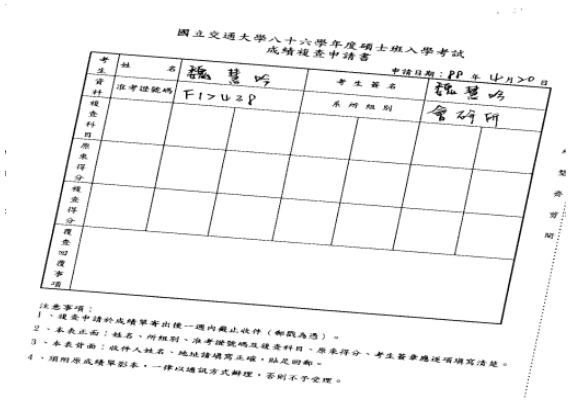


Fig. 7 Rotated image of Fig. 2.

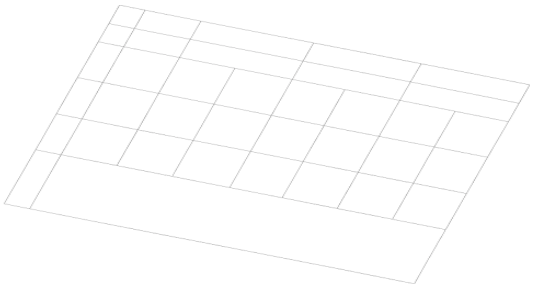


Fig. 8 Processed image of Fig. 7.

Table 5 Normalizing and sorting the vertical processing results of Fig. 7.

0.541944	0.543276	0.544607	0.727031	0.728362
0.729694	0.997337	0.998668	1.000000	
Difference from Table 2				
0.003207	0.005219	0.003888	0.002066	0.004079
0.002747	0.002663	0.001332	0	

5. Conclusion

In this paper, we proposed a new approach using high order correlation method for document processing. We utilized the correlations of neighboring pixels to successfully extract the lines and tables in a document. The texts and graphs are simultaneously removed. The results have great contribution to image segmentation and OCR, etc. HOC has great simplicity in computation and can provide accurate results under various environments. Its parallel architecture [1] also enables real-time implementation. A user-friendly system can be easily provided and minimum set of equipment is required. The property of rotation invariance makes the application even more convenient. Unlike most conventional methods that deal with segmentation and classification separately, our approach can do the job in one step and thus is more efficient.

Acknowledgement:

This work was supported in part by National Science Council under contract number NCS 88-2213-E-212-001.

References:

[1] R. J. Liou and M. R. Azimi-Sadjadi, “Multiple Target Detection Using Modified High Order Correlations”, to appear in IEEE Transaction on Aerospace and Electronic Systems, vol. 34, no. 2, pp. 553-568, April 1998.

[2] R. J. Liou and M. R. Azimi-Sadjadi, “Dim Target Track Detection Using High Order Correlation Method”, IEEE Transaction on Aerospace and Electronic Systems, vol. 29, no. 3, pp. 841-856, July 1993.

[3] B. Porat and B. Friedlander, “A frequency approach for multiframe detection and estimation of dim targets,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 12, no. 4, pp. 398-401, April 1990.

[4] I. S. Reed, R. M. Gagliardi and H. M. Shao, “Application of three dimensional filtering to moving target detection,” IEEE Transaction on Aerospace and Electronic Systems, vol. 19, no. 6, pp. 898-905, November 1983.

[5] Y. Bar-Shalom, T. E. Fortman, “Tracking and Data Association,” Academic Press, 1988.

[6] Roth, “Survey of Neural Network Technology for Automatic Target Recognition,” IEEE Transaction on Neural Networks, vol. 1, no. 1, pp. 28-43, March 1990.

[7] Ren-Jean Liou and Mu-Song Chen, 1998, “Recognition of Table-form Documents Using High Order Correlation Method”, in Proceedings of IJCNN’98, Alaska.

[8] J. Liu, C. Lee and R. B. Shu, “A Efficient Method for the Skew Normalization of a Document Image “, Proceedings of IEEE, pp. 122-125, 1992.

[9] D. S. Le, G. R. Thoma and H. Wechsler, “Automated Page Orientation and Skew Angle Detection for Document Images”, Pattern Recognition, vol. 127, no.10, pp. 1325-1344, 1994.

[10] A. Hashizume, P. S. Yeh and A. Rosenfeld, “A Method of Detecting the Orientation of Aligned Component”, Pattern Recognition Letter, vol. 4, pp. 125-132, 1986。

[11] W. Postl, “Method for Automatic Correction of Character Skew in the Acquisition of a Text Original in the Form of Digital Scan Results”, US Patent 4723297, 1988.

[12] J. M. Lu, “Automatic Form Classification by Feature Graph Matching”, Master‘s Thesis, National Central University, Taiwan, 1995.

[13] S. W. Chen, “Form Recognition for Table-form Document” Master‘s Thesis, National Central University, Taiwan, 1995.

[14] C. J. Wilson, J. Geist, M. D. Garris and R. Chellappa, “Design, Integration, and Evaluation of Form-Based Hand-print and OCR Systems”, NIST Internal Report 5932, 1996.

[15] M. D. Garris and P. J. Grother, “Generalized Form Registration Using Structure-Based Techniques”, in Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, pp. 321-344, 1996.

[16] T. Y. Zhang and C. Y. Suen, “A Fast Thinning Algorithm for Thinning Digital Patterns,” Communications of ACM, No. 27, pp. 236-239, 1984.

[17] T. Y. Zhang and C. Y. Suen, “A Modified Parallel Thinning Algorithm,” in Proceedings of ICPR’88, pp. 1023-1025, 1988.