Speed-up Line Detection Approach for Large-size Document Images by Parallel Pixel Scanning and Hough Space Minimization

H. Waruna H. Premachandra[1], Chinthaka Premachandra[2],
Chandana Dinesh Parape[3] and Hiroharu Kawanaka[4]
[1]*ICT Center,Wayamba University of Srilanka, Makadura, Srilanka*
[2]*Graduate School of Engineering, Tokyo University of Science Tokyo, Tokyo, Japan*
[3]*Graduate School of Engineering, Kyoto University, Kyoto, Japan*
[4]*Graduate School of Engineering, Mie University, Tsu, Japan*

Keywords: Line Detection, Hough Space Minimization, Large Document Image, Parallel Pixel Scanning, Speed-up Image Processing.

Abstract: Hough transform (HT) is typically used to detect lines in images, but that method is slow due to its use of voting-based parameter detection; detecting lines in large document images can take dozens of minutes. Nonetheless HT is very effective at detecting lines, so we investigate methods for fast HT-based line detection of large document images by minimizing Hough space processing and reducing the image area used for line detection with parallel pixel scanning and local image domain analysis. We conduct experiments to confirm the effectiveness of the proposed method using appropriate large documents images. The results show a significant computational time reduction as compared to conventional methods.

1 INTRODUCTION

Document image processing have widely been studied and various studies of document scanning, character recognition, and document structure analysis and understanding can be found in literature (You, 2010) (Wang, 2002) (Yip, 2001) (Manikandan, 2010) (Yang, 2000) (Borges, 2008) (Shi, 2013) (Takasu, 1995). Document structure analysis and understanding are mainly conducted by detecting lines in documents that contain them.

Effective detailed document image analysis requires a high-resolution image, but increased resolution makes images larger. In this paper, document images larger than 2000 × 3000 pixels are considered large images. Image processing time for a given target mainly depends on the image size and complexity of the detection algorithm, so large images can significantly increase processing time. Thus, studies have been conducting processing time reduction for large images through limiting the number of processed pixels and developing simpler algorithms (Premachandra, 2014) (Premachandra, 2014) (Premachandra, 2015).

Figure 1 shows a part of a large image used in this study. The image contains characters, lines, and other objects such as characters enclosed by ellipses. In this study, the objects of interest are lines, for which we develop a fast detection algorithm.

Generally, raster scanning is conducted while an image is being processed. In raster scanning, the image is scanned starting from the top-left point and ending with bottom-right point. Processing large document images generally requires considerable computational time. In this study, pixel scanning is conducted following the parallel vertical scanning (PVS) and parallel horizontal scanning (PHS) concept (Premachandra, 2013) (Premachandra, 2014). In PVS, pixels in horizontal lines are scanned keeping a constant space between two consecutive lines. With this scanning concept we can effectively ignore characters and other uninteresting objects in the image, reducing computational time. In the proposed method, whenever a black pixel is found while scanning we assume that the found pixel belongs to a line, and line availability is estimated by processing a defined local image domain (I_{d1}) over the black pixel. The approximate line inclination is also determined through this estimation. Line detection is then conducted by applying the Hough transform (HT) process to a another extracted image domain (I_{d2}) following the determined approximate

Premachandra, H., Premachandra, C., Parape, C. and Kawanaka, H.
Speed-up Line Detection Approach for Large-size Document Images by Parallel Pixel Scanning and Hough Space Minimization.
DOI: 10.5220/0005840707650769
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 765-769
ISBN: 978-989-758-175-5

line inclination. As Fig. 2 and Eq. (1) illustrate, Hough space (HS) is a conversion of (x, y) space to (r, θ) space, here with the angle θ ranging from 0 to π. In this study we reduce the HS by limiting the range of θ regarding the determined line inclination, significantly reducing computational time without loss of line detection performance as compared to the conventional HT process. This idea is one of the main novelties of this paper. As mentioned above, line availability estimation and line detection are conducted by processing only small local image domains (I_{d1}, I_{d2}) extracted from the original image. This process allows further computational time reduction, since the image area used for line detection is very limited.

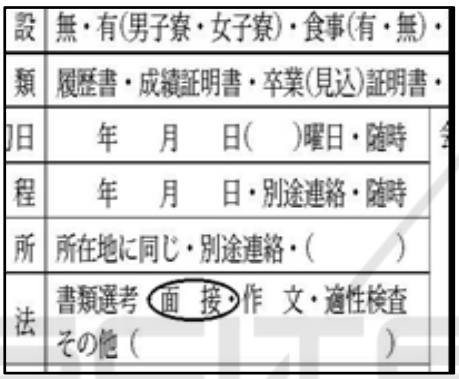

Figure 1: Part of a document image.

The new proposal was tested using appropriate large document images, and the results showed a significant computation time reduction with improved line detection performance.

The remainder of this paper is structured as follows. Section 2 introduces conventional approaches for line detection, and Section 3 details the new line detection approach following PVS and PHS, local image domain analysis, and HS minimization concepts. Section 4 details experimental results of the proposed method using large document images. Section 5 concludes the paper.

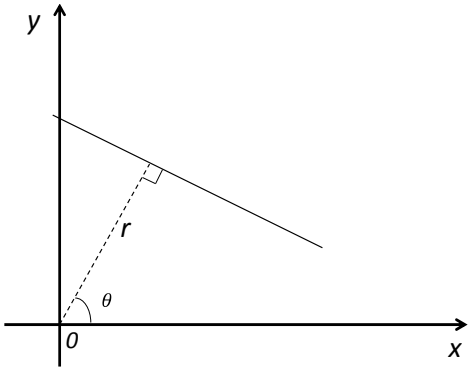

Figure 2: $(x, y) \rightarrow (r, \theta)$ conversion.

$$r = x \cos\theta + y \sin\theta \quad (1)$$

2 RELATED WORKS

Line detection is an important step in analyzing document images. Most studies in the literature related to this task implement the Hough transform (HT) (Li and Tsai, 2011) (Zhao, 2010) (Aggarwal, 2006). Classical HT is commonly used to detect regular curves such as lines, ellipses, circles, and parabolas. For applications in which simple descriptions of the desired features are not possible, such as random handwritten objects detection, a generalized HT can be used. Generalized HT is conceptually similar to template matching and has a higher computational complexity than the classical HT, which itself is similar to template matching. However, classical HT is also time consuming because of its voting-based methodology. In this study we use classical HT for line detection, to take advantage of its high line detection performance. We also aim at reduced time consumption without loss of line detection performance.

Some studies perform line detection without depending on the HT approach. Leferve et al. (Lefevre, 2002) approached fast solid line detection for scene modeling, focusing on horizontal and vertical lines, and proposed a fast local approach to line detection in binary images. In their method, pixels are analyzed using an accumulator on a block-based basis to obtain possible line segments for each block. As another non-HT approach, Kawanaka et al. (Kawanaka, 2007) conducted line detection by connected component analysis. These two approaches cannot detect lines that are connected with other objects in the image. In addition, both methods include connected component analysis, which is time consuming when large objects exist in the image.

3 PROPOSED LINE DETECTION APPROACH

This section presents the line detection algorithm in details.

3.1 Proposed Algorithm

Step 1: The document image is binarized using discriminant analysis (Otsu, 1999) and tilt correction is conducted using the LPP method (Kawanaka, 2007).

Step 2: The image is scanned by PVS as illustrated in Fig. 3. Here, the scanning lines are indicated by red. Either of PVS or PHS can be used for scanning.

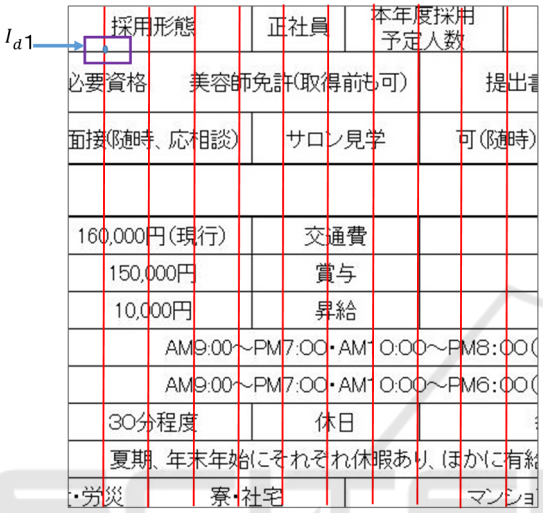

Figure 3: Parallel vertical scanning (PVS) and I_{d1} extraction.

Step 3: Whenever a black pixel of an object is found while scanning, we extract a local image domain (I_{d1}) with the black pixel as its center point (Fig. 3). The same label is set to all black pixels of that object within I_{d1}.

Step 4: The minimum bounding rectangle BR_m of a labeled object is calculated as illustrated in Fig. 4. In addition, the aspect ratio AS_{BR} of BR_m and black pixel density B_d inside BR_m are calculated following Eqs. (2) and (3). In Eq. (2), w_{BR} and h_{BR} ($w_{BR} < h_{BR}$) are aspects of BR_m. In Eq. (3), n_{BR} denotes the number of black pixels inside BR_m

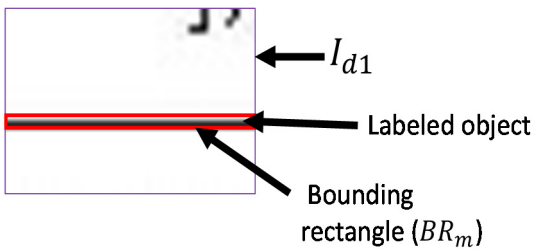

Figure 4: Minimum bounding rectangle BR_m.

$$AS_{BR} = w_{BR}/h_{BR} \quad (2)$$

$$B_d = \frac{n_{BR}}{w_{BR} \times h_{BR}} \quad (3)$$

If $AS_{BR} \leq th_1$ *and* $B_d \geq th_2$, we assume that the labeled object is part of a line and the process moves Step 5. Otherwise the process moves to Step 7. In this study, th_1 and th_2 are experimentally determined.

Step 5: The inclination θ_{BR} of BR_m is calculated. In this study, θ is not varied from 0 to π to generate the HS by Eq. (1) as in conventional HS generation. We instead determine the range of θ according to θ_{BR}. Here, θ ranges between $\theta_{BR} - \alpha$ and $\theta_{BR} + \alpha$. We set $\alpha = 10^{\circ}$. With this setting HS generation can be minimized, reducing computational time for HS generation.

Step 6: We detect the line assumed in Step 4 following HT, only generating the HS using the range for θ determined in Step 5. Line detection is conducted only when processing an image domain (I_{d2}) extracted from the original image. I_{d2} is extracted over the assumed line in Step 4, specifically considering its inclination (Fig. 5). HT is conducted for only the extracted I_{d2}, further reducing computational time. After conducting line detection, the process moves to Step 7.

Step 7: The process moves to Step 3, in which the next (i, j) pixel is scanned.

The above algorithm effectively detects horizontal and inclined lines, but not vertical lines since it is difficult to cross the pixels of a vertical line through PVS. We therefore implement the algorithm using both PVS and PHS.

Figure 5: I_{d2} extraction from the original image.

4 EXPERIMENTS

4.1 Experimental Setup

All experiments were conducted using a personal computer with a Core i7 3.4 GHz CPU and 4 GB RAM. Appropriate experimental materials were created using a digital image scanner with a resolution of 300 dpi. The size of those images was 2480 × 3508 pixels. We generated fifty document images including 771 lines written in English, Japanese, and Sinhala (the predominant language in Sri Lanka).

To clarify detection performance, detected lines are deleted by setting their pixels white. Specifically, each pixel on a detected line and its eight neighboring pixels we set to white.

We compared the proposed method with two conventional methods from the literature: a conventional HT line detection method and a line detection based on connected component analysis (CCA) (Kawanaka, 2007). The evaluation compared line detection performance and computational time consumption.

4.2 Results

Figures 6 shows an example image to which the proposed method was applied for line detection. In this figure, the upper image is the original image, and the image below is the corresponding deletion result. Here, some personal information is obscured by a brown color.

Tables 1 and 2 summarize the results of the experiments. As Tables 1 and 2 show, the proposed method significantly reduced computational time and dramatically reduced the number of false positives.

Table 1: Detection Performance Comparison.

	Detection rate	False positive rate
HT	98.7% (760/771)	5%
CCA method	95.2% (734/771)	10%
Proposed	98.4% (759/771)	**0%**

Table 2: Computational Time Comparison.

	Average computational time
HT	1065 *ms*
CCA method	765 *ms*
Proposed	**89 *ms***

Figure 6: Line detection results.

5 CONCLUSIONS

We introduced a fast line detection approach for application to large document images. In the proposed method the image is scanned along parallel vertical or horizontal lines, and line detection is conducted by the HT while minimizing HS processing by analyzing only local image domains

selected from the image. Experiments were conducted to evaluate the proposed method using appropriate large images. The method significantly increased computational speed as compared to the conventional methods. The method furthermore reduced the rate of false positives while maintaining a line detection rate similar to the conventional methods.

REFERENCES

Jin, S.,You, Y., Huafen, Y., 2010. Scanned Document Image Processing Model for Information System, *Asia-Pacific Conf. on Wearable Computing Systems*.

Wang, Q., Chi, Z., Zhao, R., 2002. Hierarchical content classification and script determination for automatic document image processing, *16th International Conference on Pattern Recognition*.

Yip, S. K., Chi, Z., 2001. Page segmentation and content classification for automatic document image processing, *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*.

Manikandan, V., Venkatachalam, V., Kirthiga, M., Harini, K., Devarajan, N., 2010. An enhanced algorithm for Character Segmentation in document image processing , *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*.

Yang, Y., Yan, H., 2000. A robust documefdotnt processing system combining image segmentation with content-based document compression, *15th International Conference on Pattern Recognition*.

Borges, P. V. K., Mayer, J., Izquierdo, E., 2008. Document Image Processing for Paper Side Communications, *IEEE Transactions on Multimedia," Vol. 10, Issue 7, pp. 1277-1287*.

Shi, Z., Setlur, S., Govindaraju, V. 2013. A Model Based Framework for Table Processing in Degraded Document Images, *12th International Conference on Document Analysis and Recognition (ICDAR)*.

Takasu, A., Satoh, S., E. Katsura, E., 1995. A rule learning method for academic document image processing, *Third International Conference on Document Analysis and Recognition*.

Premachandra, H. W.H., Premachandra, C., Parape, C. D., 2013. Parallel Scanning Based Speed-up Method for Detection of Elliptical Obstacles in High-resolution Image, *International Journal of Computer Science and Communication Networks, Vol. 3, Issue5, pp.265-270*.

Premachandra, C.,Premachandra, H. W.H., Parape, C. D., Kawanaka, H., 2014. Parallel Layer Scanning Based Fast Dot/Dash Line Detection Algorithm for Large Scale Binary Document Images, *Lecture Notes in Computer Science (LNCS), Vol. 8814*.

Premachandra, H. W.H., Premachandra, C., Parape, C. D., Kawanaka, H., 2015. Speed-up Ellipse Detection Approach for Large Document Images by Parallel Scanning and Hough Transform, *International Journal of Machine Learning and Cybernetics*.

Li, W. C., Tsai, D. M., 2011. Defect Inspection in Low-Contrast LCD Images Using Hough Transform-Based Nonstationary Line Detection, *IEEE Transactions on Industrial Informatics, Vol. 7, Issue 1, pp.136-147*.

Zhao, X., Liu, P., Zhang, M., Zhao, X., 2010. A novel line detection algorithm in images based on improved Hough Transform and wavelet lifting transform, *IEEE International Conference on Information Theory and Information Security (ICITIS)*.

Aggarwal, N., Karl, W. C., 2006. Line detection in images through regularized hough transform, *IEEE Transactions on Image Processing, Vol. 15, Issue 3, pp.582-591*.

Lefevre, S., Dixon, C., Jeusse, C., Vincent, N., 2002. A Local Approach for Fast Line Detection, *IEEE International Conference on Digital Signal Processing*.

Kawanaka, H., Sumida, T., Yamamoto, K., Shinogi, T.,Tsuruoka, S., 2007. Document Recognition and XML Generation of Tabular Form Discharge Summaries for Analogous Case Search System, *Method Inf. Med., Vol. 46, pp. 700-708*.

Otsu., N., Lopes, J., 1999. Threshold Detection Method from Grey-Level Histograms, *IEEE Trans. Systems, Man, and Cybernities,Vol. 9 No.1, pp.62-66*.