HOUGH TECHNIQUE FOR BAR CHARTS DETECTION AND RECOGNITION IN DOCUMENT IMAGES

Yan Ping Zhou, Chew Lim Tan

School of Computing, National University of Singapore, Singapore, 117543

ABSTRACT

Charts are common graphic representation for scientific data in technical and business papers. We present a robust system for detecting and recognizing bar charts. The system includes three stages, preprocessing, detection and recognition. The kernel algorithm in detection is newly developed Modified Probabilistic Hough Transform algorithm for parallel lines clusters detection. The main algorithms in recognition are bar pattern reconstruction and text primitives grouping in the Hough space which are also original. The Experiments show the system can also recognize slant bar charts, or even hand-drawn charts.

1. INTRODUCTION

In document image processing, form or table recognition has been a topic of intensive research in the last decade. There are also some commercial form or invoice recognition systems available now. But charts, the most common and powerful representation tool in technical and business data analysis, have not yet drawn much attention. Most of the work in the graphics recognition of document images has been reported in circuit diagrams [1], geographic maps [2], engineering drawings [3]. One of the major reasons for the difficulty of chart recognition is the wild variety of chart styles, such as bar chart, curve chart, pie chart, etc. Even in the same chart style, there are also a lot of variations due to position translation of graphics or text primitives. Thus it is difficult to generate a generic, style independent, chart recognition system.

In [4], Futrelle presented a diagram understanding system based on constructing graphics constraint grammars for different types of diagrams by syntactic analysis. His work focused on a high level processing. In [5], Yokokura et al put up a layout-based network which was similar to Futrelle's graphics constraint grammar to graphically describe the layout relationship information of the bar chart. He used simple vertical and horizontal projection to do segmentation and combine bar chart layout information while extracting graph and text primitives. Due to the simplicity of the segmentation method, the bar chart styles that can be recognized are constrained by a lot of assumptions.

In this paper, we present a robust system for detecting and recognizing bar charts with little assumptions. The system includes three stages, preprocessing, detection and recognition. As we know, the most salient feature in the bar chart is bar patterns. Bar patterns are composed of parallel lines pair. To detect parallel lines clusters, we develop a Modified Probabilistic Hough Transform algorithm (MPHT). It save the computation time by diminishing the number of voting points. We reconstruct the bar patterns using the output from MPHT and Grouping the text primitives in the Hough space. We also correlate the bars and their corresponding text primitives by checking the basic structural rules in the Hough domain. Experiment results show that our recognition method has an advantage of orientation immunity. Our method can recognize slant or skewed bar chart with a high correct rate. It also can read hand-drawn bar charts.

This paper is organized as follows: In section 2, The structure of bar charts detection and recognition system is presented. In section 3, our new MPHT algorithm for detecting parallel lines clusters is illustrated. In section 4, we give our experiment results performance of our system. This paper ends with a conclusion in section 5.

2. BAR-CHARTS DETECTION AND RECOGNITION

The structure of bar charts detection and recognition system is illustrated in Figure 1.

The system is composed of the following three main stages:

Stage 1: Preprocessing.

Do the traditional image segmentation operations in the image space such as connected component analysis, using size filter to separate image into graphics image and text image.

Stage 2: Detection.

It includes the following steps:

0-7803-6297-7/00/$10.00 © 2000 IEEE

1. Graphics area grouping. Group neighboring graphics elements into bar chart candidates.
2. Boundary image. Get the boundary image of bar chart candidate [6].
3. Parallel line cluster detection. Apply Modified probabilistic Hough Transform (MPHT) to detect parallel line cluster. Details will be given in section 3.
4. Corresponding text area grouping. According to the information from step 1 and 3, group neighboring text into chart text area candidate.
5. Form frame filtering. Filter those chart candidates whose line segments are all similar or whose text elements are distributed more than two lines between each parallel line pair.

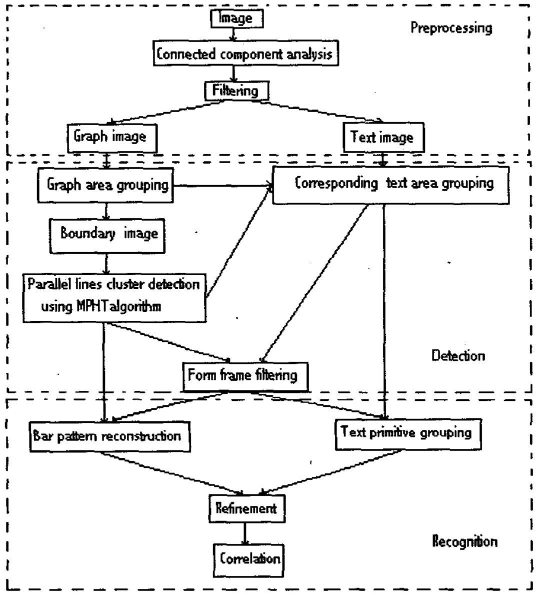


Fig.1 Overview of the bar chart detection and recognition system

Stage 3. Recognition

It includes the following four steps:

1. Bar pattern reconstruction. Using the result from parallel lines cluster detection, reconstruct the bars.
2. Text primitives grouping. Apply text primitives grouping algorithm on centroid points of text elements in Hough space.
3. Refinement. Check the basic structural rules set to refine both text and graphics primitives results.
4. Correlating. Correlate the bar patterns with their corresponding text primitives, such as bar data and tick names.

3. MPHT ALGORITHM

The Hough Transform (HT) is a robust method of extraction of geometric primitives [7,8]. The standard HT is computationally expensive. The introduction of probabilistic class of HT [9-11] is an important step for diminishing the computation cost. We develop a Modify Probabilistic Hough Transform algorithm for detecting parallel lines cluster. First we use boundary image of interested graph area to minimize the voting points. Then we order the input voting points row-wisely. The resolution of θ is set to one. We also divide θ into 18 divisions with different accessing priority. For example, division 85°-95° has the highest priority, 0°-5° together 175°-180° has the second, etc.

The MPHT algorithm can be outlines as follows:

1. Check the next priority θ division, if no division exist, then finish.

2. Update the accumulator array using a pixel successively selected from voting list. Mark it as processed.

3. Check if there are peaks in the accumulator array that is above the threshold. If not, then go to 9.

4. Verify the corridors specified by the peaks to find the line segments with a tolerant gap and write down the line segments attributes in both image and Hough spaces.

5. Remove all the pixels in the line segments from the voting lists and adjust the pointer of the next processing point.

6.Check if there are more than three line segments lie in the range $\Delta\ \theta = \pm 2$. If not then go to 9

7. Compute the minimum inter distance of found line segments, $\Delta\rho$. For each peak ρ, select $\rho \pm \Delta\rho$ as candidate peaks.

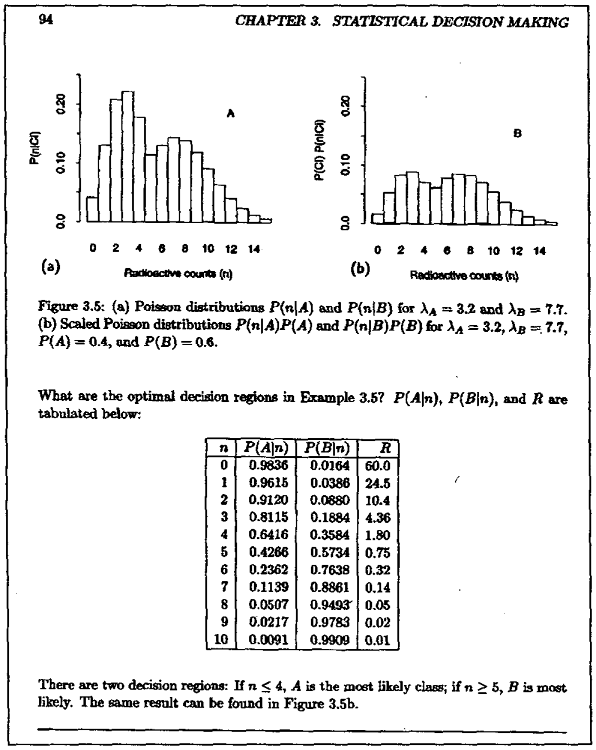
8. Repeatedly go through 4,5,7, until there is no line segment updating.

9. Check if there are unprocessed points in the voting list. If yes, go to 2.

10. Check if there are points left in the voting list. If not finish, else initialize the points as unprocessed, go to 1.

4. EXPERIMENTS

Experiments on synthetic and real images, even hand-drawn images, are carried out. The following figures 2 to 4 shows the processing result of a real image.



94 CHAPTER 3. STATISTICAL DECISION MAKING

Figure 3.5: (a) Poisson distributions $P(n|A)$ and $P(n|B)$ for $\lambda_A = 3.2$ and $\lambda_B = 7.7$. (b) Scaled Poisson distributions $P(n|A)P(A)$ and $P(n|B)P(B)$ for $\lambda_A = 3.2$, $\lambda_B = 7.7$, $P(A) = 0.4$, and $P(B) = 0.6$.

What are the optimal decision regions in Example 3.5? $P(A|n)$, $P(B|n)$, and R are tabulated below:

n	$P(A\|n)$	$P(B\|n)$	R
0	0.9836	0.0164	60.0
1	0.9615	0.0386	24.5
2	0.9120	0.0880	10.4
3	0.8115	0.1884	4.36
4	0.6416	0.3584	1.80
5	0.4266	0.5734	0.75
6	0.2362	0.7638	0.32
7	0.1139	0.8861	0.14
8	0.0507	0.9493	0.05
9	0.0217	0.9783	0.02
10	0.0091	0.9909	0.01

There are two decision regions: If $n \leq 4$, A is the most likely class; if $n \geq 5$, B is most likely. The same result can be found in Figure 3.5b.

Fig. 2 The original image (1200*939 pixels)

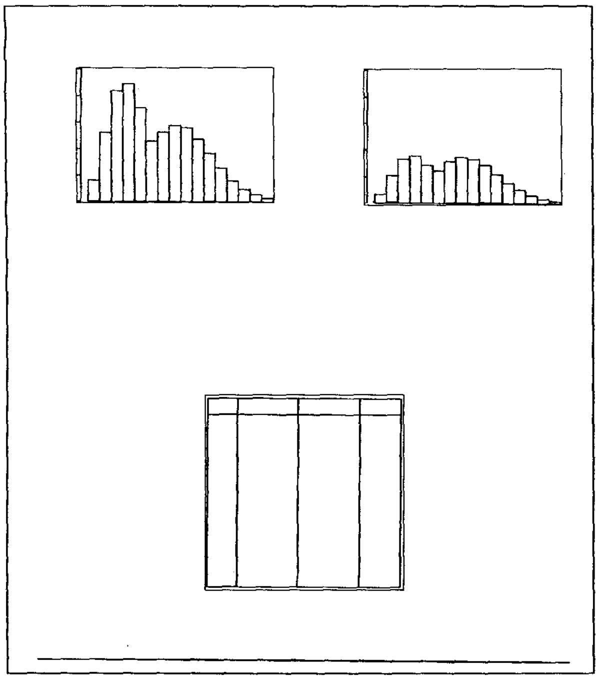


Fig. 3 The interested graph areas (boxed). The bottom line segment is filtered.

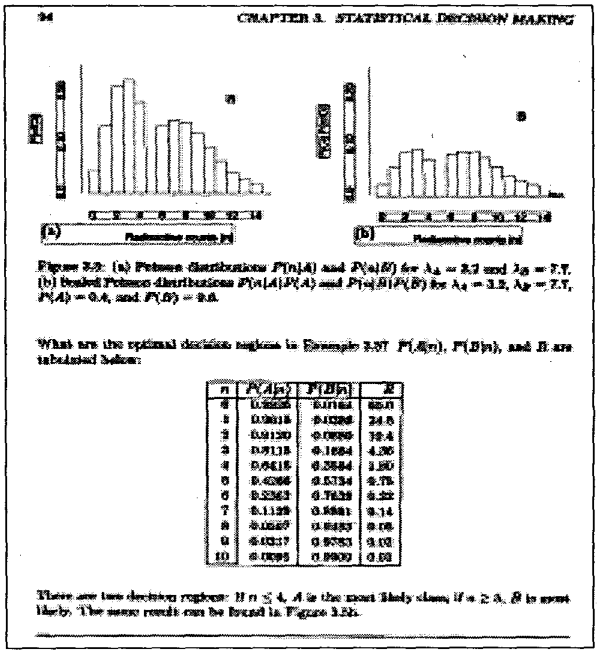


Fig. 4 Results after bar pattern reconstruction and text primitive grouping. The table is filtered after form frame filtering.

Based on the Hough Transform, the system has orientation immunity that can detect and recognize skewed charts or hand-drawn charts. Fig 5 shows an example of applying MPHT to extract bar patterns in a hand-drawn bar chart.

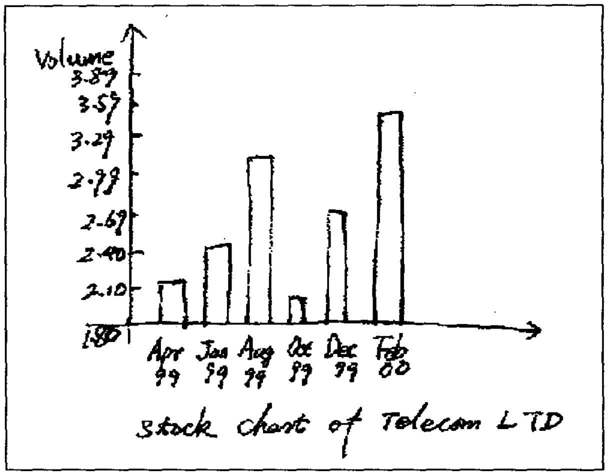


Fig. 5 The original hand-drawn bar chart (567*450)

After the connected component analysis and size filtering in preprocessing, we separate the graphics image from the text image. We use the boundary of the graphics image as the feature input of MPHT to find and reconstruct the bar patterns as shown in Fig 6.

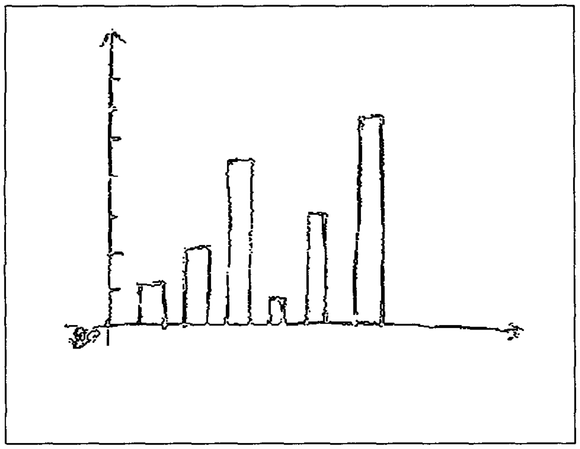


Fig 6. The input feature points for MPHT

We also use Standard Hough Transform (SHT) plus butterfly filter for peak enhancement to reconstruct bars and axes. The number of input feature points is 4153. The angle ranges of the x axis and the bars (or y axis) detected are 0°(or 180°) and 88° to 91° respectively. Both of the ranges lie in the first two priority θ divisions. It means the voting time of MPHT is at most one ninth of that of SHT. In MPHT, the number of the voting points also decreases while finding more than two lines in the same θ division. The line threshold for both SHT and MPHT is 15. The following figure shows the parametric reconstruction of the bars by MPHT. The lowest bar is missing in the reconstruction result of SHT due to the noisy voting.

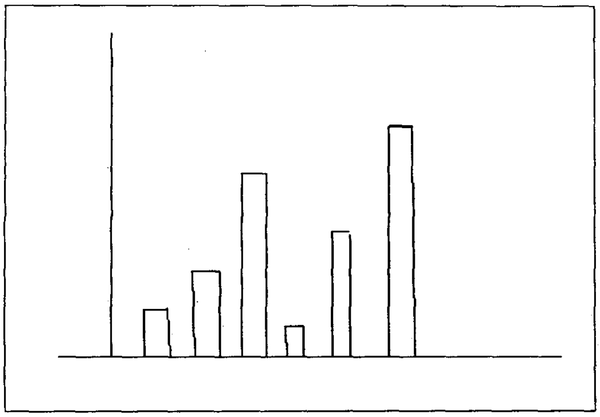


Fig 7. Bars reconstruction result by MPHT

We test 20 images scanned from books and get the following results. The form frames are all regular. It can discard these forms correctly. Among the detected charts, the bar reconstruction correct rate is 92.3%. The correlation correct rate of bar pattern and the corresponding text primitives is 87.3%.

5. CONCLUSIONS

In this paper, we present a robust system for detecting and recognizing bar charts with little assumptions. The system includes three stages, preprocessing, detection and recognition. The kernel algorithm in detection is newly developed Modified Probabilistic Hough Transform algorithm for parallel lines clusters detection. The main algorithms in recognition are bar pattern reconstruction and text primitive grouping in the hough space which are also original. The results show that the system can recognize bar charts lying in any orientations, such as slant bar charts, or even hand-drawn bar charts.

6.REFERENCES

[1] S. Lee, Recognizing hand-written electrical circuit symbols with attributed graph matching, Structured document analysis, pp340-358, 1992.

[2] E. Reiher, Y. Li, V.D. Done, M. Lalonde, C.Hayne, a system for efficient and robust map symbol recognition, Proceedings of the 13th IAPR, pp783-787, 1996.

[3] Y. Yu, A. Samal, S. Seth, Automatic segmentation of engineering drawings with symbols and connections, Proceedings of third IAPR international conference on document analysis and recognition, ICDAR'95, pp791-794.

[4] R.P. Futrelle and et al, Understanding diagrams in technical documents, IEEE Computer, Vol.25, NO.7, pp75-78, 1992.

[5] N. Yokokura and T. Watanabe, Layout-Based Approach for extracting constructive elements of bar-charts, Graphics recognition: algorithms and systems, GREC'97, pp163-174.

[6] R. Jain, R. Kasturi and B. G.Shunck, Binary image processing, Machine Vision, pp50-51, 1995.

[7] V.F. Leavers, Postprocessing, Shape detection in computer vision using the Hough Transform, pp70-75, 1992.

[8] J. Illingworth and J. Kittler, A survey of the Hough Transform, Computer vision, graphics and image processing, 44: pp87-116, 1988.

[9] J.R. Bergen and H. Shvaytser, A probabilistic algorithm for computing Hough Transforms, Journal of algorithms, 12(4): pp639-656, 1991.

[10] L. Xu, E. Oja, P. Kultanen, A new curve detection method: Randomized Hough Transform, Pattern recognition letters, 11(5): pp331-338, 1990.

[11] C.Galambos, J.Matas and J. Kittler, Progressive Probablilistic Hough Transform for line detection, IEEE proceeding of Computer Vision and Pattern Recognition '98, pp554-560, 1998.