

Bar Charts Recognition Using Hough Based Syntactic Segmentation

Yan Ping Zhou, Chew Lim Tan

School of Computing, National University of Singapore, Singapore, 117543
{zhouyanp,tancl}@comp.nus.edu.sg
<http://www.comp.nus.edu.sg/~tancl>

Abstract. *Bar charts are common data representations in scientific and technical papers. In order to recognize the printed bar chart, we present a new Hough based bar chart recognition algorithm which combines syntactic analysis into segmentation. We first detect the most salient feature in any bar chart, bar patterns, using syntactic analysis in the Hough domain. Then we group text primitives according to their centroids distribution in the Hough space. Finally, we interweave the two extracting processes to refine the recognition results. Our recognition algorithm is not dependent heavily on a priori knowledge and can recognize bar charts lying in arbitrary directions, such as oblique or skewed bar charts, or even hand-drawn bar charts.*

1 Introduction

Diagram is a very powerful representation tool in the scientific research field because people understand graphic representations faster than the corresponding text representations. In [1], Futrelle presented a diagram understanding system based on constructing graphics constraint grammars for different types of diagrams by syntactic analysis. In [2], Yokokura et al put up a layout-based network to graphically describe the layout relationship information of the bar chart. In this paper, we present a new Hough based bar chart recognition from a computational point of view. In order to recognize a bar chart, we first syntactically analyze the structure of a common bar chart and draw the common attributes into a syntactic primary rule set. Second, we present a Hough based bar pattern location algorithm which combines the syntactic primary rules. Third, we use the information obtained from the bar pattern location algorithm to group the text elements in the Hough space. Fourth, we interleave the bar pattern location and text primitive grouping with the syntactic rules for refining the recognition result.

2. Bar Chart Syntactic Analysis

Bar chart is a framed chart, in which the X-axis and Y-axis constitute its main frame and the value of each bar is the most important information. Like other charts, bar charts are composed of graphics primitives and text primitives. For a framed chart, we divide the chart into two areas: substrate area and stage area. The substrate area is composed of the chart frame and its annotations. The stage area is where the graphics

primitives that carry the most important data are active. In a bar chart, bars with variant heights are present in the stage area. The substrate area is divided into four divisions according to two axes: X axis major division, X axis minor division, Y axis major division, and Y axis minor division. Figure 1 shows the illustration of the bar chart structure. Areas and divisions are shown in dashed rectangle boxes. Graphics primitives and text primitives are indicated in ovals.

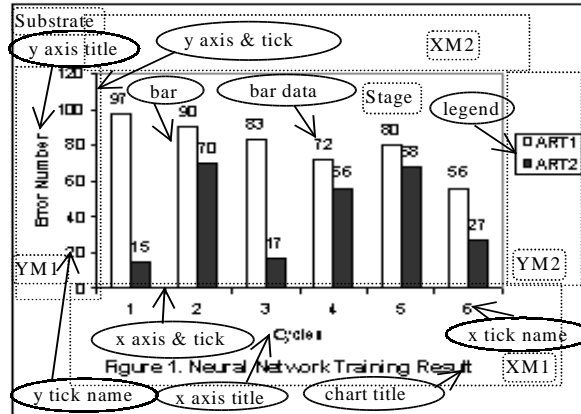


Fig. 1. General bar chart structure. XM1: X axis major division. XM2: X axis minor division. YM1: Y axis major division. YM2: Y axis minor division

3. Recognition Model and Method

The recognition model is composed of the following operation steps:

1. Preprocessing. Do the traditional image segmentation operations in the image space such as connected component analysis, using size filters to separate image into graphics image and text image, and getting the boundary image of graphics image[3].

2. Line feature extraction. Apply a Hough transform on the boundary image. Use a butterfly filter [4] to enhance the line feature points. Apply a threshold algorithm to get peak line points.

3. Bar pattern searching. Apply a hypothesis-testing bar pattern searching algorithm on peak line points set. Reconstruct the bars found.

4. Text primitives grouping. Apply a text primitives grouping algorithm on centroid points of text elements. The algorithm is similar to that in [5].

5. Refinement. Check the basic structural rules set to refine both text and graphics primitives results.

6. Correlating. Correlate the bar patterns with their corresponding text primitives, such as bar data and tick names.

Step 1 is performed in the image space. Steps 2 to 6 are performed in the Hough space.

4. Experiments

We first synthesize 10 bar charts and skew them along the X axis and the Y axis. We also scanned 20 real images from books. We also used 5 hand-drawn bar charts.

For synthetic bar charts, the correct rate of finding bars is above 90%. For the real images and hand-draw bar charts, the correct rate of bar extraction degrades to 87.3% and 78% respectively. But the text primitives grouping in the word level degrades a lot, only about 70%. The correlation rate between the bars and the text primitives is still satisfactory, above 80%. Figure 3 shows an example of processing a bar chart.

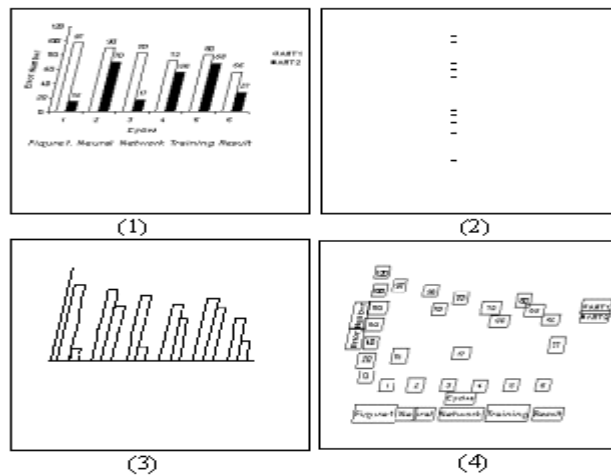


Fig.2. Results of Hough based bar pattern locating. (1) The original image. (2) The line feature points in Hough space after applying HT and butterfly filtering (3) Reconstructed result after bar pattern locating. (4) Text primitives grouping (word level)

5. Conclusions

In this paper, we present a new bar chart recognition algorithm using Hough based syntactic segmentation. It can recognize a variety of bar charts in document images including oblique bar charts. There still remain rooms for further study. The Standard Hough Transform is computationally expensive. We present a new Modified Probabilistic Hough Transform that is faster than SHT in [6]. In future, we will broaden our algorithm into a more generic diagram recognition system.

6. References

- [1] R.P. Futrelle and et al, Understanding diagrams in technical documents,IEEE Computer,Vol.25,NO.7,pp75-78,1992
- [2] N. Yokokura and T. Watanabe, Layout-Based Approach for extracting constructive elements of bar-charts, Graphics recognition : algorithms and systems ,GREC'97,pp 163-174
- [3] R. Jain, R. Kasturi and B. G.Shunck, Binary image processing, Machine Vision, pp50-51, 1995
- [4] R. Kasturi,S.T.Bow,W.El-masri,J.Shah,J.R. Gattiker and U.B. Mokate, A system for interpretation of line drawings, IEEE Trans. on PAMI,Vol,PAMI-12, No. 10,pp.978-992 ,1990
- [5] V.F. Leavers, Postprocessing, Shape detection in computer vision using the Hough Transform,pp 70-75,1992
- [6] Y.P. Zhou and C.L. Tan, Hough technique for bar charts detection and recognition in document images, IEEE international conference on image processing 2000.