Deriving Knowledge from Figures for Digital Libraries*

Xiaonan Lu, James Z. Wang, Prasenjit Mitra, and C. Lee Giles
The Pennsylvania State University, University Park, Pennsylvania, USA
xlu@cse.psu.edu, {jwang, pmitra, giles}@ist.psu.edu

ABSTRACT

Figures in digital documents contain important information. Current digital libraries do not summarize and index information available within figures for document retrieval. We present our system on automatic categorization of figures and extraction of data from 2-D plots. A machine-learning based method is used to categorize figures into a set of predefined types based on image features. An automated algorithm is designed to extract data values from solid line curves in 2-D plots. The semantic type of figures and extracted data values from 2-D plots can be integrated with textual information within documents to provide more effective document retrieval services for digital library users. Experimental evaluation has demonstrated that our system can produce results suitable for real-world use.

Categories and Subject Descriptors: [H.3.3] Information Systems: Information Search and Retrieval

General Terms: Algorithms, Design, Experimentation

Keywords: Figures, Feature Extraction, Machine Learning

1. INTRODUCTION

Figures present important information about documents. Many types of of illustrations such as pictures and non-pictorial graphics may be contained in figures. They are commonly used to illustrate key ideas and to help readers understand the content of documents. Although much attention has been devoted to summarization and retrieval of textual information within documents, relatively little attention has been given to figures that appear in documents [1]. This work attempts to address the summarization of figures within documents.

We have applied content-based approach for deriving knowledge from figures in electronic documents. We have designed methods to enable computers to classify figures into a set of predefined semantic types, and to extract data values corresponding to line curves within 2-D plots. The semantic type of a figure and data extracted from 2-D plots provide summarization about figures in documents. These information about figures can be combined with existing textual information within documents to assist digital library users in finding relevant documents more effectively.

*This work was supported in part by the US National Science Foundation under grants 0535656, 0347148, 0454052, and 0202007, Microsoft Research, and the Internet Archive.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. THE METHOD

2.1 Categorization of Figures

Our system, as shown in Figure 1, uses a machine-learning based approach to categorize figures into a set of predefined classes based on image features extracted from content of figures [4]. The defined classes consist of photograph and non-photograph, while the non-photograph is further divided into 2-D plot, 3-D plot, Diagram, and Others. The image features consist of global texture features and part object features. The global texture features measure the distribution of background, text, and picture blocks [3] within a figure, which are used to discriminate photograph vs. non-photograph figures. The object features record the existence of a specific type of objects, currently straight lines, which are used to classify different types of non-photograph figures. Finally, a supervised learning method is selected to train the computers for automatic categorization of figures.

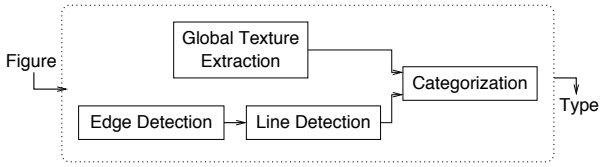

Figure 1: The process to categorize figures.

2.2 Data Extraction from 2-D Plots

Our data extraction process, as shown in Figure 2, extracts data values from line curves within 2-D plots. Currently, our system processes solid line curves and records data points of curves in database. The extracted data values may serve as the raw data for a variety of analyses, e.g., curve fitting of the raw data, analysis of published models of scientific phenomena, interpretation, and predictions.

Techniques used in our data extraction system include Hough transform based straight line detection, image thinning, Primary Chain Code(PCC) based line analysis [5], and line composite analysis. The axis detection module, as shown in Figure 2, uses a customized Hough transform technique to detect axis lines of 2-D plots. Then, objects in data regions of 2-D plots are reduced to thin lines.

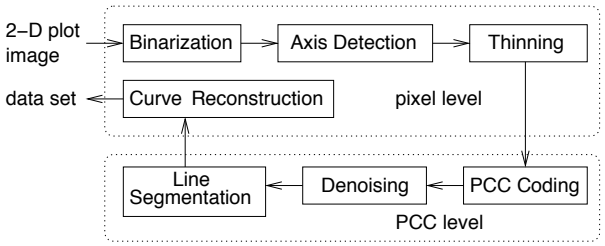

Figure 2: The data extraction process for 2-D plots.

After that, raster images are transformed into PCC based representation, which are suitable for line analysis. Noise reduction and line segment detection are conducted on the PCC level. Finally, we define a special type of line composite corresponding to solid line curves within 2-D plots. Techniques have been designed to construct solid line curves on top of detected simple line segments. After solid line curves are constructed, data points are selected from curves and data values are recorded in database.

Techniques have been designed to resolve the issue of intersections between multiple line curves. Intersections are detected based on analysis of connectivity attributes among line segments. Three types of intersection segments are defined and different handling methods are designed for different types of intersections. Finally, a derivative-matching based method has been designed to determine the continuation of curves after intersections.

3. EXPERIMENTAL RESULTS

Categorization: The performance of automatic categorization of figures in documents have been tested on the figures extracted from the CiteSeer [2] scientific literature digital library. A dataset of about two thousand scientific documents were randomly selected from the digital library and figures were extracted from those documents. The photograph vs. non-photograph and multi-class classifications were conducted and classification performance were reported [4]. The experimental results demonstrate the effectiveness of global textual features on discriminating photograph vs. non-photograph, and the effectiveness of straight line features on recognition of 2-D plots.

Data Extraction from 2-D Plots: Three datasets have been used in the evaluation of our data extraction system (1) the CiteSeer digital library, (2) Web images collected by using Google image search, and (3) MATLAB-generated plots. The performance of data extraction is measured by the correspondence between the original 2-D plots and the redrawn plots based on extracted data. Since the original data values are not available for 2-D plots from published scientific documents and from the Web, plots are redrawn based on extracted data so that we can "view" extracted data values and manually check the correspondence between extracted data values and the original 2-D plots. The ratio of matched curves for the three data sets are shown in Table 1.

In addition to the correspondence measure, the correctness of extracted data are evaluated for the synthetic dataset generated using MATLAB. Since the original data values for 2-D plots generated using MATLAB are available, it is possible for us to compare the original data values with extracted values. Specifically, for every pair of matched original curve and redrawn curve, the X and Y dimension values are normalized to 0-10 and 0-100, respectively. A set of data points are selected from each curve so that their X dimension values are uniformly distributed over the range of valid X dimension values. Finally, the Mean Squared Error (MSE) between Y dimension values of the original data sets and the extracted data sets are calculated. The average MSE between the original data sets and the extracted data sets are illustrated in Table 2.

Table 1: Correspondence between the original 2-D plots and the redrawn plots

Data Set	# Curves	# Matched	Match ratio
Synthetic	177	153	86%
CiteSeer	77	48	62%
Web	40	29	73%

Table 2: MSE between the original data sets and the extracted data sets for MATLAB-generated plots

# Curves	# Matched	MSE
177	153	1.26

The experimental results demonstrate the effectiveness of our method on extracting data from 2-D plots. Analysis of experimental results reveals various reasons for failed cases: poor image quality of scanned documents, skewness of documents, etc. Since the line analysis algorithm depends on connectivity, our method has difficulty in handling broken curves.

4. CONCLUSIONS AND FUTURE WORK

We have shown a system that can categorize figures and extract data from 2-D plots in scientific documents. We plan to extend the categories of figures and design effective image features for categorization. We aim at making the data extraction algorithm more robust to noisy figures. We will analyze more types of figures within documents.

5. REFERENCES

[1] S. Carberry, S. Elzer, and S. Demir. Information graphics: an untapped resource for digital libraries. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 581–588, 2006.

[2] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the ACM Conference on Digital Libraries*, pages 89–98, 1998.

[3] J. Li and R. M. Gray. Context-based multiscale classification of document images using wavelet coefficient distributions. *IEEE Transactions on Image Processing*, 9(9):1604–1616, 2000.

[4] X. Lu, P. Mitra, J. Z. Wang, and C. L. Giles. Automatic categorization of figures in scientific documents. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 129–138, 2006.

[5] M. Seul, L. O'Gorman, and M. J. Sammon. *Practical Algorithms for Image Analysis*. Cambridge University Press, 2000.