

Learning-based scientific chart recognition

Yanping 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

In this paper, a learning-based paradigm for scientific chart recognition is proposed. Two kinds of chart recognition methods are presented: hidden Markov model based and neural network based method. A newly developed feature extraction method is also put forward for chart images. Experiments on three kinds of charts show that the ergodic hidden Markov models achieve a satisfactory result for chart recognition. Unlike traditional primitive-based diagram recognition method, learning-based approach need not recognize the graphic primitives in charts. Thus the method bypasses the recognition error problem caused by inaccurate primitive extraction that is also a major obstacle to the construction of a general chart recognition system.

1 INTRODUCTION

Today the Internet offers us a huge amount of information. Though textual information is still the major source of data, there has been an increasing trend of introducing graphs, pictures and figures in to the information pool. It is stated that about one trillion statistical graphs are printed every year [1]. Most of statistical graphs appearing in scientific papers are scientific charts or diagrams. The scientific chart is a very powerful representation tool in the scientific research field because people understand graphs faster than the corresponding text. Graphics recognition is the conversion of information from its paper based graphical representation into computer readable data [2]. Blostein [3] segmented the whole graphics recognition system into two stages according to the different processing hierarchies: symbol recognition and symbol-arrangement analysis. In the area of graphics recognition, there is still less work reported on scientific chart recognition than table or form recognition or engineering drawing recognition.

In [4], Futelle presented a diagram understanding system by constructing graphics constraint grammars for different types of diagrams with syntactic analysis. His work focused on high level arrangement analysis. The analysis was classified into the syntactic-based framework. He assumes that the segmentation is successfully implemented before he applies graphics constraint grammars analysis. The major work he reported was on x, y data graphs and gene diagrams. In [5], Yokokura et al put up a layout-based network which is a schema-based framework to graphically describe the layout relationship information of the bar chart. In the period of symbol object recognition, he used simple vertical and horizontal projection to do segmentation and combine bar chart layout information while extracting graph and text primitives. In his work, the interweaving of segmentation and classification improves the accuracy of bar chart recognition. But due to the simplicity of the segmentation method, the bar chart styles that can be recognized are constrained by many assumptions.

In our previous work [6], we presented Hough-based bar chart detection and segmentation which combine with the syntactic analysis. Our symbol arrangement analysis was based on a blackboard framework. In the first part of the procedure, the structure of a common bar chart was syntactically analyzed. The most obvious

attribute in the bar chart is bar segments. In the second part of the procedure, a Hough-based bar pattern location algorithm which combines the syntactic primary rules was proposed to recognize the chart objects. In the third part of the procedure, the information obtained from the bar pattern location algorithm is used to group the text elements in the Hough domain. In the fourth part of the procedure, the bar pattern location and text primitive grouping are interweaved with the syntactic rules for refining the recognition result. In [7], we also reported a modified probabilistic Hough transformation technique to recognize the bar charts. The objective in the work was to speed up the standard Hough transform which is computational expensive.

If the image quality is good, all the works mentioned above can present satisfactory chart information for a particular chart type. One obstacle is that for slightly different chart types, such as dotted bar and slashed bar, a primitive-based extraction method will construct two different models for them. Thus it makes it difficult to extract scientific chart information in general.

In this paper, we propose a novel learning-based chart recognition method that focuses on the stage of symbol recognition. The graphic elements of a chart are considered as a whole object. Therefore there is no need for graphic primitive segmentation. Two kinds of learning-based chart recognition methods are proposed: hidden Markov models based method and neural network based method.

Hidden Markov Models (HMMs) are a probabilistic modeling tool for time series data. They have been successfully applied in speech recognition [8,9,10] and part-of-speech tagging [11]. In the image processing area, they also have some breakthroughs in handwritten character recognition. Several papers reported using discrete hidden Markov models to recognize handwritten words using lexicons [12,13,14]. Mohamed and Gader [15] are the first to apply continuous density hidden Markov models for a segmentation-free handwritten word recognition.

Kopec and Chou [16] proposed Markov source models in document image decoding. In their approach, document image decoding is consisted of three elements: an image generator, a noisy channel and an image decoder. The image generator is the Markov source. The noisy channel turns the ideal image into the observed image. The decoder matches the message by finding a posteriori optimal path given the observed image. Their approach achieved quite satisfactory recognition accuracy in decoding scanned telephone yellow pages.

In this paper, hidden Markov models are constructed for different general chart categories from training image data. After the optimal states path for a test chart is found, the high-level analysis is combined to understand and extract the chart information using the syntactic rule library associated with that chart category. A commonly used multi-layer feed-forward neural network with back-propagation learning algorithm is applied for comparison. A modified back-propagation algorithm is used to accelerate the speed of convergence.

We first describe the architecture of our proposed scientific chart recognition system in section 2. Then we present our hidden Markov model and neural network frame in section 3. Feature extraction is described in section 4. In section 5, we show some experiments and results of our extraction system. Finally, we conclude at section 6.

2. CHARTS RECOGNITION MODEL

Scientific charts are structured graphics representation in which the graphics elements and text elements are arranged regularly. These diagrams have simple syntactic and semantic rule constraints on their graphics elements. For example, in bar charts the bars' heights represent the values of the categories while in line charts the y coordinate of a corresponding category point represents its value. Due to the simplicity of their semantics, people find these concise graphical representations extremely useful to intuitively present their data analyses. They can suit the chart data analysis to their own aesthetic tastes by using a host of chart generation tools available. Therefore there are a lot of chart representation patterns. Thus the diversity of charts makes the general chart recognition a difficult problem. While most of the chart representations are in image format, some are in a vector format that is easier to process by computer. In our recognition system, charts are stored in image format.

Scientific chart recognition consists of two aspects. The first objective is to identify the chart type, for example, to identify whether the chart is a bar chart or a line chart. The second objective is to recognize the spatial relationship of each text and graphics element in the chart so as to get the final understanding of the meaning of the chart. Figure 1 shows the overview of the learning-based scientific charts recognition system.

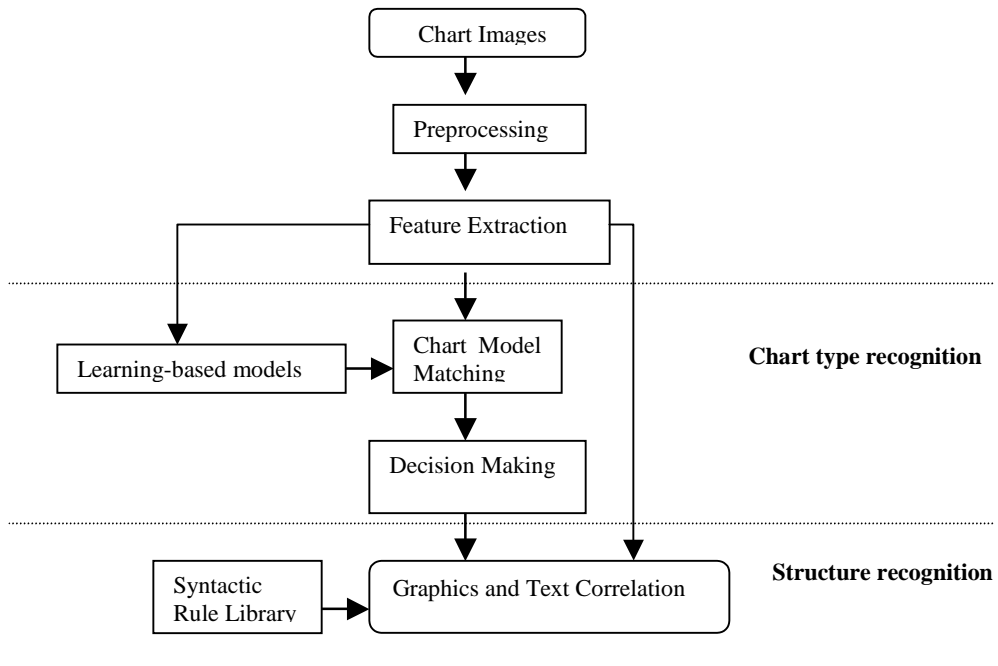


Figure1: Learning-based scientific charts recognition system

2.1 Preprocessing

Preprocessing is composed of two operations: connected component generation and size filtering for graphic chart image.

In our method, we assume the foreground and background to be black and white respectively. In this technique, we use eight-connectivity in the foreground and four-connectivity in the background to group pixels into distinctive components. The output of this operation is labeled components. Each component has a feature vector to identify its property (label, top and bottom point position of bounding box, mass, centroid position, height and width of bounding box, density).

In the connected components set, the height and width of graphic elements are usually larger than those of text elements. In our method, we use four filters to separate components into text elements and graph elements. These four filters are height filter, width filter, ratio of mass to height filter, ratio of mass to width filter.

2.2 Chart type recognition

This recognition stage is composed of three procedures: learning-based models construction, chart model matching and decision-making.

In the construction of learning-based models, data are trained to stabilize either the hidden Markov models or neural network. Details will be given in section 3.

In chart model matching of hidden Markov models, the probability $P(O|\lambda_k)$ for each different model λ_k is calculated given an observation sequence vector O . The model matching in feed-forward neural network is to compute the outputs for each input sequence. See section 3 for details.

For hidden Markov models, decision-making is to select the optimal chart model λ in K models with the highest probability using the following equation.

$$\lambda = \arg \max_{1 \leq k \leq K} [p(O | \lambda_k)] \quad (1)$$

In the neural network, we use win-take-all strategy to choose the largest output as the best match.

2.2 Structure recognition

Structure recognition is the stage to analysis the symbol arrangement in the graphic images. Graphic and text correlation is to combine the syntactic rules of different charts and the variant features extracted from original image, correlate the text with its corresponding graphic variant feature. In this paper, we adopted the same strategy as in [6].

3 LEARNING-BASED RECOGNITION FRAME

In this section, two kinds of recognition frames are described: hidden Markov models and multi-layer feed-forward neural network.

3.1 Hidden Markov Models

Hidden Markov model is a finite state automaton with probabilistic state transition. T is the length of observation sequence. N denotes the number of states in the model. M is the number of observation symbols. The state sequence is denoted as $Q=\{q_1, q_2, \dots, q_N\}$. $V=\{V_1, V_2, \dots, V_M\}$, is a symbol observation sequence. The HMM can be represented as $\lambda=\{A,B,\pi\}$, where A is the state transition probability distribution, having the form $A=\{a_{ij}\}$, $a_{ij}=\text{P}(q_j \text{ at } t+1 | q_i \text{ at } t)$. $B=\{b_j(k)\}$, $b_j(k)=\text{P}(V_k \text{ at } t | q_j \text{ at } t)$ is the observation symbol probability distribution in state j . $\pi=\{\pi_i\}$, $\pi_i=\text{P}(q_i \text{ at } t=1)$, is the initial state distribution

The left-to-right hidden Markov models are a commonly used topology in speech recognition and handwritten character recognition because this topology is more suitable to model constrained temporal order. In our processing, the chart sequences show a recurring characteristic. The ergodic topology can represent such characteristic better than the left-to-right model. In this model, it is possible to reach any state from any other state.

3.1.1 Model construction

We now have constructed three kinds of chart models, bar chart model, line chart model and high-low-close model. High-low-close chart is commonly used in stock market, in which three series of values lie in sequence on the line or bar primitives.

In our work, we use a continuous hidden Markov model with a Gaussian mixture with each state.

$$b_j(O) = \sum_{m=1}^M c_{jm} \mathfrak{R}[O, \mu_{jm}, U_{jm}], \quad 1 \leq j \leq N \quad (2)$$

Where O is the observation vector, c_{jm} is the mixture coefficient of the m th mixture in state j , \mathfrak{R} is the Gaussian density.

To train the parameters for each model, we adopt the segmental K-mean algorithm to re-estimate the parameters.

The procedure is as follow:

1. Initialize the model parameters;
2. Cluster the training vectors into segments and calculate the probable density function parameters.
3. Sample the training data to calculate the new model parameters.
4. Compare the distance of the two HMM parameters. If it's within a specified range, then stop re-estimating, otherwise go to 2.

3.1.2 Chart model matching

Chart model matching is to calculate the probability of the observation vectors given a known model λ_k , namely $\text{P}(O|\lambda_k)$. $\text{P}(O|\lambda_k)$ is calculated using the dynamic programming procedure called the forward-backward procedure in just (TN^2) time

rather than $2TN^T$ as required in direct computation. The forward variable is defined as:

$$\begin{aligned} \alpha_t(i) &= P(O_1, O_2, \dots, O_t, q_i \text{ at } t | \lambda) \\ P(O | \lambda) &= \sum_{i=1}^N \alpha_T(i) \end{aligned} \quad (3)$$

3.2 Multi-layer feed-forward neural network

Multi-layer feed-forward neural network with back-propagation learning algorithm is a widely used supervised training method for neural nets. One of the reasons for its popularity is its incredible simplicity. But the slowness in convergence of standard back-propagation algorithm is still a drawback. We adopt sigmoid function as equation 4. Equation 5 is its derivative. For weight adjustment, equation 6 is used with momentum α . In order to speed up the convergence of training, a modified back-propagation algorithm is also applied by replacing the equation 5 with equation 5'.

$$g(u) = \frac{1}{1 + e^{-\beta u}} \quad (4)$$

$$g(u)' = \beta g(u)(1 - g(u)) \quad (5)$$

$$\Delta w_{ij}(t+1) = -\eta \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(t) \quad (6)$$

$$g(u)' = \beta g(u)(1 - g(u)) + 0.1 \quad (5')$$

In the experiments, the structure of neural network is composed of 150 input nodes, 3 output nodes and one hidden layer.

4 FEATURE EXTRACTION

The training features extraction is the most important stage for a better model parameter estimation. There are two kinds of principal features, principal invariant features and principal variant features. We extract principal invariant features for training and principal variant feature for information extraction. We first define the image transition function $f(x)$ to be the transition variance function between the background and the foreground of the image in the x direction.

4.1 Principal invariant feature extraction

We use the first order derivative of transition function to segment the chart image. The segmentation function is as follows:

$$\Phi(x) = \left\| f(x+1) - f(x) \right\| \quad (7)$$

We segment images at each position that $\Phi(x)$ is greater than zero. We call such positions as feature positions. The first invariant feature is $f(x)$, the number of transitions. The second and third invariant features we select are the two transition values of the transition function before and after each feature position x . Another common feature is the gradient angle at the feature positions. We extract two gradient angles from top to bottom and from bottom to top. All these five features make up the invariant feature vector.

The principal invariant feature vectors are composed of invariant feature vectors extracted at each feature position. An observation sequence vector is composed from the principal invariant feature vectors. Each chart image is represented by 30 feature vectors in the training process regardless of the actual size of the image.

4.2 Principal variant feature extraction

The principal variant feature vectors are extracted for later information extraction or chart understanding. These vectors record the locations of the transition feature at which the principal invariant features are selected. The maximum number of location features now is six in our system. They are also selected from top to bottom and from bottom to top to be consistent with the syntactic rules.

5 EXPERIMENTS AND DISCUSSION

We use 300 bar chart, 300 line chart and 240 high-low-close chart vectors to construct three different general models. Another 120 bar chart, 120 line charts and 110 high-low-close chart vectors are used for performance testing. Some of these chart feature vectors are generated from chart images downloaded from website. Some are from images generated by chart generation software, such as Microsoft Excel. We also use the recurring characteristics of the scientific charts to sequentially shift the feature vectors to generate new observation vectors due to the shortage of chart images. The following figures 2 to 5 are some of the training and testing examples of our recognition system.

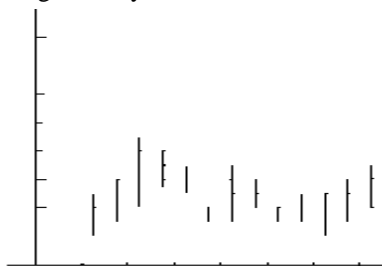


Figure 2. A high-low-close chart

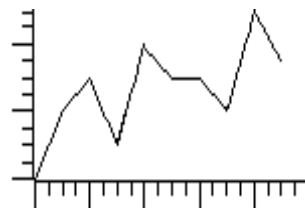


Figure3. A line chart

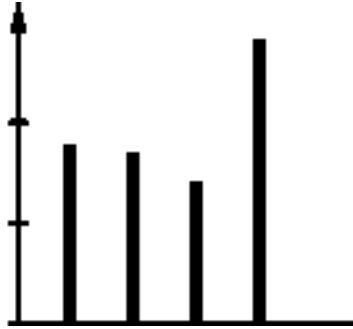


Figure 4. A bar chart

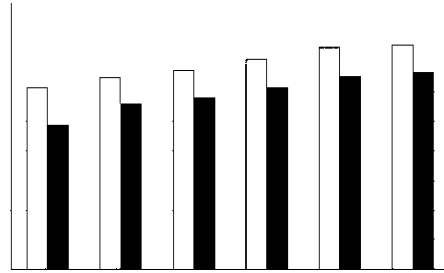


Figure 5. A combined bar chart

Experiments are done on both HMM and neural network. Experiments on HMM are done on ergodic models (EHMM) and left-right models (LHMM). The initial transition probability distribution of the EHMM is set to an equal probability. For LHMM, it has the constraint that transition from state i to state j is only allowed for $j \geq i$. Two parameters, the Gaussian mixture m and the number of hidden states n , are adjusted to get the best training result. In our experiment, m is set to 3.

In training the feed-forward neural network, the parameters α, η and the number of hidden units on the neural nets affect the performance of the neural network. Usually if the number of hidden units is larger than 10, the accuracies of testing on different network structure are similar. Table 1 shows the recognized charts for each category using different method. The parameters of two neural network are as follows: $\eta = 0.05, \alpha = 0.5$. In a whole, EHMM shows the best performance on three kinds of chart recognition. The accuracy of bar charts recognition would be improved if we classify different bar charts into smaller divisions. Modified BP does not perform better than the standard one in testing. But the speed of convergence of the former one is much faster than that of the latter one.

	Recognized bar charts (total 120)	Recognized line charts (total 120)	Recognized high-low-close charts (total 110)
EHMM	101/120	115/120	85/110
LHMM	86/120	100/120	76/110
Standard BP	93/120	97//120	82/110
Modified BP	90/120	94/120	87/'110

Table 1: Recognition accuracy of each chart category using different methods

Figure 6 shows a high-low-close chart that is misclassified as a line chart by EHMM. The two neural networks outperform the HMMs in testing the high-low-close charts. The accuracy of recognizing high-low-close charts is lower for both EHMM and LHMM. The high-low-close line segments are very close to each other. Besides

the irregularity of the X axis also lead to the observation sequence of the high-low-close chart behave like that of a line chart. In this case, the sequential property of hidden Markov model has no use to improve the accuracy of recognizing high-low-close charts.

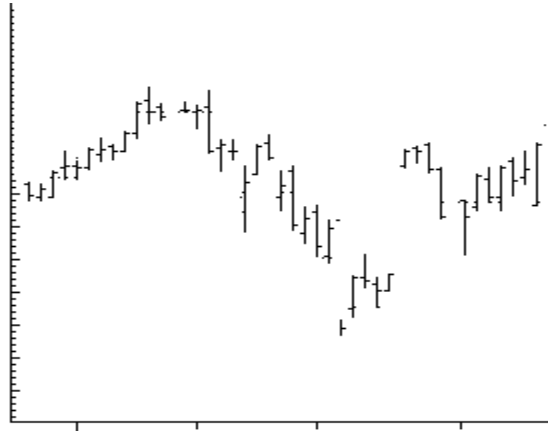


Figure 6. A high-low-close chart that is misclassified as a line chart by EHMM.

To get the computer readable meaning of charts is the final objective of chart recognition. The principal variant feature vectors store the structural information of graphic element for further analysis. As shown in the figure 7, the position at the cross signs are the structural features.

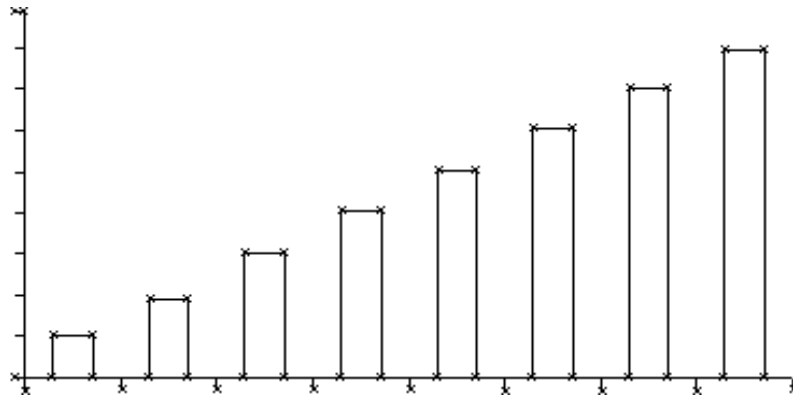


Figure 7. The location of structural features on the graphic image.(the cross signs)

6 CONCLUSION

In this paper, we focus our interest in the problem of scientific chart recognition that is still a difficult problem due the diversity of the chart types. We propose a learning-based paradigm to recognize chart images. Two types of learning-based methods are proposed. One is based on hidden Markov models. The other is based on neural network. We also present a novel principal feature extraction method. We first extract principal invariant feature vectors for model parameter estimation. Then we also encode the principal variant feature vectors and associate them with syntactical rule library for further chart information extraction. Experiments show learning-based method can get good results in recognizing different kinds of charts. To improve the recognition accuracy, combining learning-based chart type detection with graphic primitive extraction would be a good idea.

REFERENCE:

- [1] E.R.,Tuft, The visual display of quantitative information, Cheshire, CT, Graphics Press,1985
- [2] A.K. Chhabra, Graphics symbol recognition: an overview, Graphics recognition: algorithms and systems, GREC'97, pp68-79.
- [3] D. Blostein, General Diagram-recognition methodologies, Graphics recognition: methods and applications, GREC'95, pp106-122.
- [4] R.P. Futrelle *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] Y. Zhou and C L Tan, Hough-based Model for Recognizing Bar Charts in Document Images, SPIE conference on Document image and retrieval, 2001.
- [7] Y. Zhou and C L Tan, Hough technique for bar charts detection and recognition in document images, International Conference on Image Processing, 2000.
- [8] B.H. Juang and L.R. Rabiner, Mixture autoregressive hidden Markov models for speech signals, IEEE Transaction on Acoustic, Speech, Signal Processing, vol.ASSP-33,pp. 1404-1413,December, 1985
- [9] Rabiner, L.. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 1989
- [10] L. R. Rabiner and B. H. Juang, An Introduction to Hidden Markov Models, IEEE ASSP magazine, January 1986,pp 4-16

- [11] Kupiec, J. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, (6):225-242,1992
- [12] Chen and Kundu, An alternative to variable duration HMM in handwritten word recognition, in proceeding of the third international workshop on Frontiers Handwriting Recognition, Buffalo, NY, May, 1993, pp.48-54
- [13]"HMM Based On-Line Handwriting Recognition", J. Hu, M. Brown, W. Turin, *IEEE transactions on pattern analysis and machin*, **18**(10), oct 1996.
- [14] A.Gillies, Cursive word recognition using hidden markov models, in Proceeding of United States Postal Service Advanced Technology Conference, 1992, pp. 557-563
- [15] Magdi A. Mohamed and Paul Gader, Generalized Hidden Markov Models-Part II: Application to Handwritten Word Recognition, *IEEE Transactions on Fuzzy Systems*:8(1), 2000,pp82-94
- [16] G. Kopec and P. Chou, "Document Image Decoding Using Markov Source Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 16, No. 6, June 1994, pp. 602-617.