

Chart Analysis and Recognition in Document Images

Yanping Zhou, Chew Lim Tan

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

Abstract

Hidden Markov Models are a probabilistic modeling tool for time series data. It has been successfully applied to many areas, such as speech recognition, hand-written character recognition, etc. In this paper, we present a novel statistical approach using ergodic hidden Markov models to recognize scientific charts. We also present a newly developed feature extraction method for chart images. Unlike traditional primitive-based diagram recognition method, our approach need not recognize the graphic primitives in charts thus bypassing the recognition error problem caused by the 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. Pictorial representation though rich in information content is much more complex and unwieldy to process than its textual counterparts.

In this paper, we focus on one particular type of pictorial representation, namely scientific charts. People use scientific charts such as line charts and bar charts to intuitively convey a clear analysis of commercial data and research data. Our work adopts a statistical approach to chart recognition by means of hidden Markov models.

Hidden Markov Models (HMMs) are a probabilistic modeling tool for time series data. They have been successfully applied in speech recognition [1,2,3] and part-of-speech tagging [4]. 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 [5,6,7]. Mohamed and Gader [8] are the first to apply continuous density hidden Markov models for a segmentation-free handwritten word recognition.

Currently diagram recognition works such as engineering drawings recognition [9] are based on primitive extraction due to the syntax complexity of these diagrams. Scientific chart recognition systems developed by Futrelle[10],Yokokura[11] and Zhou[12] also used primitive extraction method to recognize charts. If the image quality is good, they can present satisfactory chart recognition result 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 charts information in general.

In this paper, we construct hidden Markov models for different general chart categories from training image data. After finding the optimal states path for a test chart, we combine the high-level analysis to understand and extract the chart information using the syntactic rule library associated with that chart category. Therefore our extraction method bypasses the error caused by the primitive extraction.

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

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.

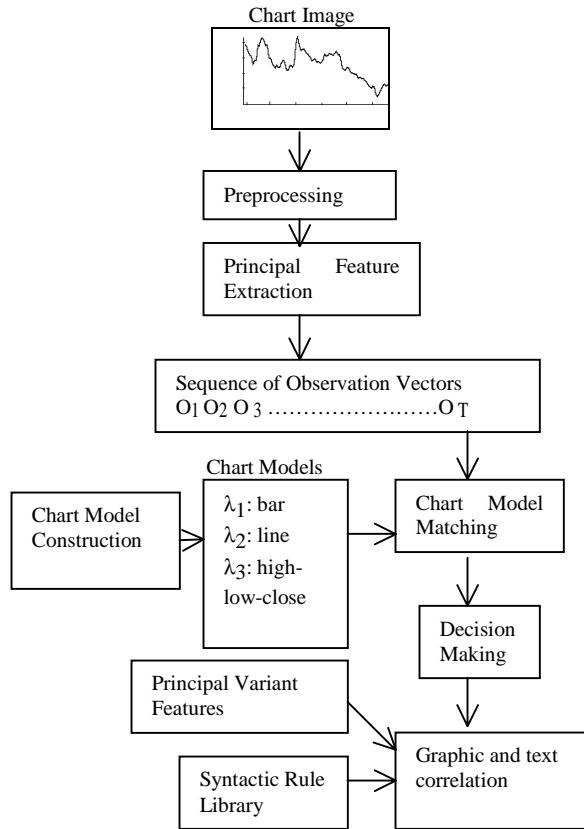


Figure1: Scientific charts recognition system

The structure overview of scientific chart recognition system is shown in figure 1. The basic processing elements are illustrated as follows:

Preprocessing: Use a size filter to sift through noise elements. Apply image rotation when it is found that positioned texts of values are in the horizontal direction.

Principal feature extraction: Extract principal invariant feature vectors for models training and testing. Extract also principal variant feature vectors for further query information extraction. Details will be illustrated in section 3.

Chart model construction: Use the training data to estimate the parameters of hidden Markov models. Details will be illustrated in section 3.

Chart model matching: Given an observation sequence vector O calculate $P(O|\lambda_k)$ for different models λ_k . See section 3 for details.

Decision making: Select the most optimal chart model λ in K models with the highest probability using the following equation.

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

Graphic and text correlation: Combining the syntactic rules of different charts and the variant features extracted from original image, correlate the text with its corresponding graphic variant feature.

3.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 $A=\{a_{ij}\}$, $a_{ij}=P(q_j \text{ at } t+1 | q_i \text{ at } t)$. $B=\{b_j(k)\}$, $b_j(k)=P(V_k \text{ at } t | q_j \text{ at } t)$ is the observation symbol probability distribution in state j . $\pi=\{\pi_i\}$, $\pi_i=P(q_i \text{ at } t=1)$, 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. Figure 2 shows the topology of our models.

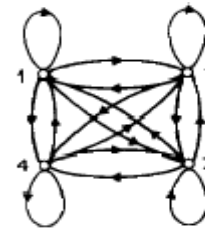


Figure2: A 4-state ergodic hidden Markov model

3.1 Chart 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.

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

Where O is the observation vector, C_{jm} is the mixture coefficient of the m th mixture in state j , \mathcal{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.

3.2 Chart model matching

Chart model matching is to calculate the probability of the observation vectors given a known model λ_k , namely $P(O|\lambda_k)$. $P(O|\lambda_k)$ 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:

$$\alpha_t(i) = P(O_1, O_2, \dots, O_t, q_i \text{ at } t | \lambda)$$

$$P(O | \lambda) = \sum_{i=1}^N \alpha_T(i) \quad (3)$$

3.3 Principal 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.

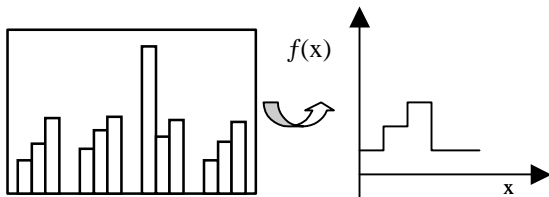


Figure 3: Feature extraction illustration

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

$$\Phi(x) = \left\| \frac{df(x)}{dx} \right\| \quad (4)$$

3.3.1 Principal invariant feature extraction. 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.

3.3.2 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.

4. 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 210 high-low-close chart vectors are used for performance testing. We also use the recurring characteristics of the scientific charts to sequentially shift to generate new observation vectors due to the shortage of chart images.

We do experiments on both the ergodic and the left-to-right models. We set the initial transition probability distribution of the ergodic model as of equal probability. As to the left-to-right model, it has its own constraints, that is transition from state i to state j is only allowed for $j > i$.

Figure 4 shows the performance of the two topology models with the same Gaussian mixture $m=3$, trained and tested both on the bar chart vectors. The line and high-low-close models used for comparison have four hidden

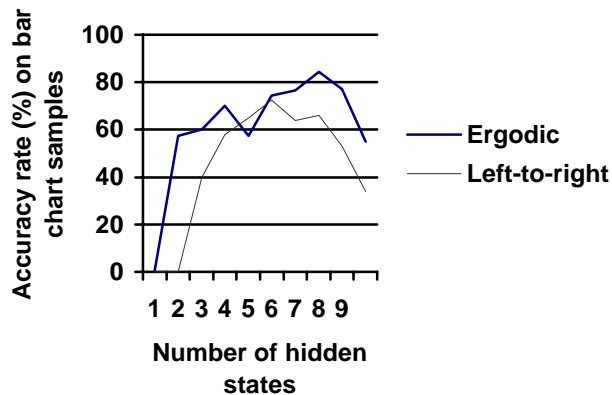


Figure4. Recognition accuracy rate of ergodic and left-to-right model tested on the 120 bar charts.

states. If the number of hidden states exceeds a particular number, the performance of both topologies degrades. This is because of the scarcity of the training data. In general, the performance of the left-to-right model is a somewhat lesser than that of the ergodic model. When we use Viterbi algorithm to find the optimal path given a model and an observation vector, the path nodes returned by the ergodic model will cycle in all the hidden states while the path nodes returned by left-to-right will end early at the last hidden state. Therefore the ergodic model is better in modeling the recurring characteristics of charts. But the ergodic model is more computationally expensive than the left-to-right model due to the complexity of its state transition matrix.

The test results for all the test samples are listed as follows:

	Bar charts	Line charts	High-low-close charts
Ergodic	84.5%	95.6%	92.7%
Left-to-right	72.3%	84.3%	82.7%

Table1: Recognition accuracy of each chart category

The results in table 1 are the best results tried in different hidden states. From the results, we can see the accuracy rate for line charts and high-low-close charts are higher than that of bar chart recognition. The former two kinds are simpler than the latter one. The above table shows that HMM based method is applicable for chart information extraction.

5. 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. Unlike the traditional primitive-based graphics recognition method, we use ergodic hidden Markov model to construct the general virtual chart models. 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. Our extraction system shows more tolerance in understanding similar chart types and makes the general scientific charts recognition more promising.

REFERENCE:

- [1] 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
- [2] Rabiner, L.. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 1989
- [3] L. R. Rabiner and B. H. Juang, An Introduction to Hidden Markov Models, IEEE ASSP magazine, January 1986,pp 4-16
- [4] Kupiec, J. Robust part-of-speech tagging using a hidden Markov model. Computer Speech and Language ,(6):225-242,1992
- [5] 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
- [6] "HMM Based On-Line Handwriting Recognition", J. Hu, M. Brown, W. Turin, IEEE transactions on pattern analysis and machin, **18**(10), oct 1996.
- [7] A.Gillies, Cursive word recognition using hidden markov models, in Proceeding of United States Postal Service Advanced Technology Conference, 1992, pp. 557-563
- [8] 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
- [9] 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.
- [10] R.P. Futrelle *et al.*, Understanding diagrams in technical documents, IEEE Computer, Vol.25, NO.7, pp75-78, 1992.
- [11] N. Yokokura and T. Watanabe, Layout-Based Approach for extracting constructive elements of bar-charts, Graphics recognition: algorithms and systems, GREC'97, pp163-174.
- [12] Y. Zhou and C L Tan, Hough technique for bar charts detection and recognition in document images, International Conference on Image Processing 2000