Computing and Informatics, Vol. 29, 2010, 165–182

IDENTIFICATION OF TECHNICAL JOURNALS BY IMAGE PROCESSING TECHNIQUES

Pei-Chun Wen

Legal Department
High Tech Computer Corp., Taipei, Taiwan, R.O.C.

Ling-Ling Wang

Department of Information Communication
Asia University
Wufeng Shiang, Taichung 41354, Taiwan, R.O.C.
e-mail: ling@asia.edu.tw

Manuscript received 25 October 2006; revised 6 March 2009
Communicated by Ivan Plander

Abstract. The emphasis of this study is put on developing an automatic approach to identifying a given unknown technical journal from its cover page. Since journal's cover pages contain a great deal of information, determining the title of an unknown journal using optical character recognition techniques seems difficult. Comparing the layout structures of text blocks on the journals' cover pages is an effective method for distinguishing one journal from the other. In order to achieve efficient layout-structure comparison, a left-to-right hidden Markov model (HMM) is used to represent the layout structure of text blocks for each kind of journal. Accordingly, title determination of an input unknown journal can be effectively achieved by comparing the layout structure of the unknown journal to each HMM in the database. Besides, from the layout structure of the best matched HMM, we can locate the text block of the issue date, which will be recognized by OCR techniques for accomplishing an automatic journal registration system. Experimental results show the feasibility of the proposed approach.

Keywords: Journal identification, automatic journal registration, hidden Markov model, layout structure, journal title

1 INTRODUCTION

1.1 Research Motivation

Many libraries subscribe for thousands of technical journals every year. Registration of current journals is time consuming for librarians. To solve this problem, what is required in a library is an automatic journal registration system which can identify an unknown journal by its title and issue date. In this study, we develop an effective method of determining the title of an unknown journal from its text-block layout structure on the cover image. Besides, we locate the text block of its issue date. Then the issue date will be recognized by optical character recognition (OCR) techniques. Hence automatic journal registration can be achieved.

1.2 Survey of Document Processing

With the advent of high-resolution, low-cost scanners and high-capacity storage devices, researchers on document processing have attracted worldwide interest. Researches on automatic document processing can be decomposed into three major subjects: document analysis, document understanding, and character segmentation and recognition.

The purpose of document analysis is to find an appropriate geometric (layout) structure to represent a document image. It decomposes a document into several coherent components, such as text lines, tables, graphics, photographs, etc. Many techniques for document analysis have been proposed. There are mainly two approaches to extracting the geometric structure from a document image [1]: the top-down [2, 3, 17, 18] and bottom-up [4, 5, 19–22] approaches. The top-down method is fast and effective for dealing with documents of specific format, but is suitable only for documents with simple geometric structures. The bottom-up method is time consuming, but possible to develop algorithms that are applicable to arbitrary document layouts. A hybrid method, combining the two approaches, may achieve a better result [6].

After the geometric structure is obtained from a document image, the document understanding process first maps the geometric structure into a logical structure by the logical relationship among components extracted in document analysis, and then distinguishes the attribute of each component. There are several mapping methods proposed for document understanding. Tree transformation techniques [7, 10] are effective in acquiring a logical relationship among components. But the tree is usually difficult to construct. The knowledge-based approaches, such as applying formatting knowledge [12] or form definition languages [11] to define layout rules, are useful for identifying the characteristics of each component. However, they are suitable for specific forms of documents, such as tables, letters, checks, envelopes, etc. They are not effective for highly complex documents.

After image understanding, the content of each text block can be obtained by OCR techniques, which convert a document image into an editable text file. Researches on automatic text segmentation and character recognition have been widely studied after the 1980's [13–16]. The OCR techniques have greatly matured in recent years.

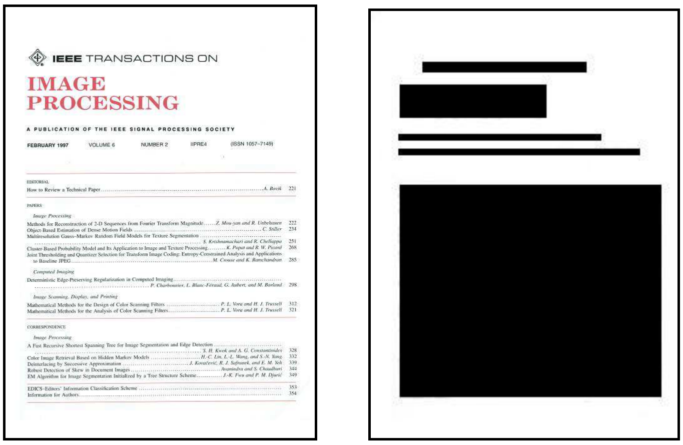
Journal cover pages may change with issues and can be considered as complex advertising documents, which are full of variety. It seems ineffective to locate the journal-title block from the cover page and then perform OCR techniques on the block to identify the journal. This is because there are many cases which lead to difficulties in performing OCR techniques on journal cover pages, such as diversity of character typefaces, overlapping of characters, various directions of character typesetting, and existence of graphics. Hence flexible characteristic structures (models) and identification techniques are desired for tackling the journal-title determination problem. In this study, we wish to find a flexible characteristic structure to represent the text-block layout of a journal cover image and to determine its title without using OCR techniques.

1.3 Overview of Proposed Method

Sometimes a reader can identify a familiar journal by catching a glimpse of the journal's cover page. In this case, the reader has no need to recognize each character of the title block because the layout structure of text blocks on the cover page is enough for identification. Based on the observation, we wish to develop a journal identification approach to reducing the load on OCR. Since emphasis of this study is on identification of an unknown journal, we assume that the input of our approach is a set of text blocks extracted from the journal's cover image [23]. Characters in the same text block are assumed to be of the same font and size, as shown in Figure 1. By storing the text-block layout structure of each kind of journal in a database, we can compare the layout structure of the unknown journal with each in the database. Then the unknown title can be determined from the one whose layout structure is best matched.

In order to achieve efficient layout-structure comparison, a model which can effectively represent the spatial relationship of text blocks on the journal cover page is necessary. Moreover, the model should also be adapted to variations in different issues of a journal. HMMs (Hidden Markov Models) are flexible statistical models in the representation of objects' relationship. They have been shown to perform very well in a variety of applications [24–30]. Especially, HMMs are tolerant of object segmentation errors in matching. In this study, one-dimensional (1-D) left-to-right Bakis HMMs [26, 28] are used to represent the layout structures. Each kind of journal corresponds to a 1-D left-to-right HMM in the database, where the layout structure of text blocks on the journal cover page is stored. Accordingly, identification of an unknown journal can be effectively achieved by comparing the layout structure of the unknown journal to each HMM in the database.

Once the best matched HMM in the database is found, the text block of the issue date of the given unknown journal can be located using the layout structure of the best matched HMM. In the future, we will apply OCR techniques to re-



IEEE TRANSACTIONS ON

IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

FEBRUARY 1997 VOLUME 6 NUMBER 2 IIPRE4 (ISSN 1057-7149)

EDITORIAL

How to Review a Technical Paper .... *A. Bovik* 221

PAPERS

Image Processing

Methods for Reconstruction of 2-D Sequences from Fourier Transform Magnitude ...... *Z. Mou-yan and R. Unbehauen* 222

Object-Based Estimation of Dense Motion Fields .... *C. Stiller* 234

Multiresolution Gauss–Markov Random Field Models for Texture Segmentation .... *S. Krishnamachari and R. Chellappa* 251

Cluster-Based Probability Model and Its Application to Image and Texture Processing .... *K. Popat and R. W. Picard* 268

Joint Thresholding and Quantizer Selection for Transform Image Coding: Entropy-Constrained Analysis and Applications to Baseline JPEG .... *M. Crouse and K. Ramchandran* 285

Computed Imaging

Deterministic Edge-Preserving Regularization in Computed Imaging .... *P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud* 298

Image Scanning, Display, and Printing

Mathematical Methods for the Design of Color Scanning Filters .... *P. L. Vora and H. J. Trussell* 312

Mathematical Methods for the Analysis of Color Scanning Filters .... *P. L. Vora and H. J. Trussell* 321

CORRESPONDENCE

Image Processing

A Fast Recursive Shortest Spanning Tree for Image Segmentation and Edge Detection .... *S. H. Kwok and A. G. Constantinides* 328

Color Image Retrieval Based on Hidden Markov Models .... *H.-C. Lin, L.-L. Wang, and S.-N. Yang* 332

Deinterlacing by Successive Approximation .... *J. Kovačević, R. J. Safranek, and E. M. Yeh* 339

Robust Detection of Skew in Document Images .... *Avanindra and S. Chaudhuri* 344

EM Algorithm for Image Segmentation Initialized by a Tree Structure Scheme .... *J.-K. Fwu and P. M. Djurić* 349

EDICS–Editors' Information Classification Scheme 353

Information for Authors 354

Fig. 1. Example of text-block extraction: a) a journal cover image; b) the corresponding text-block layout

cognize characters in this text block. Hence registration of current journal can be accomplished automatically by computers.

The rest of this paper is organized as follows. Section 2 describes the proposed journal identification approach based on HMMs. Section 3 presents several experimental results. Section 4 contains conclusions of the paper and our future work.

2 THE PROPOSED APPROACH

The input of our proposed approach is an image scanned from an unknown journal cover page, whose text blocks are assumed to have been extracted. In the database we store a set of 1-D left-to-right HMMs, each of which corresponds to the layout structure of a kind of journal. Comparing the layout structure of text blocks on a given unknown journal cover page to each HMM in the database, we can find the best matched HMM and then identify the unknown journal. Exhaustive matching with HMMs in the database is time consuming. Instead, pre-classifying the set of journals (or HMMs) in the database is a feasible way to reduce the computational cost. In the following, we will explicate each component of the proposed approach, including

1. construction of an HMM for each kind of journal,
2. pre-classification of journals,
3. matching of the input with HMMs,
4. post-classification of journals.

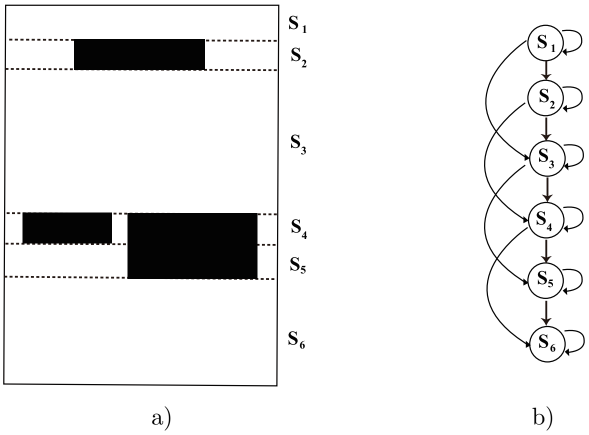
2.1 HMM Construction

2.1.1 HMM Construction for a Certain Issue of a Journal

In this subsection, we introduce how to construct a 1-D Bakis HMM [26, 28] for a certain issue of a journal. First, the image of its cover page is segmented into a set of text blocks, as shown in Figure 1 b), by an adaptive segmentation method [23]. Next, the corresponding HMM is to be constructed. We present our 1-D "top-to-down" HMM construction approach through a simple example shown in Figure 2. In this example, a segmented image is partitioned into a set of horizontal sections, as shown in Figure 2 a), according to the layout of text blocks on its cover image. All horizontal profiles in a horizontal section are identical. Each horizontal section corresponds to a state in the HMM, as shown in Figure 2 b). The number of states in the HMM equals the number of horizontal sections in the segmented image. The states corresponding to sections of large height in the segmented image are considered to have low probabilities of transitions to other states, and conversely for the states corresponding to sections of small height. Hence the initial state transition probabilities $a_{i,i}$ (from state S_i to itself), $a_{i,i+1}$ (from state S_i to state S_{i+1}), and $a_{i,i+2}$ (from state S_i to state S_{i+2}) are heuristically set as follows:

$$\begin{aligned} a_{i,i} &= \frac{\log(h_i)}{\log(N)} \\ a_{i,i+1} &= [1 - a_{i,i}] \times a_{i+1,i+1} \\ a_{i,i+2} &= [1 - a_{i,i}] \times [1 - a_{i+1,i+1}] \end{aligned}$$

where h_i is the height of the i^{th} horizontal section, and N is the total number of horizontal profiles in the segmented cover image.



a) b)

Fig. 2. a) A segmented journal cover image; b) the corresponding HMM

The character size (or fontsize) and spatial relationship of text blocks in each horizontal section are both good features for journal identification, which are recorded in the corresponding state of the HMM. We define layout feature, $l(i)$, to characterize the spatial information about text blocks of the i^{th} horizontal section as follows:

$$l(i) = \sum_{j=0}^{n-1} m_i(j) \tag{1}$$

with

$$m_i(j) = \begin{cases} j & \text{if } P(i,j) \text{ is on a text block} \\ 0 & \text{if } P(i,j) \text{ in on the background} \end{cases}$$

where $P(i,j)$ is the j^{th} pixel in a horizontal profile of the i^{th} horizontal section, and n is the total number of pixels in the profile. Figure 3 shows several horizontal sections with their layout features. In Figure 3 a), there is only a single text block in each of the three horizontal sections. The three text blocks are of the same width but located in different positions. Their layout features are distinct. Therefore block locations can be characterized by the layout features. However, we may have different horizontal sections which have the same layout feature values, as shown in Figure 3 b). Therefore we use another feature, $c(i)$, which is the character size of the text block in the i^{th} horizontal section, to characterize this section. If there are more than one text block in a horizontal section, the character size of the widest text block is chosen. The pair of layout feature and character size is then defined as the representative observation of the i^{th} state in the HMM. They are used for calculating observation probabilities in the HMM matching process.

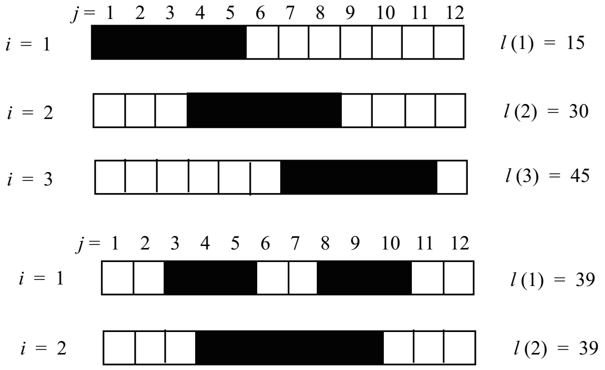


Fig. 3. Horizontal sections with their layout features

2.1.2 HMM Construction for Each Kind of Journal

Storing an HMM for each issue of a journal is costly both as to storage space and matching time. In this study, a single HMM is constructed for each kind of journal. Hence it is necessary that the HMM can appropriately characterize the journal. However, only selecting a certain issue to construct the HMM of a journal is not reliable. In the study, more than one issue of a journal are used to construct the HMM of the journal. In HMM construction, we first select for each kind of journal a released issue which has the maximum number of horizontal sections. Then, an HMM is constructed for the issue by the method mentioned in Section 2.1.1. Next, several of the remaining issues are randomly selected to estimate some parameters in the constructed HMM. Each of the selected issues (called the training issues) is to be matched with the constructed HMM by the Viterbi algorithm [31]. The detailed HMM matching process is given in Section 2.3. After matching, we obtain an optimal state sequence for each training issue. These sequences are then used to estimate some parameters in the corresponding states of the constructed HMM.

For HMM matching to be flexible in the journal identification process, in each state of the HMM, a Gaussian distribution is used for estimating the probability that an input layout feature and character size are observed in the state (a detailed description is given in Section 2.3). The optimal state sequences of the training issues are used to estimate the means and standard deviations of the Gaussian distribution in each state. Let the optimal state sequence of the i^{th} training issue be $y_{i1}, y_{i2}, \ldots, y_{iN}$, where y_{ij} denotes a state number and N is the number of horizontal profiles in the cover image. Let $l(i, j)$ denote the corresponding input layout feature of y_{ij}. Then the mean $\mu_l(k)$ and the standard deviation $\sigma_l(k)$ of the layout feature observed in the k^{th} state are estimated as follows:

$$\mu_l(k) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} l(i, j) \times f_k(y_{ij})}{\sum_{i=1}^{M} \sum_{j=1}^{N} f_k(y_{ij})} \tag{2}$$

$$\sigma_l^2(k) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} [l(i, j) - \mu_l(k)]^2 \times f_k(y_{ij})}{\sum_{i=1}^{M} \sum_{j=1}^{N} f_k(y_{ij})} \tag{3}$$

where M is the number of training issues and f_k is a function defined as follows:

$$f_k(y_{ij}) = \begin{cases} 1 & \text{if } y_{ij} = k \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, the mean $\mu_c(k)$ and the standard deviation $\sigma_c(k)$ of the character size observed in the k^{th} state are estimated as follows:

$$\mu_c(k) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} c(i, j) \times f_k(y_{ij})}{\sum_{i=1}^{M} \sum_{j=1}^{N} f_k(y_{ij})} \tag{4}$$

$$\sigma_c^2(k) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} [c(i, j) - \mu_c(k)]^2 \times f_k(y_{ij})}{\sum_{i=1}^{M} \sum_{j=1}^{N} f_k(y_{ij})}. \tag{5}$$

Besides, by using the obtained optimal state sequences of the training issues, the state transition probabilities $a_{k,k}$ (from state S_k to itself), $a_{k,k+1}$ (from state

S_k to state S_{k+1}), and $a_{k,k+2}$ (from state S_k to state S_{k+2}) can be represented as follows:

$$\begin{aligned} a'_{k,k} &= \frac{\log(h'_k)}{\log(N)} \\ a'_{k,k+1} &= [1 - a'_{k,k}] \times a'_{k+1,k+1} \\ a'_{k,k+2} &= [1 - a'_{k,k}] \times [1 - a'_{k+1,k+1}] \end{aligned}$$

where $a'_{k,k}$, $a'_{k,k+1}$, and $a'_{k,k+2}$ are the reestimated state transition probabilities of $a_{k,k}$, $a_{k,k+1}$, and $a_{k,k+2}$, respectively, and h'_k is defined as

$$h'(k) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} f_k(y_{ij})}{M}.$$

2.2 Pre-Classification of Journals

It is clear that the computational cost of HMM matching can be reduced if we classify the models in advance. The physical size of a journal's cover page is a good feature to classify journals. Hence the two size features, height and width of the cover page, are used for pre-classification. Based on the size features of a cover page, only a few candidates (HMMs) in the database are selected for matching. To achieve this purpose and to refrain from exhaustive comparison of the size features for each journal, a size-checking method is used and described as follows.

First, we define a size array. The size array records the cover page sizes of each kind of journal. For example, the size array for journals in Table 1 is given in Figure 4. Each entry (i, j) in the size array records the journals whose width is i units and height is j units. As to our measurements, the width of most technical journals is from 10 to 25 cm and the height is between 20 and 35 cm. Note that the values of width and height are quantized to reduce the storage space of the size array. In our experiments, the quantized unit is 0.5 cm. Let $s_1 = (x_1, y_1)$ and $s_2 = (x_2, y_2)$ be two entries of the size array. The size difference between s_1 and s_2, denoted by $d(s_1, s_2)$, is given below:

$$d(s_1, s_2) = |x_1 - x_2| + |y_1 - y_2|$$

Given an unknown journal, we select only the journals which have a small size difference (e.g., $d \leq 2$) with the unknown one to do HMM matching.

2.3 Matching

We apply the Viterbi algorithm [31] to measure the similarity of the input and each HMM. A multivariate Gaussian distribution is used in the computation of the observation probability of each state during the Viterbi matching process. After comparing the input with each HMM, we obtain a similarity measure. By the

journal	cover page size	
	width (cm)	height (cm)
1	17.0	25.1
2	15.1	22.2
3	16.6	23.4
4	16.0	23.0
5	13.2	20.9
6	14.8	24.2

Table 1. Cover page size of several journals in the database

	21		22		23		24		25
13	5								
14									
15			2				6		
16					4				
						3			
17									1

Fig. 4. The corresponding size array of journals in Table 1

similarity measures, the best matched HMMs can be selected. Detailed description is given in the following.

Given a journal to be identified, we scan its cover page and segment its image to extract text blocks on the cover page [23]. Using Equation (1), we then calculate the layout feature and character size for each horizontal profile in the image. The set of layout features and character sizes are used as the input sequence to the Viterbi matching algorithm. Since the 2-D image is converted to a 1-D sequence, much computational cost can be reduced in the statistical matching process. It is noticeable that we do not compute the observation probability distribution for each state in the HMM until HMM matching is performed. An advantage is that matching of the input with HMMs in the database becomes flexible.

The probability $b_i(x)$ of observing in the i^{th} state the layout feature and character size of the x^{th} horizontal profile in the input image is computed as follows:

$$b_i(x) = \frac{1}{2\pi\sigma_l\sigma_c} \exp\left\{ -\frac{1}{2} \left[\frac{(l(x) - l(i))^2}{\sigma_l^2} + \frac{(c(x) - c(i))^2}{\sigma_c^2} \right] \right\}$$

where $l(x)$ and $c(x)$ are the layout feature and character size of the x^{th} horizontal profile in the input image, respectively, and $l(i)$ and $c(i)$ are the layout feature and character size recorded in the i^{th} state of the HMM, respectively. In the journal identification process, we compute the standard deviations σ_l and σ_c using Equations (3) and (5), respectively; however, they are set as 1 000 and 5 in the HMM construction process (i.e., when HMM matching is performed for a training issue). The initial state probability π_i at state S_i is heuristically set as

$$\pi_i = \begin{cases} 0.5 & \text{if } i = 1 \text{ or } 2 \\ 0 & \text{elsewhere.} \end{cases}$$

We assume that HMM matching starts from the first or second state of an HMM. Hence, the values of initial state probabilities at the top two states are larger than those at the bottom ones. The statistical matching between the input image and each 1-D HMM can be achieved by the Viterbi algorithm.

After matching the input with HMMs in the database, we select the journals whose layout structures on the cover pages are much similar to those of the input. These are considered as candidates. To still reduce the number of candidates obtained, the journals with similar layout structures can be combined to the same class. Then, instead of constructing an HMM for each journal, we construct a generalized HMM for the layout-similar journals. The method described in Section 2.1.2 can also be used to construct the generalized HMM. As a result, the number of HMMs in the database is reduced and the computational cost in matching decreases.

2.4 Post-Classification and Post-Processing of Journals

After HMM matching, we obtain several candidates which are best matched with the given input. These candidates have similar layout structures with the input. To choose the correct one among these candidates, some other features should be used for further checking. For each candidate, we can find in the input a text block which is best matched with the title block in the candidate from the HMM matching result. A technical journal usually has the same character and background colors in the title block for all issues. Hence the two colors can be used as features to check if a candidate should be remained. If there is a great difference between colors of a candidate and the input, we remove the candidate. However, the color checking is not done for journals whose colors on the cover pages are changed with issues.

When the best matched candidate is selected, we can locate the issue-date block using the HMM matching result of the input with this candidate. In the future, we will use OCR techniques to recognize characters in the text block for automatic registration. The researches concerning OCR techniques are our future work.

3 EXPERIMENTAL RESULTS

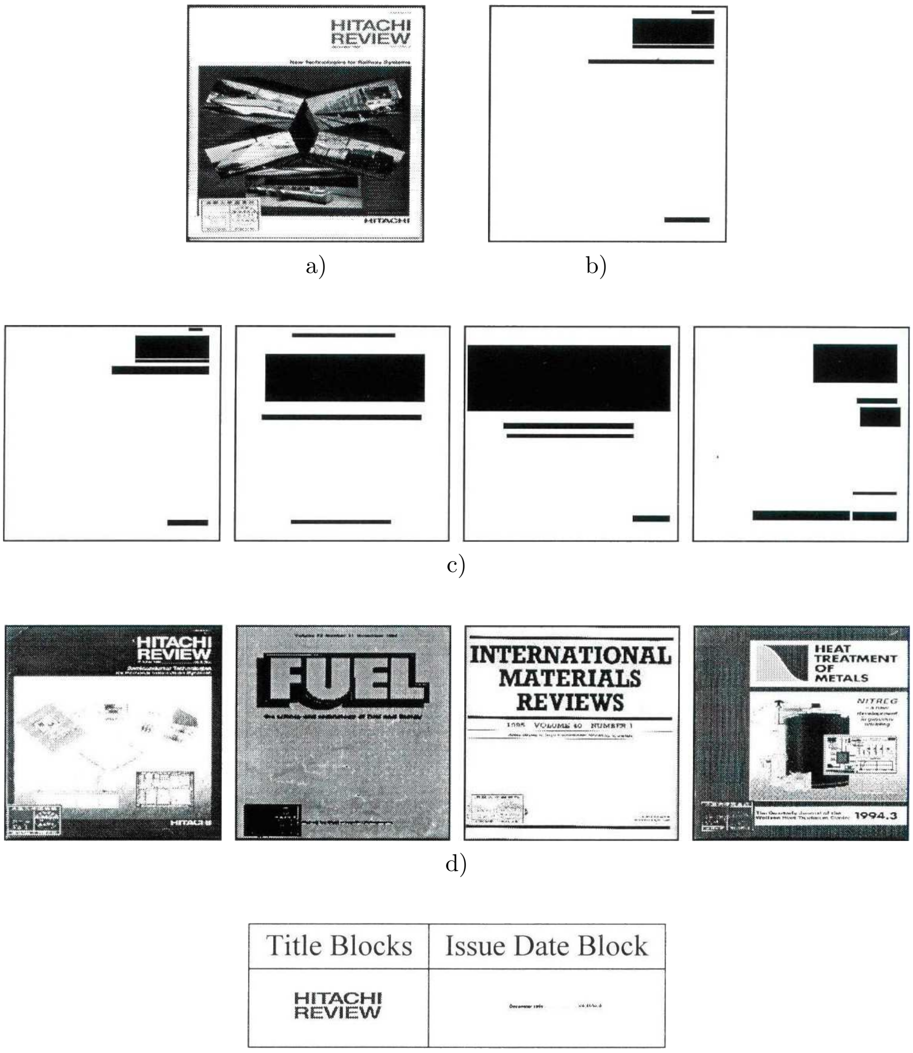
The effectiveness of the proposed method is demonstrated with several examples in this section. There are 255 kinds of technical journals used for testing. The corresponding HMM of each kind of journal is stored in the database. In journal identification, a given unknown journal cover image is first segmented into text blocks, and then the layout feature and character size of each horizontal profile are used as input to the Viterbi matching algorithm. There are several experimental results shown in Figures 5–8. Note that the mark pasted in the left bottom of each cover page is not processed in our experiments. They are pasted when received by the librarians. Four images which best match the unknown journal cover page are displayed in descending order in terms of their matching scores in these figures from left to right. Text blocks of the title, issue date, and table of contents on the given cover image, which are extracted by the HMM matching result of the input with the best matched one, are also shown in these figures.

In Figures 5 and 6, the best matched image is the desired one though its background is different from that of the input. Especially in Figure 6, the layout structures of the input and the best matched one are different. For the journal Circuits and Devices, the locations of top three text blocks on the cover page of each issue are invariant, but the others may change a lot. Hence in HMM constructions, its HMM states corresponding to the text blocks which change with issues are assigned larger standard deviation values for the layout features (by Equation (3)). Therefore the desired one can obtain a high matching score and can be retrieved. In Figure 8 an example is shown that illustrates the proposed method works well when the title characters are overlapped.

Numerous experiments have also been performed to verify the effectiveness of the proposed approach. For each input, if the desired image is one of the first three retrieved ones, the retrieval is considered as successful, else it is acceptable if the desired image is one of the first six. However, the retrieval is unacceptable elsewhere. In the experiments where no training issue is used for HMM parameter reestimation, about 85 per cent are successful, 10 per cent acceptable, and 5 per cent unacceptable. If we select 5 different issues foe each kind of technical journal for HMM parameter reestimation, the successful retrieval rate becomes 90 per cent and the unacceptable retrieval rate becomes 2 per cent. If color checking is applied for further checking, the successful retrieval promotes to 93 per cent. A main reason of fail retrieval is that with some journals the cover pages change a lot with issues.

4 CONCLUSIONS

An effective approach to identifying an unknown journal based on HMMs is proposed in this study. From the experiments, HMMs are shown insensitive to little variation of the character size and spatial relationship of text blocks on the journal cover page. Using HMMs results in effective matching. The matching between an input unknown journal with each HMM in the database can be easily achieved by computing the probability that the input is generated from the HMM. The feasibility of the proposed approach has been illustrated by the experimental results. The journals whose HMMs are best matched can be retrieved with high accuracy. Using the HMM matching result, we can also locate the title block, issue-date block, and table-of-contents block of the input (if they are on the cover page). In the future, we will use OCR techniques to recognize characters in these text blocks. Finishing an automatic



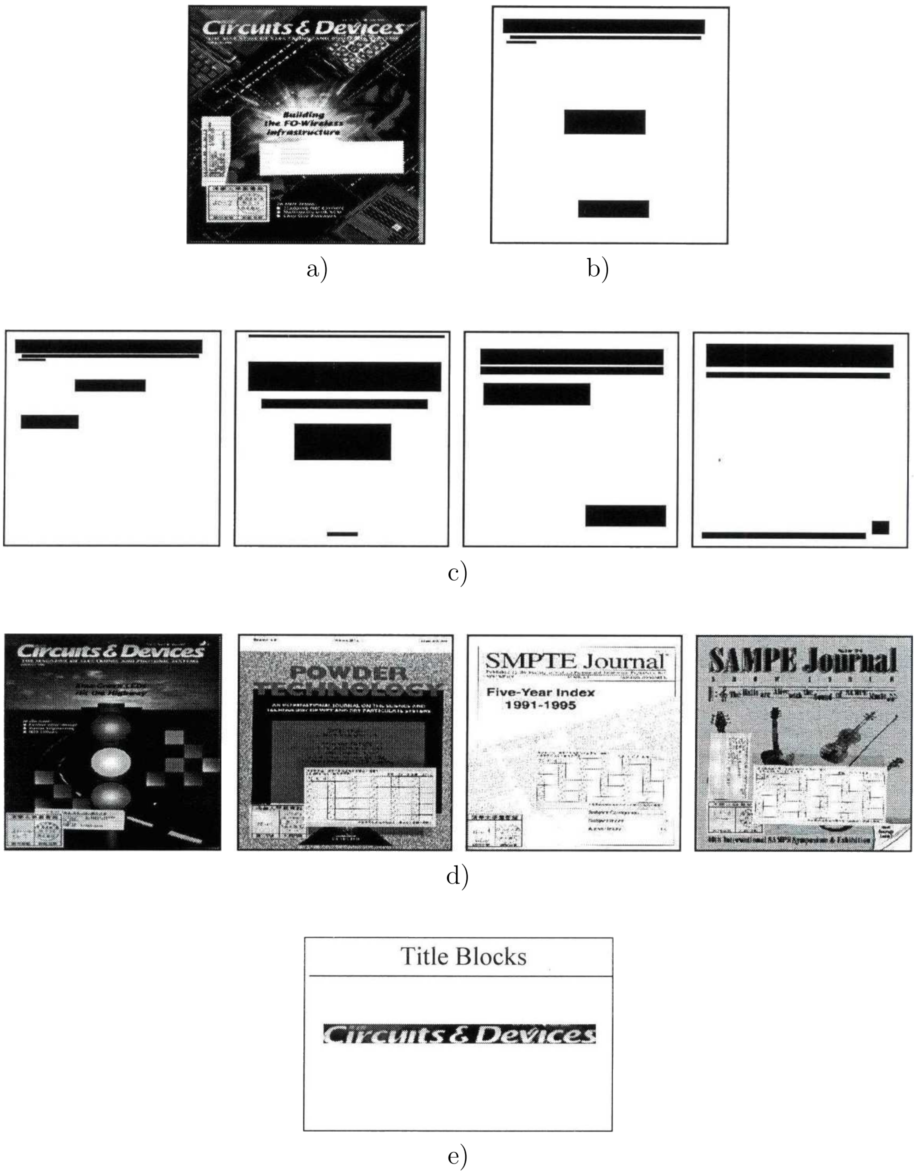
a) b)

c)

d)

e)

Fig. 5. Matching results: a) given unknown journal cover image; b) text-block layout of a); c) four best matched layouts; d) corresponding cover images of c); e) text blocks of the title and issue date extracted from the given input



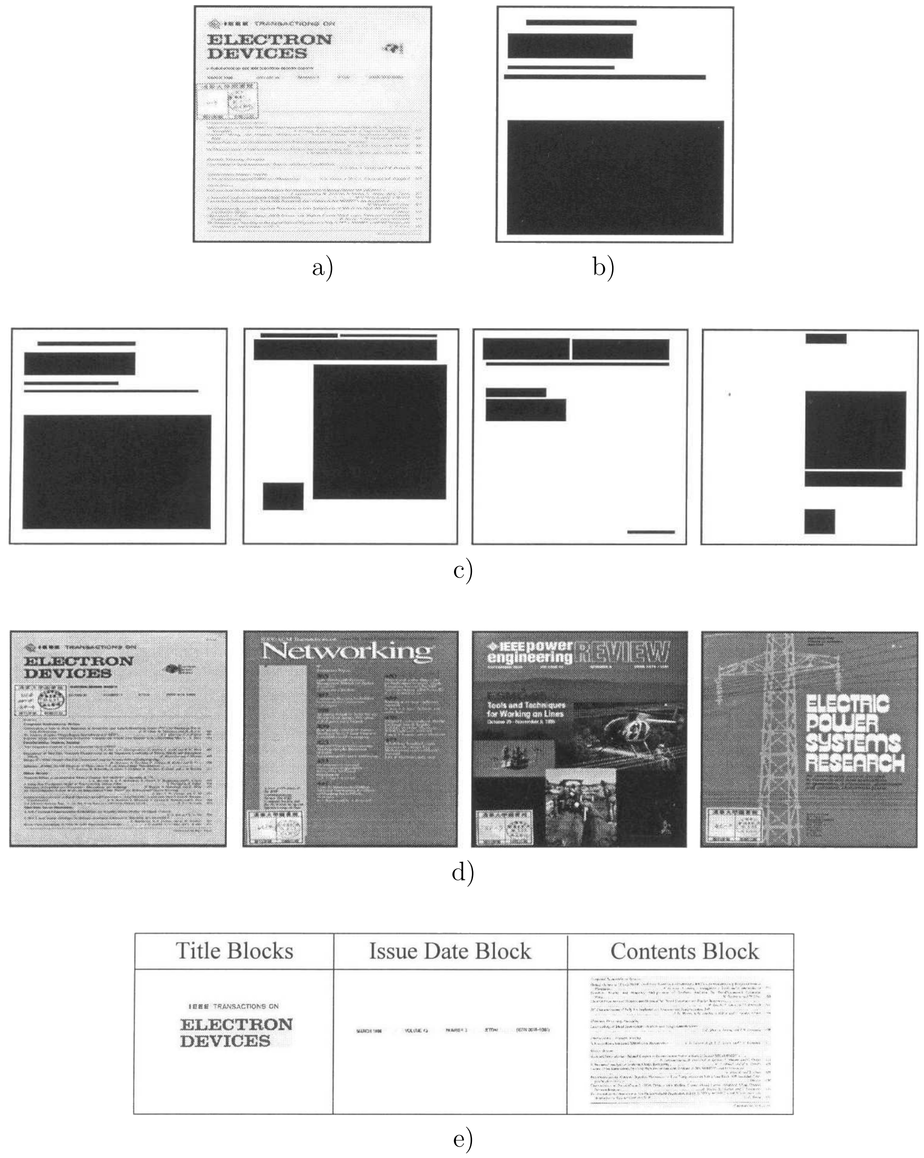
a) b)

c)

d)

e)

Fig. 6. Matching results: a) given unknown journal cover image; b) text-block layout of a); c) four best matched layouts; d) corresponding cover images of c); e) text blocks of the title extracted from the given input


IEEE TRANSACTIONS ON
ELECTRON DEVICES
Networking
IEEE power engineering REVIEW
Tools and Techniques for Working on Lines
ELECTRIC POWER SYSTEMS RESEARCH
Title Blocks
Issue Date Block
Contents Block

a) b)

c)

d)

e)

Fig. 7. Matching results: a) given unknown journal cover image; b) text-block layout of a); c) four best matched layouts; d) corresponding cover images of c); e) text blocks of the title, issue date, and table of contents extracted from the given input

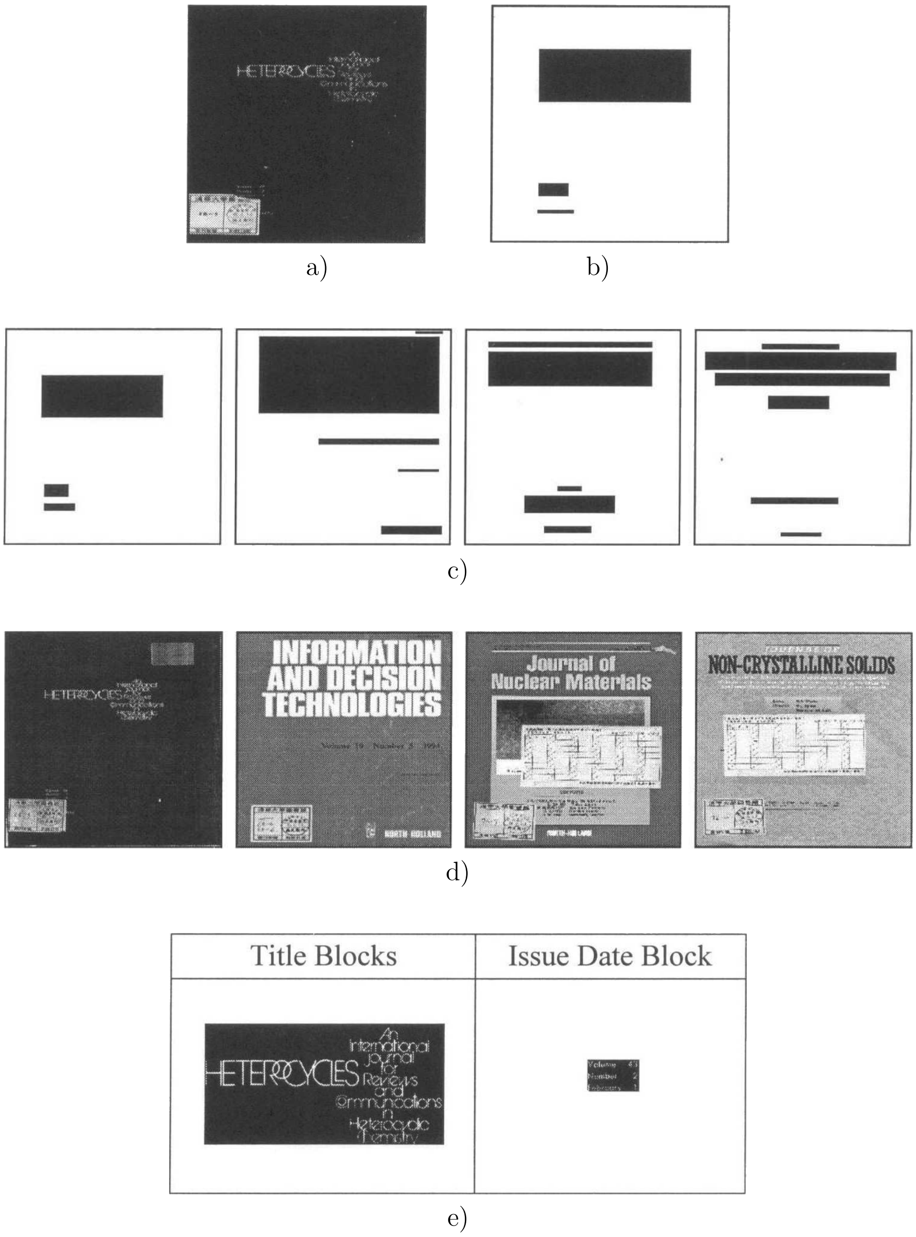

HETEROCYCLES
a)
b)
c)
INFORMATION AND DECISION TECHNOLOGIES
Journal of Nuclear Materials
NON-CRYSTALLINE SOLIDS
NORTH-HOLLAND
d)
Title Blocks
Issue Date Block
An International Journal for Reviews and Communications in Heterocyclic Chemistry
e)

Fig. 8. Matching results: a) given unknown journal cover image; b) text-block layout of a); c) four best matched layouts; d) corresponding cover images of c); e) text blocks of the title and issue date extracted from the given input

table-of-contents recognition system for technical journals is also the direction of our further research.

REFERENCES

[1] Tsay, N. Y.—Yan, C. D.—Suen, C. Y.: Document Processing for Automatic Knowledge Acquisition. IEEE Transactions on Knowledge and Data Engineering, Vol. 6, 1994, No. 1, pp. 3–21.

[2] Krishnamoorthy, M.—Nagy, G.: Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, 1993, No. 7, pp. 737–747.

[3] Nakano, Y.—Fujisawa, G.—Kunisaki, O.—Okada, K.—Hananoi, T.: A Document Understanding System Incorporating With Character Recognition. Proceedings of International Conference on Pattern Recognition, Paris 1986, pp. 801–803.

[4] Schurmann, J.—Bartneck, N.—Bayer, T.—Franke, J.—Mandler, E.—Oberlander, M.: Document Analysis from Pixels to Contents. Proceedings of the IEEE, Vol. 80, 1982, No. 7, pp. 1101–1149.

[5] Fletcher, L. A.—Kasturi, R.: A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, 1988, No. 6, pp. 910–918.

[6] Nagy, G.—Kanai, J.—Krishnamoorthy, M.: Two Complementary Techniques for Digitized Document Analysis. Proceedings of ACM Conference on Document Processing Systems, Santa Fe 1988, pp. 169–176.

[7] Tsujimoto, S.—Asada, H.: Major Components of a Complete Text Reading System. Proceedings of the IEEE, Vol. 80, 1992, No. 7, pp. 1133-1149.

[8] Wang, L. S.—Wang, Y. K.—Fan, K. C.: Page Segmentation and Identification for Intelligent Signal Processing. Proceedings of IPPR Conference on Computer Vision, Graphics, and Image Processing, Taiwan, ROC 1994, pp. 143–152.

[9] Lin, Y. S.—Tsai, W. H.: Image Segmentation for Color Document Analysis. Proceedings of IPPR Conference on Computer Vision, Graphics, and Image Processing, Taiwan, ROC 1994, pp. 135–142.

[10] Watanabe, T.—Luo, Q.—Sugie, N.: Layout Recognition of Multi-Kinds of Table-Form Documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, 1995, No. 4, pp. 432–445.

[11] Higashino, J.—Nakano, Y.—Fujisawa, H.—Ejiri, M.: A Knowledge-Based Segmentation Method for Document Understanding. Proceedings of International Conference on Pattern Recognition, Paris 1986, pp. 745–748.

[12] Toyoda, J.—Noguchi, Y.—Nishimura, Y.: Study of Extracting Japanese Newspaper Article. Proceedings of International Conference on Pattern Recognition, Munich, Germany 1982, pp. 1113–1115.

[13] Govindan, V. K.—Shivaprasad, A. P.: Character Recognition Review. Pattern Recognition, Vol. 23, 1990, No. 7, pp. 671–682.

[14] Liu, C.-L.—Sako, H.—Fujisawa, H.: Performance Evaluation of Pattern Classifiers for Handwritten Character Recognition. International Journal of Document Analysis and Recognition, Vol. 4, 2002, No. 3, pp. 191–204.

[15] Rahman, A. F. R.—Fairhurst, M. C.: Multiple Classifier Decision Combination Strategies for Character Recognition: A Review. International Journal of Document Analysis and Recognition, Vol. 5, 2003, No. 4, pp. 166–194.

[16] Chellapilla, K.—Simard, P.—Nickolov, R.: Fast Optical Character Recognition Through Glyph Hashing for Document Conversion. Proceedings of the Eighth International Conference on Document Analysis and Recognition, Seoul, Korea 2005, pp. 829–834.

[17] Conway, A.: Page Grammars and Page Parsing A Syntactic Approach to Document Layout Recognition. Proceedings of International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan 1993, pp. 761–764.

[18] Nagy, G.—Seth, S.: Hierarchical Representation of Optically Scanned Documents. Proceedings of International Conference on Pattern Recognition, Montreal, Canada 1984, pp. 347–349.

[19] O'Gorman, L.: The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, 1993, No. 11, pp. 1162–1173.

[20] Dengel, A.—Barth, G.: High Level Document Analysis Guided by Geometric Aspects. International Journal of Pattern Recognition and Artificial Intelligent, Vol. 2, 1988, No. 4, pp. 641–655.

[21] Pavlidis, T.—Zhou, J.: Page Segmentation and Classification. CVGIP: Graphical Models and Image Processing, Vol. 54, 1992, No. 6, pp. 484–496.

[22] Akiyama, T.—Hagita, N.: Automated Entry System for Printed Documents. Pattern Recognition, Vol. 23, 1990, No. 11, pp. 1141–1154.

[23] Chen, W. Y.—Chen, S. Y.: Adaptive Page Segmentation for Color Technical Journals' Cover Images. Image and Vision Computing, Vol. 16, 1998, Nos. 12–13, pp. 855–877.

[24] Sigletos, G.—Paliouras, G.—Karkaletsis, V.: Role Identification From Free Text Using Hidden Markov Models. Proceedings of the Panhellenic Conference in Artificial Intelligence (SETN), Lecture Notes in Artificial Intelligence No. 2308, pp. 167–178, Springer Verlag 2002.

[25] Skounakis, M.—Craven, M.—Ray, S.: Hierarchical Hidden Markov Models for Information Extraction. Proceedings of the18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico 2003, pp. 427–433.

[26] Rabiner, L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, Vol. 77, 1989, No. 10, pp. 257–286.

[27] Bose, C.—Kuo, S.: Connected and Degraded Text Recognition Using Hidden Markov Model. Pattern Recognition, Vol. 27, 1994, No. 10, pp. 1345–1363.

[28] Agazzi, O. E.—Kuo, S. S.: Hidden Markov Model Based Optical Character Recognition in the Presence of Deterministic Transformation. Pattern Recognition, Vol. 26, 1993, No. 12, pp. 1813–1826.

[29] Samaria, F.—Young, S.: HMM-Based Architecture for Face Identification. Image and Vision Computing, Vol. 12, 1994, No. 8, pp. 537–543.

[30] Lin, H. C.—Wang, L. L.—Yang, S. N.: Color Image Retrieval Based on Hidden Markov Models. IEEE Transactions on Image Processing, Vol. 6, 1997, No. 2, pp. 332–339.

[31] Forney, G. D.: The Viterbi Algorithm. Proceedings of the IEEE, Vol. 61, 1973, No. 3, pp. 268–278.

Pei-Chun Wen received the B. Sc. and M. Sc. degrees in computer science from National Tsing Hua University, Hsinchu, Taiwan, ROC in 1994 and 1996, respectively. In 2004, she received the LL. M. degree in Institute of Laws for Science and Technology of National Tsing Hua University. From 1996 to 2000, she was a software engineer (1996–1997), senior software engineer (1998), and principal engineer (1999–2000) in the Advanced Technology Division of Ulead System, Inc. From 2003 to 2005, she was a principal patent engineer in Ulead Systems, Inc. Currently, she was a senior deputy patent manager in the legal department of High Tech Computer Corp.

Ling-Ling Wang received the B. Sc., M. Sc., and Ph. D. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, ROC in 1984, 1986, and 1990, respectively. From 1986 to 1987, she was an associate engineer in the System Software Department of ERSO, ITRI, Hsinchu, Taiwan. From 1991 to 1997, she was an Associate Professor of computer science at National Tsing Hua University, Taiwan. Currently, she is a Professor of information communication at Asia University, Taichung, Taiwan. Her current research are in image processing, pattern recognition, computer vision, and artificial intelligence.