Design of an Automated Data Entry System for Hand-Filled Forms

Lim Woan Ning, Yap Keem Siah, Marzuki Khalid, Rubiyah Yusof*
Centre for Artificial Intelligence and Robotics (CAIRO)
Faculty of Electrical Engineering,
Universiti Teknologi Malaysia,
Jalan Semarak,
54100 Kuala Lumpur
e-mail: marzuki@utmnet.utm.my
(all correspondence should be sent to)*

ABSTRACT

In this new informative era, data and information is the most important asset to the organizations. A large amount of money and manpower have been spent in data gathering, data entry, and storage every year. In Malaysia, data gathering is still largely done through manually filled forms. This data is then entered and stored into databases in government and private organizations manually. Such mode of data entry and storage requires a lot of manpower and is time consuming. At the *Centre for Artificial Intelligence and Robotics (CAIRO) of Universiti Teknologi Malaysia*, research is being carried out to design a system for automated data entry of handwritten-filled forms. The system consists of a high-speed scanner with an auto-feeder and a computer. In the first phase a software is developed that allows the user to design the template of existing forms such that only the regions of interests are captured. The next phase involved the design of a software to capture handwritten characters in the regions of interest through the scanned forms. Image processing techniques are then used to filter and improve the image of the scanned handwritten characters before they are recognized using a neural network algorithm. Once the characters are identified and verified they are automatically stored into a database. This system can be used efficiently in many organizations that involves gathering and processing of a large number of data such as the National Registration Department, Survey Research Malaysia, *Kementerian Pendidikan* and *Lembaga Hasil Dalam Negeri.*

Keywords
Form Processing, Automated Data Entry, and Handwritten Character Recognition.

INTRODUCTION

The Malaysian government is moving towards an e-government system with the use of sophisticated technology beyond the 21st century. This is evident with the setting up of the government complex of Putra Jaya and the Multimedia Super Corridor. As more and more employees are needed for more demanding jobs, we would need to develop better machines to replace manual labor-intensive jobs.

As computer database systems are expanding at an incredibly increasing rate in Malaysia, government agencies and the private sectors would have to find new solutions in data entry. Currently, almost all data entry jobs in the country are being done manually. Such a technique is very laborious and slow which is not only counter-productive but also a waste of useful manpower for other more demanding jobs.

At the Centre for Artificial Intelligence and Robotics (CAIRO) a system for automatic data entry of manually filled forms of very high volume is currently being developed. The system comprises of a fast scanner with an auto-feeder and a computer. A software with the appropriate algorithms is currently being developed for communication between the scanner and the computer and also for processing the forms. Image processing and neural network techniques are used to identify and recognize handwritten characters in the forms.

Once completed the system would be able to process handwritten-forms at high speed and store the information extracted from the forms automatically into existing databases. Such a system would eliminate the need for manual data entry as is the case currently being used. Hundreds of workers in this labor-intensive manual data entry job could be trained for higher value-added jobs in manufacturing or elsewhere to meet the country's acute labor shortage (when the economy recovers). Moreover, the system would be cheaper in the long run and more importantly forms would be able to be processed at a much faster rate to meet more demanding e-government style of administration in the next millenium.

Handwritten character recognition has been one of the fields of pattern recognition which has received a great deal of interest over the past decades. Nowadays character recognition has been implemented in numerous applications such as address and zip code recognition, signature identification, and forms processing to name a few. With the advent of neural networks, more successful results have been achieved and there has been a lot of interest in applying neural networks for such application. Moreover, several commercial OCR products incorporating neural networks as their core recognition tools are already available.

In the system that we designed, we used the self-organizing neural network – Fuzzy Adaptive Resonance

0-7803-6355-8/00/ $10.00 © 2000 IEEE

Theory Mapping or Fuzzy ARTMAP as the recognition algorithm. This neural network is relatively a new neural network paradigm which have a number of advantages when compared to the popularly-used conventional back-propagation algorithm.

Forms processing can be encountered as a large OCR task which differs from the common OCR applications, for example, the ones that are used in occasional scanning of a small number of individual pieces of text [1]. Form processing involves the recognition of several thousands of documents per day. The large number of documents involved implies that big savings are possible even for recognition rates well below 100%. The main requirements that must be fulfilled in such a system is to have a relatively low failure rate of recognition which does not lead to costly checking and remedying operations, and also a relatively fast processing speed that is equivalent to a trained typist.

In many forms processing systems, frame recognition and segmentation techniques are used to extract the characters in the form. However, since forms maybe designed in many different ways, this will cause an increase in the error rate, for example, mis-processing of some characters and also incorrect extraction of data. Thus, in our system, we designed a module that allows the user to re-design existing forms such that only the regions of interest are extracted once the form is scanned.

The rest of this paper has been organized as follows. The next section presents an overview of the proposed system. We next discuss the software modules in the AUDESY, include the techniques we employed. The next section discuss the stated the system specifications and the last section concludes the paper.

OVERVIEW OF THE PROPOSED SYSTEM

The automated data entry system (AUDESY) consists of a high-speed scanner with an auto-feeder, a high-speed computer and software. In recent years, scanner technology has improved tremendously such that the cost has reduced greatly. However, for this system to be efficiently utilized we recommend high-speed scanners that cost in the range of RM30,000 to RM50,000 each. A personal computer can be used, however, a workstation is recommended due to its speed as a number of algorithms are needed to be executed. The processing time for one form depends on the size of information in the form that needs to be extracted. If more information that is needed to be extracted from the form the processing speed will be longer and, thus, a fast computer is necessary in order to match the speed of humans.

The software is the major component of the system. We design the software into six modules which are as follows:

(a) Manager module,
(b) Composer module,
(c) Scanner module,
(d) Verifier module,
(e) Converter module, and
(f) Reporter module.

Figure 1 shows the modules of the system. Note that these modules shall have the extendibility to become a multi-pages forms module.

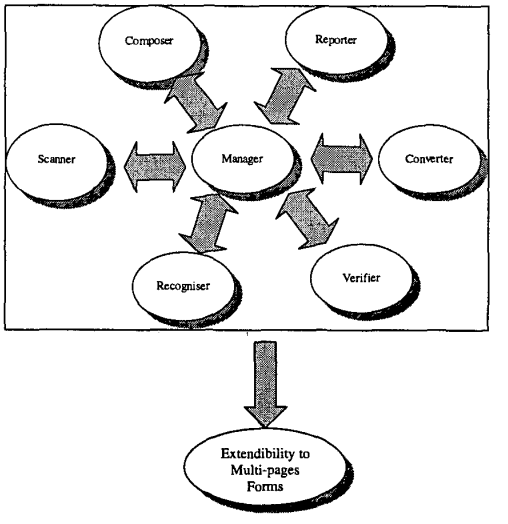

Fig. 1. Modules in the AUDESY

MANAGER MODULE

The Manager module is the main module of the system that interacts with users. It provides user-friendly interface so that the users can do the configurations of the whole system easily. This module also able to launch and communicate with other modules. In order to make the use interface as friendly as possible, as well as rapid development, we developed this module by using Visual Basic. This module has to be developed in very early stage to building up the communications between modules.

COMPOSER MODULE

The Composer module allows the users to configure the template form. The configuration parameters to be used as references in the recognition process in this module are the type and the co-ordinates for the indicators, fields and characters. The indicators are used to detect the form position so that the rotation and transition scales of a real form to its template form can be calculated for alignment purpose. There are two indicators, one at top-right corner and another at the bottom-left corner as shown in Figure 2.

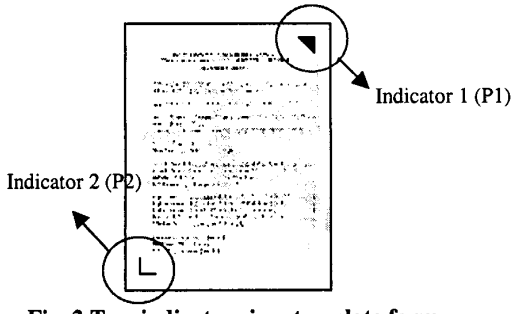

Fig. 2 Two indicators in a template form.

In this module, a user-friendly interface is developed to let the users configure the template form. The template form which had been scanned in the Scanner module will be loaded into the main screen in order for the users to identify the locations of the indicators, fields and characters and to specify the types of the fields. Copy and paste facilities are available in this module for easier and faster form configuration. The GUI of this module is shown in Figure 4.

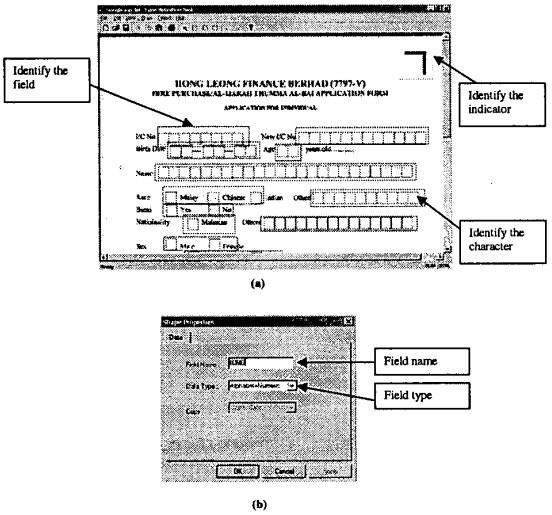

Fig. 3. (a) Template form configuration screen (b) Field's property screen.

SCANNER MODULE

The Scanner module allows users to scan the forms in grey scale with the size of 200 dots per inches (dpi). The forms are scanned in this scale to enhance the images quality so those more obvious features can be extracted. Although 300dpi is normally used by researchers for the testing of hand-written character recognition system, it is not suitable to be used in AUDESY. This is because, a too large image will slowdown the recognition speed, which is always the concern for a form recognition system.

RECOGNIZER MODULE

The Recogniser module is the core module in the system. It utilises the state-of-the-art technology of the Artificial Intelligence to mimic human operator for recognising hand-written characters. This module consists of three main components, which are form alignment, hand-written character recognition and automated word verification. The processing steps of the Recogniser module are shown in Figure 4. Firstly, a real form image will be aligned to it appropriate position and the images for each character will be extracted and recognised. The final results will be grouped by words and verification at word level will be done.

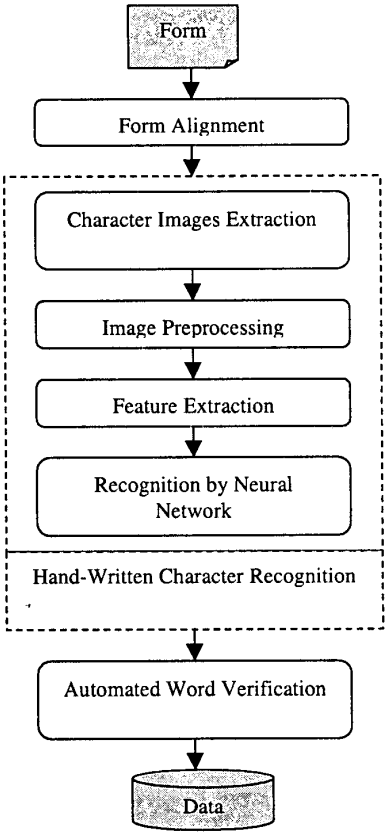

Fig. 4. Processing steps of Recognizer module.

The forms must be aligned before the characters can be extracted. In the form alignment process, template matching techniques is used to locate the indicators in the real form in order to calculate the rotation and transition scales of the form. The form will then be aligned such that the boxes are in the correct positions to enable the extraction of the characters. Figure 5 shown the screen snapshots of the form before and after alignment.

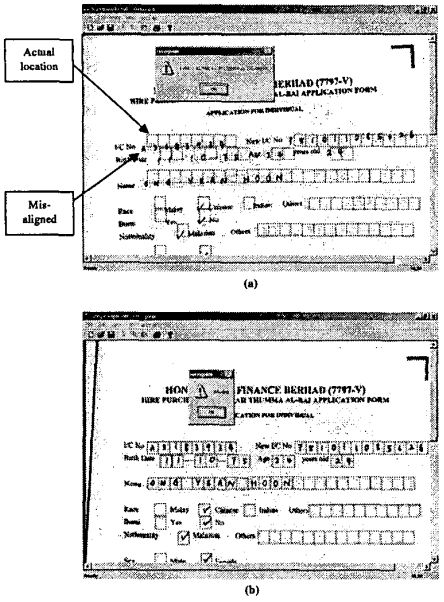

Fig. 5. Screen snapshots of aligning of the form. (a) before alignment (b) after alignment.

After the form being aligned, each character in the forms will be extracted based on the co-ordinates specified in the Composer module. The extracted character images will then be processed using various digital image-processing techniques. After the image pre-processing stage, feature extraction techniques will be used to extract the features from the noiseless character images. The extracted features will then be passed to the Fuzzy ARTMAP neural network for recognition or classification. The GUI of hand-written character recognition is shown in Figure 6.

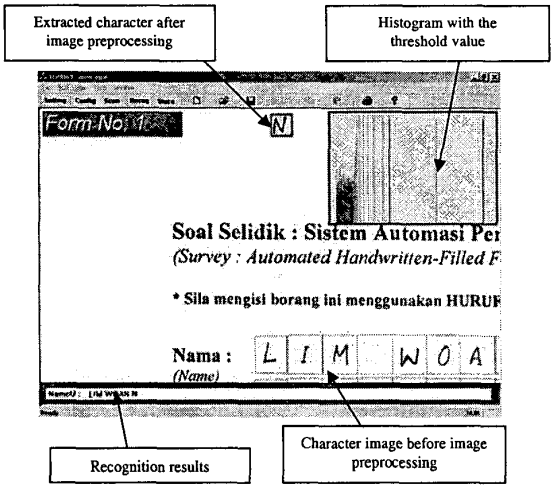

Fig. 6. Screen snapshot of the hand-filled forms in AUDESY.

The recognition engine in AUDESY is able to recognise four types of fields, which are digit, upper case, lower case and combination of digit and upper case. The combination of digit, upper case and lower case will not be considered because the recognition accuracy is too low, only 70.12% that will affects the whole system performance. Furthermore, by processing the combination field of digit and upper case is already appropriate for AUDESY as a form recognition system.

In order to recognise different types of character sets, four Fuzzy ARTMAP neural networks are used. The first, second and third neural networks are trained with only digit, lower case and upper case characters respectively, while the forth neural network is trained with both digit and upper case characters. The characters from NIST Special Database 19 is used to trained the neural network. In further discussion, the term of 'combined characters' will be used to represent the combination of digit and upper case characters.

VERIFIER MODULE

The Verifier module allows the users to verify the recognition results generated in the Recogniser module. This module involves human inspection, where a clerk is needed to compare the results with the original form and make correction to wrongly recognised data. Double checking the recognition results by human is needed to maintain the high reliability of the system. The Verifier module is divided into two small modules, which are local verification module and global verification module.

Local verification module is a character level verification. In this module, the users will only be prompted to enter the correct answer for the unrecognised or confusing character. The confusing character is character which the recognition engine is not sure about the answer. This module allows real-time training of the recognition engine - Fuzzy ARTMAP neural network. After the correct answer being entered, the users can choose whether to train or not train the neural network to learn this character.

Global verification module involves page-by-page verification. In this module, users are able to verify the forms one by one. In manual data entry process, forms are processed twice. In the first phase, a user will key in all the data into the computer, and in the second phase, another user will verify the data. In AUDESY, this module is designed to mimic the second phase of the manual process. It provides the flexibility for data checking in order to ensure the reliability of the data. This process will in fact slow down the system very significantly. However, it can be skipped anytime if the users think the recognition results are reliable enough.

CONVERTER MODULE

The Converter module allows the users to convert the recognition results which are in the text format into specified database format. Currently, the database format for AUDESY is Microsoft Access. Figure 7 shows the data in text format before the conversion and the data in Microsoft Access database format after the conversion. Figure 8 shows a user-friendly GUI for data navigation of Microsoft Access database.

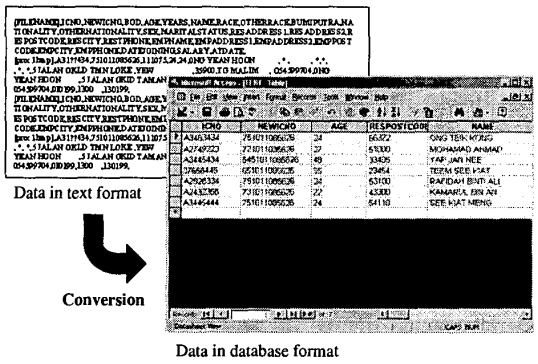

Fig. 7. Screen snapshot of data converted from text format to MS Access database format.

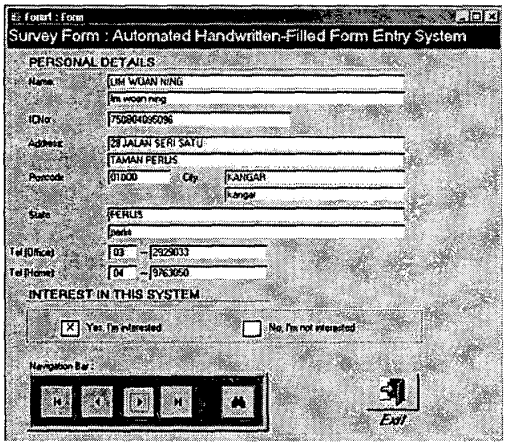

Fig. 8. Screen snapshot of a user interface for data navigation of MS Access database.

Converter module plays an important role in helping the retrieval of useful information from the data. This is because data in database format is easier to be manipulated. Various manipulations can be done on the database, such as finding a record, doing statistical analysis and generating report.

SYSTEM SPECIFICATIONS

AUDESY is developed under the Microsoft Windows 98 environment and the hardware platform is Intel's Pentium III 500 MHz with 128 Megabytes of Random Access Memory (RAM). Two types of scanner are used in this system for the analysis of scanning speed, which are Kodak DS3000 (a commercial high-speed scanner using SCSI card) and HP ScanJet 6250C with auto-feeder (a scanner for home user using USB port).

The scanner module, composer module, recognition module and verification module are developed using Microsoft Visual C++ 6.0 and Matrox Image Library 5.1. While, the manager module and converter modules are developed using Microsoft Visual Basic 6.0 and Microsoft Access 97.

CONCLUSION

Automated data entry through handwritten-filled forms can be viably applied in many organizations that handles form processing in a large scale for fast and efficient data storage. The trend now is towards the design of better recognition algorithms with higher accuracy. The automated data entry system that we are currently developing uses a relatively new neural network paradigm, the Fuzzy ARTMAP, that has the advantage of incremental learning and fast convergence capability [7, 8]. However, to be successful, it has to be tightly coupled with good image processing and feature extraction techniques. Once the system can be effectively designed and commercialized, labor-intensive manual data entry can slowly be eliminated in many organizations in the near future. The use of this system is in line with the government's policy of using IT and automation.

REFERENCES

[1] Dorronsoro, J., Fractman, G., Santa Cruz, C., "Large scale neural form recognition," Industrial Applications of Neural Networks, 1998, pp.354-362.

[2] Myler, H.R., Weeks, A.R., Computer Imaging Recipes in C, London : Prentice-Hall, 1993.

[3] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, Vol.9, No.1, 1979, pp.62-66.

[4] Gorman, L.O., Kasturi, R., Document Image Analysis, California : IEEE Computer Society Press, 1995.

[5] Carpenter, G.A., Grossberg, S, et al., "A Self-Organizing Neural Network For Supervised Learning, Recognition, and Prediction," IEEE Communications Magazine, 1992, pp.38-46.

[6] Carpenter, G.A., Grossberg, S, et al., "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Map," IEEE Transactions on Neural Networks, vol.3, 1992, pp.698-712.

[7] Tay, Y.H., Handwritten Character Recognition by Fuzzy Artmap Neural Network with Application to Postcode Recognition, Thesis of University Technology Malaysia, 1997.