APPLICATION OF TEXT TRANSFER BASE ON OPTICAL CHARACTER RECOGNITION

Aye Mya Thu , Mie Mie Tin
University of Computer Studies, Mandalay, Myanmar
htwe.mama99@gmail.com

ABSTRACT

This technology allows a machine to automatically recognize characters through an optical mechanism. Human beings recognize many objects in this manner our eyes are the "optical mechanism." But while the brain "sees" the input, the ability to comprehend these signals varies in each person according to many factors. By reviewing these variables, OCR can understand the challenges faced by the technologist developing a system. Paper is to develop the Internal Remittance Banking System using OCR in fax transfer, telegraph transfer and bank draft transfer. This paper is based on optical character recognition to recognize the characters and test the keys for verification.

1. OPTICAL CHARACTER RECOGNITION

There are many applications using OCR technology. By using OCR, Internal Remittance system will be saved costing and time. It is also more reliable than manual system because human data entry errors will not occur. OCR has been used to enter data automatically into a computer for dissemination and processing. The earliest of systems was dedicated to high volume variable data entry.

Any standard form or document with repetitive variable data would be a candidate application for OCR. In this system, fax transfer form, telegraph transfer form and bank draft form are fixed, and are therefore OCR technology is applicable.

There are many challenges OCR need to face. First, if OCR read a page in a language, OCR may recognize the various characters, but be unable to recognize words. However, on the same page, usually able to interpret numerical statements - the symbols for numbers are universally used. This explains why many OCR systems recognize numbers only, while relatively few understand the full alphanumeric character range [1].

Second, there is similarity between many numerical and alphabetical symbol shapes. For example, while examining a string of characters combining letters and numbers, there is very little visible difference between a capital letter "O" and the numeral "0." As humans, we can re-read the sentence or entire paragraph to help us determine the accurate meaning. This procedure, however, is much more difficult for a machine [1].

Third, OCRs rely on contrast to help us recognize characters. OCRs may find it very difficult to read text which appears against a very dark background, or is printed over other words or graphics. Again, programming a system to interpret only the relevant data and disregard the rest is a difficult task for OCR engineers [1].

There are many other problems which challenge the developers of optical character recognition systems. In this paper, we will review the history, advancements, abilities and limitations of existing systems [1].

2. IMPLEMENTATION PROGRAM

A The implementation program can be categorized fax transfer, telegraph transfer, bank draft transfer, drawing schedule encashment schedule, summary.

Text capture is a process to convert analogue text based resources into digitally recognizable text

resources. These digital text resources can be represented in many ways such as searchable text in indexes to identify documents or page images, or as full text resources.

An essential first stage in any text capture process from analogue to digital will be to create a scanned image of the page side. This will provide the base for all other processes. The next stage may then be to use a technology known as Optical Character Recognition to convert the text content into a machine readable format.

Optical Character Recognition (OCR) is a type of document image analysis where a scanned digital image that contains either machine printed or handwritten script is input into an OCR software engine and translating it into an editable machine readable digital text format (like ASCII text). [2]

OCR works by first pre-processing the digital page image into its smallest component parts with layout analysis to find text blocks, sentence/line blocks, word blocks and character blocks. Other features such as lines, graphics, photographs etc are recognized and discarded.

The character blocks are then further broken down into components parts, pattern recognized and compared to the OCR engines large dictionary of characters from various fonts and languages.

Once a likely match is made then this is recorded and a set of characters in the word block are recognized until all likely characters have been found for the word block. The word is then compared to the OCR engine's large dictionary of complete words that exist for that language [2].

These factors of characters and words recognized are the key to OCR accuracy by combining them the OCR engine can deliver much higher levels of accuracy.

Modern OCR engines extend this accuracy through more sophisticated pre-processing of source digital images and better algorithms for fuzzy matching, sounds-like matching and grammatical measurements to more accurately establish word accuracy [2].

3. DIFFERENT USES FOR OCR

There are many uses for the output from an OCR engine and these are not limited to a full text representation online that exactly reproduces the original. Because OCR can, in many circumstances, deliver character recognition accuracy that is below what a good copy typist would achieve it is often assumed it has little validity as a process for many historical documents. However, as long as the process is fitted to the information requirement then OCR can have a place even when the accuracy is relatively low (see Accuracy below for more details). Potential uses include:

Indexing – the OCR text is output into a pure text file that is then imported to a search engine. The text is used as the basis for full text searching of the information resource. [2]

Full text retrieval – in this mode the OCR text is created as above but further work is done in the delivery system to allow for true full text retrieval. The search results are displayed with hit highlighting within the page image displayed. [2]

Full text representation – in this option the OCR'd text is shown to the end user as a representation of the original document. In this case the OCR must be very accurate indeed or the user will lose confidence in the information resource [2].

4. KEY ISSUES FOR WHETHER TO USE OCR

There are several key issues to consider in deciding whether to use OCR at all or choosing between different possible appropriate uses for the text output. The main factors to consider are a combination of accuracy, efficiency and the value gained from the process. If the accuracy is below 98% then considerations of the cost in terms of time and effort to proof read and correct the resource would have to be accounted for if a full text representation is to be made. If the OCR engine is not capable to delivering the required accuracy then re-keying the text may become viable, but only if the intellectual value to be gained from having the re-keyed text matches the projects goals and budgets. Otherwise, OCR for indexing and retrieval may be the most viable option [2].

5. THE COMPLEXITY OF LEARNING

So network learning involves choosing from, possibly an exponential number of network structure a giving values to, possibly, an exponential number of real values. Basic results

from computational learning theory show how difficult this can be, both in terms of the number of cases required for training and the time or space required for the optimization. These two aspects are referred to as sample complexity and computational complexity respectively [3].

In learning, there are roughly three distinct phases as more cases are obtained to learn from the small sample, medium sample and large sample phases. Initially with small sample, learning corresponds to going with one's biases or priors [3].

6. CHARACTER MODELING

There are two steps in OCR

- Training and
- Recognition

6.1. Training

In training, models of the characters are constructed. Training steps are built by merging input signals equivalent characters. But the same character may have different character weight. Therefore training step has the following sub steps.

1. Initialization (initialize input signals)
2. Character merging
3. Building models

After initialization, OCR system will get the initial pattern model for each character. Depending on input bitmap image, there must be a merge between image and character. That step is second step and important. If there is a miss-match, the built training data would not be used in recognition.

After these two steps, the final process will take place, called Building models. After building models, the training data can be used in recognition.

6.2. Recognition

Recognition reuses these models to determine the input character. The final goal in OCR (Optical Character Recognition) is recognition itself to determine the input character.

7. CLASSIFICATION

The task of classification occurs in a wide range of human activity. At its broadest, the term could cover any context in which some decision or forecast is made on the basis of currently available information, and a classification procedure is then some formal method for repeatedly making such judgments in new situations.

The problem concerns the construction of a procedure that will be applied to a continuing sequence of cases, in which each new case must be assigned to one of a set of pre-defined classes on the basis of observed attributes or features.

The construction of a classification procedure from a set of data for which the true classes are known has also been variously termed pattern recognition, discrimination, or supervised learning (in order to distinguish it from unsupervised learning or clustering in which the classes are inferred from the data).

7.1. Perspective on Classification

A wide variety of approaches has been taken towards this task. Three main historical strands of research can be identified: statistical, machine learning and neural network.

These have largely involved different professional and academic groups, and emphasized different issues. All groups have, however, had some objectives in common.

7.2. Statistical approaches

Two main phases of work on classification can be identified within the statistical community.

The first, "classical" phase concentrated on derivatives of Fisher's early work on linear discrimination.

The second, "modern" phase exploits more flexible classes of models, many of which attempt to provide an estimate of the joint distribution of the features within each class, which can in turn provide a classification rule.

7.3. Machine learning

Machine Learning is generally taken to encompass automatic computing procedures based on logical or binary operations learn a task from a series of examples.

Here we are just concerned with classification, and it is arguable what should come under the Machine Learning umbrella.

8. MAKING SUCCESSFUL OCR SYSTEM

1. It takes a complimentary merging of the input document ~ stream with the processing requirements of the particular application with a total system concept that provides for convenient entry of exception type items with an output that provides cost effective entry to complete the system[1].
2. To compensate for this problem, the processing system permitted direct key entry of the fail to read items at a fairly high speed.

Directly keyed items from the misread document were under intelligent computer control which placed the proper data in the right location for the data record [1].
3. The output of these early systems provided a "country club" type of billing. That is, each of the credit card sales slips was returned to the original purchaser.

This provided the credit card customer with the opportunity to review his own purchases to insure the final accuracy of billing [1].

9. SYSTEM PROCESSING

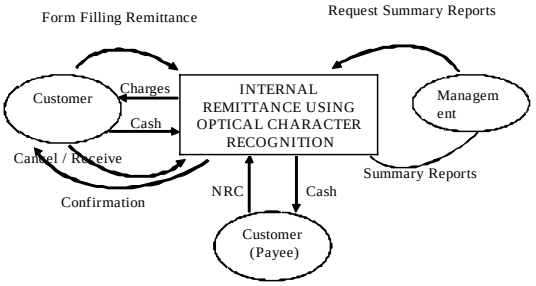

Figure 1. Context dataflow diagram of the system

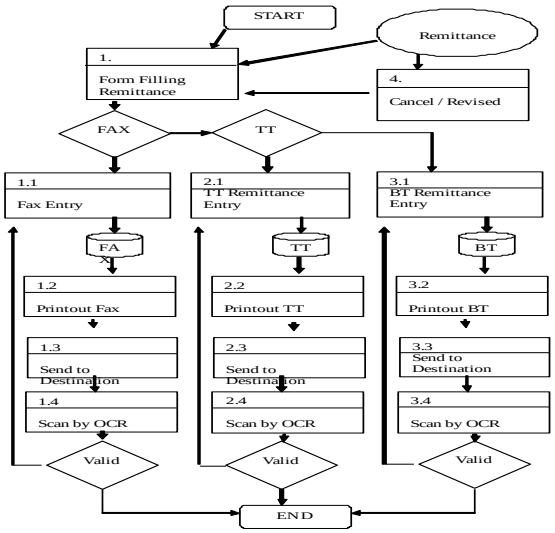

Figure 2. Detail dataflow diagram of the Internal Remittance System

10. CONCLUSION

The interface of this system is graphical user interface, the windows user can understand easily how to use even high technology (OCR) is used in the system. This OCR module is also applicable for other application. But the system needs some improvement to eliminate error rate of OCR.

In training, models of the characters are constructed. Training steps are built by merging input signals equivalent characters.

After initialization, OCR system will get the initial pattern model for each character. Depending on input bitmap image, there must be a merge between image and character. That step is second step and important. If there is a miss-match, the built training data would not be used in recognition. After these two steps, the final process will take place, called Building models. After building models, the training data can be used in recognition.

Recognition reuses these models to determine the input character. The final goal in OCR is recognition itself to determine the input character.

10.1. Advantages of the proposed system

Using the OCR, the user does not need to do double entry processes. Therefore, data entry error can be reduced. The processing time of the system will be speeded up because of the man - machines system.

The rapid reporting system is also available for the Management in any time and any criteria.

REFERENCES

[1] The Association for Automatic Identification and Data Capture Technologies "Optical Character Recognition"AIM, Inc.634 Alpha Drive Pittsburgh, PA 15238-2802,USA(2000) http://www.aimglobal.org/technologies/othertechnologies/ocr.pdf

[2] Tanner, S. "Deciding whether Optical Character Recognition is feasible" King's Digital Consultancy Services (2004) On-line document http:// www. odl. ox. ac.uk/ papers/OCRFeasibility_final.pdf

[3] Buntine, W. "A Guide to the Lecture on Learning Probabilistic Networks from Data" IEEE Trans. On Knowledge and Data Engineering (1996)