INTEGRATION OF OPTICAL CHARACTER RECOGNITION (OCR) TECHNOLOGY WITH A DATABASE APPLICATION

By:

Annalie Faith P. Abella
Geraldine Gail A. Callos
Chryshyll T. Ti

SCHOOL OF ARTS AND SCIENCES
ATENEO DE DAVAO UNIVERSITY

MARCH 2006

INTEGRATION OF OPTICAL CHARACTER RECOGNITION (OCR) TECHNOLOGY WITH A DATABASE APPLCATION

An Independent Research

Presented To

The Faculty of the Computer Studies Division

Ateneo de Davao University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science Major in Information Technology

by

Abella, Annalie Faith P.

Callos, Geraldine Gail A.

Ti, Chryshyll T.

SCHOOL OF ARTS AND SCIENCES

ATENEO DE DAVAO UNIVERSITY

MARCH 2006

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Nowadays, everything must be finished before or within the deadlines. People are doing so many things in so little time. Developers see this trend and decided to focus on building applications that will make the people's works easier. They began introducing and continuously developing new gadgets that can help the society finish their tasks in a quicker manner. In these past years, employees are spending most of their time manually inputting important information from different forms to the database. There are new technologies that are available and useful to these employees but they are not fully utilized by them. These technologies, like the Optical Character Recognition, can help enhance the efficiency on working environments. This document discusses the proponents' desire to expand the usage of Optical Character Recognition so that there will be no need to manually enter all information from a document individually. It discusses the integration of an optical character recognition to an application database which will lead to the efficiency of storing data electronically.

Keywords:
Optical character recognition, Application database

CHAPTER I

INTRODUCTION

1.1 Background of the Study

Over the years, database applications have been extensively used by companies in storing and acquiring accurate data in an efficient and effective manner. It helps companies organize their data to reduce inconsistency and redundancy. In addition, it also provides security restrictions causing information to be inaccessible to unauthorized people. At present, business establishments are relying more on technology in handling their business transactions. Database applications are used to reduce physical storage of information by allowing users to enter data into the database. However, database management system users still perceive a need to continuously improve the role of database system applications in a company's working environment.

Besides database applications, there are other technologies that can improve data entry and maintenance such as optical character recognition. The use of Optical Character Recognition is one of the thousands of software developed to help people in this time of rapid execution. Optical character recognition (OCR) is the translation of optically scanned bitmaps of printed or written text characters into character codes that can later be used by the computer. This is an efficient way to turn hard-copy materials into data files that can be edited and otherwise manipulated on a computer. Suppose you want to digitize a certain book overnight. You could stay up all night typing and still not

finish or you could use a high-end scanner and in minutes scan the book into a computer using optical character recognition (OCR) technology. This is the technology long used by libraries and government agencies to make lengthy documents quickly available electronically. Nowadays, OCR technologies have proven their usefulness in efficient data storage.

1.2 Application Context

Even though Optical Character Recognition and Database Applications prove to be extremely useful to its users, they are still not fully utilized by users and developers. Since developers and users want efficient data storage and retrieval, the proponents came up with the idea of integrating the two technologies to achieve this goal. This leads the proponents' study to focus on building an application that integrates Optical Character Recognition into a database application. The study's focal point will be storing the required data or information details into the database from a source document or an image file using the coordinates specified by the user. The user scans the document and then the application will then recognize the data and store the specific document details into the database.

1.3 Objective of the Study

The study aims to attain the following project objectives:

- Expand or develop the usage of a database applications by integrating it with an Optical Character Recognition technology that can make it more functional to its users
    - The technology-based application aims to integrate the Optical Character Recognition technology into a database application
    - The technology-based application aims to store accurate recognized characters into the database
- The proponents aim to familiarize and understand strategies and methods used in the OCR technology's text recognition processes that are utilized in the Optical Character Recognition Software Development Toolkits.

1.4 Scope and Limitation of the Study

1.4.1 Scope

The study will primarily focus on integrating an optical character recognition (OCR) into a specific database application by storing specific data taken from the recognized image.

1.4.2 Limitations

The study and application will not focus on resolving the existing database and optical character recognition technology's limitations. The study will simply focus on utilizing their capabilities by developing an application that would be able to recognize and store specific texts from an image file to a database.

1.5 Significance of the Study

This study is significant because it expands the usage and capabilities of the database application. It is significant because using the OCR technology in entering data into the database would hasten the encoding process. Through integrating optical character recognition into a database application, the user will be able to directly store specific data to a database either from a saved image file or an image file created after scanning an original physical document. Through this application, the process of storing data to the database will be faster because the user will not need to encode all of the information needed.

This application is useful for offices that handle various forms that need to be stored electronically. An example of these forms and offices are registration forms in the Registrar's office.

Lastly, this study can serve as a stepping stone in applying an Optical Character Recognition technology with a database application. It will not only emphasize a new way of storing data but it can open new doors for other

developers in enhancing the application's capabilities and resolving some limitations of the project.

1.6 Glossary of Terms

1. **Optical Scanner** – is a device that can read text or illustrations printed on paper and translate the information into a form the computer can use.
2. **ICR (Intelligent Character Recognition)** - is the computer translation of hand printed and written characters. Data is entered from hand-printed forms through a scanner, and the image of the captured data is then analyzed and translated by sophisticated ICR software.
3. **OMR (Optical Mark Recognition)** – is the technology of electronically extracting intended data from marked fields, such as checkboxes and fill-infields, on printed forms
4. **MICR (Magnetic Ink Character Recognition)** – is a character recognition system that uses special ink and characters. When a document that contains this ink needs to be read, it passes through a machine, which magnetizes the ink and then translates the magnetic information into characters.
5. **ASCII Code (American Standard Code for Information Interchange)** – Assigns a number to each character that can be typed on a computer keyboard. The code aids programming and communication between computers.

6. **TWAIN** - defines a standard software protocol and application programming interface (API) for communication between software applications and image acquisition devices.
7. **Database** – A collection of information organized in such a way that a computer program can quickly select desired pieces of data.
8. **Relational Database Model** – A database system in which any database file can be a component of more than one of the database's tables.
9. **Zoning** – is a partition of the control box of the pattern (i.e. the smallest rectangle containing the pattern); the elements of such partition are used to identify the position in which features of the pattern are detected. It is the process of distinguishing graphic materials from text blocks
10. **Pattern Recognition** – is an important field of computer science concerned with recognizing patterns, particularly visual and sound patterns. It is central to optical character recognition (OCR), voice recognition, and handwriting recognition.
11. **.NET** – is the Microsoft Web services strategy to connect information, people, systems, and devices through software. Integrated across the Microsoft platform, .NET technology provides the ability to quickly build, deploy, manage, and use connected, security-enhanced solutions with Web services.
12. **MICR Toner** – is a special magnetic toner used in printing bank checks on plain paper.