Adoption of an Open Source Optical Character Recognition (OCR) for Database Buildup of the Students' Scholastic Records

Milleth M. Bautista and Benilda Eleonor V. Comendador

***Abstract*—Various ways were employed by higher education institution in order to ensure the accuracy of collected students' scholastic records. Students' grades are now collected and stored in digitized form, enabling a faster, more reliable safe keeping. Nonetheless, these organizations found the conversion of printed out scholastic records accumulated through the years to be both tedious and time consuming. Furthermore, manual encoding of students' grades can often result to inaccuracy. Thus, the paper focuses in assisting Higher Education Institutions by using Optical Character Recognition (OCR) in automatically recognizing and storing students' grades from printed out grade sheets. With the use of this tool, each grade on the scanned grade sheet would be stored and indexed to the respective student, lessening the tedious task of manually encoding the grades. In addition the system would also allow the digital storing of the scanned grade sheet.**

***Index Terms*—Optical character recognition (OCR), Tesseract, open source, grade sheet, scholastic records, database buildup.**

I. Introduction

The transition from paper-based transaction to computerization and centralization has always been considered a complexity in system implementation. The educational system has not been exempted from this difficulty. High conversion cost, from manual or legacy systems into centralized systems, has always been a factor to consider in applying computerization [1]. But for educational institutions, conversion of their records, specifically scholastic records of students, is of great importance.

A students' scholastic record is the collection of all final grades or ratings of the students which they earned during their stay in an institution. In an organization without centralized database system for managing students' scholastic records, students' final grades are reported by the teachers by accomplishing a grade sheet form by course manually. The said document presents a list of all students enrolled in a particular course and the corresponding final rating of each student. These records are submitted to the registrar's office for processing and record keeping. Consequently, with the implementation of the computerized system, educational institutions are faced with the tedious job of converting these records into digital ones. Some educational institutions nowadays require their instructors to print the grade sheets and then encode the final grade of each student into their system.

In the encoding of grades, keyboarding remained the most common way of inputting data into computers. Since the computerized storage of grades was mostly implemented recently, in order to consolidate the students' grades, the registrar manually checks all submitted grade sheets for each student. Automated generation of certification of grades per semester is very difficult.

Several techniques had been applied in order to lessen the manual encoding of data. Optical Character Recognition engines are used in order to scan and recognize printed out characters and turn it into editable text. Optical Character Recognition (OCR) was defined as a system that performs full alphanumeric recognition of printed or handwritten characters through document scanning [2]. Various engines both commercial and open source are available for users in order to assist them in record conversion. Such optical character recognition engines are utilized not only in document scanning but also license plate recognition and in extracting text from natural scene images [3]. It can be utilized both by computer based and mobile based systems [4], [5]. One of the most popular open source OCR is Tesseract. Tesseract was first developed by HP and afterwards improved by Google, releasing it as open source in 2005 [6]. Because of its nature being open source, Tesseract was often used for integrating character recognition capabilities into a system [3]. In the past, experimental testing were performed to compare Tesseract's recognition accuracy against both proprietary and open source OCR [3], [7]. Although Tesseract has shown a high level of accuracy, processing of images before scanning greatly helps its accuracy techniques such as applying luminosity and linearization which increase the scanning results accuracy [8].

The Database Buildup for Students' Scholastic Records aimed to address the problems encountered by the faculty and registrar of the educational institution. Through this system, the laborious and tedious task of manually encoding grades is eliminated by the use of the Optical Character Recognition (OCR).

II. The Developed System

A. System Architecture

The system is composed of a software tool that employs various open source applications in order to properly convert characters into data that will then be stored directly to the

Manuscript received December 9, 2015; revised March 10, 2016.

M. Bautista is with Cavite State University, Philippines (e-mail: millet.bautista@gmail.com).

B. E. V. Comendador is with Polytechnic University of the Philippines, Philippines (e-mail: bennycomendador@yahoo.com).

database (see Fig. 1). ImageMagick, an open source image processing software, was utilized to enhance the quality of the scanned image in preparation for scanning [9]. Tesseract was used to recognize printed out characters into editable text.

It primarily requires its users to produce an image of the grade sheet or scan a printed out grade sheets, saving it as an image file. The system will then require the users to upload the scanned grade sheet into the system. The system will then perform various image processing techniques to enhance the characters in the image and allow the Optical Character Recognition Engine to scan and convert the characters properly. The system enhances the image for OCR Engine recognition through the use of Image Magick. After enhancing the image, removing noise and grids, it will then be converted into an editable text file using the open source Optical Character Recognition Engine Tesseract. It is essential to use Image Magick since Tesseract has difficulty reading characters inside grids and tables.

Once Tesseract has converted the image text into editable text file, the system will put the recognized string into a series of Regular Expression processing that will filter the text so that the end result would only display the student number of each student and their corresponding grades separated by a comma for CSV conversion. Afterwards, the system will then convert the text into a Comma delimited (.csv) file for database storage.

Once each student number and grade had been stored in the database as records, the system will display the grade sheet of the subject, using each recorded student number and matching it with the data stored in the database to display the student name. The system also references the student grade to display the respective remarks for the grade.

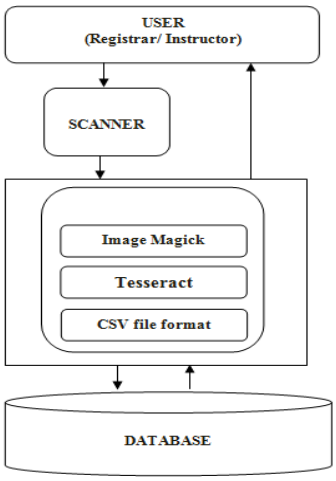

Fig. 1. System architecture of database buildup of the students' scholastic records.

B. Software

The study developed a system that assists both the instructors of the university and the registrar and his/her staff by scanning printed out grade sheets submitted by the instructors and turning them into digital copies.

The grades of the students on the digital copies will be extracted and saved in the database. The extraction will be done by using Tesseract OCR. The system will be reading each student number, using a technique where the system searches for the numbers on a specific location and fetching the numbers that fits the criterion of the ID number. This technique was used in order to further increase the accuracy of the record matching [10]. After doing so it matches the student numbers to the student number stored in the database. Each student number together with the corresponding student grade tagged with the schedule code for uniqueness will then be stored individually in the database as a record that will allow future users to query for further use. A display of the scanned student grades will be shown after the scanning. If there are errors in the scanning and recognition of characters, the users may then modify the said error. The users can search for the grades by individual student or by class on a given course.

All scanned images of the grade sheets are stored in the system with the following naming convention: the subject code, the student course year and section, followed by the school year and the semester separated by dashes and ending with the file name. Once it is saved into the system the user can retrieve the file and view the uploaded scanned image of the grade sheets. The developed tool has provision that facilitates the registrar to add new users of the system. Furthermore, it allows the registrar to revoke an instructor's user access to the system. The existing users can change their passwords for security purposes. Reports such as the copy of the grade sheet can be generated from the system as well as the individual copy of student's grades for a specific semester. The developed prototype provides a grade sheet template and can generate the certification of grades per student which was tested in the Cavite State University.

C. Research Design

The authors developed a prototype tool that may assist educational institutions in digitally storing student scholastic records. Afterwards, they conducted an experiment to test the accuracy of the said tool, depending on certain document condition in converting printed out records into digital ones. In addition, the study dealt with the concerns and problems encountered by the registrars and instructors of a university regarding the storing and consolidation of student grades.

The study used the descriptive method and utilized the Non-Probabilistic Purposive Sampling in the selection of 80 faculty members and 6 registrars as respondents.

A software tool was developed by following the Feature Driven Development (FDD) as its software development methodology. FDD is an agile development methodology for implementing software functionality. It is based on breaking the requirements down into small client-valued pieces of functionality [11]. FDD was used in order to develop the software tool which will go through experimental testing for its accuracy, speed and usability regarding the scanning of the students' grades and turning them into editable, indexed data stored in a database. Furthermore, factors such as the condition of the source material — photocopy or original, the Dots Per Inch of the scanned image and document character spacing were tested in order to find the effects on the scanning and OCR reading accuracy.

III. Results and Discussion

The first objective of the research is to find the challenges

encountered by instructors and registrars in manually encoding the students' grades. Though the survey it was found that the manual encoding of students' grades is prone to discrepancies and errors. Since most instructors and staff would have to encode more than 100 student grades at a time, typographical errors are unavoidable. Another problem often encountered in the use of the manual process is the lost of grade sheets that results to the use of various sources of grades such as class cards or course checklists. After taking in consideration the various problems encountered by the user, the tool was designed and developed in order to assist them in storing the scholastic records of students. Through experimental testing, the tools conversion accuracy on various factors such as the condition of the source materials specifically photocopied and originally printed, the Dots Per Inch of the scanned Image and the document character spacing.

Table I shows the summary of significant findings during the testing of Accuracy for scanning Original Documents

TABLE I: SUMMARY OF FINDINGS DURING THE TESTING FOR ACCURACY AND SPEED FOR SCANNING ORIGINAL DOCUMENTS

Document Sample	Number of Errors Due to Image Noise	Speed of Reading (Seconds)
25-000	0	24
25-003	1	24
25-023	3	19
25-024	0	16
22-011	4	22

As presented in Table I, original grade sheets produced less errors due to the absence of image noise that prevented the Optical Character Recognition from reading the grade sheets accurately. It was also found that the original images were scanned and read faster by the Optical Character Recognition engine. The naming convention or image number of the document sample was auto-generated by the scanner used in scanning.

Table II shows the summary of significant findings during the testing of Accuracy for scanning Photocopied Documents.

TABLE II: SUMMARY OF FINDINGS DURING THE TESTING FOR ACCURACY FOR SCANNING PHOTOCOPIED DOCUMENTS

Document Sample	Number of Errors Due to Image Noise	Speed of Reading (Seconds)
22-001	9	41
22-003	3	32
22-027	4	36
22-029	6	29
22-000	1	30

It can be deduced from the data presented in Table II that the photocopied documents produced image noise that affected the performance of the OCR. The said noise was recognized as character by the tool, specifically special characters such as periods and commas. As such, the captured noise can affect the reading accuracy of the OCR. Based on the sample document, it can be deemed that liquid or powder photocopier machine was used to photocopy the document.

Table III shows the summary of significant findings during the testing of Accuracy and Speed for scanning Documents at Various Levels of Dots per Inch.

TABLE III: SUMMARY OF FINDINGS DURING THE TESTING FOR ACCURACY AND SPEED FOR SCANNING DOCUMENTS SCANNED IN VARIOUS LEVELS OF DOTS PER INCH

Dots Per Inch of the Scanned Image					
100 DPI		200 DPI		300 DPI	
Scanning Speed (Seconds)	No. of Errors	Scanning Speed (Seconds)	No. of Errors	Scanning Speed (Seconds)	No. of Errors
25	2	25	2	25	0
25	0	24	0	23	0
25	16	23	16	27	1
23	2	21	2	24	0
22	1	28	1	27	1

It shows the importance of using a standard Dots Per Inch in scanning documents for Optical Character Recognition. The same documents were scanned using three different Dots Per Inch, 100, 200 and 300 respectively. The results described that the documents scanned in lesser DPIs produced more errors. On the other hand, documents scanned in 200 DPI scanned faster but produced more inaccurate results further lengthening the process by a couple of seconds.

Table IV shows the summary of significant findings during the testing of Accuracy for scanning Images Depending on the Character Structure

TABLE IV: SUMMARY OF FINDINGS DURING THE TESTING FOR ACCURACY FOR SCANNING IMAGES DEPENDING OF CHARACTER STRUCTURE

Document Sample	Number of Errors Due to Character Spacing
22-003	1
25-002	1
25-005	0
25-017	1
25-021	0
25-027	2
25-003	0
25-029	3
22-000	0
25-024	0

Table IV shows the effects of character spacing to the scanning of documents. Since Tesseract Engine reads characters individually and then compares the distance of each character, inaccuracies in OCR reading were found when characters were spaced irregularly. The table above shows the inaccuracies found when such characters were irregularly spaced. The research used an algorithm where spaces were expressed in characters. It was found that characters irregularly spaced, especially those with high character kerning and tracking made the OCR recognize it as character spacing. This resulted to inaccuracies specially for the alpha characters.

After the experimental testing, the tool was then evaluated for its functionality. In order to do so, the tool underwent an evaluation, with the respondents consist of select instructors and registrars of a State University. The respondents' perception of the functionality of the tool was assessed in terms of its accuracy, usability and speed.

Table V shows the summary of findings regarding the respondents' perception on the functionality of the tool in terms of the aforementioned factors.

TABLE V: SUMMARY OF FINDINGS REGARDING THE LEVEL OF EFFECTIVENESS OF THE TOOL

Variables Tested	Mean
Accuracy	3.85
Usability	4.35
Speed	4.29

The accuracy of the system earned a general mean of 3.85 which clearly shows that the respondents agreed that the system presented accurate results. Through the use of human intervention, data misread by Tesseract OCR was then modified to ensure the accuracy. Moreover, the respondents found that the system was error free. With regards to the system's usability, the tool earned a general mean of 4.35, showing that the respondents found that information presented in the system was clear, concise and informative to the user. Still, respondents found that the operational and error messages were easy to read and understandable. Lastly, the respondents found that the software was easy to navigate. The general mean of 4.29 regarding the tools speed clearly shows that the use of the OCR to scan previous grades of students and directly store them in the database was found to greatly increase the speed of processing with regard to the encoding of grades. Furthermore, the respondents found that the software consolidated past grades of students faster than the usual method.

IV. CONCLUSION AND RECOMMENDATION

Based on the acquired results of the study, the research came up with several conclusions:

Through series of testing, it was found that using of photocopied materials for OCR scanning and storing data into the database produced less accurate results because of the image noise present on photocopied materials. On the other hand, original materials are scanned faster and produce more accurate results because of the absence of image noise. Apparently, Dots Per Inch (DPI) setting in scanning also affect the speed and accuracy of scanning. Image scanned at 300 DPI produces more accurate results. On the other hand, images scanned at 200 DPI are read faster but produce less accurate results. Images scanned at 100 DPI are less accurate and are read slower by the software.

Thus, the use of the Database Buildup of Students' Scholastic Records would greatly assist the office of the university registrar since they would not have to manually search for each old grade sheets. The use of Tesseract OCR engine in converting text from images into editable data greatly increased the speed of encoding students' grades. Through it, the manual encoding of grades by instructors and registrars alike can be eliminated.

It was found out that the recommended type of grade sheet to be scanned is that of the original copies of the grade sheets. Conversely, a DPI of 300 is the optimal dots per inch setting for images to be scanned by the optical character recognition engine Tesseract.

Furthermore, the research recommends that the organization keeps a standardized format, font and text layout for all grade sheets in order to increase the accuracy of scanning and reading of the OCR.

REFERENCES

[1] J. A. Hoffer, R. Venkataraman, and H. Topi, *Modern Database Management*, Prentice Hall, 2012.

[2] P. Charles, V. Harish, M. Swathi, and C. H. Deepthi, "A review on the various techniques used for optical character recognition," *International Journal of Engineering Research and Applications*, vol. 2, no. 1, pp. 659-662, January-February 2012.

[3] C. Patel, A. Patel, and D. Patel, "Optical character recognition by open source OCR tool Tesseract," *International Journal of Computer Application*, vol. 55, no. 10, pp. 50-56, October 2012.

[4] S. Charjan, R. Mante, and P. N. Chatur, "Comparing Tesseract results with and without character localization for smartphone application," *International Journal of Research in Computer and Communication*, vol. 2, no. 5, pp. 298-302, May 2013.

[5] S. Z. Zhou, "Open source OCR framework using mobile devices," *Multimedia on Mobile Devices*, vol. 6821, 2008.

[6] R. Smith, "An overview of the Tesseract OCR," in *Proc. IEEE Ninth International Conference on Document Analysis and Recognition*, 2007.

[7] S. Dhiman and A. Singh, "Tesseract Vs Gocr, a comparative study," *International Journal of Recent Technology and Engineering*, pp. 80-83, 2013.

[8] S. Badla, *Improving the Efficiency of Tesseract OCR Engine*, 2014.

[9] *ImageMagick Tricks*, Packt Publishing.

[10] S. Lu, Y. Qu, Y. Check, and Y. Xie, "ID numbers recognition by local similarity voting," *International Journal of Andvance Computer Science and Applications*, *Special Issue in Image Processing and Analysis*, pp. 54-63, 2010.

[11] L. Westfall, *The Certified Software Quality Engineer Handbook*, USA: Quality Press, 2010.

Milleth M. Bautista obtained her undergraduate degree in information technology in 2011 at Cavite State University. She earned her master's degree in information technology majored in management information system from Polytechnic University of the Philippines in 2015. Presently, she is an instructor in Cavite State University.

Benilda Eleonor V. Comendador was a grantee of the Japanese Grant Aid for Human Resource Development Scholarship (JDS) from April 2008 to September 2010. She obtained her master of science degree in global information and telecommunication studies (MSGITS), majored in project research at Waseda University, Tokyo, Japan, in 2010. She was commended for her exemplary performance in completing the said degree from JDS. She finished her master of science in information technology at Ateneo Information Technology Institute, Philippines in 2002. Presently, she is the chief of the Open University Learning Management System (OU-LMS) and the program chair of the Master of Science in Information Technology (MSIT) of the graduate school of the Polytechnic University of the Philippines (PUP). She is an associate professor and was the former chairperson of the Department of Information Technology of the College of Computer Management and Information Technology of PUP. She attended various local and international computer related trainings and seminars. She was the country's representative to the Project Management Course in 2005, which was sponsored by the Center for International Computerization Cooperation (CICC) in Tokyo, Japan together with other 9 representatives from various ASEAN countries.