ISSN 2278-3091
Volume 9, No.1.1, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4891.12020.pdf
https://doi.org/10.30534/ijatcse/2020/4891.12020

Digital Data Classification and Extraction for Records Management of PAPS and PACS Documents

Jennifer E. Sabugaa[1], Glenn S. Lahayon[2]
1Faculty, Caraga State University, Ampayon, Butuan City, Philippines, jenniferespanola@gmail.com
2Student, Caraga State University, Ampayon, Butuan City, Philippines, glennlahayon@gmail.com

ABSTRACT

The Department of Social Welfare and Development – Adoption Resource and Referral Unit is in need of emerging technologies to store records of the Alternate Parental Care program. A technology that would digitize data, case visualize and notify managers and Server-Based digital storage. To hasten the storing process, the system does the classification and extraction of data from a PDF document using Tesseract Optical Character Recognition (OCR). OCR extracts the text from an image by converting the whole image into binary pixels and from binary pixels and compare the recognized sets to its library to predict the text and set of words. It uses Keras to train the model with tensorflow on the backend to classify data. Keras works by defining a sequence of layers in a network by creating a sequential class and adding new layers. This will be done by creating an array of layers and pasting it to the constructor of the sequential model. The researchers produced 4 four documents with ten pages of the same content in different resolutions. Based on the result gathered, 75 DPI resolution has an accuracy percentage of 94.8548778, for 150 DPI is 99.23619895, while 98.91023618 for 300 DPI and 98.7416335 for word application generated document. For 75 DPI, the error contains almost English words, while in 150 DPI and 300 DPI the error usually came from punctuation marks and in word application generated document, the error usually came from a capitalized letter.

Key words: Data Classification, Data Extraction, Case Visualization, Optical Character Recognition, Tesseract

1. INTRODUCTION

In today's era, government offices are mandated to make the business transaction easier and faster. President Rodrigo Roa Duterte signed into law the Republic Act No. 11032 or the "Ease of Doing Business and Efficient Government Service Delivery Act of 2018". This new law is an amendment of the Republic Act 9485 also known as "The Anti-Red Tape Act of 2007". One of the most important features of this law is the standardized deadline for government transactions. The new law has a big impact particular in the Alternative Parental Care (APC) program handled by the DSWD Field Office Caraga. APC is any arrangement, formal or informal, temporary or permanent, for a child who is living away from his or her parents. This DSWD program has two (2) documents essential to facilitate the adoption process: the HOME STUDY REPORT for Prospective Adopted Parents (PAPs) and CHILD STUDY REPORT for Prospective Adoptive Child (PACs). These documents are prerequisites in showcasing each adoption timeline. The adoption timeline must be strengthened to comply with the new law. Timeline defined as sequence of events that is visualized by plotting them on an axis at the instant of an interval [7]. Timeline visualization for the APC will be good if match with notification to ensure the case handler about the case via email and push notification. The short message administration (SMS) innovation is a standout amongst the most steady and most broadly utilized portable specialized techniques after telephone calls [8]. While the email notification focusses on retention, increasing the lifetime value of a customer and serves as the best marketing tool [9].

With the advent of the digital world, knowledge is divided into two parts the binary strings of 0s and 1s, which make up the genetic code of digital data which people, created, manipulated and shared [1]. The consumption of data is essential to its longevity. Over 2.5 million bytes of data are generated every single day, and will only expand from there. It is projected that 1,7 MB of data will be produced every second by 2020 for every person on Earth [28]. The reason why one goal of practitioners in recent developments in information and communications technology is to provide accurate information to the right user at the right time [3] by transforming non-digitized data into digitized one. Digitization improves the efficiency of a business's process, consistency, and quality. Integrating conventional records into a digitized system removes redundancies and shortening the communications chain. It will Improve accessibility and facilitate better information exchange for staff and users.

Digitized materials contain confidential data that needs to be safe and secured. There are two types of storage that is to be considered upon storing digitized data, which is the cloud storage and local storage. In terms of security and safety, local storage will be considered as an advantage. Upload and download speed also must be considered in storing data. However, the maintenance of local storage is expensive (4].

For an institution like the Department of Social Welfare and Development, it is more important to keep the files safe and secure and a server was readily established with the use of the Regional Information Communication and Technology Management Unit (RICTMU). According to the study of Prabhaka & Rani [6] institutional repositories are in need due to technological change, significant increase in the overall volume of research, increasing need for proper storing and access to unpublished information, increase demand to use knowledge objects from anywhere at any time and increasing uncertainty over who will handle the preservation archiving of materials.

Because the manual inputs of APC cases cause a time delay and affect the productivity of the user to perform other tasks, there is a need to integrate emerging technology to facilitate the fast and reliable data extraction and data classifications from a printed format. The volumes of PAPs and PACs reports are to be kept safe and secure, that can be rapidly extracted and classified for storage.

With this, the researchers have used Tesseract Optical Character Recognition (OCR) to extract data from a PDF and Keras to train the model with Tensorflow on the backend. This is to classify the types of records being extracted from the PDF file and automates the storing of the digitized copy of the documents to server-based storage management.

2. RELATED LITERATURES

Form case managers who personally interview the PAPs and child, the data they gathered will be recorded in the form of the Home Study Report and Child Study Report. This report will be submitted to the Regional office so that the case will be accounted for. Once the case will be verified, it will be saved into the caseload inventory. The caseload inventory will serve as the database of all cases that the region is holding.

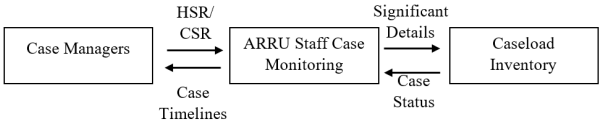

Figure 1: Current Flow of Adoption Cases

A study compares tesseract with other commercial Optical Character Recognition (OCR) application, considering the plate number as a parameter of text extraction. OCR is converting the printed document into editable text for further processing [10]. The technology was designed to allow the machine to recognize text automatically. Accordingly, Tesseract was one of the open-source OCR engines. The process flow of Tesseract OCR starts from the input image and passes to Adoptive Threshold which converts the input image into a binary image. Then, the associated segment examination which is utilized to remove character plots. The next thing is to find lines and words which find character outlines and organize into words.

The processing of image and how it traverses into the engine with a final result of extracted text from an image. This architecture tested with the use of windows command prompt with images that contain English text. Tesseract OCR supports various languages, however, the researchers used images with English content in testing the accuracy of system.

Two systems were evaluated using two datasets with a small dataset of thermal images in low resolution and a large dataset of thermal images in high resolution. YOLOv2 has a precision rate of 52% which is a good start for different datasets. After fine-tuning, YOLOv2 increases its precision of up to 84%. On the other hand, Keras initially gives a 26% accuracy rate and after fine-tuning, the precision rate raised to 87% [11].

"Real-Time Mailbox Alert System via SMS or Email" revolutionize the use of short messaging technology and electronic mail. Upon receiving a mail, a sensor from a mailbox will communicate to GSM Modem then the GSM Modem will automatically send a notification to the owner of the mailbox via SMS or email. The system uses Zelio Logic Smart Relay (ZLSR) which is a programmable logic controller and designed for a small automated system [12].

3. METHODS

The system uses a pdf file that serves as input passes through an OCR engine that extracts text from the document. With extracted text, the significant data will be saved to the database and the document will be uploaded to the server. With data from the database, the case notification and visualization will be generated.

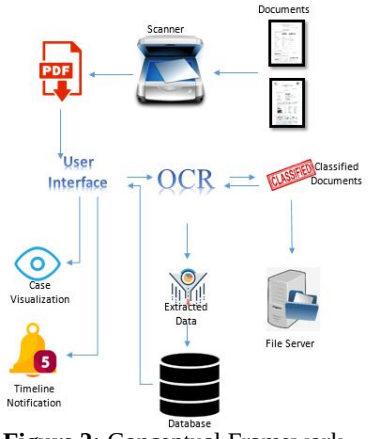

Figure 2: Conceptual Framework

The researchers finds two ways to extract data from the document: the use of the Convolutional Neural Network (CNN) and the Optical Character Recognition (OCR). CNN recognizes the text or digit by treating the document as an image and converting the image into a 2-dimensional array. CNN will create a 2x2 or 3x3 matrix and it will traverse to the whole image and the output will pass to a Neural Network to predict the text. While OCR extracts the text from an image by converting the whole image into binary pixels and from that binary pixels and

compare the recognized sets to its library to predict the text and set of words.

To extract data from the APC document, the system will allow the user to upload an APC document and it will be placed in a canvas. From the canvass, the user can view what page of the document will be extracted. From the selected document, the system will transform a pdf page into an image. Then with the use of Tesseract with optical character recognition, the text will be extracted from the data from the document. The extracted data will then be saved to a database, while the pdf file will be uploaded to the server for safekeeping and possible use in the future.
Optical Character Recognition is the technology used to convert text from images into a machine-encoded format. The flow starts from an input image is being processed converting the original image into a binary image and put into an OCR. After OCR processing the text from the image will be extracted. For the OCR engine to support a wide range variety of image formats, the leptonica library was used. This library is an open source library used for image processing applications. Feature detection is one of the OCR approaches which analyzing the lines and strips of the character. Another approach is the pattern recognition which analysis the lines of text and converts the image into a binary matrix where white pixels are 0's and black pixels are 1's.

Document classification is the prediction of what type of document is being processed in the system. The documents need to be converted as an image to train the model and to predict the type of image that is currently processed. To train a model the researcher converts the documents into a group of images. Scanned documents in the form of pdf, the selected page of a document was saved as a jpg file with the use of Photoshop. After creating a group of images, the researchers uses Keras to train the model with tensorflow on the backend. Keras works by defining a sequence of layers in a network by creating a sequential class and adding new layers. This will be done by creating an array of layers and pasting it to the constructor of the sequential model. The first layer of the system must characterize the number of inputs to expect, the way that it is specified defer on the type of network. The layer can also be added with an activation function. This layer will define the format of prediction. After defining the network, it will be compiled which means it will be transformed from a simple sequence into a highly efficient series of matrix intended to be executed in a CPU or GPU depending on the configuration setting.

To establish a server-based digitized records management the system will utilize the use of the OCR tool to determine what kind of document is being uploaded at the moment. The system will then segregate the document's base on their classification. After document classification, the system will upload the document to a certain folder in the server for safe keeping. To obtain this the system PHP should be configure to accept data size of 40M.

Before transferring the file to the server, the system will get the metadata of the file.

To make a timeline visualization the system will need a JavaScipt and CSS embedded in the codes. The JavaScript needed will be the timeline.js and the timeline.css, a vanilla JavaScript plugin to render a responsive, horizontal/ vertical timeline component from plain HTML. This script is copyright of Mike Collins.

To build a real-time notification the system will need AJAX, jQuery, and Bootstrap. The system will fetch records from the backend of the system. Comparing it to the timelines set, once the case will lapse base on the timeline a notification will be viewed on the user interface.

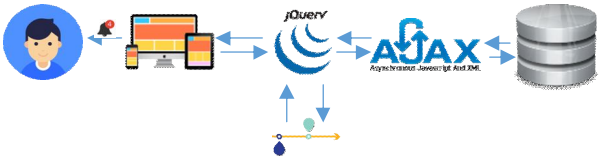

Figure 3: Case Notification

Data Validation and Accuracy

To evaluate the system-generated output or the text extracted from the document, the researchers obtained a copy of the document using word format with the exact set of words as on the extracted text from the document. After extracting the text from the document, the extracted text will then cross-matched with the original text. The number of failed words will be counted and it will be subtracted to the total number of words. The total number of words will serve as the 100% accuracy minus the total words fail it will generate the percent of accuracy. The accuracy test will run through sets of DPI's with ten pages as its parameter at different dots per inch (DPI). Given DPI are 75, 150 and 300 with a document generated from Microsoft word. To count the number of words the researchers utilizes the Microsoft Word application and compare the extracted text to its original text. Then the failed words were count and subtracted to the total number of words.

Data Extraction from APC Document

Convolutional Neural Network (CNN) and Optical Character Recognition (OCR) share the same characteristics, however the major difference is CNN is mostly applied to the analysis of visual imagery while Optical Character Recognition (OCR) is the conversion of images into an electronic text. Per the definition of the two, they almost got the same function and used in terms of extracting text from a document like pdf and image file. Base on the researcher's experience, the most convenient to use is the optical character recognition which requires fewer dependencies and system hardware requirements. CNN is implemented with tools like Tensorflow, a free and open-source library that is used for machine learning applications. The installation of Tensorflow requires python installed on the system, upon training Tensorflow requires Keras installed on the python. To make the training faster, Tensorflow requires a high-end Graphics Processor Card with CUDA toolkit installed a computing platform model created by Nvidia. While ORC is

implemented with the use of a Tesseract an engine that is considered as one of the most accurate OCR engines available.

The researchers decides to use a Tesseract OCR because when it comes to deployment, a system with tesseract engine can be easily deployed without installing other dependencies. A server-based system with a live running system needs extra care in terms of deploying a new system.

Based on the confusion matrix generated with Keras, the accuracy rate for Keras falls to 50% with four predictions for CSR while the true value is HSR, predicted 4 CSR with the true value of 4 CSR. While on Tesseract OCR the document will be detected based on the extracted text. If the system finds a Case Study Report or Home Study Report it will decide whether the document is a CSR or HSR.

Data Extraction using Tesseract

The data was extracted from the APC document using tesseract.js, a javaScript format of Tesseract OCR. The APC document was uploaded to the system and an HTML canvas will hold temporarily the document. The canvas will also serve as a page selection for a multiple page document. The user will select to what page of the document will be extracted. After selecting the page, the user will trigger the extract command to start the extraction process. On the backend of the extraction process, the selected page will be transformed from pdf into an image. The data from the image will be extracted with the use of tesseract, after extracting the text, the results will be split to make new lines using jquery split function. To identify the needed data for data storing will be determined with the use of jquery indexOf function and substring function.

Figure 5 shows the extracted data from the APC document. On the pop-up modal displays a form with corresponding data that was needed to be stored on the database. On the right side is the console where all the extracted data was displayed. The console display is the programmer's determiner if the system runs accordingly.

Department of Social Welfare and Development

Field office Caraga, Butuan City

December 10,2015

CHILD STUDY REPORT

I. IDENTIFYING INFORMATION

Name	: Allie Flores Maghuyop
Sex/Age	: Male/ 10 years and 8 months old
Date of Birth	: March 20,2005
Place of Birth	: Bagumbayan, Dinagat, Surigao del Norte
Birth Status	: Legitimate
Legal Status	: Surrendered
Date Surrendered	: October 1 l, 2010
Date of Placement	: February 2. 2006
Present Whereabouts	: At the custody of the Prospective Adoptive Parents, Mr. 8t Mrs. Alan and Nancy Ebuña in Bagumbayan, Dinagat, Province of Dinagat Islands

Family Composition

Name	Age	Relationship	Educational Attainment	Occupation
Antonio 5. Maghuyop, *Jr*, (deceased)		Father	Elementary Level	

Figure 4: Sample Image for Data Extraction

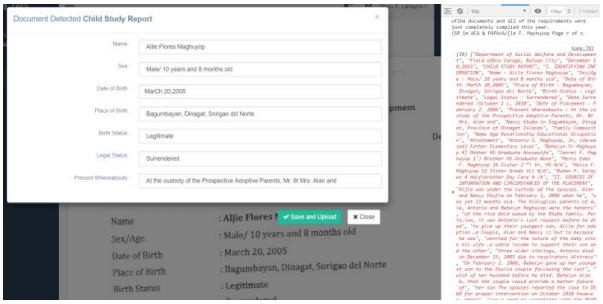

Figure 5: Extracted Data

Document Classification

Documents are classified to determine the specific data being extracted and stored to the database. Document classification follows the data extraction from the APC document. The system is enabled to determine a Child Study Report and Home Study Report. After extracting the data, an argument code was called on a jquery to determine whether the Child Study Report or Home Study Report words were present. With this, the system can determine what type of document was being processed.

Figure 26 shows the part of the form where the name of the detected document was shown. The name of the document was automatically detected during the process of data extraction.

Server Based APC Digitized Records Management

Digitized APC records refer to the scanned document of APC in the form of pdf or image file. After extracting the data from the APC document, important data will be viewed on the browser for editing and finalization. After finalizing the form, the data will be sent to the database while the document will be uploaded to the server using ajax file upload, the form will be sent to a URL using post request.

Timeline Visualization of Successful APC

The data that was stored on the database will be selected and retrieve in order base on the id of the case. The retrieved data will be displayed on a table for inventory. With a button for timeline visualization, the case will be displayed in the form of a timeline. Timeline visualization was made using timeline JavaScript. The package contains CSS for timeline styles.

Case Notification

System notification is in the form of visual notification on the user's dashboard. This notification will show what is lacking in the case for the case manager to respond and comply with the case timelines. The notification was made with the use of font-awesome icons with a badge. This badge will display the number of notifications that needs attention. Notification will be automatically form using jquery document ready function with a query regarding the timeline of the case.

4. RESULTS

To further test the accuracy of Tesseract OCR, the researcher plots the accuracy rate with ten pages of the document at different dots per inch (DPI). Given DPI are 75, 150 and 300 with a document generated from Microsoft word. To count the number of words the researcher utilizes the Microsoft Word application and compares the extracted text to its original text. Then the failed words were count and subtracted to the total number of words.

Table 1 shows the percent accuracy of the document with 75 DPI. The test document contains 10 pages of the different words found in an HSR. Base on the result of the testing, the average accuracy of 75 DPI is 94.8548778.

Table 1 75 DPI Accuracy Percentage

75 DPI			
Page	Number of Words	Failed Words	Percent Accuracy
1	96	10	89.58333333
2	310	19	93.87096774
3	388	7	98.19587629
4	393	16	95.92875318
5	385	12	96.88311688
6	312	22	92.94871795
7	364	17	95.32967033
8	374	17	95.45454545
9	338	15	95.56213018
10	288	15	94.79166667
Average			**94.8548778**

The researchers also tested documents with 150 DPI resolution. Of all the tested resolution, 150 DPI has the highest average accuracy which marks up to 99.23619895.

Another test was conducted on a document with 300 DPI. The document has the same content as the mentioned DPI's before. As shown in table 3, the average accuracy for 300 DPI is 98.91023618.

To have a clean document, the researcher also testes the accuracy of a document that is generated through word application. Table 2 shows the result of accuracy with a clean document. The average accuracy is 98.7416335.

Table 2 Word Processor Generated Result

Word Processor Generated Result			
Page	Number of Words	Failed Words	Percent Accuracy
1	96	1	98.95833333
2	310	6	98.06451613
3	388	3	99.22680412
4	393	3	99.23664122
5	385	3	99.22077922
6	312	3	99.03846154
7	364	1	99.72527473
8	374	6	98.39572193
9	338	8	97.63313609
10	288	6	97.91666667
Average			**98.7416335**

During the testing, it is highly noticeable the number of punctuation marks with error in prediction. The result shows that the higher the resolution, the lesser error in punctuation marks.

5. CONCLUSION

Based on the results and findings of the study, the text from Alternative Parental Care documents can be extracted using Tesseract Optical Character Recognition. This text can be filtered to get the significant words needed to be stored in the database for the timeline visualization and notification. APC document can also be classified parallel to the extraction of text from the APC document.

The system can also upload and save the classified document to the server. A large size document will be accepted by the system. The Saved document will be accessible to APC concerned personnel for future use.

Case visualization and notification were the special features of the system. The case visualization will be generated based on the data about the cases saved in the database. In connection to the case visualization, a timeline notification was also made to notify the user of what is lacking about the case.

Based on the results during testing, the document with 150 DPI resolution scores the most accurate with 99.23619895 %. A Minimal error was noticed during the testing of 150 DPI resolution. It implies that volumes of printed format records can be digitized, data extracted and can be classified without manually entering each character.

Though Tesseract has already a high accuracy rate, it is highly recommended to use other tools or technology that would collaborate Tesseract to obtain much higher accuracy. A higher specification computer is recommended to obtain a faster rendering of image processing results.

For future studies, the use of non-conventional techniques in application development should be explored. This should come with a test that would prove a significant improvement in the output when these technologies are used. Another is the comparative analysis between the systems against those which are available in the market.

REFERENCES

[1] B.U. Kannappanavar; S. T. Rajanikanta, Satish K. Tandur, "Importance Of Digitization Of Library Materials", February 2010

[2] Satyabati Devi & T. A.V. Murthy, The Need for Digitazation. UBS Publisher's Distributors Pvt.Ltd, New DelhiEditors: Singh, Chilana Rajwant, 2005.

[3] Saima Khan, Dr. Shazia Khan, Mohshina Aftab, "Digitization and its Impact on Economy" June 2015

[4] Melissa Hedge, "Cloud Storage VS Local Storage – Which is Right for Your Business?", Feb. 3, 2016

[5] Lingbanan, Importance of Digitization and Content Management, 2016

[6] S. V. R. Prabhakar, S.V. Manjula Rani, "Benefits and Perspectives of Institutional Repositories In Academic Libraries", Jan 2018 https://doi.org/10.21922/srjhsel.v5i25.10948

[7] Phong H Nguyen, Kai Xu, Rick Walker, and BL William Wong, "TimeSets: Timeline visualization with set relations", October 2015 https://doi.org/10.1177/1473871615605347

[8] Oludare Olaleye, Ayodele Olaniyan, Olalekan Eboda1, Adeleke Awolere," SMS-Based Event Notification System", 2013

[9] Wesley Yu, “Why the Notification is the Most Important Email You Can Send for Growth”, November 19, 2014
[10] Chirag Patel, Atul Patel, PhD., Dharmendra Patel, "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study", October 2012 https://doi.org/10.5120/8794-2784
[11] Abbott, R., Del Rincon, J. M., Connor, B., & Robertson, N., "Deep object classification in low-resolution LWIR imagery via transfer learning",2017
[12] Siva Kumar a/l Subramaniam, Siti Huzaimah binti Husin, Yusmarnita binti Yusop,
[13] Prabhu, “A Simple Tutorial to Classify Images Using TensorFlow—Step by Step Guide”, April 27, 2018
[14] Ray Smith, Ray Smith, Dar-Shyang Lee "Adapting the Tesseract Open Source OCR Engine for Multilingual OCR", July 25, 2009
[15] Sagar Sharma, “Activation Functions in Neural Networks”, Sep 6, 2017
[16] Sajid, “Send Email Using Gmail SMTP Server from PHP Scrip”, Jan 21, 2019
[17] Sam Deering, “jQuery String Contains Functions”, Mar 29, 2011
[18] Scott LeeTiernan, Edward Cutrell, Mary Czerwinski, and Hunter Hoffman, “Effective Notification Systems Depend on User Trust”
[19] Segit Prasetya Nugroho, “How to make A Web Push Notification in PHP, JQuery, Ajax and MySQL”, Feb 8, 2017
[20] Sumit Saha “A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way”, Dec 15, 2018
[21] Suncong Zheng, Jiaming Xu, Peng Zhou, Hongyun Bao, Zhenyu Qi, Bo Xu, A neural network framework for relation extraction: Learning entity semantic and relation pattern”, 2016 https://doi.org/10.1016/j.knosys.2016.09.019
[22] Swadhin Pradhan, Lili Qiu, Abhinav Parate, and Kyu-Han Kim, “Understanding and Managing Notifications”, 2017 https://doi.org/10.1109/INFOCOM.2017.8057231
[23] Yash Agarwal, “Create your first Image Recognition Classifier using CNN, Keras and TensorFlow backend”, Jul 8, 2018
[24] How to train Tensorflow models, Jul 18, 2017. Retrieve from https://towardsdatascience.com/how-to-traine-tensorflow-models-79426dabd304
[25] Jamaica B. Lingbanan, “Importance of Digitization and Content Organization”, May 25, 2016. Retrieved from https://www.slideshare.net/jamaicalingbanan1/importance-of-digitization?from_action=save
[26] Responsive Horizontal/Vertical Timeline in Vanilla JavaScript – timeline.js, Nov 5, 2018. Retrieve from https://www.cssscript.com/responsive-horizontal-vertical-timeline/
[27] Tensors and operations, Retrieve from https://www.tensorflow.org/js/guide/tensors_operations
[28] Irfan Ahmad, “How Much Data Is Generated Every Minute?” Retrieve from https://www.socialmediatoday.com/news/how-much-data-is-generated-every-minute-infographic-1/525692/, June 15, 2018