The Seybold Report

ISSN 1533-9211

STUDENT MEDICAL CERTIFICATE VALIDATION USING OPTICAL CHARACTER RECOGNITION

[1]MUHAMMAD AMIRUL ZAQWAN ABDUL RAHMAN, [3]NUZULHA KHILWANI IBRAHIM, [4]NOORAYISAHBE MOHD YAACOB AND [5]MOHAMED DOHEIR

[1,3,4,5] Center for Advanced Computing Technology(C-ACT),

Fakulti Teknologi Maklumat dan Komunikasi

Universiti Teknikal Malaysia Melaka,

76100 Durian Tunggal, Melaka, Malaysia.

[2]ABD SAMAD HASAN BASARI,

[2]Faculty of Computer Science and Information Technology,

Universiti Tun Hussein Onn Malaysia (UTHM)

P.O. Box 101, 86400 Parit Raja, Batu Pahat, Johor Darul Takzim, Malaysia

DOI 10.5281/zenodo.6553561

ABSTRACT

A student medical certificate validation is developed in web-based application. Based on the problem statement, medical certificate is university used to as a record that a student unable to attend to a class. It is simply too easy to obtain a sick certificate and to stay off class. Forged medical certificate has been used by students to absent their class. However, due to lack of integrity among the citizens, people tend to purchase or create forged medical certificates from various website that offers these documents. The objectives of this project are to study the optical character recognition (OCR) technology and document authentication process, and to validate the functionality of the system. The OCR technology method is applied in this project for performing an easy way to check the verification of a medical certificate. The method that have been used is for encode the information data to protect the data from being forge. Thus, the lecturer can verify the medical certificate by using web-based application to get the real information from it.

Keywords: Optical Character Recognition (OCR), Image Processing, Student Leave, Medical Certificate, Open Computer Vision Library (OpenCV)

The Seybold Report
ISSN 1533-9211

1. INTRODUCTION

2. Developing new system and technology for administrations who are receiving medical certificate (MC) to prevent fraudulence, increase accuracy information and harmonies relationship between two parties which for student to lecturer. In order to avoid from people, trick the top management by faking MCs, OCR algorithm will help to solve this situation by verifying the MCs' serial number from database's server either it is valid or invalid. It also allows an instant communication to both parties through FTMK system for MCs submission and makes sure the validity of the MCs as well as increase accuracy and quality of data which can achieve higher productivity and decrease traditional communication way. I believe this system is very important to an organization in managing and approving the MCs because it can prevent frauds and build transparent, productivity, efficiency as well as harmony culture from all parties.

This project uses optical character recognition (OCR) which is an electronic or mechanical conversion of text-typed, handwritten or printed images into machine-encoded text, whether from a scanned document, a document photograph, a scene photograph or a subtitle text superimposed on a picture.

This technique is widely used to generate data entry by processing MC samples and scanning the image using an OCR algorithm. Then, the data will verify the serial number of samples either it is valid or invalid. In organizations, admins may take a few days to process the MCs because usually patient will inform the medical leave using phone calls, messages, or emails even though the MCs are not originally proven and submit the hard copy in the next working day or class. In this circumstance, employees or students will take advantage to the management as long as they give the hard copy and their problems solved.

3. LITERATURE REVIEW AND METHODS

According to Emanuel Goldberg (1914), optical character recognition may be traced to technologies involving telegraphy and creating reading devices for the blind. The machine that machine that read characters and converted them into standard telegraph code. In the late 1920s and into the 1930s, he developed the new technology that called a "Statistical Machine" for searching microfilm using an optical code recognition system.

As a result, Optical Character Recognition (OCR) is beneficial technology for the user because by implementing OCR it can make a document can easily be edited and has low-cost processing, hence it improves how your business operates.

2.1 Problem Analysis

Currently, it is necessary to authenticate the number series printed on the medical certificate by contacting the doctor or clinic to ensure that the medical certificate is valid, in order for the administration to check the authentication of certain medical certificates. It will take time to check the medical certificate too.

The Seybold Report
ISSN 1533-9211

Besides this, there has been an increase in the use of fake medical certificates in students based on the work done in the last few years. Furthermore, colleges require a medical certificate, and they neglect the method of obtaining a medical certificate, so they take a shortcut. The lecturer expects students to present a medical certificate under current working conditions if they have been absent from work due to illness. For the lecturer, medical certificates are used as identifiers. The incidence of fraud in the form of forged medical certificates is rising at the workplace, though. If an applicant cannot prove the validity of a medical certificate, disciplinary proceedings may be brought against him or her. The lecturer will make an intentional effort to ensure that the use of fraudulent medical certificates is halted or stopped.

2.2 Technique

There are two core OCR algorithm core types, which may produce a ranked list of candidate characters. Matrix matching involves pixel-by - pixel comparison of an image to a stored glyph; it is also called pattern matching, pattern recognition, or image correlation. This depends on the input glyph being isolated correctly from the rest of the image, and on the glyph being stored in a similar font and at the same scale. This technique works best with the typewritten text and when new fonts are found it does not work well. That is the technique implemented, rather directly, by the early physical photocell-based OCR.

2.3 System Architecture

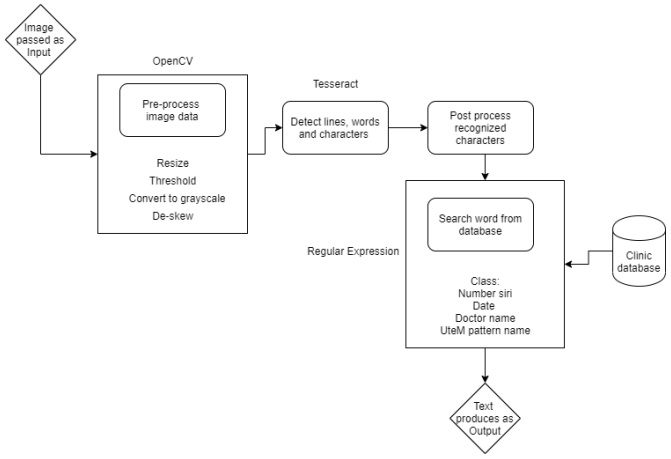


Figure 1 refers to an artificial intelligence technique, the program will start with an input image that is a medical certificate from the UTeM Clinic. Essentially, this device engine is designed to provide high accuracy for word and line detection. The new edition of Tesseract is 4. Adds a new neural net Long Short-Term Memory based OCR engine that focuses on the line recognition but also supports the existing Tesseract OCR engine that operates by identifying character patterns

The Seybold Report
ISSN 1533-9211

2.2.2 Line and Word Finding

The line finding algorithm is designed to allow recognition of the skewed page without having to de-skew, saving the loss of image quality. Blob filtering and line building are the principal parts of the cycle. Where a page layout analysis has already provided roughly uniform text sizes for text regions, a simple percentile height filter removes drop-caps and vertically touching characters. The median height is about the size of the text in the region, so filtering out blobs smaller than a fraction of the median height, most likely punctuation, diacritical marks, and noise is safe.

Filtered blobs are more likely to match the pattern of a non-overlapping, parallel yet sloping line. The x-coordinate sorting and processing of blobs allows the assignment of blobs to a unique text line while tracking the slope across the page, with a significantly reduced risk of assigning an incorrect text line in the presence of skew. When the blobs filtered are assigned for the lines, the basic lines are estimated with a minimum of the squares fit and the blobs filtered are returned to the respective lines. The finale phase includes blobs that overlap at least half horizontally, which together with a correct basis put diacritical marks and which correctly match sections of certain broken characters.

The basic lines are mounted more precisely using a quadratic spline once these lines are found. This was a further first for the OCR system, which allowed Tesseract to handle pages that contain curved baselines commonly used for scanning, not just book connexons. Baselines are equipped to separate the blobs in groups which displace the initial straight base line with a relatively continuous effect.

The quadratic division has an average square size fit to the most popular section. The quadratic spline benefits from the relatively stable calculation but the disadvantage of discontinuities when more than one spline segment is required. A more conventional cubic spline that function better.

To determine if they are fixed pitches, Tesseract tests the text lines. When fixed pitch text is found, Tesseract copies words into pitch characters and deactivates the chopper and the associator to recognize the sentence. An extremely non-trivial feature is unfixed pitch or relative spacing text. By measuring gaps in a limited vertical range between the base line and the middle line, Tesseract resolves most of these issues. Spaces like a threshold are made blurred at this point so that after a word of acknowledgement a final decision can be made.

2.2.3 Word Recognition

The identification of how a word should be divided into characters is part of the recognition process for any character recognition engine. The original output is first classified for the line finding segmentation. The other step in the word recognition is for the unfixed pitch text only. Tesseract tries to change the result by cutting the blob off with the worst confidence of the classifying character. Candidate chop points are

The Seybold Report
ISSN 1533-9211

found in polygonal contour approximation concave vertices and can either have a concave opposite vertex or line segment.

If the chops were exhausted, if the word was still not good enough, the associate would be given it. The associate searches for possible combinations of the highest blobs in the candidate characters in the A * (best first) segmenting chart. It does this without building a segmentation graph but maintains instead a hash table of the visited countries. The A * search is further deleted from the preferential list and tested by the classification of unclassified fragment combinations.

2.2.4 Static Character Classifier

A two-step process is to continue classification. The class pruner produces a shortlist of classes which the unknown can correspond to. Each function has an ill-quantified 3-dimensional search table, a little vector of groups that it can be fitted, and all the features have bit vectors summarized. The classes with the highest counts are the shortlist for the next step after correcting the expected number of features. The function of the unknown searches a small vector of the prototypes of that class which might fit, and then determines the actual similarity between them. For each concept called a configuration, that character class prototype is represented by a logical sum-of-product statement that records the overall similitude demonstration of each function in each configure as well as each prototype during the distance measuring process. Best of all stored class configurations is the best combined distance to be calculated from the summarized feature and prototype proof.

2.2.5 Lingustic Analysis

In Tesseract, there is very little language study. When considering a new division of languages, in each one of the following categories the language module selects the best word string available: Top frequently, Top numeric word, Top UPPER case word, Top less than the optional original upper word, Top classifier choice. For a certain segmentation, the final decision is simply a word with the lowest total distance rating, where each of the above categories is multiplied by a different constant. Words of various segmentations may contain different numbers of character. These words can hardly be compared directly, even if Tesseract claims to produce probabilities. This problem is resolved by generating two numbers for each Tesseract character classification. The first, called confidence, is the less the standardized prototype distance. This allows "confidence" because larger numbers but also a distance are stronger, because the further from zero, the greater the distance. The second output, called the rating, divides the normalized prototype length by the unknown character. The second output is called the rating. Character ratings can be meaningfully summarized within the word since the overall length of the outline is always the same for all characters in one word.

2.3 Implementation Architecture

ISSN 1533-9211

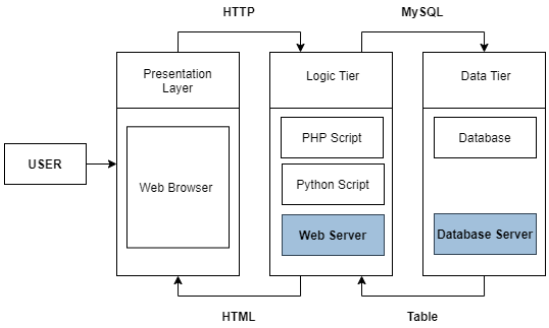
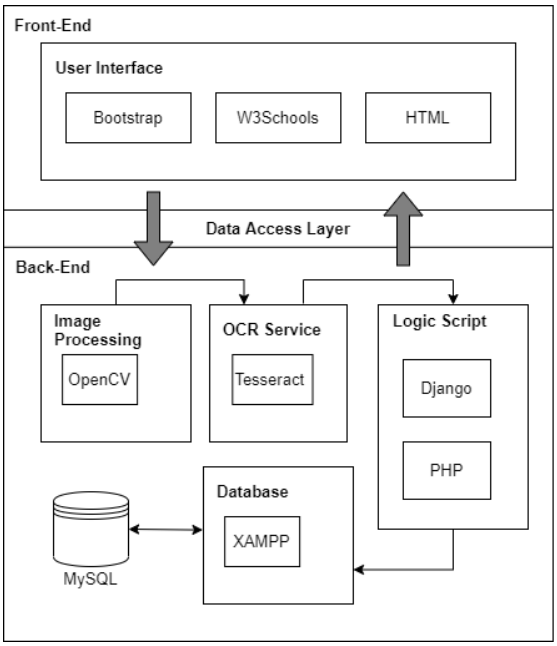


Figure 2: System Architecture

Figure 2 demonstrates the three-layer device tier architecture used to build this framework. The interface layer reflects the user's client device and applications

used to communicate with the framework like Google Chrome. After that, the logic tier reflects the language of the programming and the service used to allow the user to access the web site, such as Apache. The XAMPP framework is used in this program since it comes with the kit containing the language PHP and Python, and the framework Apache. The third tier is the data tier contained in the database server, such as the MySQL database, with both data and tables.

This will consist of W3Schools, Bootstraps and another section for the front-end compartment to design the user interface for display to the end-user. The Django and PHP Services are the back-end mechanism for building web application business process and logic. The model feature is a getter setter to bring the value to the back-end cycle between the front-end components.



The Seybold Report
ISSN 1533-9211

Figure 3: System Architecture

The web application is created with the use of the programming language PHP and Django. PHP and Django are chosen to build the framework backend, since XAMPP database is supported by PHP and Django. Model, View, and Controller (MVC) are implemented as a design pattern in Django architecture. As a design pattern, PHP architecture follows the Model, View, and Controller (MVC). Rather of integrating all the components in a single folder, the separate folder is less safe and difficult to maintain source code.

2.4 Technical Design

Firstly, grayscale is a monochromatic variety of shades, between black and white. Thus, the grayscale image has only grey shades and no color. Even color images contain grey scale material, while digital images may be saved as black and white greyscale images. This is because each pixel has a color-independent luminance value. Brightness or intensities that can be measured from null and black to complete white can also be described as light.

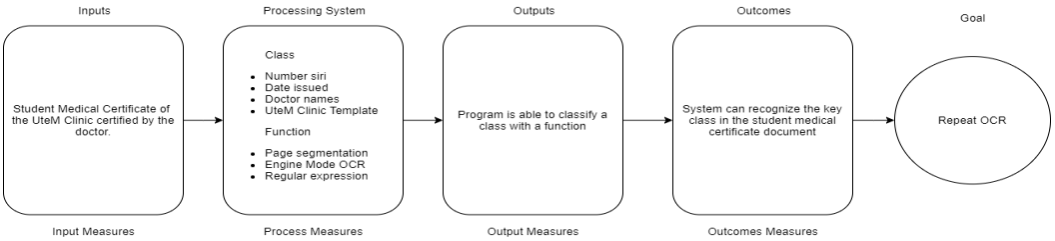
Next, thresholding is an image segmentation process used for the generation of binary images. There are two types of thresholds, namely fundamental and adaptive thresholds. It consists of a non-linear translation that transforms a grey image into a binary image, where the pixels below or above the defined threshold value are assigned to both levels. A simple threshold method is selected in this OpenCV. In operation, a standard value is assigned to pixels with values greater than the threshold value specified.

Furthermore, it also helps to read all types of images supported by pillow and the Leptonica image libraries such as jpeg, png, gif, bmp or tiff, and others as a self-contained invoking script for Tesseract. Furthermore, if used as a script, the known text may be printed by Python-tesseract instead of written to a file. The function used in this system is image by string that returns the result of the OCR on the image to the string

The standard expression is a string sequence that allows you to fit or locate certain strings or strings using a separate template syntax. The UNIX world uses regular expressions. The Python re module provides full support for Perl-like regular expressions in Python. The re-module increases the derogation for re.error when an error occurs when a regular expression is compiled or used. Two important functions to handle regular expressions would be covered. But first, a little: when used in regular expressions, different characters have special significance. To avoid confusion while dealing with regular expressions, we would use Raw Strings as an r'expression.'.

Overall, the result will come out with a logical function by referring to the text of the image.

The Seybold Report
ISSN 1533-9211



3.RESULT

Based on the observation in Figure 5, there are two forms of sample method used to define whether post-training and testing improvements are required. Total research sample that test 10 and total testing sample that train 20. From this dataset, the percentage of accuracy in the algorithm can be verified by extracting the requirements info from the medical certificate clinic.

The process in PSM phase 1 of the graph reveals that 73.33% of success results with 29.67% loss. After analysis, the method of image processing is the key factor in the percentage of accuracy. In this algorithm sequence, which uses a binary, scaled and threshold method in phase 1.

Firstly, is for a multi-color image (RGB), which converts a binary phase into a black and white image. Some algorithms are used to convert the color image to a more complex zonal analysis from a simple threshold. However, most OCR engines deal internally with monochrome pictures and convert it as a first step in a monochrome color.

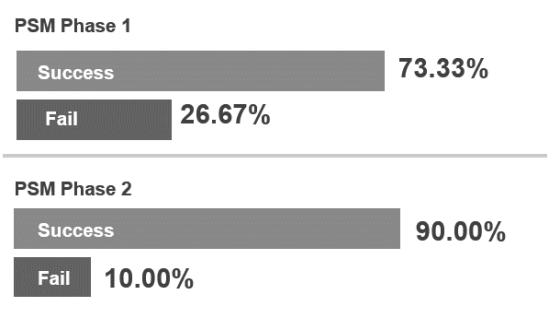
The second one, the scaled approach used is to ensure that the images are scaled to the correct size. The scale factor work along the horizontal and vertical to adjust the size of the image pixel and also the interpolation in the scale process used by the inter cubic increases the number by 4x4 by transferring the images to large pixels, helping the OCR process to identify the text in the image of small characters.

Third, the threshold is an important way of extracting important, pixel-coded information while reducing background noise. The method will display the mask and pick the foreground by using it. The two threshold functions used are a simple threshold, the first argument of which should be a grayscale, the second classifying the pixel values, and the third the maximum value applied to the pixel values that surpass the threshold, combined with an adaptive threshold, which is to be centered in lighting conditions in various areas where the algorithm sets the threshold for a pixel based on a specific region around it. Different thresholds for different regions of the same image that provide better results to improve accuracy by masking the text character.

The Seybold Report
ISSN 1533-9211

Next, the process in PSM Phase 2 of the graph increase to 90.00% of the performance results with a decrease of 10.00% of the loss. After the analysis, the improvement comes with the shift due to the modification of the algorithm sequence functions in the image processing where the scale, smoothing, binary and threshold method is to be used in phase 2. The same function as scale, binary process, and threshold only one function has been added to this algorithm is a smoothing method to minimize blurring in the image. This smoothing technique visual effect is a smooth blur that is close to the image viewed through a transparent screen. The role is performed by a Gaussian blur. Edge detection is typically used for gaussian smoothing. The standard deviation value of the function is 3, which means that a small amount of Gaussian solution can help to smooth the noise phase.

Overall, the process results in an excellent percentage of accuracy of the medical certificate through an improvement of 16.67% and reduction by the loss of 19.67%.



Type	Fail	Success
Training	6 of 20	14 of 20
Testing	2 of 10	7 of 10

Type	Fail	Success
Training	2 of 20	18 of 20
Testing	1 of 10	9 of 10

Accuracy Percentage = Training Sample + Testing Sample / Total Sample * 100%

Figure 5: Result Accuracy Percentage

Table 1: Testing Sample

Training Sample				
Sample	Class Detect		Result	
	Phase 1	Phase 2	Phase 1	Phase 2
1	4	4	Good	Good
2	4	4	Good	Good
3	4	4	Good	Good
4	4	4	Good	Good

The Seybold Report
ISSN 1533-9211

5	4	4	Good	Good
6	2	2	Bad	Bad
7	3	3	Good	Good
8	4	4	Good	Good
9	4	4	Good	Good
10	4	4	Good	Good
11	4	4	Good	Good
12	4	4	Good	Good
13	2	2	Bad	Bad
14	2	3	Bad	Good
15	4	4	Good	Good
16	4	4	Good	Good
17	3	3	Good	Good
18	0	2	Bad	Good
19	0	4	Bad	Good
20	0	3	Bad	Good

Note:

Criteria to detect best result at least 3 of 4 class (UTeM Clinic Pattern, Doctor Name, Number Series,

Date Issued)

Table 2: Training Sample

Testing Sample				
Sample	Class Detect		Result	
	Phase 1	Phase 2	Phase 1	Phase 2
1	3	3	Good	Good
2	4	4	Good	Good
3	4	4	Good	Good
4	4	4	Good	Good
5	4	4	Good	Good
6	0	3	Bad	Bad

The Seybold Report
ISSN 1533-9211

7	4	4	Good	Good
8	4	4	Good	Good
9	3	3	Good	Good
10	0	4	Bad	Good

Note: Criteria to detect best result at least 3 of 4 class (UTeM Clinic Pattern, Doctor Name, Number Series, Date Issued)

4. DISCUSSION

Improvements can be made to the web application framework a precision and speed to extract the text. The student will ably get better process of extracting good quality of the medical certificate document.

Next, because it does not have password protection, this device has an unreliable user authentication misbehavior. This device therefore requires the password encryption to protect from approved use all user personal information such as e-mail, and matrix number to get the password and log in to the device using the appropriate username and password.

At the end, this system should have session and database auditing to audit all user login and logout session and the user activities throughout accessing the system. Admin can access all the changes data, the activities, and the operations in the database by using database auditing procedure.

5. CONCLUSION

The conclusion that can be concluded after completing this system is the developed system has eased the student and the lecturer to apply and tracking leave that has been applied. This system has met its objective and solve the main problem that has been identified previously in this report but improve for the better performance and for the future use is still needed. All the proposition of improvement needs to be implemented to make the system more reliable and secure.

ACKNOWLEDGEMENT

This paper is part of research work under the Fundamental Research Grant Scheme (FRGS) number FRGS/2018/FTMK-CACT/F00395. The research is conducted in the Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM).

REFRENCES:

[1] Zelic, F. (2020, July 31). [Tutorial] OCR in

Python with Tesseract, OpenCV and Pytesseract.

AI & Machine Learning Blog.

The Seybold Report
ISSN 1533-9211

[2] Foong, N. W. (2020, August 4). A Beginner's Guide to Tesseract OCR - Better Programming. Medium.

[3] Rosebrock, A. (2020, July 16). Tesseract OCR: Text localization and detection. PyImageSearch.

[4] The Web framework for perfectionists with deadlines | Django. (2005). Django Is a Python-Based Free and Open-Source Web Framework.

[5] PHP: Hypertext Preprocessor. (2020, August 21). PHP: Hypertext Preprocessor

[6] PHP: MySQLi - Manual. (2005). MySQLi – Manual - PHP.

[7] RegExr: Learn, Build, & Test RegEx. (1950). RegExr

[8] AI Courses by OpenCV.org. (2020, August 4). OpenCV

BIBLIOGRAPHY

[9] Digital image processing techniques. (1985). Computer Vision, Graphics, and Image Processing, 29(3), 394. https://doi.org/10.1016/0734-189x(85)90134-3

[10] reh. (2012). OCR — Optical Character Recognition. Orthopädie & Rheuma, 15(1), 58. https://doi.org/10.1007/s15002-012-0032-x

[11] Ramesh, N., Srivastava, A., & Deeba, K. (2018). Improving Optical Character Recognition Techniques. International Journal of Engineering & Technology, 7(2.24), 361. https://doi.org/10.14419/ijet.v7i2.24.12085

[12] Lin, T., Qiang, B. H., Long, S., & Qian, H. (2013). Deep Web Data Extraction Based on Regular Expression. Advanced Materials Research, 718–720, 2242–2247. https://doi.org/10.4028/www.scientific.net/amr.718-720.2242

[13] Monaco, M. (2018). Regular Expressions 101. Technical Services Quarterly, 35(3), 305–306. https://doi.org/10.1080/07317131.2018.145686

The Seybold Report

ISSN 1533-9211