www.rspsciencehub.com Volume 02 Issue 12S December 2020

INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB

e-ISSN : 2582 - 4376
Open Access

RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Science, Technology & Management (ICSTM-2020)

Analyzing and Experimenting Open Source OCR Engines in RPA with Levenshtein Distance Algorithm

Malathi T[1], Diwaan Chandar C S[2], Nithish S[3], Niranjan V[4], Swashthika A K[5]

[1]Assistant Professor, Department of Computer Science & Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu - 638401

[2,3] Second Year, Department of Information Science & Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu - 638401

[4, 5] Second Year, Department of Computer Technology, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu - 638401

niranjan.ct19@bitsathy.ac.in[4]

Abstract

Robotic Process Automation is a platform used to automate boring and repetitive computer processes using software bots so that humans could involve in tasks which include creativity and decision making which could not be done by robots. Optical Character Recognition takes out printed characters in an image and converts it to text. Google Tesseract OCR and Microsoft OCR were the commonly used OCR engines available in UiPath, a tool for Robotic Process Automation. In Previous, research on comparing those two open source OCR engine, there we made comparison on basic factors which included speed, hardware requirements, accuracy ,but in that case, accuracy was been calculated manually which gave us results but with less precise, as it was a manual process to substitute scraped data to that formulas, In this research we've made results with more precision by performing a String comparison algorithm named, "Levenshtein Distance Algorithm" which is deployed in UiPath.

Keywords: Optical Character Recognition(OCR); Robotics Process Automation(RPA); Google Tesseract OCR; Microsoft OCR

1. Introduction

Robotics and automation has stepped into reality a few years ago and is evolving so rapidly around the world in areas such as industrial automation, space engineering, stellar space engineering, even in urban and rural areas all over the world. I always wonder how these programs work seamlessly 24 by 7 hours,[1]as it was designed to do that actually, As we use this technology to overcome bored repetitive tasks and human risky jobs, we often deploy them daily to do that kinda tasks, However they depend on humans to get engaged with that technology,[7] there are many myths in this century regarding future predictions about these technologies (mainly robots or Artificial Intelligence) that they'll overrule the humans by making their own decisions and works which involve their own creativity. [4]But these theories clearly states that machines can't stand a chance against human intelligence cause they're still competing with our skills,[3]but there are many tasks which humans seek for technology help like repetitive tasks as mentioned above to make the result more efficient and with more precision, and there's many investors currently investing billions of dollars in these kind of technologies to seek more compound profit in future. Microsoft OCR, which is a built in OCR engine in Microsoft windows 10 and Tesseract OCR,[2]an open source OCR engine developed by Google were the two available open source OCR engines in UiPath, a tool for Robotic Process Automation. In the previous paper[1]research made is by checking the accuracy of Tesseract OCR and Microsoft OCR by using some manual methods, which is not precise. Also, we had used different sets of images for testing the accuracy and had also used systems of different specifications for this research which may result in error for time taken and accuracy percentage. Hence, to propose a more valid result, we had planned to improve our results by using a string comparison algorithm named, “Levenshtein algorithm”, which is used to calculate the similarity between two input strings and returns it's accuracy in percentage. We had also used the same set of images for testing both OCR engines. And executed the workflow on the same system in order to calculate the time taken for the execution error-free.

2. String comparison - levenshtein distance algorithm

The operation accepts two strings and returns the percentage of similarity between two strings using the Levenshtein Algorithm in the System.Single form. The Levenshtein algorithm (also referred to as Edit-Distance) calculates the minimum number of editing operations needed to change one string in order to obtain another string. The dynamic programming approach is the most prevalent way of measuring this. A matrix is initialized and the Levenshtein distance between the m-character prefix of one is calculated in the (m,n)-cell with the n-prefix of the other term. From the upper left to the lower right corner, the matrix can be filled. An insert or a delete, respectively, corresponds to each hop horizontally or vertically. The cost is typically set at 1 for each of the tasks. If the two characters in the row and column do not match or 0, if they do the diagonal jump will cost either one. The expense is often reduced locally by each cell. This way, the Levenshtein gap between both terms is the number in the lower right corner.

If it is needed to compare a string to some sample data, this could effectively be used. For example, the status of an application needs to be updated on the basis of a statement from Approvers. "The request is approved "Application is approved"The request raised is approved yesterday"Yesterday's request is approved. The traditional string comparison methods would not operate in such instances, but the operation will give a percentage of similarity.

3. Methodology

In this research proposal, an string comparison algorithm plays an important role to give more accurate results than our previous study, the whole sequence of the execution will be like; unzipping and feeding the data from our local machine storage to the workflow ocr engines(either Microsoft OCR engine or Tesseract OCR Engine first) ; and redirecting the extracted data to the string algorithm analysing container as show at *figure 2.1* and there the major part plays on comparing the extracted data with the original data from the images which was used to feed the OCR engines, and eventually the accuracy(data) will be saved to local storage or can also use cloud storage for purpose, if we’re deploying this workflow to the UiPath Orchestrator.

The main upgrades from the previous paper[1] is about:

Same set of source data is supplied to both ocr machines to expect better comparison results

Both workflows for ocr has been executed with same hardware equipments whereas used different hardwares for previous comparison[1]

Images with lighten backgrounds are used to extract more data to obtain more precise information.

3.1 What does container refer to?

Images with fancy or decorable fonts are used to test the algorithm [1]Containers or Blocks which is often used in uipath studio to classify set of activities or program in an order to execute in a sequential manner, if a container is set to top level node, the activities which is under that container executes first and then the workflow further moves to next container which holds set of instructions readily to run followed by the previous block.

4. Architecture and workflow

4.1 Architecture

Some people might be good at programming by nature. But it's not necessary to be everyone. There are many people, who always struggle to grasp the basics of programming or simply they can't program. [1]Hence UiPath studio offers a no-code environment where we can enter with minimal or zero programming background using some visual code blocks.[1]There were several boring and repetitive tasks in Information Technology and Business Process which may be mundane for many employees. Hence, we could deploy a bot so that humans could involve in some other creative activities and other activities which include human decisions in order to make an accurate process in a very low time.[1]The primary architecture of UiPath software consists of three components; UiPath Studio, UiPath Robot, UiPath Orchestrator which plays a vital role in automating a task:

[1]UiPath Studio is a design tool that enables a user to create programs.[1]It has many activities(pre-defined functions) and repositories that are predefined. To model the workflow for the automation process, users can drag and drop activities. In simple terms, UiPath Studio is a method used to model the automation workflow to automate repetitive processes using predefined activities and libraries. And, UiPath Robots is a software that hosts the process installed in the UiPath studio that allows us to carry out our projects with or without human monitoring (or) supervision on any computer.UiPath Orchestrator is a web application that allows us to manage the development and deployment of our resources in our machine. It allows us to launch and schedule our bots on our or other desktops, and also control the bot's status and evaluate the effects of their work.

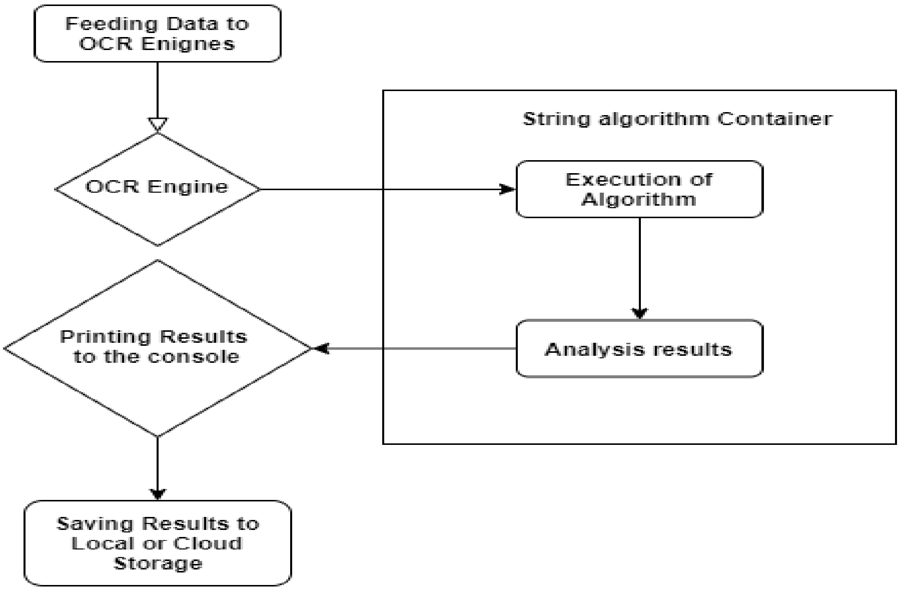

Feeding Data to OCR Enignes
OCR Engine
String algorithm Container
Execution of Algorithm
Analysis results
Printing Results to the console
Saving Results to Local or Cloud Storage

Fig.1. OCR Flow chart

4.2 Workflow

This research consists of two workflows, one for Tesseract OCR and another for Microsoft OCR. A folder which consists of 100 images is given as input. An excel sheet which consists of Original Text data of the image is used for the calculation of the results. The original text data is read as string 1 for comparison on both workflows and it extracts the text data using Tesseract OCR on Workflow 1 and Microsoft OCR on Workflow 2. The extracted data is given for String 2 for the comparison on respected workflows. Now, the extracted data by Tesseract OCR and Microsoft OCR were fed into the columns C and D of the excel sheet.[1]And the similarity between the string 1 and 2 (String 1 is same for both OCR and String 2 is the extracted text respected to the OCR engine) were calculated and the percentage is returned in columns E and F respectively. Similarly the time taken for the execution has been calculated on the UiPath Studio for both OCRs. Finally, the average of accuracy for the 100 samples and average time taken for the extraction of a single image of each OCR has been calculated.

5. Result analysis

a.) Mean Accuracy- There is a measure of the similarities between the original text and the extracted text calculated by using Levenshtein Distance Algorithm-String Comparison in uipath, and the value is determined between them, and finally the mean accuracy is calculated for the determined values b.) Overall Execution Time (or) Time taken - The time taken for 100 images to be extracted including the time taken for the string comparison is calculated and tabulated. c.) Mean time taken(per images) - The average time taken to extract each picture is calculated and calculated and tabulated from the total time taken, and mean time taken value will be calculated from the total time taken divided by total no of images fed.

Table 1. Characteristics/ Engine

Characteristics/ Engine	Tesseract OCR Engine	Microsoft OCR Engine
Mean accuracy(in percentage)	71.04549606	76.68340455
Time taken(in hour format)	00:19:11.00	00:27:08.00
Mean time taken(s)per image	00:00:11.51	00:00:16.28

6. Error analysis

Some datas has been classified below acc to their accuracy percentage,

Table 2. Tesseract OCR

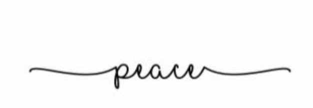
Data/accuracy	Original	Extracted text
0 - 20%	"STAY HOME STAY SAFE"	"m xfiom ? smv gm"
20 - 40%	"Never forget 3 types of people in your life: 1. Who helped you in your difficult times. 2. Who left you in your difficult times. 3. Who put you in your difficult times."	"Never forget 3 types of p_eople In your life: LVMNVIMJMEOMM 2.V\m mmhmdmmm mummywindrrmnm"
40 - 60%	"FASTER THAN A MERCEDES MY HEART BEATS FOR YOU Loesje P.O.-BOX 1045 6801 BA ARNHEM HOLAND"	"FASTER THAN A MERCEDES . MY HEART BEATS FOR YOU 1922/;"
60 - 80%	HEAVEN, YOUR HANDS TOOK ME THERE."" -Lenon Hodson"	"""YOUR EYES PROMISEO ME HEAVEN, YOUR HANDS TOOK ME THERE.""""
100%	BLACKLANE	BLACKLANE

Table.3 Microsoft OCR

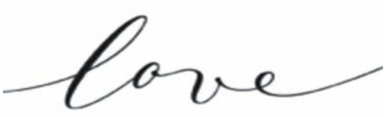
Data/accuracy	Original	Extracted Text
0 - 20%	"STAY HOME STAY SAFE"	<null>
20 - 40%	"Never forget 3 types of people in your life: 1. Who helped you in your difficult times. 2. Who left you in your difficult times. 3. Who put you in your difficult times."	"Never forget 3 types of people un your life: CCu I. In attest b•tm 3. in dTÄE Sa•"
40 - 60%	"To heal a wound you need to stop touching it."	"To heal woun yoti toucl?inæ it."
60 - 80%	Hello World!	t lollo World!
100%	Blacklane	Blacklane

Note: All images were taken in Joint Photographic group format (.JPG or JPEG).

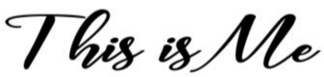
6.1 Decorative font outcomes



peace



love

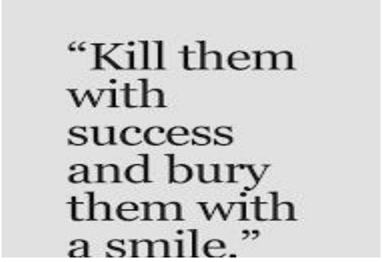


This is Me

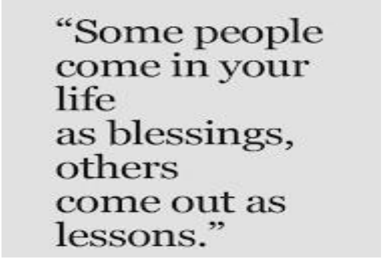
For these types of fonts, OCR Engines returned null values and also random alphabets such as, Microsoft ocr returned null and Tesseract returned “fı‘lﬂl/” for the value “**love**” as in the above image.

6.2 Images with grey background

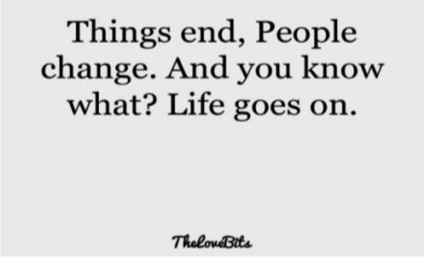
Images with grey background provide (attached above) an accuracy of nearly 80 - 90% in both Google Tesseract and Microsoft OCR.



“Kill them
with
success
and bury
them with
a smile.”



“Some people
come in your
life
as blessings,
others
come out as
lessons.”



Things end, People
change. And you know
what? Life goes on.

TheLoveBits

7. Future works

[1]So, the overalls analysis and this comparison research has been done under three major factors which includes, velocity, accuracy, time taken, using distance algorithm, majorly this research was the improvised version of the previous research which gave more precise results over the first one,[1]but there are still some portions of this research can be upgraded for better purposes such as like deploying storage area from local storage to cloud storage which will be used to manipulate data over worldwide collaborators for further analysis,[1] so,moreover the future work will be on storing data with cloud support, Which'll be also used to run other OCR Engines available in the UiPath framework?

Conclusion

By making calculations at different factors, such as precision, time taken, to put the experiment to an end.[1] In certain cases the results of the Microsoft OCR are more reliable compared to the results of the Tesseract OCR. But Tesseract OCR also gives better results in certain cases as well.This provides different accuracy values.But when considering the time it took to identify the characters in the images compared to Microsoft OCR, Tesseract got efficient result,[1]So by performing these comparison with the use of String comparison algorithm

Acknowledgements

The Authors sincerely thank Ms.Malathi T, Assistant Professor department of Computer science and Engineering, Bannari Amman Institute of technology, Sathyamangalam, Erode- 638 401 for guiding us for the successful completion of this research.

10. References

[1] Malathi T, Diwaan Chandar C S, Nithish S, Niranjan V, Swashthika A K , An experimental performance analysis on robotics process automation(RPA) with open source OCR engines: microsoft OCR and tesseract OCR. ICMMM2.0, 2020.

[2] S.V. Rice, F.R. Jenkins, T.A. Nartker, The Fourth Annual Test of OCR Accuracy, Technical Report 95-03, Information Science Research Institute, University of Nevada, Las Vegas, July 1995.

[3] R. Smith, “A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation”, Proc. of the 3rd Int. Conf. on Document Analysis and Recognition (Vol. 2), IEEE 1995, pp. 1145-1148.

[4] A.C. Kak e M. Slaney. Principles of Computerized Tomography. IEEE Press, 1988.

[5] S.V. Rice, G. Nagy, T.A. Nartker, Optical Character Recognition: An Illustrated Guide to the Frontier, Kluwer Academic Publishers, USA 1999, pp. 57-60.

[6] P.J. Schneider, “An Algorithm for Automatically Fitting Digitized Curves”, in A.S. Glassner, Graphics Gems I, Morgan Kaufmann, 1990, pp. 612-626.

[7] L.R. Franca Neto, C.A.B. Mello and R.D. Lins. Filtering Techniques for Digital Images of Historical Documents. XV Brazilian Symposium of Telecommunications, Recife, Brazil,September, 1997.

[8] B.A. Blesser, T.T. Kuklinski, R.J. Shillman, “Empirical Tests for Feature Selection Based on a Psychological Theory of Character Recognition”, Pattern Recognition 8(2), Elsevier, New York, 1976.

[9] R.D. Lins, M.S. Guimar~aes Neto, L.R. Franca Neto and L.G. Rosa. An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming, pp. 111-121,North-Holland, January, 1995.

[10]R.J. Shillman, Character Recognition Based on Phenomenological Attributes: Theory and Methods, PhD. Thesis, Massachusetts Institute of Technology. 1974.

[11]C.A.B.Mello, L.R.Franca Neto e R.D.Lins. A New Technique for Compressing Static Images. Proceedings of the XVI Computation Brazilian Society Congress, Brazil, 1996.

[12]R.W. Smith, The Extraction and Recognition of Text from Multimedia Document Images,

PhD Thesis, University of Bristol, November 1987.

[13]P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley-IEEE, 2003.

[14]K.Sayood. Introduction to Data Compression. Morgan Kauffman Publishers, Inc., 1996.

[15]I.H. Witten, A.Moffat and T.C. Bell. Managing Gigabytes - Compressing and Indexing Documents and Images. Van Nostrand Reinhold, 1994.