A Comparative Study of Optical Character Recognition in Health Information System

Ribeiro, Mário R. M.
Algoritmi Research Centre
Department of Informatics, University of Minho
Braga, Portugal
mario.rmr.1337@gmail.com

Duarte, Júlio
Algoritmi Research Centre
Department of Informatics, University of Minho
Braga, Portugal
jduarte@di.uminho.pt

Vasco, Abelha
Algoritmi Research Centre
Department of Informatics, University of Minho
Braga, Portugal
id6616@alunos.uminho.pt

António, Abelha
Algoritmi Research Centre
Department of Informatics, University of Minho
Braga, Portugal
abelha@di.uminho.pt

José, Machado
Algoritmi Research Centre
Department of Informatics, University of Minho
Braga, Portugal
jmac@di.uminho.pt

***Abstract*— Most Health Institutes are transitioning between documents in physical format and digital format. It is pertinent and important to develop applications that helps health professionals on this transition. An application that would aid the process of digitalization of documents was developed using a Python library. To help with the decision of which library to use, a study was made regarding the precision and speed of execution of PyOCR, PyTesseract and TesseOCR.**

Keywords—: OCR, Wrapper, Python, HIS

I. Introduction

For the effective functioning of any health entity, whether hospitals or clinics, public or private, a division is required responsible for the reception, classification, conservation and availability of documents associated with clinical activity. This division is usually referred as the Clinical Archive. We are currently in a period where most of these divisions are transitioning between documents in physical format and digital format, working with both formats simultaneously. It is pertinent and important to develop applications that facilitate this transition to obtain the highest rentability from this hospital division. In partnership with the Clinical Archive of the Hospital da Senhora da Oliveira in Guimarães, an application that would aid the process of digitalization of the documents was developed. The destination of these documents is AIDA platform. To achieve this goal a Python platform was developed that uses the technology of Optical Character Recognition, namely the open source engine Tesseract.

A. AIDA

Agency for Integration, Diffusion and Archive of Medical Information (AIDA) is a platform that tries to overcome the difficulty of integration of all clinical systems, as well as support the medical and administrative complexity of different Hospital information sources [1, 2]. AIDA is currently installed at some major Portuguese hospitals. It is an electronic platform that provides employees with intelligence featuring a pro-active behavior in its main functions: communication between heterogeneous systems, storage management and hospital information; response to requests in time; sending and receiving information from hospital sources like laboratories, medical reports, images, prescriptions, and others. AIDA establishes connection with all Systems of medical information: EHR; Administrative Information System (AIS); Medical Information System (MIS); and Nursing Information System (NIS) [3, 4]. AIDAS's covers all tasks needed to execute a medical examination. At the same time, AIDA agents ensure that information is shared with other hospital subsystems. Therefore, clinical professionals can also access all information through their specifics systems of record. The information will still be available in other platforms like MIS, NIS or AIS but the AIDA importance is to assemble and to provide patient health record at one place.

B. OCR Technology

OCR, the acronym for "Optical Character Recognition" refers to the concept of recognition, analysis and understanding of characters through an optical mechanism. In the human being, this concept is represented by the ability to read, the eyes being the optical mechanism and the brain, namely the Wernicke area [6], the analysis and understanding of the input provided. In the scope of technology, OCR is the electronic or mechanical conversion of text, be it manuscript or typography, in machine language. The first concept of OCR was patented in 1929 by Tausheck in Germany, while in 1933, Handel did the same in the United States of America. These are the first known OCR records. However, it was only in the 1950s, with the arrival of computers, that this technology went from theory to practice.

The workings of OCR technology can be understood in five phases. These phases are Scanning, Segmentation, Preprocessing, Character Extraction and Recognition. In the first step, a digital image of the original document is obtained through a camera or scanner. These devices convert the received light intensity to gray levels. Normally, since most of the documents that are to be scanned are composed of information represented by black color on a white background, the digital image will be converted to a black and white image. This conversion is achieved through the thresholding method where pixels with gray levels that are below a certain number are converted to white and those above that number are converted to black. In the second step, segmentation, the distinction between written text and images is made. It is also at this stage that all text is segmented into the most basic components, isolating each word and each character. The scanned image may contain some noise which may resolve to errors in the character recognition step. In the third step we intend to eliminate this problem through a preprocessing of the image. The resolution of this problem involves the smoothing and normalization of characters, where "holes" in the characters are corrected through fill techniques and the size, angle and rotation of the characters are corrected. In the fourth stage, considered the most difficult, a search is made regarding the characteristics that allow the identification of a symbol, ignoring the rest. In the last phase, the raised characteristics are compared to a set of known characteristics to be able to identify the corresponding character, thus ending the image to text conversion. [7, 8, 10]

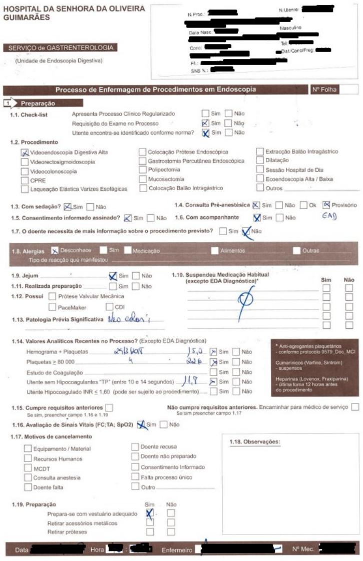
C. *Tesseract*

Tesseract is an open source OCR software developed by Hewlett Packard between 1984 and 1994. In 1995 it was featured in the UNLV Annual Test of OCR Accuracy where it obtained excellent results when compared to other available software. Its development began as a PhD project and grew as a possible addon to the HP product line, namely the scanners. Motivated by the fact that OCR technologies are still underdeveloped and after a collaboration with HP Labs Bristol and HP's Scanner Division, Tesseract has gained a leading edge in recognition accuracy over other commercially available software. Despite this leadership Tesseract would only be available in open source in 2005.

The Tesseract works through a series of traditional steps. In the first step the input image is converted into a binary image containing only the black and white colors. In the second step, there is an analysis of the components where their contours are stored. This phase has a very high computational cost, but it brings a significant advantage to the process: it becomes much simpler to detect text with inverted colors (white text on a black background), making it as easy as recognizing black text on a white background. This phase distinguishes Tesseract as the first software to be able to handle inverted-color text in such a trivial way. At the end of this phase, the contours are converted into Blobs. Blobs are organized into lines of text that are later parsed to detect anomalies in the standard size of the contours. The lines of text are then divided into words using the space between the characters as a reference. The stage of recognition occurs in two phases. In the first phase an attempt is made to recognize the previously separated words. Each word that is successfully recognized is added to the reference data. With this addition of data, a second recognition attempt is made, which corresponds to the second phase. Finally, a step occurs to correct the less obvious spaces and check alternatives to the vertical axis to locate lowercase text. [5, 9, 11]

D. *Resources*

In partnership with the person in charge of the Clinical Archive of Hospital da Senhora da Oliveira, a survey was made of the documents that enter this department. A sorting was then carried out with regarding the type of document to satisfy two conditions. The first would be the existence of such a volume of documents necessary to carry out the tests. The second condition refers to the model of the document, as it was crucial that they present the information that is to be extracted in a visible and clear way. From this screening came two types of documents ideal for the study in question. Then, a quality screening was carried out for the documents, eliminating any copies that contained information illegible to the human eye. The two types of document selected are shown in the images below.



HOSPITAL DA SENHORA DA OLIVEIRA
GUIMARÃES

SERVIÇO de GASTRENTEROLOGIA
(Unidade de Endoscopia Digestiva)

N.Proc.: [redacted] N.Utente: [redacted]
[redacted]
Data Nasc: [redacted] Masculino
[redacted]
Tel: [redacted]
Conc: [redacted] Dist/Conc/Freg: [redacted]
[redacted]
Fl.: [redacted]
SNS N.: [redacted]

Processo de Enfermagem de Procedimentos em Endoscopia Nº Folha

1 Preparação

1.1. Check-list
Apresenta Processo Clínico Regularizado □ Sim □ Não
Requisição do Exame no Processo ☒ Sim □ Não
Utente encontra-se identificado conforme norma? ☒ Sim □ Não

1.2. Procedimento
☒ Videoendoscopia Digestiva Alta
□ Videorectosigmoidoscopia
□ Videocolonoscopia
□ CPRE
□ Laqueação Elástica Varizes Esofágicas
□ Colocação Prótese Endoscópica
□ Gastrostomia Percutânea Endoscópica
□ Polipectomia
□ Mucosectomia
□ Colocação Balão Intragástrico
□ Extracção Balão Intragástrico
□ Dilatação
□ Sessão Hospital de Dia
□ Ecoendoscopia Alta / Baixa
□ Outros ____

1.3. Com sedação? ☒ Sim □ Não
1.4. Consulta Pré-anestésica ☒ Sim □ Não □ Ok ☒ Provisório
1.5. Consentimento informado assinado? ☒ Sim □ Não
1.6. Com acompanhante ☒ Sim □ Não CAD
1.7. O doente necessita de mais informação sobre o procedimento previsto? □ Sim ☒ Não

1.8. Alergias ☒ Desconhece □ Sim □ Medicação ____ □ Alimentos ____ □ Outras ____
Tipo de reacção que manifestou ____

1.9. Jejum ____ ☒ Sim □ Não
1.11. Realizada preparação ____ □ Sim □ Não
1.12. Possui □ Prótese Valvular Mecânica
□ PaceMaker □ CDI
1.13. Patologia Prévia Significativa Não colar;

1.10. Suspendeu Medicação Habitual (excepto EDA Diagnóstica)*

	Sim	Não
Ø	□	□
	□	□
	□	□
	□	□
	□	□

1.14. Valores Analíticos Recentes no Processo? (Excepto EDA Diagnóstica)

Hemograma + Plaquetas	25/3/2018	15,0	☒ Sim	□ Não
Plaquetas ≥ 80 000	4	226	☒ Sim	□ Não
Estudo de Coagulação			□ Sim	□ Não
Utente sem Hipocoagulantes "TP" (entre 10 e 14 segundos)		11,8	☒ Sim	□ Não
Utente Hipocoagulado INR ≤ 1,60 (pode ser sujeito ao procedimento)			□ Sim	□ Não

* Anti-agregantes plaquetários
- conforme protocolo 0579_Doc_MCI
Cumarínicos (Varfine, Sintrom)
- suspensos
Heparinas (Lovenox, Fraxiparina)
- última toma 12 horas antes do procedimento

1.15. Cumpre requisitos anteriores □
Se sim, preencher campo 1.16 e 1.19
Não cumpre requisitos anteriores. Encaminhar para médico de serviço □
Se sim preencher campo 1.17

1.16. Avaliação de Sinais Vitais (FC;TA; SpO2) ☒ Sim □ Não

1.17. Motivos de cancelamento
□ Equipamento / Material
□ Recursos Humanos
□ MCDT
□ Consulta anestesia
□ Doente falta
□ Doente recusa
□ Doente não preparado
□ Consentimento Informado
□ Falta processo único
□ Outro ____

1.18. Observações:

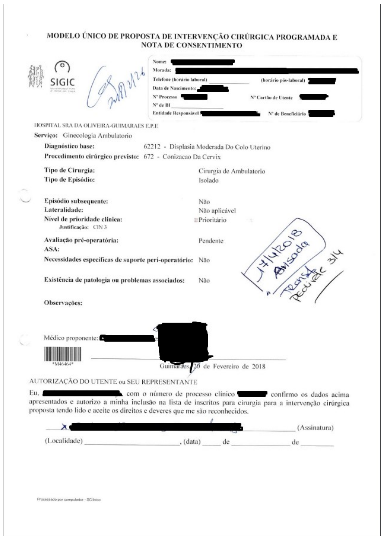
1.19. Preparação

	Sim	Não
Prepara-se com vestuário adequado	☒	□
Retirar acessórios metálicos	□	□
Retirar próteses	□	□

0546_Reg_AOP

Data [redacted] Hora [redacted] : [redacted] Enfermeiro [redacted] Nº Mec. [redacted]

Figure 1 - Type 1 Document



MODELO ÚNICO DE PROPOSTA DE INTERVENÇÃO CIRÚRGICA PROGRAMADA E NOTA DE CONSENTIMENTO

SIGIC

21/2/18

Nome: [redacted]
Morada: [redacted]
Telefone (horário laboral) ____ (horário pós-laboral) [redacted]
Data de Nascimento: [redacted]
Nº Processo [redacted] Nº Cartão de Utente [redacted]
Nº de BI ____
Entidade Responsável [redacted] Nº de Beneficiário [redacted]

HOSPITAL SRA DA OLIVEIRA-GUIMARAES E.P.E
Serviço: Ginecologia Ambulatorio
Diagnóstico base: 62212 - Displasia Moderada Do Colo Uterino
Procedimento cirúrgico previsto: 672 - Conizacao Da Cervix

Tipo de Cirurgia: Cirurgia de Ambulatorio
Tipo de Episódio: Isolado

Episódio subsequente: Não
Lateralidade: Não aplicável
Nível de prioridade clínica: Prioritário
Justificação: CIN 3

Avaliação pré-operatória: Pendente
ASA:
Necessidades específicas de suporte peri-operatório: Não

Existência de patologia ou problemas associados: Não

17/4/2018
Avisada
Transf
Pediatric 3/4

Observações:

Médico proponente: [redacted] [redacted]

M46464

Guimaraes, 20 de Fevereiro de 2018

AUTORIZAÇÃO DO UTENTE ou SEU REPRESENTANTE

Eu, [redacted], com o número de processo clínico [redacted] confirmo os dados acima apresentados e autorizo a minha inclusão na lista de inscritos para cirurgia para a intervenção cirúrgica proposta tendo lido e aceite os direitos e deveres que me são reconhecidos.

X [redacted] ____ (Assinatura)
(Localidade) ____, (data) ____ de ____ de ____

Processado por computador - SClínico

Figure 2 - Type 2 Document

II. Development

In this phase the development and execution of tests regarding the performance of the chosen wrappers using the documents and the materials already mentioned were carried out. Since the goal would be to extract the process number, an eight-digit number that exists as an identifier, as soon as possible, 4 different tests were performed that vary in the area of the analyzed document for each combination of library and document. In the first test the entire document was analyzed and in the second test only the vignette where the process number is found is analyzed. In the third and fourth tests a horizontal and vertical bar is analyzed which contain the process number to be extracted.

The parameters chosen for evaluation are speed and accuracy. To evaluate the accuracy a system was created that detects four types of errors. In the cases where the number extracted differs from the original by a maximum of 1 or 2 characters it is considered Error Type 1. When more than one number is extracted, one of which is the correct one, it is considered Error Type 2. When the number extracted contains 3 or more wrong digits, it is considered Error Type3. Finally, if no number is extracted, it is considered Error Type 4. To evaluate the speed, a counter has been implemented that records the time that the area of the document in question takes to be analyzed.

The test algorithm is divided into four phases. In the first phase the document is prepared for analysis. Through the ImageMagick library this process begins by transforming the pdf document type to the highest quality document type possible, considering library compatibility. In the case of the PyOCR and PyTesseract libraries the .tiff was chosen and in the case of TesseOCR the .jpeg was chosen. Then the image resolution is set to 300. The next step corresponds to the appropriate cropping of the image. After this process, the image is ready for phase two. In this phase the methods of the libraries that perform OCR in the image obtained in the previous phase are executed. It is at this stage that the time is recorded that will be used to evaluate the parameter of speed. After extracting the information, it is necessary to filter it to make the parsing of the relevant information, filtering the unnecessary. This goal is achieved using regular expressions. This process maintains any join of eight and only eight consecutive digits, discarding everything else and corresponds to phase three.

Finally, at phase four, the results obtained are compared the intended value. The success of the analysis or the type of error are then recorded. The time obtained in the information extraction phase is also recorded.

III. RESULTS

Table 1. Precision data for Type 1 Document regarding the total area

Library	Success	Type 1	Type 2	Type 3	Type 4
PyOCR	38,46%	15,38%	3,85%	7,69%	38,46%
PyTesseract	46,15%	23,08%	19,23%	0,00%	11,54%
TesseOCR	30,77%	38,46%	26,92%	0,00%	3,85%

Table 2. Precision data for Type 2 Document regarding the total area

Library	Success	Type 1	Type 2	Type 3	Type 4
PyOCR	28,21%	0,00%	48,72%	12,82%	10,26%
PyTesseract	10,26%	0,00%	71,79%	10,26%	7,69%
TesseOCR	12,82%	2,56%	74,36%	2,56%	7,69%

Table 3. Precision data for Type 1 Document regarding the vignette area

Library	Success	Type 1	Type 2	Type 3	Type 4
PyOCR	34,62%	26,92%	7,69%	3,85%	26,92%
PyTesseract	42,31%	26,92%	23,08%	3,85%	3,85%
TesseOCR	38,46%	30,77%	23,08%	3,85%	3,85%

Table 4. Precision data for Type 2 Document regarding the vignette area

Library	Success	Type 1	Type 2	Type 3	Type 4
PyOCR	25,64%	0,00%	30,77%	17,95%	25,64%
PyTesseract	20,51%	0,00%	35,90%	28,21%	15,38%
TesseOCR	28,21%	0,00%	33,33%	20,51%	17,95%

Table 5. Precision data for Type 1 Document regarding the horizontal bar

Library	Success	Type 1	Type 2	Type 3	Type 4
PyOCR	61,54%	11,54%	0,00%	0,00%	26,92%
PyTesseract	80,77%	11,54%	0,00%	0,00%	7,69%

TesseOCR	65,38%	26,92%	3,85%	0,00%	3,85%

Table 6. Precision data for Type 2 Document regarding the horizontal bar

Library	**Success**	**Type 1**	**Type 2**	**Type 3**	**Type 4**
PyOCR	38,46%	2,56%	15,38%	20,51%	23,08%
PyTesseract	33,33%	0,00%	25,64%	20,51%	20,51%
TesseOCR	35,90%	5,13%	28,21%	17,95%	12,82%

Table 7. Precision data for Type 1 Document regarding the vertical bar

Library	**Success**	**Type 1**	**Type 2**	**Type 3**	**Type 4**
PyOCR	53,85%	19,23%	0,00%	0,00%	26,92%
PyTesseract	73,08%	3,85%	11,54%	0,00%	11,54%
TesseOCR	69,23%	11,54%	7,69%	0,00%	11,54%

Table 8. Precision data for Type 2 Document regarding the vertical bar

Library	**Success**	**Type 1**	**Type 2**	**Type 3**	**Type 4**
PyOCR	48,72%	0,00%	2,56%	0,00%	48,72%
PyTesseract	41,03%	5,13%	17,95%	0,00%	35,90%
TesseOCR	64,10%	2,56%	10,26%	0,00%	23,08%

Table 9. Speed results regarding document type 1

Library	**Total Area**	**Vignette Area**	**Horizontal Bar**	**Vertical Bar**
PyOCR	24,07s	6,62s	2,54s	5,03s
PyTesseract	25,18s	7,49s	2,84s	5,89s
TesseOCR	22,53s	5,83s	2,39s	5,06s

Table 10. Speed results regarding document type 2

Library	**Total Area**	**Vignette Area**	**Horizontal Bar**	**Vertical Bar**
PyOCR	14,55s	5,70s	3,68s	4,55s
PyTesseract	15,01s	6,32s	3,88s	4,69s
TesseOCR	12,85s	5,44s	3,01s	3,86s

IV. DISCUSSION

As for the precision metrics in document type 1, the library that showed the best results was PyTesseract, constantly obtaining a higher success rate in all tests performed. The remaining libraries presented very similar results, with slight advantage for the TesseOCR library. However, the PyOCR library presents a less varied distribution in the type of error, being predominant the Error Type 4, whereas the TesseOCR library presents greater variety. As for the second typology of documents, the results obtained allow us to conclude that the PyOCR library presents a better performance when the original image edition is minimal. In contrast, the TesseOCR library performs best when the information to be extracted is concentrated in one area.

As for the metric of speed, it is concluded that the TesseOCR library is clearly the fastest to perform the information extraction, followed by the PyOCR and PyTesseract libraries. Since the horizontal area and the vertical area analyzed contains the same number of pixels, it is concluded that the vertical area encompasses more information in the type 1 document than in type 2, and the reverse is true for the horizontal area. This means that the ideal area of analysis of the document will vary according to the typology. That is, it is not possible to obtain an area of analysis that behaves in an ideal way for any document.

V. CONCLUSION

By conducting these tests and subsequent analysis of the results it is possible to draw some conclusions about the performance of the three libraries under study. The PyTesseract library stood out in the precision metric, sacrificing runtime. It would be the most appropriate library in cases where time is not an important factor. The TesseOCR library stands out for the fast execution with better success rates than the PyOCR library when the area of analysis is more restricted, that is, when the image quality is lower. This would be the library to use when speed is the most relevant factor in the process. Finally, the PyOCR library presented better execution times than the PyTesseract library, but worse than the TesseOCR library. However, it showed a better performance when the area of analysis is larger. This library would be indicated when the scanning process does not allow image preprocessing.

VI. REFERENCES

[1] Duarte, J., Salazar, M., Quintas, C., Santos, M., Neves, J., Abelha, A., Machado, J.: Data quality evaluation of electronic health records in the hospital admission process. In: 2010 IEEE/ACIS 9th International Conference on Computer and Information Science (ICIS). (2010) 201–206.

[2] Duarte, J., Portela, C.F., Abelha, A., Machado, J., Santos, M.F.: Electronic health record in dermatology service. In Cruz-Cunha, M.M., Varajão, J., Powell, P., Martinho, R., eds.: ENTERprise Information Systems. Volume 221 of Communications in Computer and Information Science. Springer Berlin Heidelberg (2011) 156–164.

[3] Pereira, R., Duarte, J., Salazar, M., Santos, M., Abelha, A., Machado, J.: Usability of an electronic health record. In: 4ht IEEE International Conference on Industrial Engineering and Engineering Management, Hong Kong (2012).

[4] Duarte, J., Pontes, G., Salazar, M., Santos, M., Abelha, A., Machado, J.: Stand-alone electronic health record. In: Proceedings of the 2013 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM2013), Bangkok, Tailand (2013).

[5] R. Smith, “An Overview of the Tesseract OCR Engine,” 2005.

[6] K. R. Pugh, W. E. Mencl, A. R. Jenner, L. Katz, S. J. Frost, J. R. Lee, S. E. Shaywitz, and B. A. Shaywitz, “Neurobiological studies of reading and reading disability,” J. Commun. Disord., vol. 34, no. 6, pp. 479–492, 2001.

[7] S. Mori, C. Y. Suen, and K. Yamamoto, “Historical Review of OCR Research and Development,” Proc. IEEE, vol. 80, no. 7, pp. 1029–1058, 1992.

[8] A. Verma, S. Arora, and P. Verma, “OCR-OPTICAL CHARACTER RECOGNITION,” pp. 181–191.

[9] C. Patel, A. Patel, and D. Patel, “Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study,” Int. J. Comput. Appl., vol. 55, no. 10, pp. 50–56, 2012.

[10] R. Mithe, S. Indalkar, and N. Divekar, “Optical Character Recognition,” Int. J. Recent Technol. Eng., vol. 2, no. 1, pp. 72–75, 2013.

[11] S. Rakshit and S. Basu, “Recognition of Handwritten Roman Script Using Tesseract Open source OCR Engine,” Natl. Conf. NAQC, pp. 141–145, 2010.