JOSÉ CARLOS GUTIÉRREZ MENÉNDEZ

End-to-end system for extracting and interpreting Textual Information of Interest from Identity Documents Images

Sistema de ponta a ponta para extração e interpretação das Informações de Interesse Textuais a partir de Imagens de Documentos de Identidade

São Paulo 2019

JOSÉ CARLOS GUTIÉRREZ MENÉNDEZ

End-to-end system for extracting and interpreting Textual Information of Interest from Identity Documents Images

Sistema de ponta a ponta para extração e interpretação das Informações de Interesse Textuais a partir de Imagens de Documentos de Identidade

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências.

Área de Concentração:
Engenharia de Computação

Orientadora:
Profª. Dra. Graça Bressan

São Paulo 2019

Catalogação-na-publicação

Gutiérrez, José Carlos
Sistema de ponta a ponta para extração e interpretação das informações de interesse textuais a partir de imagens de documentos de identidade / J. C. Gutiérrez -- São Paulo, 2019.
124 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Reconhecimento de texto 2.Documentos de identificação I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

DEDICATORY

To my parents and grandparents for giving me their love and support accepting the difficulty of being so far away and separated for so long. I owe everything to them, and I promise we will be together soon.

To my wife for being always at my side helping me to build our future and for being just as she is, perfect. Her love makes me capable of everything.

To my brother for being my lighthouse and my example to follow. Now that we are together, we will change the world.

To Rodolfo, for being like another brother to me and always being there when I need it. Difficult moments become easier thanks to his friendship.

ACKNOWLEDGEMENTS

I would especially like to thank the supervisor of my research, Professor Graça Bressan. I thank her for accepting me as her student and giving me one of the best opportunities of my life. In addition, for her patience and help, her understanding for both academic and personal issues, and for guiding me through a good path that opens the doors to new opportunities. Sincerely, I will always appreciate and will be infinitely grateful for all the help that Professor Graça gave me.

Also, I would like to thank the University of São Paulo, especially LARC and my colleagues and professors who received me and gave me the opportunity to increase my knowledge and develop this research project. Moreover, to the teachers who taught me throughout my life and formed the basis of my knowledge.

I thank all my family and all the people I have met in my life for making me who I am. Especially to the following:

- To my mom for being the sweetest and best person in the world, for the great amount of love, dedication and education that she has always given me, for all the daily effort she makes to take care of our family.
- To my dad for all his advice, education and father's love, and for being the smartest person I know, for being my example and showing me that what really matters is to be professional in everything I do.
- To my brother Armandito for being literally as an angel, for teaching me the ways I should go and that our goal is to be like dad, for being with me and supporting me as when we were children.
- To all my grandparents, each one of them has given me a lot of love. I also owe my grandmother Sonia everything in this world, she raised us, took care of us and keeps doing it and always enjoys doing it. To my grandmother Lidia, who God carried close to him for her tenderness and love, thanks for teaching me the word of God.

To my grandfather Manolito, for always showing me how much he loves us, for all his advice and teachings, for being my example to follow by showing the courage he has to keep working forever. To my grandfather Guti, for the education he transmits to me, for always being so funny, for showing me that the years do not matter when you really love a person.

- To Sandra, my greatest treasure in this world, for being the love of my life, for being with me always giving me happiness and making me feel loved, for all the love and support she gives me, for making all my worries disappear when we are together. Also, I would like to thank *her* mother and grandmother for becoming part of my family *and all the help they gave to me*.
- To Rodolfo, whose friendship has been unconditional, for all the help he has given me and for being like another brother to me. Many of my achievements are thanks to him.
- To my Japanese-Brazilian brother Marcelo, since from the first time we met, he was always willing to help me. I thank him for always helping to make my stay in Brazil better.
- To Baby for supporting me and helping me in everything. I thank her for being like a mother to me here in Brazil.
- To my best friends Roger and Yosvany for their unique friendship and let me know that I can always count on them.
- To my Cubans friends here in Brazil, to name a *few, José* Enrique, Nelson, Claudia, Heydi, Victor, Manuel, Fernando, Loubrys and Patricia; that together we have made this new stage in our lives *more fun and happier*.
- To God for everything and for being the energy that allows me to continue

Finally, to all people who in one way or another have been kind enough to give me their help and who contributed directly or indirectly to the conclusion of this work. Thanks to all...

RESUMO

Os documentos de identidade (ID) são uma das principais fontes para obter informações sobre um cidadão. O centro de muitas aplicações nos setores administrativos e de serviços é a extração dos dados contidos nos cartões de identificação. Portanto, nesta pesquisa é proposta a implementação de um sistema automatizado capaz de extrair e interpretar as informações textuais a partir de imagens de documentos de identidade. O sistema de ponta a ponta proposto permite a automação de um processo de registro ou verificação que requer a aquisição de informações sobre um cidadão usando seus documentos de identidade. O sistema obtido através desta pesquisa é considerado como um sistema de ponta a ponta, uma vez que abrange todas as etapas do processo de extração das informações de interesse a partir de imagens de IDs.

Diferente dos sistemas baseados em modelos, o sistema proposto usa um algoritmo de atribuição semântica que permite classificar e atribuir significado às informações dos IDs baseado nas semânticas destas. Esta pesquisa é a primeira descrição abrangente de um sistema completo de extração de informações para processar IDs que descreve desde o processamento da imagem até o reconhecimento da entidade nomeada. Para avaliar o desempenho da pesquisa, foram propostas diferentes métricas baseadas nas funções internas do sistema. A avaliação final mostra resultados satisfatórios indicando que o sistema de ponta a ponta é capaz de extrair e interpretar informações textuais de imagens de documentos de identidade sem conhecimento prévio de seus layouts.

Keywords: Documentos de Identificação; Reconhecimento e Classificação da Entidade Nomeada; Reconhecimento de Texto.

ABSTRACT

Identity documents (ID) are one of the primary sources for obtaining information about a citizen. The center of many applications within the administrative and service sectors is the extraction of the data contained in ID cards. Therefore, in this research is proposed the implementation of an automated system able to extract and interpret the textual information from identity documents images. The proposed end-to-end system allows the automation of a registration or verification process that requires the acquisition of information about a citizen using his identity documents. The system obtained through this research is considered as an end-to-end system since it covers every stage of the information of interest extraction process from IDs images.

Different to the template-based systems, the proposed system uses a semantic attribution algorithm that allows to classify and attribute meaning to the information from IDs according to its semantics. This research is the first comprehensive description of a complete information extraction system to process IDs that describes from image processing to named entity recognition. To evaluate the performance of the research were proposed different metrics based on the internal functions of the system. The final evaluation shows satisfactory results showing that the end-to-end system is capable of extracting and interpreting textual information from identity documents images without prior knowledge of their layouts.

Keywords: Identification Documents; Named Entity Recognition and Classification; Text Recognition.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

2D	Two-Dimensions
3D	Three-Dimensional
BB	Bounding Boxes
BTH	Date Of Birth
BoW	Bag-Of-Words
CCs	Connected Components
CD	Candidate Definition
CO	Consolidation Rules
CR	Candidate Refinement
DIP	Digital Image Processing
EM	Electromagnetic Spectrum
ER	Extremal Regions
FD	Feature Definition
ID	Identity Document
IDN	ID Numbers
IE	Ie Information Extraction
IT	Information Technology
KEY	Keywords
MSER	Maximally Stable Extremal Regions
NER	Named Entity Recognition
NERC	Named Entity Recognition And Classification
NLP	Natural Language Processing
OCR	Optical Character Recognition
PER	Person Names
RoI	Regions Of Interest
SE	Structural Element
SSL	Semi-Supervised Learning
SWT	Stroke Width Transform

Chapter 1

1 Introduction

Processing of large volumes of data in a few seconds has been made possible by the rapid growth of technology. Cisco reports show that in just one second the global Internet traffic reached more than 20,000 GB in 2016, the company predicts that by 2021 the value will increase to 105,800 GB per second (Cisco, 2017). Therefore, for economic and business reasons, the transformation of all the information of physical documents into digital data stored in computers is highly demanded by companies. Thereby, the digital data can be accessible to any information processing system. However, textual information from documents is often typed and processed manually by human operators, which implies a time-consuming and prone to errors process (Cheriet, Kharma, Liu, & Suen, 2007; Kaur & Garg, 2015).

In order to reduce both the associated human error and the time involved in the entire process of transcribing the data, the textual information contained in the documents could be extracted by automatic processing systems. Systems with the ability to recognize and extract text from images are also known as Optical Character Recognition (OCR) system and they are the core of many industrial computer vision applications (Cheriet et al., 2007; Islam, Mondal, Azam, & Islam, 2016; Lukás Neumann & Matas, 2013b; Simon, Rodner, & Denzler, 2015).

Computer vision systems are oriented to understand and process one or a set of images in the development platform in which they are executed. The detection and recognition of Regions of Interest (RoI) are some of the main tasks performed within any computer vision system. In order to achieve these objectives, multiple digital image processing (DIP) techniques must be employed. Documents to be processed by a computer vision application may contain, besides text, graphics, and images that overlap. Therefore, the DIP techniques to be used to process the RoI depends on the objective of the application.

1.1 Research context

Identity documents (ID) are one of the primary sources for obtaining information about a citizen. The center of many applications within the administrative and service sectors is the extraction of the data contained in ID cards. Over the years, the personal information required to perform an enrollment or registration process has been manually extracted from these documents. Nevertheless, thanks to the advancement of technology, manual work has been minimized by the application of computer systems capable of extracting this data using different types of scanners (flatbed scanner, barcode scanner, etc.). Moreover, in the last decade, the performance of mobile devices has been improved every year, allowing them to be able to capture an image, processing it and get results similar to those of a flatbed scanner (Chabchoub, Kessentini, Kanoun, Eglin, & Lebourgeois, 2016; de las Heras, Terrades, Llados, Fernandez-Mota, & Canero, 2015; Kaur & Garg, 2015; Sharma & Sharma, 2016). Features such as the size, mobility and ability to instantly capture images in an easy way, compared to a scanner, have made smartphones receive more attention in their use as a mobile scanner (Bai, Yin, & Liu, 2013; Jacobs, Simard, Viola, & Rinker, 2005; Sharma & Sharma, 2016).

The text data present in ID cards, such as the full name of the bearer, date of birth, identification number, etc. are part of the information required by the person responsible for completing the admission form during an enrollment or registration process. Nevertheless, a smartphone user can speed up the process since a photo of its ID can be taken by himself. Thus, an OCR system able to process the ID image can extract all the textual information, thereby reducing the amount of human work involved.

Usually, pictures taken by smartphone cameras are captured in natural scenes. Distortion of perspective, different styles of text and light sources that introduce shadows and reflections are some of the drawbacks present in the images captured in natural scenes that hinder the process of text recognition (Bai et al., 2013; de las Heras et al., 2015; Zhu & Zanibbi, 2016). Therefore, in order to efficiently segment the text of the rest of the objects, an integration of different

DIP algorithms must be performed. Moreover, the information extracted must be accurate since a recognition error can generate a mistake during the registration.

An automated system in charge of performing a registration process in addition to extracting all the information must be able to classify it by its meanings. In the literature, there are some systems able to extract the information already classified based on the template of the identity documents processed (de las Heras et al., 2015; Ryan & Hanafiah, 2015; Simon et al., 2015). The use of templates to extract the textual information from IDs guarantees a fast processing since the location of this one is previously known. However, documents with different formats varies the position in which is located the information contained in an identity document. Thus, systems that use templates to classify information based on their positions are limited by the number of templates they could recognize. Moreover, a person must previously select the regions of interest that identify where is located the information of the citizen.

In the interest of automatically identify each information in an ID without using templates, the system must be able to classify textual information of interest by its meaning. Based on the pattern of identity documents (not in the layout), there are some specific words used as headings that attribute semantic value to the information to which they refer (Figure 1-1). Therefore, finding these Keywords and using them as classifiers allows the system to identify, for instance, which line refers to the name or ID number.

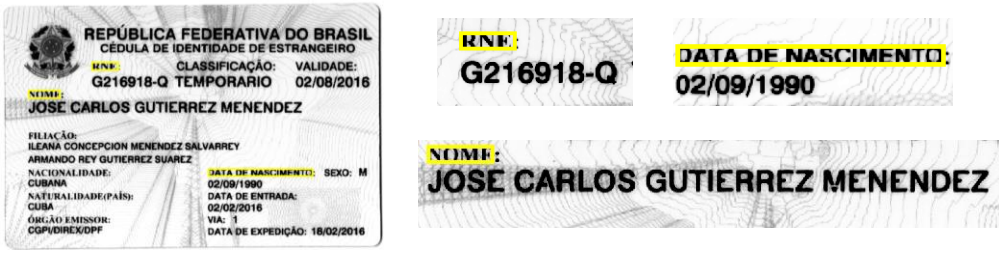

Figure 1-1 Image of an identity document of a foreign resident in Brazil. Marked in yellow, examples of Keywords that identify and classify information in an ID. Source: The Author.

The proposed system must perform different digital image processing algorithms to achieve detect and extract the textual data of the citizen from an ID. Further, a mechanism able to assign semantic value to information referred by Keywords have to be implemented in order to automate a registration process.

1.2 Motivations

In modern times, almost all information is processed digitally to take advantage of the computational power of digital systems and the high scalability offered by the Internet. Therefore, there is a growing demand aimed at developing real-time automatic systems capable of emulating human abilities, thus reducing possible manual errors and processing time.

A registration process is intended to establish the identity of a person with an institution. Admission to enter a company building, rent a car, open a bank account, travel to another country, all are situations in which it is always required to prove the identity of a citizen for security reasons or add it to a database of an organization. Usually, the way to prove the identity of a citizen during a registration process is using its ID (national ID, driver license, passport, etc.). Nevertheless, the action of extract all the information provided by the ID is performed manually. In these modern times of the digital revolution, perform a manual work is a perceptible limitation since it implies delay in the accomplishment of tasks and possible errors by human mistakes.

An automated system able to extract information from images of IDs and interpret their contents is a possible solution. Notwithstanding, template-based DIP systems are limited since there are a large number of different identity document designs. Therefore, the methods of digital image processing to be used must provide a high-quality result since the correct extraction and interpretation of the information contained in an ID depend on the efficient integration of them.

1.3 Research Objectives

This research proposes the design and implementation of an end-to-end system capable of extracting and interpreting textual information from identity documents images without prior knowledge of their layouts to overcome the limitations of automated systems, in terms of classification of information from unknown IDs.

1.4 Contributions

Two main contributions are reached in this work. The combination of them allow a robust extraction and classification of the information of interest contained in an identity document. These contributions consist of:

- An implementation of a semantic attribution algorithm that allows to classify and attribute meaning to the information;
- An integration of the semantic attribution algorithm with a text recognition system to obtain an end-to-end system for interpreting and extracting textual information of interest.

Both contributions allow the automation of a registration or verification process that requires the acquisition of information about a citizen using his identity documents.

As collateral contributions are three publications, two related to the integration of text recognition methods to obtain the better OCR results based on a software approach and the other related to the semantic attribution algorithm.

Valiente, Rodolfo; Sadaike, Marcelo T; Gutiérrez, José C; Soriano, Daniel F; Bressan, Graça; et al. ***Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV); Athens*** *: 142-146. Athens: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). (2016)*

Rodolfo Valiente, José C. Gutiérrez, Marcelo T. Sadaike, and Graça Bressan. 2017. Automatic Text Recognition in Web Images. In Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web (WebMedia '17). ACM, New York, NY, USA, 241-244. DOI: https://doi.org/10.1145/3126858.3131570

José C. Gutiérrez, Rodolfo Valiente, Marcelo T. Sadaike, Daniel F. Soriano, Graça Bressan, and Wilson V. Ruggiero. 2017. Mechanism for Structuring the Data from a Generic Identity Document Image using Semantic Analysis. In Proceedings of the 23rd Brazilian Symposium on Multimedia and the Web (WebMedia '17). ACM, New York, NY, USA, 213-216. DOI: https://doi.org/10.1145/3126858.3131594

1.5 Research Methodology

In the interest of obtaining the desired contributions and validate the proposed research, the methodology used is based on:

- A literature review of the main methods used for automatic text recognition and classification in images, in the interest of identifying which algorithm could be improved or adapted to use it in the process;

- Study of identification document analysis systems in order to detect some opportunities for improvement;
- Implementation of a semantic attribution system capable of classifying the information by its meaning;
- Integration of different algorithms of interpretation and extraction of text in the interest of implementing an end-to-end system capable of extracting and understanding the textual information contained in Identity Documents;
- Conducting computing simulations of identity documents with different formats and layouts;
- Analysis of the results obtained to evaluate the performance of the system in different scenarios taking into account objective metrics.

1.6 Structure of Work

This research is composed of this first introduction chapter along with four more chapters and the bibliographical references used. Next, the objectives discussed in each chapter are detailed:

Chapter 1 describes the context in which the research takes place in a real work scenario related to registration/enrollment processes. Furthermore, details the reasons why this study was carried out, the research objective and the research methodology employed.

Chapter 2 defines the theoretical framework and concepts needed to understand the research. Topics related to DIP systems intended to text recognition and interpretation, computer vision algorithms, natural language processing and named-entity recognition are also detailed.

Chapter 3 describes studies related to text recognition in images and those intended to extract information from identity documents in order to correlate and contextualize these works with the research.

Chapter 4 defines the proposed architecture, describing and explaining each stage that integrates it.

Chapter 5 describes the results of the research based on formal evaluation metrics. Moreover, this chapter develops analytic and critical thinking on primary results and analysis with reference to theoretical arguments grounded in the literature review.

In the last Chapter are reported the conclusions obtained so far, showing a statement on the extent to which each of the aims and objectives has been met, the limitations of the research and some interesting points to be addressed in futures works.

Chapter 2

2 Theoretical Framework

This chapter addresses the theoretical basis of computer vision and digital image processing algorithms often used in text recognition systems. Additionally, this chapter describes how natural language processing algorithms are used to automatically extracting structured information from documents.

2.1 Computer Vision Systems

Vision is considered one of the most important and influential sense in the human perception. The capability to interpret and acquire information, for the most part, is based on visual input, thereby the processing of these visual data is of significant importance (Bailey, 2011; Cyganek, 2013; Marques, 2011; Number, 2016).

Thanks to the rapid advances in digital technology, efforts aimed at developing systems able to emulate human capabilities have increased. Those computer systems designed to emulate the performance of human vision are known as computer vision system (Fisher et al., 2014). Transferring these capabilities to a machine using computer methods and algorithms for scene analysis make it possible augment or enhance human vision (Sarfraz, 2005; Shih, 2010).

In the interest of better understanding of the research proposal, in this section are presented basic definitions related to digital image processing and computer vision system.

2.1.1 *Image: Definition*

Computer vision systems designed to emulate human vision need to process a visual input. This input can be an image or a set of images (video). An image is a visual representation in two-dimensions (2D) of a three-dimensional (3D) object or scene (Crist, 2011; Marques, 2011; Shih, 2010).

Imaging sensors have been developed to cover much of the electromagnetic spectrum (EM), unlike the human eye that is limited to the visible

portion of the EM spectrum. Computer systems use these imaging sensors to capture an image and then process it using computer methods and algorithms for scene analysis.

Nevertheless, once an image is processed by a computer vision system, an additional definition describes it better.

2.1.2 *Digital Image: Definition*

A digital image is a discrete representation of both spatial quantization (pixels) and intensity (brightness or color) of an image. It can be represented as a two-dimensional function of real numbers (grayscale images), or a set of 2D functions (color images). As shown Figure 2-1, the value of the numbers represents the intensity of the pixels in grayscale images, or the amount of red, green and blue in color images (Marques, 2011; Petrou & Petrou, 2011; Shih, 2010).

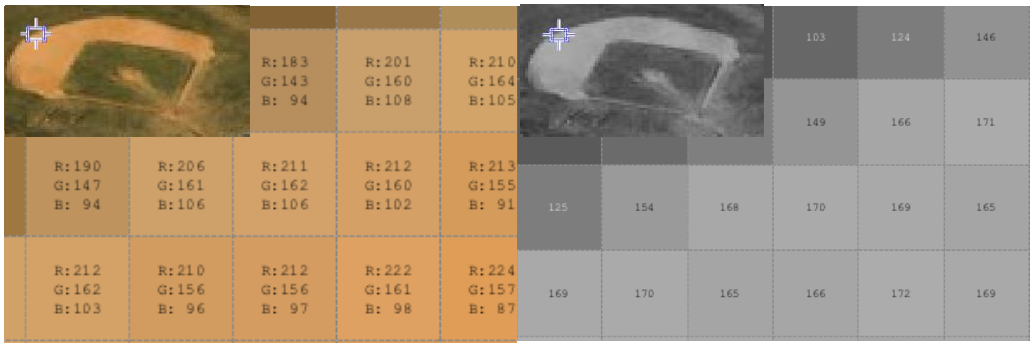

Figure 2-1 Discrete representation of a color image (left) and a grayscale image (right). The boxes delimit the pixels, while the numbers show intensity value of each pixel. Source: The Author.

Computer systems process an image of size $M \times N$ as a 2D function $f(x, y)$ (Equation 2-1), where the intensity value of a pixel of coordinates (x_0, y_0) is denoted as $f(x_0, y_0)$.

Equation 2-1 Digital Image representation. Source: (Marques, 2011)

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & \cdots & f(0,N-1) \\ f(1,0) & f(1,1) & \cdots & f(1,N-1) \\ \vdots & \vdots & & \vdots \\ f(M-1,0) & f(M-1,1) & \cdots & f(M-1,N-1) \end{bmatrix}$$

Once in this format, an image can be manipulated by digital operations using computer algorithms.

2.1.3 Digital Image Processing: Definition and Components

Digital image processing (also referred as image processing in terms of computer vision), can be defined as the process of modifying a digital image using mathematical operations. The result of performing a digital image processing depends on the objective of the application and the image processing techniques used (Bailey, 2011; Marques, 2011; Shih, 2010).

A generic digital image processing system is built around a computer that runs most of the image processing algorithms. As shown in Figure 2-2, some hardware and software elements for image acquisition, storage, and display are also part of a generic DIP system.

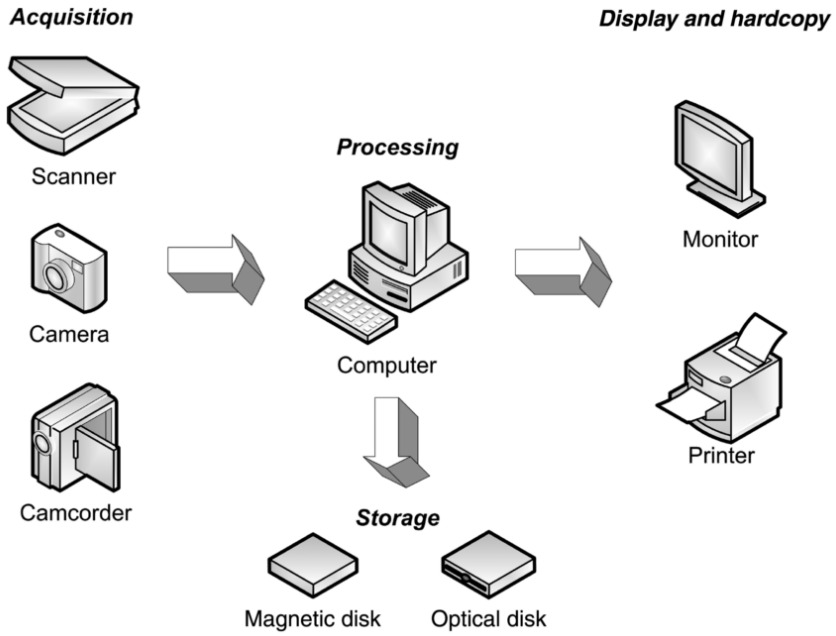

Figure 2-2 Components of a digital image processing system. Source: (Marques, 2011).

The acquisition stage consists of physical devices capable of capturing an image or video and sending it in digital format to the computer. Once the digital image is received, the computer runs the software that performs the image processing algorithms. The processing results could be displayed on a monitor or printed for analysis by human operators. Furthermore, these results are stored on magnetic or optical disks for long-term storage or to be accessible to other services (Gonzalez & Woods, 2008; Marques, 2011; Shih, 2010).

Enhancing an image, classifying an object, or detecting some critical feature within an image could be the focus on different digital image processing systems. Even being the principal focus different, many of the techniques used at fundamental levels remain the same and thus it is possible to group them according to the type of data they process (Bailey, 2011).

2.1.4 *Image Processing Operations*

According to the processed data and complexity of the algorithms used in image processing operations, it is possible to classify them as Low-Level, Intermediate-Level, and High-Level operations as shown in Figure 2-3 (Downton & Crookes, 1998; Ratha & Jain, 1999).

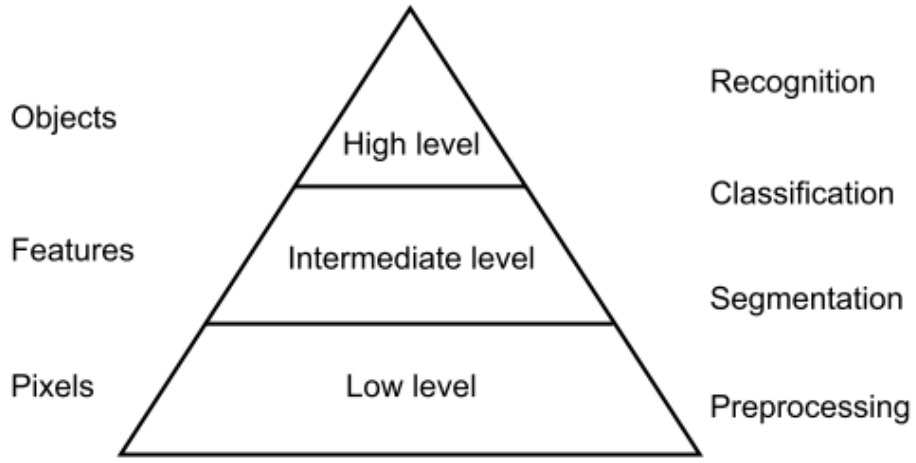

Figure 2-3 Classification of Image Processing Operations. Source: (Bailey, 2011).

In the Low-Level group are the preprocessing operations. Algorithms in this category are characterized by directly process the pixels in an image (using relatively simple operations such as multiplication and addition) and generate a new image with the relevant information highlighted. Examples of operations in this group are contrast enhancement and filtering for noise reduction or edge detection.

Operations designed to group pixels into regions with common properties are known as segmentation techniques. Methods such as thresholding, color detection, region growing and connected components labeling are some of the most commonly used segmentation techniques for detecting regions in an image. Some researchers group these operations at the Intermediate-Level (Downton &

Crookes, 1998; Marques, 2011; Ratha & Jain, 1999), while others consider that these are at the boundary between the Low and Intermediate Levels (Bailey, 2011). These operations also process pixels but use more complex operations, thus allowing a set of regions with common features to be generated instead of an image (thereby being an image-to-region transformation).

Classification methods are also located at the intermediate level. After the segmentation operations, classification tasks use the features of each region to identify or classify objects. The output of these methods is no longer based on an image since classification transforms the data from regions to features, and then to labels.

At the highest level are the Recognition operations which perform the analysis and interpretation of the contents of an image. These techniques process a set of features or labels and then generate, based on specific recognition algorithms, a set of objects. Algorithms used in recognition tasks are characterized by being nondeterministic and complex.

2.1.5 *Computer Vision Systems: Definition and Components*

Computer vision systems or machine vision systems (this last term is used to refers to the practical vision systems, such as industrial vision) are designed to capture images and analyze them using digital image processing (Bailey, 2011; Fisher et al., 2014; Gonzalez & Woods, 2008; Marques, 2011). Alike DIP systems, computer vision systems are a combination of hardware devices and software capable of processing visual scenes using computer algorithms. Processing results are used to control or perform a specific task depending on the image processing operations employed.

Computer vision systems solutions to emulate the performance of the human visual system, follow a sequential processing scheme based on the use of the image processing operations and intelligent algorithms. Figure 2-4 shows the main components of a computer vision system (Gonzalez & Woods, 2008; Marques, 2011; Ratha & Jain, 1999).

The problem domain defines which are the desired results and the general objectives to obtain them using a computer vision system. This component

usually proposes an automatic recognition of the information of interest in images without the need for human intervention to obtain the desired results.

The second component realizes the acquisition of one or more images that contain the scene to be analyzed and transforms them into digital images. Low resolution, poor lighting or blurred images are some factors that affect the quality of the resulting images and system performance. Therefore, during the design of this block, these factors must be considered.

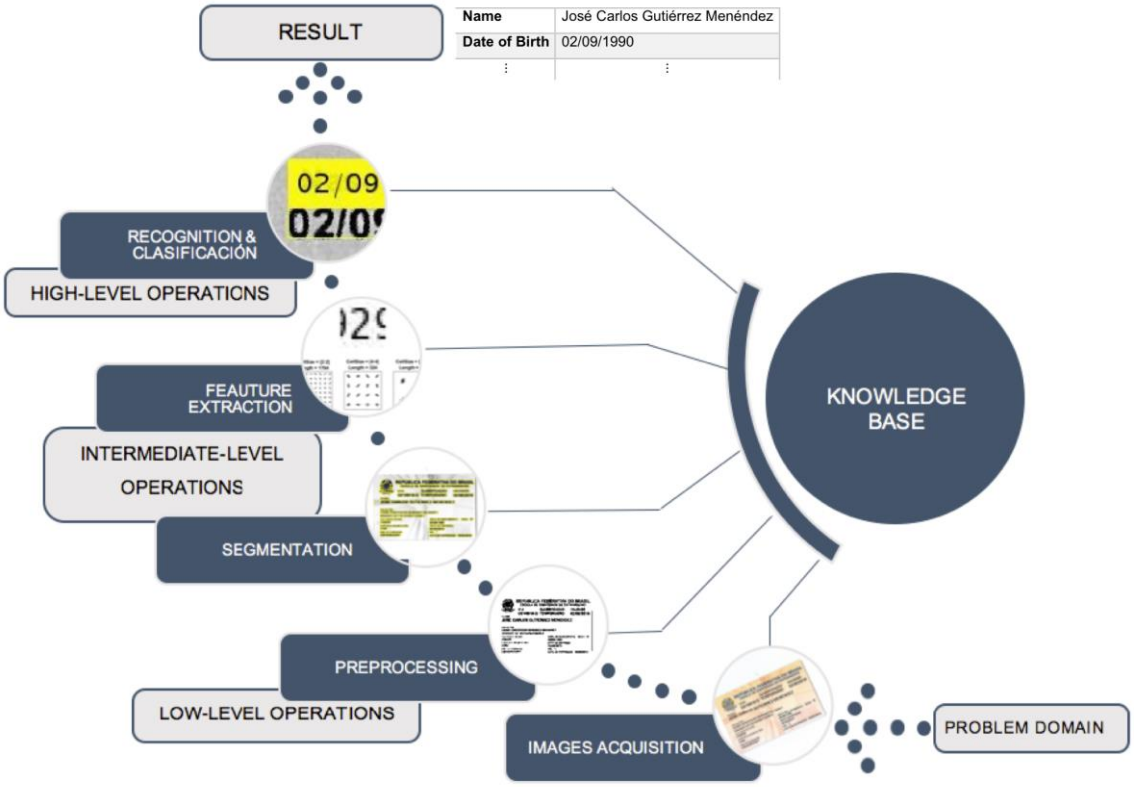

Figure 2-4 Computer vision system components. Source: The Author.

Next stage performs the preprocessing tasks to enhance the acquired image. Algorithms commonly used during this stage are based on contrast improvement, brightness correction, and noise removal.

The segmentation block creates labeled regions once it divides foreground objects of the image background. Achieving an automatic image segmentation is one of the most challenging tasks in a computer vision system.

In the feature extraction block, each of the segmented areas is analyzed and encoded according to its statistical data. Features such as color (or intensity)

distribution, texture, and geometric dimensions of the segmented objects are usually grouped into a feature vector.

In the last block is performed the recognition of the regions of interest of an image. The feature vector obtained in the previous stage is a numerical representation of the image contents. Therefore, it can be used by classifier algorithms to achieve recognition of the regions of interest. Minimum distance classifiers, probabilistic classifiers, and neural networks are some of the more used techniques to pattern recognition. The ultimate goal of this block is to classify and recognize the object of interest among all the objects within the image.

As shown in Figure 2-4, there is a module called knowledge base linked to all stages. The connections between this module and the stages are to represent that achieving a successful solution of a given problem depends on how much knowledge about the problem domain has the computer vision system.

2.2 Text Recognition System

A text recognition system is a typical application where a computer vision system is used to extract and recognize text information from images. Systems capable of extracting textual information from images apply OCR algorithms to regions with text in an image. These regions are the result of performing preprocessing and segmentation techniques.

Image processing covers a wide and diverse array of techniques and algorithms; therefore, this section provides a brief description of the most representative image preprocessing and segmentation operations used in text recognition system, and a general description of classification and recognition approaches.

2.2.1 Text Recognition System: Preprocessing Operations

Enhancing and cleaning up images using preprocessing algorithms allows extracting the pixels that belong to the text characters. Color image processing is time and space consuming, hence to avoid cost concern, most text recognition

systems process gray-scale or binary images (Cheriet et al., 2007; Hornberg, 2007; Shih, 2010).

Grayscale transformations are used to improve the contrast between characters and background. This transformation can change the gray levels of the entire image or modify the gray levels within a defined window by a mapping function.

Preprocessing algorithms based on smoothing and noise removal operations are used to reduce high-frequency noises (salt-pepper) and also remove unwanted details (Figure 2-5(a)).

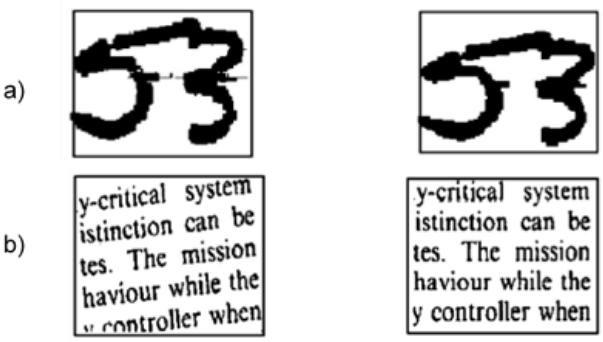

Figure 2-5 Examples of the uses of different preprocessing algorithms. a) smoothing and noise removal operations; b) skew correction operations. Source:(Cheriet et al., 2007)

In text recognition systems is commonly used a preprocessing filter to detect skew and rectify slanted characters. Depending on the language, text lines are horizontal or vertical; hence the segmentation is much easier if the text is in one of these directions (Figure 2-5(b)).

2.2.2 *Text Recognition System: Segmentation Operations*

Once regions of interest (text character in text recognition systems) in an image have been enhanced by the use of preprocessing filters, the next step is extracting them. This extraction is performed using segmentation techniques as thresholding, morphological operations and/or connected component analysis (Cheriet et al., 2007; Hornberg, 2007).

Thresholding operations are based on the gray level histogram of the image (global thresholding operation), or based on the gray level of the local background (dynamic thresholding operation). Thresholding operations define as object pixels to those that lie within a specified range of gray level. Otherwise,

pixels that are out of the range of gray level known as the threshold level it defines them as background pixels (Figure 2-6).

Figure 2-6 Result of applying an adaptive threshold operation over an image with text. Source: The Author.

Connected components are groups of connected pixels that represent the different objects in the image. There are two ways to consider if two pixels are connected and depend on the connectivity used, 4-connectivity or 8-connectivity. In text recognition systems, once performed the connected components algorithm, each character is a component (Figure 2-7). Moreover, noisy pixels are also segmented as separate components. Hence, based on their area they could be removed.

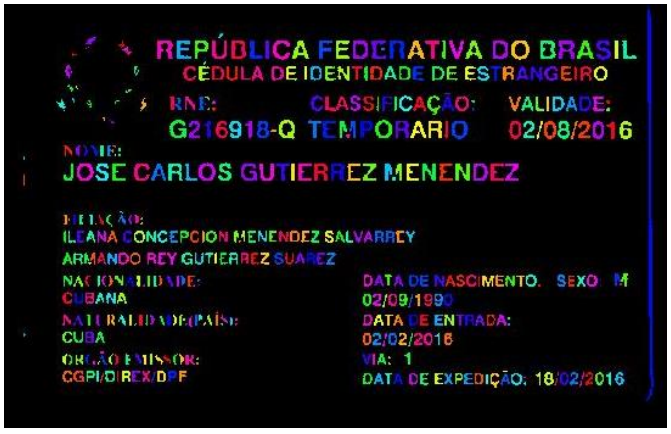

Figure 2-7 Connected components obtained using an 8-connectivity. Each color represents a component. Source: The Author

Morphological operations are algorithms that analyze the geometric structure of objects using a structural element (SE) to segment them. A change in the SE allows different results to be obtained using the same operation. In text recognition systems, they are used to remove unwanted noisy elements as well as to connect separate parts of the same character.

2.2.3 Text Recognition System: Classification Operations

Classification in text recognition system is based on the features extracted from the segmented characters (Cheriet et al., 2007; Gonzalez & Woods, 2008; Hornberg, 2007). The features to use must allow discerning different classes of character and maximize the effectiveness of recognition.

There are a large number of features commonly used for character recognition. Feature extraction methods can be categorized into geometric features (moments, histograms, direction features), structural features (registration, line element features, Fourier descriptors, topological features), and feature space transformations methods (principal component analysis, linear discriminant analysis, kernel).

2.2.4 Text Recognition System: Recognition Operations

The last stage of text recognition systems performs the recognition operations using the features extracted. These operations assign to each set of features with a class label or membership scores to defined classes (Cheriet et al., 2007).

Methods for character recognition are classified as statistical methods (parametric and nonparametric), structural methods (string and graph matching), machine learning methods (supervised and unsupervised learning), and the combination of multiple classifiers. Once all the characters are detected, holistic or analytical (based on segmentation) methods are performed for the recognition of words and strings.

2.3 Natural Language Processing

At the moment when an OCR system finishes its processing, the textual data are recognized and extracted but in an unstructured way. These raw data do not imply any textual or grammatical meaning for a machine, thus it is when natural language processing is necessary in order to an automatic system can interpret and structure them as a human does.

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate text or

speech to do some specific task or automated a process (Chowdhury, 2005). NLP is composed of different tasks capable of transforming a raw and unstructured text into a structured and meaningfully text. Each task has a different working way with different rules, approaches, and methods for doing the respective processing in order to achieve its objective (Dudhabaware & Madankar, 2014).

2.3.1 Structured and Unstructured data

Structured Data means basically databases of data, where each data has a very well-defined meaning (Figure 2-8). The information specified by its hypernyms such as a person name, address, and phone number are all examples of structured data. For instance, cell phone is a hypernym of iPhone, if iPhone is a type of cellphone.

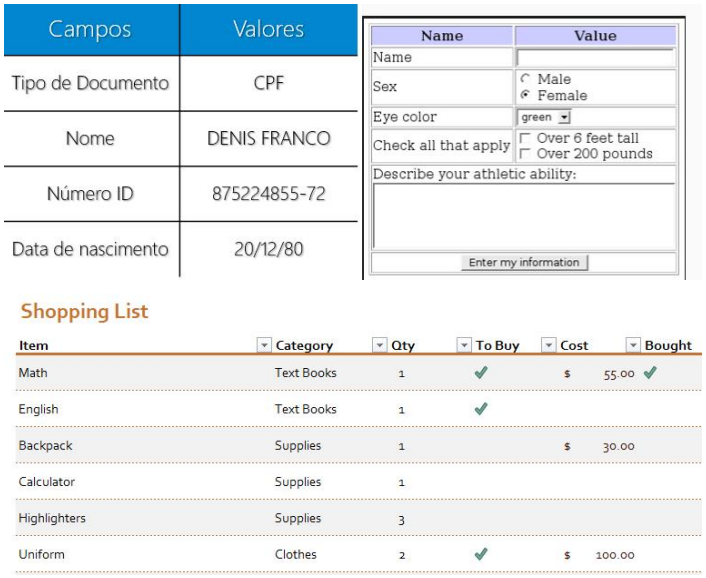

Figure 2-8 Example of structured data (forms, database). Source: The Author

In contrast, unstructured data are not as easily characterized, since they lack a recognizable format (the word "format" refers to a well-defined meaning or classification). Unstructured data refers to data like complete corpus, raw audio, or images (Figure 2-9) where it is hard to recognize what is in the image or what entities are in the text. The management of unstructured data is recognized as one of the major problems in the information technology (IT) industry. The main reason being that it has been much harder for computers to make sense of

unstructured data compared to structured data (Arasu & Garcia-Molina, 2003; Blumberg & Atre, 2003; McCallum, 2005).

Figure 2-9 Examples of unstructured data (images, audio, corpus). Source: The Author

With the purpose of automating an Information Extraction (IE) process, people noticed that it is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money, and percent expressions. Therefore, identifying references to these entities in text is recognized as one of the important tasks of NLP and is called Named Entity Recognition and Classification (NERC or just NER) (Nadeau & Sekine, 2007).

2.4 NERC - Named Entity Recognition and Classification

The term “Named Entity” was established on the Sixth Message Understanding Conference (MUC-6). These group of conferences was focusing on IE tasks where one of their objectives was structuring information from unstructured text, such as newspaper articles (Grishman & Sundheim, 1996; Nadeau & Sekine, 2007).

In the sixth edition of the conference, the committee developed the "named entity" task, which basically involves identifying names and numeric expressions. Figure 2-10 shows a sample sentence with named entity annotations. The tag ENAMEX ("entity name expression") is used for both people and organization

names, whereas the tag NUMEX ("numeric expression") is used for currency and percentages.

```
Mr.  <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON">Martin
Puris</ENAMEX>, president and chief executive officer of <ENAMEX
TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX>, about <ENAMEX
TYPE="ORGANIZATION">McCann</ENAMEX>'s acquiring the agency with billings of <NUMEX
TYPE="MONEY">$400 million</NUMEX>, but nothing has materialized.
```

Figure 2-10 Sample named entity annotation used in MUC-6. Source: (Grishman & Sundheim, 1996)

2.4.1 *NERC: Techniques*

Several techniques are proposed in the literature to perform NERC tasks. Based on their approaches, they can be classified into three categories: rule-based, machine learning, and hybrid (Chiticariu, Krishnamurthy, Li, Reiss, & Vaithyanathan, 2010; Nadeau & Sekine, 2007).

Early systems made use of handcrafted rule-based algorithms and heuristics definitions. Most recent ones rely on machine learning as a way to automatically induce rule-based systems using a collection of training examples. Nevertheless, when training examples are not available, handcrafted rules remain the preferred technique (Nadeau & Sekine, 2007).

Machine learning approaches used to NERC are divided in supervised, semi-supervised and unsupervised learning. The idea of supervised learning is to study the features of positive and negative examples of Named Entity over a large collection of labeled documents and thus create rules to detect entities in new documents. Its performance depends on the number of annotated corpus examples available to train the system, being this number the main shortcoming of this approach.

Semi-supervised learning (SSL) put around that just a small degree of supervision is carried out. The main technique for SSL is called “bootstrapping” and involves the use of a small number of example (set of seeds), for starting the learning process. The system searches for texts that contain these initial examples and tries to identify some contextual clues and patterns around the entities common among them. Then, it tries to find other instances of the same category as initial examples that appear in similar contexts. The newly found examples are used to perform a new search in order to discover new relevant

contexts. After several searches, a large number of examples of the same category and a large number of contexts are eventually obtained. Notwithstanding, the performance of that algorithm can deteriorate rapidly when noise (perturbation) is introduced in the entity list or pattern list.

The last learning approach, unsupervised learning, rely on lexical resources, on lexical patterns and on statistics computed on a large unannotated corpus. The most common technique used in this approach is clustering, based on detecting the similarity of the context of the words to gather named entities.

Since all these approaches are focused on features of the words and characters, a correct selection of which features used to detect and classify named entities is essential for obtaining a good NERC system (Nadeau & Sekine, 2007).

2.4.2 *NERC: Features Detected*

In NERC systems, the features are used to apply the rules that allow classifying a word or a phrase as a named entity. For instance, a system might have two rules, a recognition rule: “capitalized words are candidate entities” and a classification rule: “the type of candidate entities of length greater than x letters is a *name*”. To make these rules effective, each word is represented by a features vector that contains Boolean, numeric and nominal values. These values describe specifics features of the words, for example, its length and if the word is capitalized (Carreras, Màrquez, & Padró, 2003; Chiticariu et al., 2010; Nadeau & Sekine, 2007).

The features most often used for the recognition and classification of named entities can be grouped into three different categories: Word-level features, List lookup features and Document and corpus features. The first category focuses on the composition of the character of the words, describing word case, punctuation, a numerical value, and special characters. The second is based on lists with previous classifications of words, where it is checked the probability of a word belong to a specific list (for example, if the word Paris is an element of a list of cities, then the probability of this word to be a city, in a given text, is high). The last category analyses the content and the structure of

documents, basically, it considers the number of occurrences of words, the frequency of specifics words in different documents, and the position of the word in the document (Nadeau & Sekine, 2007).

Nonetheless, there are some limitations that have to be considered in NERC systems when the features are chosen. For example, although capitalization can aid in recognizing named entities in some languages, in others like Chinese or Arabic that do not have any capitalization at all or the German that capitalizes all nouns is insufficient.

2.4.3 NERC: General Limitations

As mentioned in the previous paragraph, some factors such as Language limit the scalability of a NERC system. Others like the textual genre (journalistic, scientific, informal, etc.) or domain factor (gardening, sports, business, etc.) can be reasonably supported, however porting a system to a new domain or textual genre and produce accurate results remains a major challenge (Chiticariu et al., 2010; Nadeau & Sekine, 2007). Finally, another limiting factor is the Entity type since based on the application requirements, a NERC system must classify in the most studied types (ENAMEX and NUMEX) or subtypes and subcategories of them (city, country, “politician”, “entertainer", etc.).

2.5 Final conclusions

After performing the analysis of the theoretical basis, it is possible to determine the aspects to consider for the design of the proposed system and the functionalities that it will have. As part of a computational vision system, it will be composed of different image processing operations that allow the correct text extraction. Once extracted, the text has to be processed using natural language processing, specifically a named entity recognition process with the aim of detecting in the text, those data that correspond to the information of interest.

Chapter 3

3 Literature Review

Information extraction from images and the interpretation of their contents has motivated the development of computer vision for years. Different image processing systems have been developed with the aim of simulating the human capacities of interpretation of visual information.

This chapter describes a selection of studies related to the development of image processing systems focused on text detection and extraction, named entity recognition and documents classification. An adequate analysis of these works helps in the selection of the necessary techniques to be used to achieve the objectives of this research.

The first part of this chapter describes a selection of DIP works related to text detection. Since the list of works related to image processing for the extraction of text is extensive, those that propose methods that are relevant to reach the objective of this research are selected. The analysis of these works is done to select methods and functions that can be used to extract textual information from identity documents. Subsequently, named entity papers are analyzed to select the adequate NERC approach able to identify the information of interest from the unstructured text extracted. It also discusses how a combination of rules can be used to recognize and differentiate entities. Afterward, some papers are analyzed in which the documents are processed by a system based on templates. The systems described in the latter papers use different approaches to classify documents aim to facilitate the process of extracting information. Finally, some conclusions are described based on the analysis and results of the related works with the purpose of using them as references for the research.

3.1 Text detection

Research on scene text detection and recognition has been a continuous interest in the field of digital image processing. Several approaches and techniques have been employed to accomplish this task. To benchmark

performance of the different scene text segmentation and recognition algorithms, a Robust Reading Competition is organized at International Conference on Document Analysis and Recognition (ICDAR). The series of these competitions addresses the need to quantify and track progress in this domain. Hence, researchers worldwide accept them as the standard for evaluation (Karatzas et al., 2015, 2013; Shahab, Shafait, & Dengel, 2011). Therefore, in this research, those works with relevant results in these competitions are taken as references.

The Robust Reading competitions are structured in different challenges addressing text extraction in diverse application domains (Figure 3-1). During the different editions, new tasks and challenges have been introduced. The last ICDAR Robust Reading competition published (Karatzas, Gomez, Nicolaou, & Rusinol, 2018) consisted of ten Challenges: Reading Text in Born-digital Images (Challenge 1), Reading Text in Focused Scene Text Images (Challenge 2), Reading Text in Videos (Challenge 3), Reading Text in Incidental Scene Text Images (Challenge 4), and the 2017 edition of the Competition introduced six new challenges COCO-Text, FSNS, DOST, DeTEXT, MLT and IEHHR. The challenges are organized around specific tasks covering text localization, text segmentation, and word recognition. Since the 2015 edition of ICDAR, tasks assessing End-to-End system performance have been introduced to all Challenges.

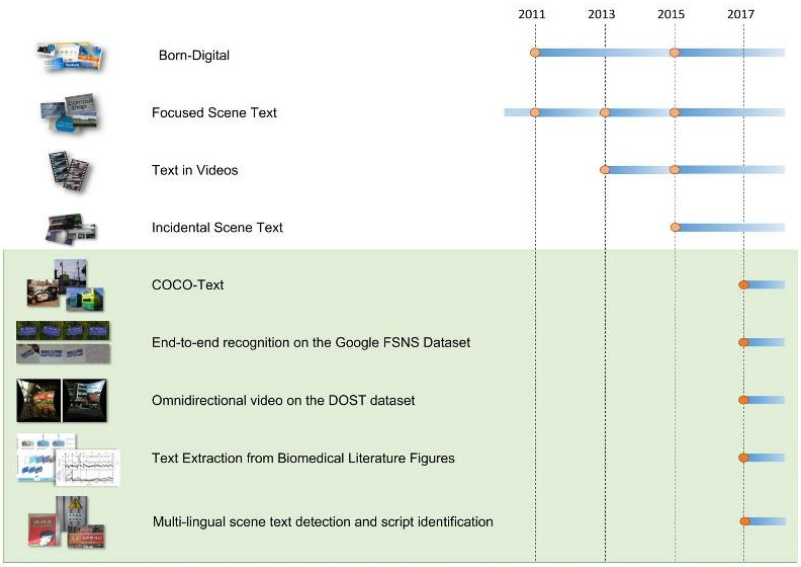

Figure 3-1 Robust reading competitions and challenges introduced in different editions. Source: ("ICDAR 2017 RobustReading Competition," n.d.)

Each challenge has its own dataset (Figure 3-2). The Challenge 1 dataset consist of born-digital images used in electronic documents (Web and email) with embed textual information. Challenge 2 focused on reading texts in real scenes. Its dataset consists of text images that are explicitly focused on the text content of interest and in most of the cases in a horizontal position. Challenge 3 dataset contains various short video sequences that include text in different languages. Challenge 4 focused on real scene images similarly to Challenge 2. Contrary to Challenge 2 which is based on images focused on the text content, Challenge 4 dataset consist of images of text scene where the focus is not on the text content of interest.

a) b) c)

Figure 3-2 Examples of images of a) Born-Digital, b) Focused Scene Text, and c) Incidental Scene Text challenges. Source: ("ICDAR 2017 RobustReading Competition," n.d.).

The new challenges introducing in the 2017 ICDAR edition are named by their dataset (Figure 3-3). COCO-Text challenge (R. Gomez et al., 2018) contains images of complex everyday scenes not collected with text in mind and thus contain a broad variety of text instances (similar to Challenge 4). DeTEXT competition focuses on extracting (detecting and recognizing) text from biomedical literature figures (Yang, Yin, Yu, Karatzas, & Cao, 2018). Another competition based on video is the DOST Challenge, which shows video scenes in the real environment of Downtown Osaka (Iwamura et al., 2017). Images of street name signs of France cropped from Google Street View images are the focus on the FSNS challenge (Raymond Smith et al., 2016). Information extraction in historical handwritten records and to move towards these document understanding are the objective in the IEHHR competition (Fornes et al., 2018). The last competition is the MLT Challenge, this is similar to the Challenge 2 but

facing various scripts and languages in order to perform a multilingual text scene detection. The main idea of this challenge is to try to answer the question of whether text detection methods could handle different scripts without fundamental changes in the used algorithms/techniques, or if script-specific methods are really needed (Nayef et al., 2018).

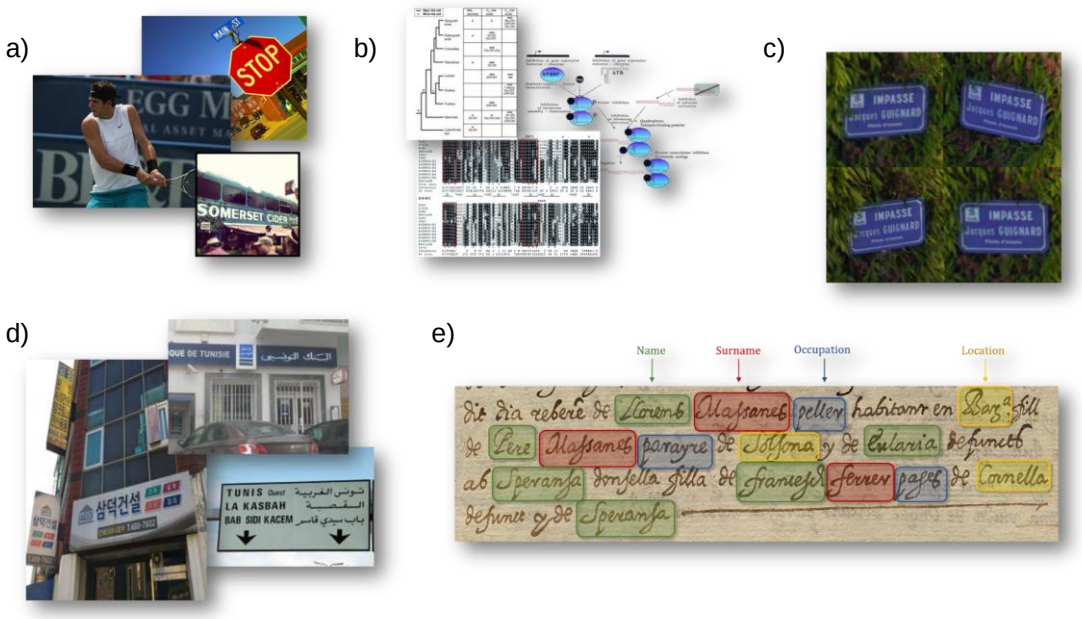

Figure 3-3 Examples of images of a) COCO-Text, b) DeTEXT, c) FSNS, d) MLT and e) IEHHR challenges. Source: ("ICDAR 2017 RobustReading Competition," n.d.)

Among all the mentioned challenges, the dataset of Challenge 2 (*Reading Text in Focused Scene Text Images*) is the most similar to the scenario represented by this research, where the ID Documents images are taken, focusing explicitly on their textual information often written in only one language and in one direction. Moreover, it has been shown that the performance of text recognition algorithms and the overall system performance depends critically on the effectiveness of their text localization algorithms (Epshtein, Ofek, & Wexler, 2010; Huang, Lin, Yang, & Wang, 2013; Islam et al., 2016; Uchida, 2014). The last published winning method of Challenge 2 in the ICDAR 2015 contest was able to localize only 66% words correctly (Karatzas et al., 2015, 2013). Thus, the qualitative results of the competitions demonstrate that there is still room for improvement. Therefore, works that address this challenge are examined for using them as references.

In this challenge, the ICDAR contest evaluation protocol follows the proposed by Wolf et al. (Wolf & Jolion, 2006) for the text localization task. It evaluates both quantity and quality of rectangle matches through all images in the database, and considers not only one-to-one matching, but also one-to-many and many-to-one matchings. The evaluation is computed by $Precision$, $Recall$, and $f-measure$ which are defined in the Equation 3-1:

Equation 3-1 Evaluation protocol proposed by Wolf et al. Source: (Wolf & Jolion, 2006)

$$Precision = \frac{\sum_i^N \sum_j^{|Di|} M_D(D_j^i, G^i)}{\sum_i^N |D^i|}$$

$$Recall = \frac{\sum_i^N \sum_j^{|Gi|} M_G(G_j^i, D^i)}{\sum_i^N |G^i|}$$

$$f-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

N is the total number of images in a dataset; $|Di|$ and $|Gi|$ are the number of detection and ground truth rectangles in the i-th image; $M_D(D_j^i, G^i)$ and $M_G(G_j^i, D^i)$ are the matching scores for detection rectangle D_j and ground truth rectangle G_j. Two rectangles are considered as matched when their overlapping ratio is higher than a defined threshold, which controls the quality of the matching. Ground truth is a term used to refer to information provided by direct observation as opposed to those inferred by prediction.

3.1.1 *Text localization approaches*

In general, text localization methods can be divided into two major categories: Texture-based and Region-based (Bai et al., 2013; Epshtein et al., 2010; Huang et al., 2013; Huang, Qiao, & Tang, 2014; Lukás Neumann & Matas, 2013a; J. Zhang & Kasturi, 2014). Both categories consist of identifying whether a specific region of an image contains a text or not.

Texture-based methods (Jaderberg, Vedaldi, & Zisserman, 2014; Kai Wang, Babenko, & Belongie, 2011; Mandal, Roy, Palz, & Blumenstein, 2015; Wang, Wu, Coates, & Y. Ng, 2012; Zhu & Zanibbi, 2016) consist on scan the image at different scales using sliding windows and classify the contents of the

window in text and non-text regions. To confirm the classification, they use a binary classifier to score the text confidence of candidate regions of variable scales. These methods assume that the text regions in images have distinct textural properties from the background. Machine learning methods are often used in this approach. While these methods are generally more robust to noise in the image, using sliding windows is computationally expensive since the algorithms need to search in rectangles with different geometrical conditions.

The majority of recently published methods for text localization falls into the Region-based category (Bai et al., 2013; Epshtein et al., 2010; Islam et al., 2016; Lukás Neumann & Matas, 2016; Phan, Shivakumara, & Tan, 2012; Shi, Wang, Xiao, Zhang, & Gao, 2013; Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, & Hong-Wei Hao, 2014). In this approach, characters are detected using local properties (color, stroke-width, etc.) where those pixels exhibiting similar properties are grouped into components using a low-level filter, such as Stroke Width Transform (SWT)(Epshtein et al., 2010) or Maximally Stable Extremal Regions (MSER) (Matas, Chum, Urban, & Pajdla, 2004). The resulting connected components (CCs) are then filtered geometrically and using texture properties to exclude CCs that certainly cannot be letters. This approach is able to detect texts with different scales in the same image and is not limited to horizontal texts. Moreover, it provides a character segmentation, which makes easy to group them into text lines for the OCR stage. Nevertheless, this approach is sensitive to image noise and distortions since it uses low-level operations. In addition, the use of heuristic rules to filter components might not work well with different datasets. Therefore, these disadvantages could lead to incorrect component grouping.

Thanks to the recent advance of technology, the use of methods based on deep learning are more common than a few years ago. Many deep learning-based methods are proposed to directly detect words in images. Approaches based on deep learning have become dominant both in the detection of texts and in the recognition of texts. In text detection, usually a convolutional neural network is used to extract feature maps from a scene image, and then different decoders are used to decode the regions (He, Zhang, Yin, & Liu, 2018; Liao, Shi, & Bai, 2018; Xuebo Liu et al., 2018; Ma et al., 2018; Zhou et al., 2017). The use of this

approach is conditioned on having a large image dataset and powerful computational hardware during the training stage. Nevertheless, considering the novelty of this type of approach and at the time this research was carried out, in this study, they are considered as contributions for future works.

- *Maximally Stable Extremal Regions (MSERs)*

The MSERs fall in the Region-based category as mentioned above, and it has been one of the most widely used region detectors. MSER-based methods are based on detecting extremal regions, which stay nearly the same through a wide range of thresholds (Figure 3-4). An extremal region is a connected component where all pixels inside it have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary (Matas et al., 2004).

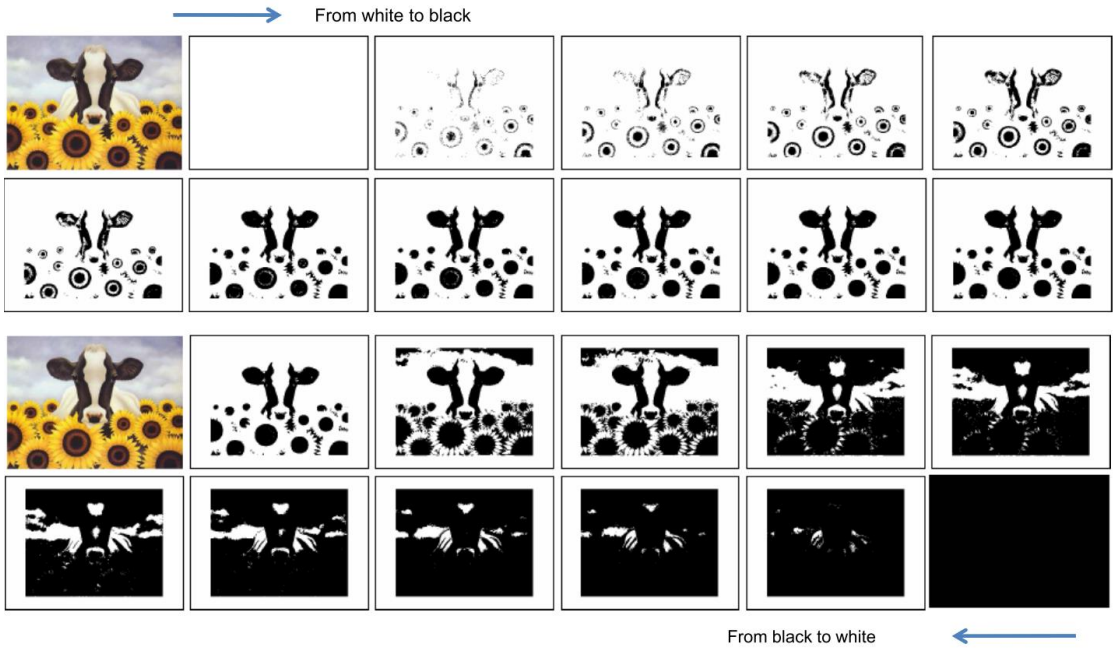

Figure 3-4 By sweeping the thresholds of the image, from one extreme (all white) to the other (all black), the image shows blobs that grow and merge. When these blobs remain stable between a range of different thresholds, they are considered extremal regions. Source: (Bimbo, 2011).

To obtain the MSERs, an image is processed using incremental steps of threshold values. In the sequence of images obtained, pixels whose intensities are below a threshold are considered as "black" and those above or equal as "white". The first image is completely white, but as the threshold value increases, black spots corresponding to the minimum local intensity begin to appear and grow. Eventually, the regions corresponding to two local minimums are merged

until the last threshold value is reached. The set of all connected components, which stay nearly the same (same area, size, etc.) in the sequence, is the set of all maximal stable extremal regions extracted (Figure 3-5).

Figure 3-5 Example of the use of MSER technique to detect text in an image. The letter "K" is identified as an MSER because the size of the connected region does not change significantly in the gray level (g) range from 135 to 195. Source: (Donoser, Riemenschneider, & Bischof, 2010).

Two promising properties make the MSER detector effective in text detection. First, the MSERs detector is computationally fast and can be computed in linear time of the number of pixels in an image. Second, it is a powerful detector with high capability for detecting low quality texts, such as low contrast and low resolution.

3.1.2 *Related Works*

As mentioned above, the list of works related to image processing for text extraction is extensive. Therefore, two works corresponding to two of the best text localization results in the Challenge 2 of the contest ICDAR 2013 (Karatzas et al., 2013) were selected to be analyzed, since they helped in the choice of the appropriate methods and functions to be implemented during the research.

The main contributions of these works to this research were the use of the MSER as a text detector, and the analysis of the geometrical and stroke widths features common to the text characters for their correct detection. Moreover, the selection of the best character recognition confidence and the detection of all the text lines to calculate their orientations and put them in a horizontal alignment were also contributions to improve the recognition accuracy.

- *On Combining Multiple Segmentations in Scene Text Recognition (Lukás Neumann & Matas, 2013a)*

This paper describes the published technique that occupies the second place in the ICDAR 2013 contest in text localization (Karatzas et al., 2013).

Neumann & Matas propose a text localization and recognition method based on improvements of their previous works (Lukas Neumann & Matas, 2011; Lukás Neumann & Matas, 2011, 2012). The authors present the benefits of keeping multiple segmentations of each character until the last stage of the processing and the impact of preprocessing with a Gaussian pyramid (P. J. Burt & E. H. Adelson, 1983). Additionally, they propose an algorithm to select the final sequence of characters based on finding an optimal path in a directed graph where nodes correspond to labeled regions and edges represent the relations between two regions.

Neumann & Matas propose a region-based method where individual characters are detected as Extremal Regions (ER). These ERs refer to regions whose outer boundary pixels have strictly higher values than the region itself. Extremal regions are obtained from a parametrized character detector with values of threshold, adjacency and color space projection used to segment the foreground information. Different values of these parameters generate ERs that can represent either a complete character or a part of it. Nonetheless, the authors propose that the selection of the final value for these parameters be made at a later stage by a trained cost function, which exploits context of the character in its text line.

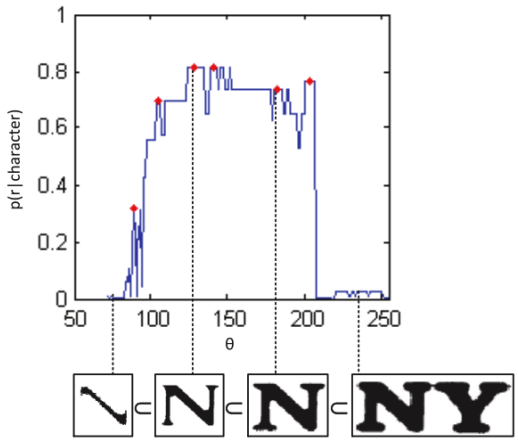

Figure 3-6 Influence of the use of different threshold values (horizontal axis) to obtain an Extremal Region. As shown the figure, the use of a small threshold value could detect an ER that not correspond to a complete character. On the other hand, a high value could join two characters in a single region. Source: (Lukás Neumann & Matas, 2013a).

In their proposal, the set of values of the parameters used to obtain all regions representing possible characters is based on the results of previous

works by the authors (Lukás Neumann & Matas, 2012). Since it is not possible locally determine a unique threshold value unambiguously for each character and with the intention of reducing the number of thresholds to use, the authors use a CSER classifier to select a set of threshold values that most likely correspond to a valid character segmentation (Figure 3-6). Moreover, their previous experimental validations showed that the combination of intensity, intensity gradient, hue and saturation projections yield the best results in the detection of characters.

In the case of adjacency, they propose a preprocessing of the image with a Gaussian pyramid to determine which pixels belong to a single component. In a Gaussian pyramid, an image is weighted down using a Gaussian average (Gaussian blur) and scaled down. Thus, characters formed by smaller elements or a single element consisting of multiple joint characters are merged or divided respectively to obtain individual regions that correspond to individual characters (Figure 3-7). To achieve this, at each level of the pyramid only a certain interval of character stroke widths is amplified. This last process does not represent a significant time cost since the image size is reduced 4 times at each level.

Figure 3-7 Results of processing with a Gaussian pyramid to determine which pixels belong to a single component. a) Characters formed by multiple small regions are merged and a single region corresponds to a single character. b) A single region represented by the characters "ME" is broken into two elements. Source: (Lukás Neumann & Matas, 2013a).

In this proposal, the authors first group all the regions in text lines by a pruned exhaustive search proposed by themselves in one of their previous work (Lukás Neumann & Matas, 2011). This search is based on geometrical characteristics of the regions to find similarity among them and obtain the text direction. At this point, the text lines obtained contain regions which correspond

to characters, as well as regions that represent only a portion of a character or clutter regions. Next, each region in a text line is labeled by the character recognition module proposed by the authors in (Lukas Neumann & Matas, 2011). This module provides confidence values that allow eliminating the previously included clutter regions.

For each text line, the authors propose to create a directed graph where the nodes correspond to labeled regions, and the edges indicate the predecessor relationship between a sequence of regions of the same text line. This relationship is approximated by a heuristic function that selects the nearest neighboring region in the left-to-right direction, once several geometric conditions are satisfied.

The constructed graph for each text line represents all the possible combinations of words or sequence of words based on the set of regions detected with the different parameter values. The selection of the correct combination (path in the directed graph Figure 3-8) determines the final value of the parameters used during the segmentation of the characters and returns the word corresponding to the regions analyzed. To achieve this selection, each node and edge has an associated score determined by trained cost functions (Neumann & Matas, 2013a) based on parameters such as region text line positioning, character recognition confidence, threshold interval overlap and transition probability. A standard dynamic programming algorithm is used to select the path with the highest score.

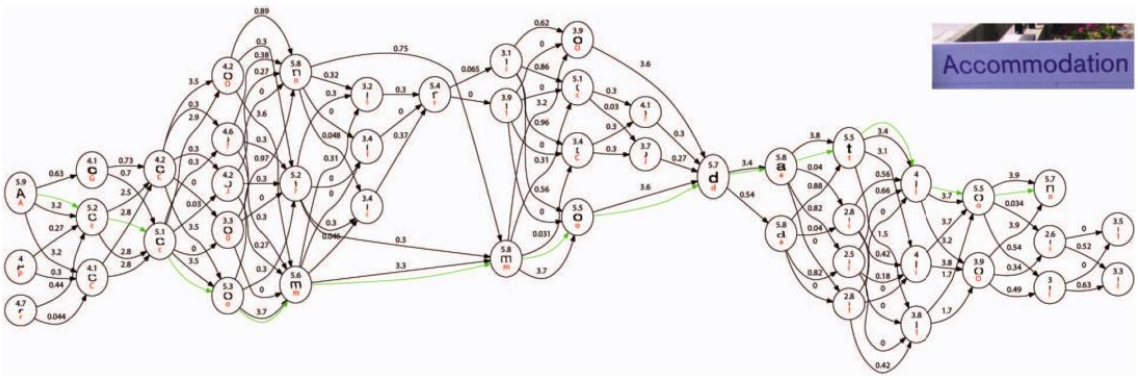

Figure 3-8 Example of a directed graph constructed from the detected regions in the image with the text "Accommodation". The nodes correspond to the labeled regions and the edges to the connection relation between regions. The optimal path is represented in green. Source: (Lukás Neumann & Matas, 2013a).

The authors present the results of different experiments and configurations to compare them with the performance of its proposal (Table I - Neumann & Matas, 2013a). All the experiments were run on a single core on a standard PC. They show in this paper, how selecting the final segmentation at a later stage yields better localization results than making the final decision in the text detector. Furthermore, they show how the preprocessing with a Gaussian pyramid achieves better localization results than a system evaluated on a single-scale for a cost of approximately 25% of execution time additional. The authors mention that the processing of the proposal is near real-time on a standard PC, consuming an average of $3.10s$ per image.

The proposal is also evaluated using the ICDAR 2011 competition evaluation scheme and dataset. The Neumann's method achieves $recall$ 67.6%, $precision$ 85.4% and $f-measure$ 75.4% in text localization in ICDAR 2011 edition. This represents a significant 4 percentage point improvement over the winner of the competition (Kim's Method - Shahab, Shafait, & Dengel, 2011). The authors also compare their proposal with the work presented by Shi et al. (Shi et al., 2013) and they show that their proposal achieves better results (Table II - Neumann & Matas, 2013a). In the ICDAR 2013 contest dataset, their proposal named Text Spotter achieves $recall$ 64.84%, $precision$ 87.51% and $f-measure$ 74.49% in text localization, which allows them to reach the second place in this edition of the contest.

- *Robust Text Detection in Natural Scene Images (Xu-Cheng Yin et al., 2014)*

This final paper related to text detection describes the proposal winner of the ICDAR 2013 contest in the text localization task. The authors propose a MSER-based scene text detection method (Xu-Cheng Yin et al., 2014). This method is based on four stages: character candidate extraction, text candidate construction, text candidate elimination and text candidate classification

For the character candidate extraction, Yin et al. propose a fast and effective pruning algorithm able to extract the possible regions of characters based on a strategy that minimizes the variations between MSERs. They rely on

the hierarchical structure of MSERs and in regions features to reduce the number of character candidates.

As characters cannot contain or be contained by other characters in the real world, authors propose to eliminate the children of a hierarchical structure of MSERs if it is known that the parent is a character and vice versa. To achieve a reasonable accuracy, they design two algorithms that perform the parent-children elimination operation (Xu-Cheng Yin et al., 2014). Both algorithms are based on the comparison of the intensity variance between the detected regions, selecting between the parent or children according to who has the lowest variation. The variance is previously regularized (Xu-Cheng Yin et al., 2014) to penalize variations of MSERs with aspect ratios too large or too small to avoid improper disposal, since MSERs corresponding to characters may not necessarily have the lowest variations. An example of this situation is shown in Figure 3-9. The children of the MSERs tree in Figure 3-9(a) correspond to characters while the parent of the MSERs tree in Figure 3-9(b) corresponds to a character.

a) b)

Figure 3-9 Examples of MSERs trees. a) The children correspond to the characters. b) The father corresponds to a character. Source: (Xu-Cheng Yin et al., 2014).

In the next step, character candidates are grouped into text candidates by the single-link clustering algorithm. This algorithm merges the two clusters whose two closest members have the smallest distance until a specific threshold is exceeded. The resulting clusters form a hierarchical cluster tree if termination threshold is specified.

Authors use the bases of the single-link algorithm to represent each character candidate as data points, and the text candidates as the top-level clusters in the final cluster tree. The distance function and the threshold are the parameters that define how to cluster the regions into text candidates. Therefore, to determine these parameters, Xu-Cheng Yin et al. propose to use the weighted sum of features as the distance function (Xu-Cheng Yin et al., 2014). The feature vector used in the function describes similarities between data points (character

candidates) according to their geometrical characteristics, colors and stroke widths. On the other hand, the feature weight vector together with the threshold is determined using the distance metric learning algorithm proposed by the authors. For the learning task, the authors decide to use the logistic regression as their model and from this; they define the objective function to be minimizing in order to learn the desired parameters. Thus, the character candidates are clustered into text candidates by the single-link clustering algorithm using the learned parameters.

After text candidates are detected, many of these corresponding to non-text candidates. Thence, the authors propose to estimate the posterior probabilities of text candidates corresponding to non-text and remove those with high probabilities of being a non-text before training the classifier. To achieve this, a character classifier (trained with geometrical and stroke widths features common to the characters) analyzes the character candidates that compound each text candidates. Additionally, the prior probability that a text candidate is a text is estimated based on its length. Then, the authors propose to apply the Bayes rule to estimate the posterior probabilities of text candidates based on their prior probability and their probability of being a text according to their character candidates (Xu-Cheng Yin et al., 2014).

A candidate region is rejected if its posterior probability is greater than or equal to a certain threshold. Such elimination helps to train a more powerful text classifier for identifying text. To find the appropriate threshold, the authors tested the performance of their proposal with different values of the threshold. They found that as the threshold increases, the recall value increase, which is beneficial for the scene text detection. The threshold valued defined by the authors on a scale from $0 - 1$ is 0.995.

Text candidates corresponding to true texts are identified by the text classifier. In the proposed system, the authors use an AdaBoost classifier trained on the training set of ICDAR 2011 database as the text candidate classifier.

To extend their system to a multi-orientation text detection system, the authors propose to use a forward-backward algorithm to detect all the text lines,

and once they are detected, their orientations are calculated to put them in a horizontal alignment. In this orientation, text lines can be processed by the system described above.

The proposed system is evaluated on the benchmark ICDAR 2011 Robust Reading Competition database (Challenge 2) and achieves $recall\ 68.26\%$, $precision\ 86.29\%$ and $f-measure\ 76.22\%$. Furthermore, the method of Xu-Cheng Yin et al. is also evaluated on the ICDAR 2013 contest dataset, under the name of USTB_TextStar, their proposal achieves the best result with $recall\ 66.45\%$, $precision\ 88.47\%$ and $f-measure\ 75.89\%$. As it is possible to compare, their proposal produces a better $recall$, $precision$, and $f-measure$ over other methods in the databases of these editions (Table V - Karatzas et al., 2013; Table 1 - Xu-Cheng Yin et al., 2014). Authors mention that the processing of the proposal is profiled on a Linux laptop with a 2.00 GHz processor consuming an average of $0.43s$ per image, being faster than the presented by Shi et al.(Shi et al., 2013) and Neumann and Matas (Lukás Neumann & Matas, 2013a).

3.2 NERC

As mentioned before, Named Entity Recognition and Classification is one of the important tasks of NLP since it allows recognize information units like names and numeric expressions. NERC is capable of structuring information from unstructured text by identifying references to entities in text (Nadeau & Sekine, 2007).

In this research, it is essential the use of NERC techniques capable of interpreting the textual information previously extracted from identity documents. Information such as the name of the bearer, date of birth, document validation date, etc. are some examples of the data that should be automatically classified to achieve the objective of this research.

Although there are different approaches to perform a NERC task, only those works that carry out a NERC process based on rules are analyzed. This decision is based on the limitations of each approach and considering that one of the contributions of this research is the ability of extract and interpret the textual information of identity documents without prior knowledge of their templates.

Therefore, this research is not limited to a certain number of IDs templates, and it cannot consider use IDs images examples as training data to supervised learning. While it is generally recognized that manually building and customizing rules is a complex and labor-intensive process, when training examples are not available to perform a supervised machine learning, handcrafted rules remain the preferred technique (Nadeau & Sekine, 2007; Santos & Guimarães, 2015).

3.2.1 *Rule-based NERC*

Although most state-of-the-art results for NER tasks based on machine learning (Nadeau & Sekine, 2007; Santos & Guimarães, 2015) , the rule-based approach is extremely appealing due to the associated transparency of the internal system state, which leads to better understanding of errors. Rule-based systems may achieve a high degree of precision, but one of the major challenges of these in practice is producing accurate results in new domains. In machine learning-based systems, adapting to a new domain has traditionally involved acquiring additional labeled data and learning a new model from scratch. However, when a large amount of these additional data is not available to train a new model, the balance to perform a NERC task is inclined to use a rule-based system (Jiang, Banchs, & Li, 2016; Nagesh et al., 2012).

The most basic method for named-entity detection is based on the use of dictionaries. This method consists of comparing a set of pre-stored named-entities with a word or a group of words candidate to named-entities (Lee, Kang, Kim, & Rim, 2014). Nevertheless, a conventional dictionary-based method cannot fully process named-entities in different domains or when the information continuously changes between different documents (for example, the name of the bearer in the IDs).

Domain customization for rule-based NERC typically requires a significant amount of manual effort to identify the explicit semantic changes needed for the new domain and implement the customization rules (Xiaohua Liu, Zhang, Wei, & Zhou, 2011). Although in the scope of this research Identity Documents of different templates (nationality, organizations, etc.) are processed, domain customization is not a limitation since textual information from IDs are often

structured in a similar pattern. This affirmation stems from how the information in IDs is composed (as shown in Figure 1-1), where there are usually fields representing titles (Keywords) with fields representing their respective values. Therefore, this limitation is not an inconvenience to use a rule-based NERC system in this research to find these Keywords and the information of interest.

- *Rules Categories*

A rule-based NERC system often consists of a combination of four categories of rules (Chiticariu et al., 2010):

  - Feature definition (FD) rules identify basic or low-level features from text or components of an entity. For instance, they can be used to identify a candidate to first name or last name based on dictionaries or by detecting words in caps.
  - Candidate definition (CD) rules identify complete occurrences of an entity by combining the output of multiple FD rules. E.g., the first name followed by the last name is a person candidate.
  - Candidate refinement (CR) rules refine candidates generated by CD rules and share different annotation types. E.g., discard candidate persons also identified as candidate locations (“Washington”).
  - Consolidation rules (CO) resolve overlapping candidates generated by multiple CD and CR rules. For instance, a CD rule may identify ‘Dr. Jones’ as a person, while another CD rule may identify ‘Jones’ as a candidate person. A consolidation rule is then used to merge these two annotations to produce a single annotation for ‘Dr. Jones’.

Some rules frequently used are based on regular expressions, dictionary with references to words in the processed corpus, sequences candidate analysis, filters to discard/retain certain intermediate annotations based on predicates on the annotation text and its local context, etc. (Chiticariu et al., 2010). Thus, the combination of rules from different categories allows detect named-entities in text and structured the information by its meaning.

3.2.2 Related Work

A paper related to rule-based NERC systems is selected to analyze it considering how the named-entities are identified. Some interesting aspects are analyzed considering its possible adaptation in the proposed system to interpret all the information of interest. The main contribution of the studied work to this research is the approach employed to define rules and use them, specifically the feature definition rules.

- *Towards Efficient Named-Entity Rule Induction for Customizability (Nagesh et al., 2012)*

Nagesh et al. propose an approach that facilitates the process of building a domain invariant NERC system with the use of customizable rules via rule induction. Their goal is generating a set of CD and CR rules (two of the four rules categories) from an annotated document corpus, a set of FD rules, and a default CO rule for each entity type (Nagesh et al., 2012). Although using a domain invariant NERC system is not the focus of the research developed in this document, the study performed by Nagesh et al. proposes interesting ideas related to how they define the rules to identify the entities.

The authors propose to create features definition rules in the form of dictionaries and regular expressions and used them to identify components of an entity such as first name and last name. These rules specify basic character-level extraction primitives such as regular expressions that identify uppercase words or numbers and dictionaries that match over text, creating a tuple for each match (Figure 3-10).

```
create view Caps as
   extract regex /[A-Z](\w|-)+/ on D.text as match from Document D;

create view First as
   extract dictionary 'FirstNameGazeteer' on D.text as match from Document D;

create view Last as
   extract dictionary 'LastNameGazeteer' on D.text as match from Document D;
```

Figure 3-10 Example of FDs rules used to identify named-entities. Source: (Nagesh et al., 2012)

The candidate definitions (CD) and the candidate refinement rules (CR) are created by induction using clustering algorithms and propositional rule

learners. CRs are used to discard outputs of the CD rules that may be incorrect. These CRs are defined by conditions that select the adequate entity for an annotation identified as two or more different entities. Finally, default CO rules for each entity type are given as input and they are used to remove overlapping spans from CRs rules.

The analysis of this paper shows one of the possible techniques to identify entities using rules. An extrapolation of the use of these rules to the scenario where the system proposed in this research acts, could be a solution to identify which of the textual information extracted from an ID corresponding to the name of the bearer, the ID number, and the rest of the interesting data contained on an ID. The creation of features definition and candidate definition rules based on regular expressions and dictionaries could be used to identify the information of interest, and the information contained in IDs referred as “Keywords” in the scope of the system proposed in this research work (as shows Figure 1-1). Furthermore, the use of candidate refinement rules considering the Keywords detected, could be used to verify if the information of interest was classified in the adequate entity category.

3.3 Processing of Identity Documents

Identity documents are considered complex documents that contain a combination of different fonts, images, background artifacts and words. As there is no international IDs template, achieving a system capable of locating the text within the complex background for any ID is a difficult challenge. The use of one of the methods described above to extract the text information from ID images and then classify it, could automate the registration process that is common in the administrative and services sector.

Many researchers have worked on document processing systems to identify and extract the information of interest. Nevertheless, systems used to classify the information extracted from the Identity Documents are analyzed mainly in Patents compared to public publications, since their use is of great interest to the private sector (ABBYY, 2019; GB Group Plc, 2019; GmbH Anyline,

2019; ICAR VISION SYSTEMS, 2019; US 10,084,606 B2, 2018; Microblink, 2019).

Two research works are analyzed, one is based on the classification of identity documents according to the type of document and the origin, and the other is based on extract the semantic information from complex documents. In both papers are proposed templates-based systems, since the authors determine that extracting the textual information is facilitated by the classification of the document.

3.3.1 *Use case visual Bag-of-Words techniques for camera-based identity document classification (de las Heras et al., 2015)*

De las Heras et al. present a real application for identity document classification of images taken from mobile devices (de las Heras et al., 2015). On images acquired with a mobile device, the identification process must deal with perspective distortions, image blurring, highlights, etc.

The proposed method (Figure 3-11) relies on the traditional Bag-of-Words (BoW) framework, where a SURF descriptor (Bay, Tuytelaars, & Van Gool, 2006) is used to detect the local features and then, a k-means algorithm clusters these features into a vocabulary of representative words. Finally, each image is represented as a histogram of quantized local features. The authors use this histogram representation to learn an SVM classifier for the different document classes. In addition, to make the proposed method more robust, the authors train with three modifications that introduce blurry images in the dataset, an approximate segmentation of the input document and the inclusion of spatial information in the representation of the image.

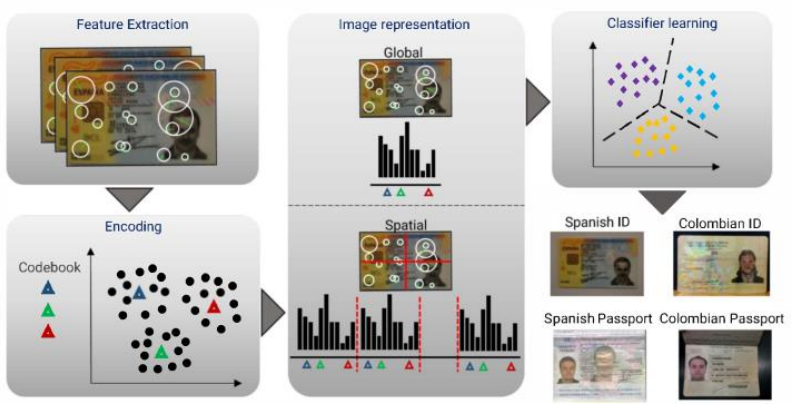

Figure 3-11 Identity document classification pipeline proposed by De las Heras et al. Source: (de las Heras et al., 2015).

For the approximate segmentation, they rely on prior knowledge that ID documents contain text, and some of them include a facial image of the owner. Therefore, when detecting the text and the portrait image, the document is segmented in the approximate limits of the document. The algorithm of text detection used is based on the one proposed by Neumann.

The authors evaluate their proposal with different configurations in three datasets containing more than 2000 images from 129 different document classes. The datasets are divided into mobile-taken images, scan images and synthetic images (de las Heras et al., 2015). The results (Table 1 - de las Heras et al., 2015) show how the performance of the methods is increased once adding blurring to the training and approximate segmentation modifications are used. This increase occurs, since adding blurred images to the training set leads the classifier to discriminate better between similar document classes. Moreover, the classification of a segmented document is higher once the background is removed and the resulting image is similar to those used for training.

The approach presented in this paper is limited to how many document templates are known. However, although the proposed system in this research work is based on identity documents processing without prior knowledge of their template, a template of an ID can be created once process it. Templates can be an output of the proposed system, and use them could accelerate the processing stage of the same kind documents. Thus, the proposed system in this research work more than extract and interpret the information of interest from IDs, could

automatically create templates from the first ID processed and use the approach of this paper to classify other documents of the same type.

3.3.2 *Semantic information extraction from images of complex documents (Peanho, Stagni, & da Silva, 2012)*

On the other hand, in order to process electronically the contents of printed documents, Peanho et al. propose a semi-automated system capable of extracting information from digital document images (Peanho et al., 2012). Their proposal is based on the classification of the values extracted from electricity bills. Although the focus of this paper is not IDs, methods, and techniques presented in this work can be adapted to this kind of document.

The system is based on a document model created from a scanned image whose semantic contents will be extracted. This document model is used to manually provide the semantic interpretation of each field and employ it as a template. Accordingly, all scanned images of other instances of the same kind of document match with this model, thus the interpreted contents can be extracted.

In order to improve the matching process, the authors previously perform some DIP methods over the scanned images. These methods aim to detect the text regions of the scanned documents and use them, together with the previously registered document model, as inputs of the matching process to detect the best match between the document model fields and each text bounding box of the instance image. To obtain these text regions, they follow a linear workflow as shown in Figure 3-12.

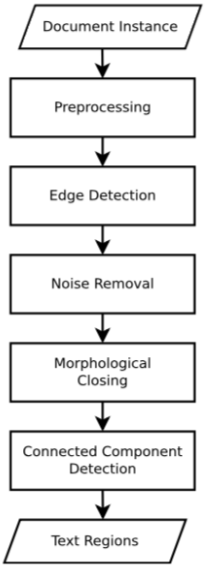

Figure 3-12 Text regions segmentation approach proposed by Peanho et al. Source: (Peanho et al., 2012)

One of the preprocessing techniques they employed to enhance the text regions detection is the well-known and widely used Hough transform (Illingworth & Kittler, 1988). This transformation is used to detect the text baseline skew angle θ of the image, and then rotate the scanned image by −θ in order to align the text with the reference axes.

To create the document model, the authors use an Attributed Relational Graph. Their Attributed Relational Graph is a directed graph in which the vertices represent document fields and the edges represent relationships between fields (Figure 3-13).

Figure 3-13 Example of how the relationship between fields are established using Attributed Relational Graph. Source: (Peanho et al., 2012)

They take a registration image, obtained as any document instance, and generate an Attributed Relational Graph for the document model, based on a semi-automated process. They used the same text detection process to register new fields. Nevertheless, all the process is supervised by users since the segmentation process can miss some text regions or detect wrong. Moreover, users can also add parent relations between registered fields to the Attributed

Relational Graph. This semi-automated application avoids the need to force the user to enter every field region and insert every fixed field content by hand.

Another interesting technique employed by the authors is the use of simple heuristic rules to simplify the search for the best match between the two Attributed Relational Graphs. They create rules that check if some fields correspond to the expected structure. For example, in the case of electricity bills, it is natural to expect that a numerical value is always near the fixed text field which explains its meaning, and its relative position to this instruction field does not change significantly when compared with other instances of the same document. Furthermore, they use features such as position, width, contents, etc. to increase the possibility of possible matches.

3.4 Final Discussion

Once reviewed the principal works related to some aspects of the research proposal, some conclusions can be obtained based on their proposed approaches and results. Next, the conclusions are divided into three sections based on their relation to the research proposal (text localization, name-entity classification and documents processing).

3.4.1 *Discussion about the works related to text localization and recognition*

The following conclusions are based on the study carried out of different works published in the ICDAR 2013 competition (Karatzas et al., 2013). It is good to remember that researchers from around the world accept these contests as the standard for evaluating the performance of different scene text recognition and segmentation algorithms. Moreover, the selected works are those that specifically obtain the best results in Challenge 2, since the conditions of the dataset of this challenge (focused scene text) are similar to those presented in the images to be processed by the research proposal.

The main conclusion obtained is that, despite the different approaches that exist for the text localization, the region-based methods are the most attractive to be used for this task according to their results. As shown in the ICDAR 2013 contest (Table V - Karatzas et al., 2013), three of the best results are MSERs-based methods.

Despite the works described are MSER-based methods, both use a different prune MSERs algorithm to classify the regions corresponding to text and non-text. It is interesting to note that these proposed methods are based on the geometric characteristics of the characters to classify the detected regions. Nevertheless, as Xu-Cheng Yin et al. describe in their work, their proposal achieves a better result for characters detection in multi-language images since they use simple regularized variations, contrary to the method proposed by Neumann & Matas of using a complex classifier, where features like Euler number and horizontal crossing are language-dependent. In a comparison according to the time consuming on processing an image, the method proposed by Xu-Cheng Yin et al. offer a speed advantage over the other work described under similar technology profiles.

3.4.2 *Discussion about the works related to the NERC systems*

As mentioned above, a named-entity classification task is essential when an interpretation of information of interest is needed. NERC task is capable of structuring information from unstructured text. As after a text recognition process, all the textual data is unstructured, in order to classify the data contained in IDs the use of a NERC task become fundamental.

While existing different approaches to performing a NERC task, the rule-based method is chosen as the most appropriate to be used in this research. This selection is based on the conditions and limitations of the system proposed, where the system is not restricted to a certain number of IDs templates, and it is domain-specific since the structure of IDs does not change significantly.

A combination of different rules could allow classifying the information of interest previously extracted from IDs. An adequate use of regular expressions and dictionaries as features definition rules, followed by candidate definition and refinement rules can be employed to identify those fields called as Keywords and the information referred by them.

3.4.3 Discussion about the works related to the documents processing techniques

The works used as references in this domain describe systems capable of classifying the information contained in documents according to the type of document and its model. Although in the related works the textual information of ID is not extracted, which is one of the objectives of this research proposal, the authors affirm that the previous classification of the document facilitates the process of extracting information.

In both papers to classify the documents, the authors propose template-based methods. The use of templates to extract the information from documents guarantees a fast processing since the location of this one is previously known. However, the position where the information contained in an identity document is located varies between documents with different formats. Thus, systems that use templates to classify the information of interest based on their positions are limited by the number of templates they could recognize.

Nevertheless, there are many interesting aspects in the last paper analyzed (Peanho et al., 2012) that can be used in this research. More than describe a whole text region detection process, they propose the use of a directed graph in which the vertices represent document fields and the edges represent relationships between fields (Attributed Relational Graph). These relationships are based on features such as position, width, expected structure, etc. Although these techniques are used in a template-based system, a similar approach can be adapted to the proposed system. This can be used as rules (CD and CR) that establish the relationship between the *Keywords* and the information of interest.

Chapter 4

4 Research Methodology

When dealing with complex documents, such as Identity Documents, in which the contents of different regions, fields, and layout can be highly heterogeneous, the recognition of the information of interest can be a difficult problem. Therefore, in this chapter is described the research used to implement a system capable of extracting and attributing semantics to the textual information of identity documents. The chapter is structured in three different section. First, the scenario where the proposed system operates is explained. Subsequently, the computer vision methods to be used to achieve a successful text extraction in the proposed scenario are mentioned. Finally, it is described how the named entities are identified and the created algorithm to attribute semantics to the information extracted.

4.1 System operation scenario

The proposed system is intended to extract text data present in the ID, such as the full name of the bearer, date of birth, identification number, etc. Once these data are extracted, the next objective is to structure them according to their meaning. Nevertheless, to classify all the information of interest in their adequate entities, the proposed system uses a set of specific words capable of attributing semantics to other sentences or to the document (section 4.3.2). From now on in this work, the term Keywords is used to refer to these specific words. As shown in Figure 4-1, words or sentences such as "Nome", "No. de Inscrição" are examples of Keywords that identify the name and the ID number of the bearer.

Figure 4-1 Marked in yellow, Keywords that identify and classify information in an ID. Source: The Author.

Identity documents commonly contain these Keywords to refer to the citizen information. Without the use of these Keywords, it would be almost

impossible to distinguish certain information when there are many with the same format, such as the document date of issue and the date of birth of the bearer. Since the proposed system uses these Keywords to assign semantics, the input image must contain them in order to achieve a successful structuring of the information.

In the proposed system, the input image is taken with a smartphone to take advantage of their mobility. Nevertheless, since image processing operations could be too intensive to run on a mobile phone, with basis on a cloud computing approach, the image is sent to a server for intensive processing. In addition, considering the objective of the proposed system, the image taken by the smartphone must have as the main focus of the captured scene an identity document. In this way, it is guaranteed that the conditions of the input images are similar to those of the ICDAR contests (Challenge 2), being reasonable to use an approach similar to the techniques described in the related works.

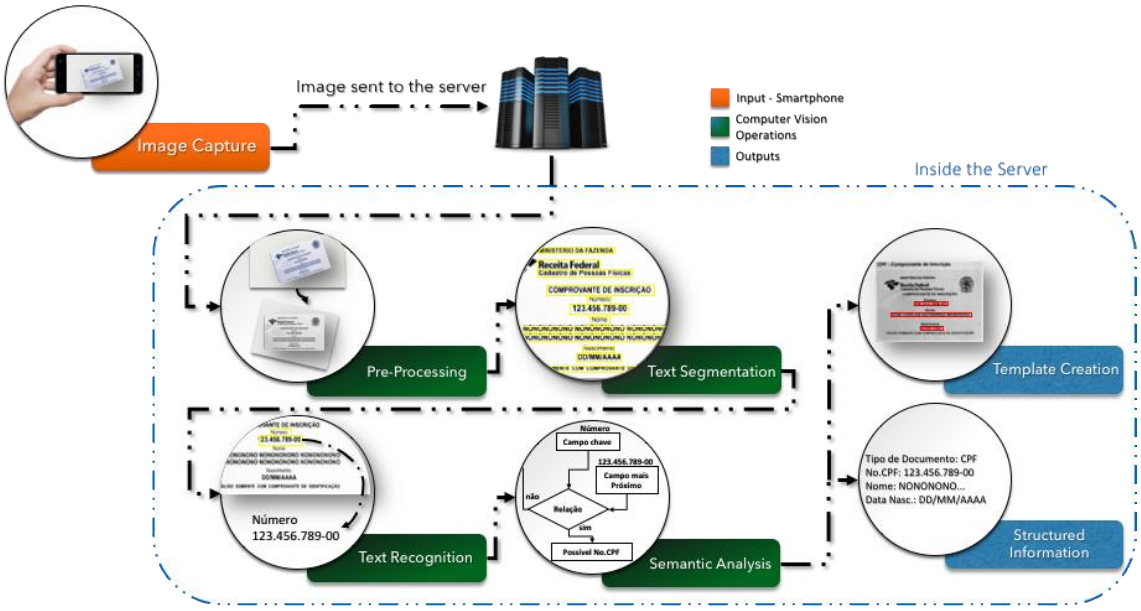

Figure 4-2 Inside the server, the Global Architecture of the proposed system. Source: The Author.

Figure 4-2 shows the global architecture of the proposed system. Inside the server, the image taken by a smartphone is processed with Low, Mid and High-level computer vision operations to finally obtain the information structured according to its meaning.

4.2 Text extraction method

Digital image processing systems intended to extract textual information from images are divided into two main tasks: text localization and text recognition. The former finds and identifies the pixels that make up the individual regions within an image that belong to textual information, while the latter recognizes each character and text lines of the identified regions and returns them as text data.

This research focuses on text detection rather than recognition since due to the challenges of scene text, several works have shown that text localization knowledge is required to achieve good recognition accuracy (Epshtein et al., 2010; Huang et al., 2013; Islam et al., 2016; Uchida, 2014; Xu-Cheng Yin et al., 2014). The result of the localization stage is a set of text regions candidates. However, to find the Keywords and the information of interest, these regions first need to be interpreted and classified as text by text recognition algorithms. Therefore, in this research, they are sent to the off-the-shelf Tesseract OCR engine (R. Smith, 2007).

As shown in Figure 4-2, the architecture of the system is compound by four computer vision operations (green blocks). They are the core of the system and the first three of them are implemented considering the text extraction techniques of the related works. As the proposed text extraction process is divided into two stages (first localization and next recognition), this is separated into the following two subsections. All the text extraction process detailed here is based on a previous research described on (Valiente, 2018). This work was developed as the first step to obtain this research

4.2.1 *Text extraction: text localization method*

As it was concluded in

Chapter 3, the use of the region-based methods for text localization seems to be an adequate selection according to their results obtained in the ICDAR contests (Table V - Karatzas et al., 2013). Therefore, in this work is proposed to use a method based on Connected Components to obtain, in a similar way to the related works, the MSERs of the image. The proposed method optimizes the

conditions for text recognition before providing the image to Tesseract OCR engine.

Images preprocessing helps to improve the performance of the MSER-based methods, since images can be taken under different lighting conditions and at various distances from the smartphone. Preprocessing methods such as contrast adjustment and the use of a median filter allow to enhance the image and reduce its noise once it is transformed to grayscale (Figure 4-3).

Figure 4-3 Results of the proposed preprocessing methods from an input image. Images are zoomed in to observe the effect of the methods. Source: The Author.

In addition, during the process of acquiring the ID image, the text lines could have a certain degree of tilt respect to the horizontal, thus causing errors in the results of the OCR process (Table 2 - Valiente, Sadaike, Gutiérrez, Soriano, & Bressan, 2016). Thus, image rectification is necessary and is executed by performing an improved Hough transform (Cheriet et al., 2007; Shih, 2010). Hough transform is not applied to the original image, but to a filtered image.

As proposed in the previous research described on (Valiente, 2018; Valiente, Gutiérrez, Sadaike, & Bressan, 2017; Valiente et al., 2016), the images are filtered using an adaptive threshold method combined with morphological operations. The adaptive threshold performs the conversion of a grayscale image to a binary image after classifying as background pixels those whose intensities are less than their respective local threshold, contrary to the foreground pixels (Figure 4-4(a)). On the other hand, morphological operations are used with square structuring elements of different sizes for two reasons. First, to filter those regions formed by foreground pixel with a size "too small" respect to a gaussian distribution of the sizes of all the regions. The other reason is to connect the

remaining regions, grouping those that are close to each other. This last operation is intended to group the regions corresponding to the text (Figure 4-4(b)).

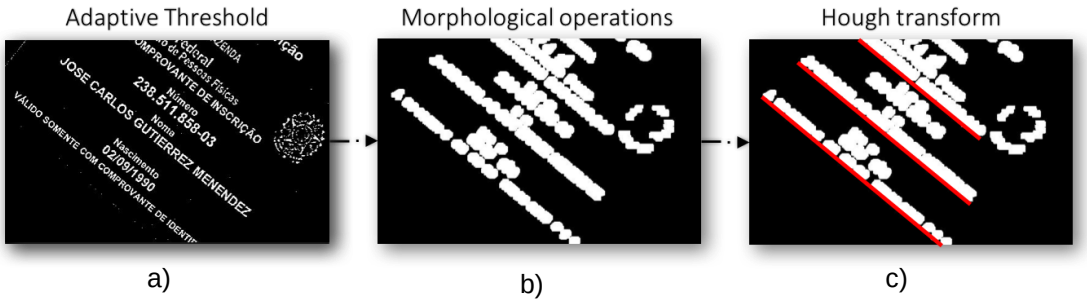

Figure 4-4 Results of the proposed preprocessing methods to rectify the image. a) Binary image obtained as result of the Adaptive Threshold filter, b) Output of the morphological operations, some noise regions are eliminated while other regions are merged, c) Text orientation baseline detected and used to determine the skew angle. Source: The Author.

At this moment, the Hough transform is performed on the filtered image. The horizontal and vertical line segments of the regions (text area) are extracted and used as image features for image rectification (Figure 4-4(c)). Thence, the transformation to be applied is computed and with the skew angle, the image is rectified.

From the rectified image, the proposed system performs an approximate segmentation, similar to the one proposed in (de las Heras et al., 2015). Nevertheless, the criteria used for segmentation are not the same. In this proposal, the image is cropped to the size that fence the area where the greatest amount of information with similar sizes is concentrated in the filtered image. The purpose of this segmentation is to separate the identity document area from the rest of the image, in this way, all the following processing operations only operate on the area of interest.

Figure 4-5 From the rectified filtered image, it is assumed that the regions of interest are in the area where the greatest amount of information with similar sizes is concentrated. Therefore, the image is cropped to only contains the information of interest. Source: The Author.

This delimitation of the text candidate regions during the location stage provides a reduction of the computational cost for the subsequent processing stage, subjecting only those regions to a structural analysis, classifying them as textual or non-textual.

Once the image has been cropped, the next operation is to detect the MSERs. MSER-based methods detect a set of connected regions from an image, where each region is defined by an extremal property of the intensity function in the region and its outer boundary. Although the MSER algorithm detects most of the text, it also detects many other stable regions in the image that are not text (Figure 4-6).

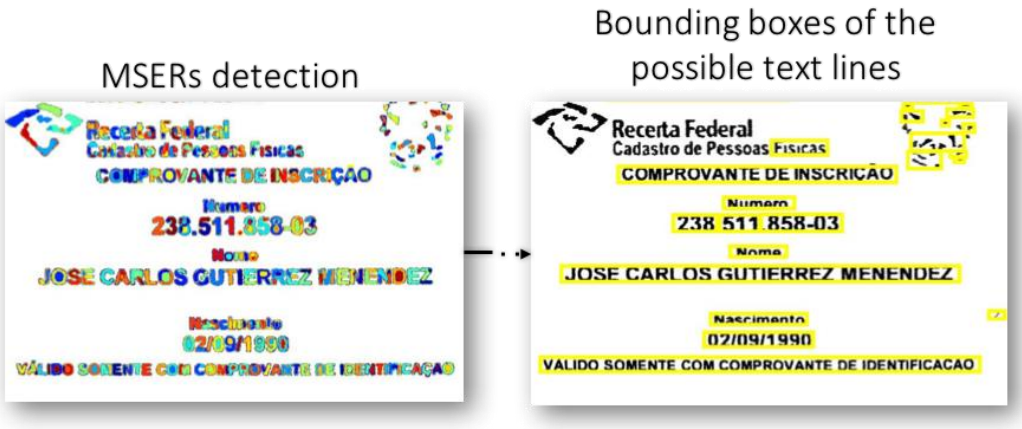

Figure 4-6 In the left image, the MSERs detection result where are detected most of the text, but there are also detected many other stable regions in the image that are not text. In the image on the right, the resulting bounding boxes after the filtering based on geometrical features. Source: The Author.

In order to eliminate the non-text detected regions, the system uses several geometric properties that are good for discriminating between text and non-text regions. Stroke width is a common metric used to discriminate between text and non-text. It is defined as the length of a straight line from a text edge pixel to another along its gradient direction. The stroke width remains almost the same in a single character. However, there is a significant change in stroke width in non-text regions as a result of their irregularity. Text regions tend to have little stroke width variation, whereas non-text regions tend to have larger variations (Epshtein et al., 2010; Phan et al., 2012); thus, regions with larger variations are removed.

At this point, all the detection results are composed of individual text candidates (regions). To use these results for recognition tasks, such as OCR, the individual text characters must be merged into words or text lines, which carry more meaningful information than just the individual characters. The system proposes grouping the characters into words based on distance, orientation, and similarities between characters. To achieve it, the bounding boxes (BB) of each individual character are horizontally expanded based on their aspect ratio. Therefore, the overlapping BBs of similar regions can be merged to form a chain of horizontal single bounding boxes around individual words or text lines. Moreover, as in the related works, the proposed system assumes that each text region is generally found in groups (words and sentences) (Epshtein et al., 2010; Lukás Neumann & Matas, 2013a). Thus, those isolated regions represented by BBs with a single region inside are eliminated.

Once this point is reached, the candidate regions for text (words or sentences) are segmented by bounding boxes (Figure 4-6). Therefore, these BBs can be sent to an OCR engine to determine their textual content.

4.2.2 *Text extraction: text recognition method*

The proposed text recognition process is based on the thesis work of Valiente (Valiente, 2018; Valiente et al., 2017) using the Tesseract OCR engine (R. Smith, 2007). Tesseract is an open-source library currently supported by Google used for machine-printed recognition featuring line, word, and character segmentation and character recognition. A useful capability of this library is that it can be trained for newer language and scripts. Tesseract OCR Engine is considered as one of the most efficient open source OCR engines currently available and it has demonstrated competitive OCR accuracy (Doermann & Tombre, 2014; Mishra, Patvardhan, Vasantha Lakshmi, & Singh, 2012).

In simple images, Tesseract provides results with a high accuracy rate. Nevertheless, in the case of complex images as identity documents images, the vast number of different layouts, languages, and fonts make it hard to robustly read all information using plain an OCR engine. Thus, Tesseract provides better precision results if it is used in those areas where the candidate text is found

(Patel, Patel, & Patel, 2012; Valiente et al., 2016). This latter is the main reason why it is necessary to perform a previous text localization process.

When an image is filtered, its information is modified. Therefore, if a series of different filters are applied to an image, a set of filtered images representing the same scene is obtained, but the information they contain is not represented in the same way.

As explained in the previous research (Valiente, 2018), once detected the regions of interest (RoI) of the image, these are sent to the OCR engine. In this work, the term RoI and bounding boxes (BBs) are used interchangeably to refer to the area where the candidate regions for text are located. The result of the recognition confidence depends on the level of text classification in these regions. That means how the Tesseract recognizes the text based on its classifier. Hence, as it is known the location of the BBs, these can be extracted from the set of filtered images and sent to the OCR. As is to be expected, the recognition confidence results for each bounding box vary in relation to the image that is processed.

The confidence value is a percent of accuracy given by the OCR function (R. Smith, 2007), that can be used to select between each word processed the most accurate OCR result. Concisely, it is developed an iterative algorithm in MATLAB capable of selecting the best word recognition result once a set of filtered images that correspond to the same scene is processed (Figure 4-7). In addition, it is proposed to eliminate regions with low confidence values and those where a single character is found. The output of the proposed algorithm improves the final word recognition process.

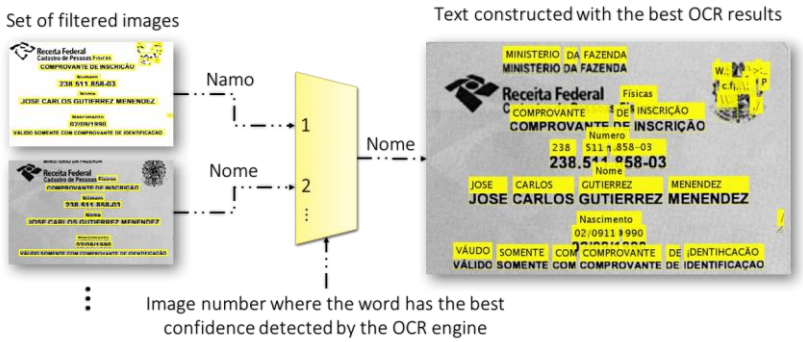

Figure 4-7 Proposed algorithm to improve the final text recognition process. Source: The Author.

Figure 4-7 shows an example for the bounding box of the word “Nome”. For instance, in the first image of the set, the OCR output is “Namo”, with 87% of accuracy, and for the next image, the OCR output is “Nome”, with 89% of accuracy; then, the algorithm selects the second, which is the better result. The same process is performed for all the bounding boxes. Experimental results of this algorithm confirm that the process enhances the result of OCR, allowing it to obtain better accuracy of words recognition (Valiente et al., 2016).

Once the text is formed with the best words, the text extraction process concludes. Then, all the textual information of the identification documents is sent to the algorithm of semantic attribution and NERC method.

4.3 Semantic Analysis Process

There are few papers related to the extraction of IDs information. The found ones are based on the use of templates to extract the information of interest (de las Heras et al., 2015; Ryan & Hanafiah, 2015; Simon et al., 2015). The use of templates to extract the information from the ID guarantees a fast processing since the location of this one is previously known. However, the position where the information contained in an identity document is located, also varies between documents with different formats. Thus, systems that use templates to classify information based on their positions are limited by the number of templates they could recognize. Moreover, a person must manually select the region of interest where is located the information of the citizen that must be extracted. Therefore, the main difference between these works and the one proposed, is the implementation of an automatic process that use a novel mechanism to obtain the semantics of the text using Keywords like classifiers; hence, it is not necessary the use of templates or a prior knowledge of where the information of interest is located to performs the ID recognition.

The term Semantic describes the meaning or interpretation of the linguistics symbols as words or expressions. Therefore, the Semantic proposes how the meanings are assigned to the words (Lyons, 1995; Y. Zhang & Liu, 2007). In this work, the use of the term semantics refers to the classification or assigned meaning that sentences could acquire.

Once the proposed text extraction process ended, all text data of the ID have been detected by the OCR. Nevertheless, the system is not able to identify the meaning of each detected word or the relationship between them, as it does not have a semantic analysis process. Therefore, this research proposes a Semantic Analysis mechanism as a new layer of a text recognition system.

The mechanism is designed to process images of identity cards or documents containing personal information referenced by specific words (Keywords). The set of these specific words is stored in a dictionary previously created.

The first stage of this process, uses the distances between each word detected by the OCR to recognize sentences and relationship between them. Once all the words have been related, the information is structured as sentences. Then, in order to automatically identify each information in an ID and thus avoid using templates, the system must be able to differentiate lines of text by their meaning. Based on the pattern of identity documents (not in the layout), there are some Keywords used as headings or titles fields that attribute semantic value to the information to which they refer (Figure 4-1). Therefore, finding these Keywords and using them as classifiers allows the system to identify, for instance, which line refers to the name or ID number, this being the second stage of the process

4.3.1 *Semantic Analysis Process:* ***Sentence detection***

In the proposed system, once finished the text extraction, some sentences or words could be divided into separated fields, causing the loss of their semantic values. Figure 4-8 shows examples of such text-parts carry little semantic value when viewed separately but become semantically relevant and easily identifiable when perceived as a group. Thus, relating all fields of words detected and handling them as sentences, improves the way in which the system recognizes the semantics of information.

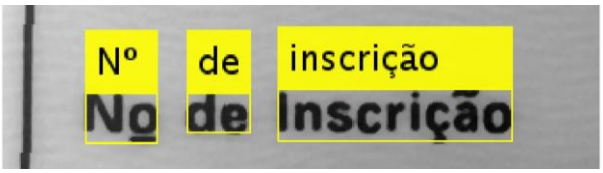

Figure 4-8 Marked in yellow, the detected fields that correspond to the same sentence but they are separated. Source: The Author.

The sentences of the Indo-European grammar are in horizontal direction. Therefore, the procedure used by the system to recognize sentences, is based on all words must be in the same horizontal line or close to this to be considered as words of a same sentence; additionally, adjacent words must have a similar height (L. Gomez & Karatzas, 2013; Kok Loo & Lim Tan, 2002; Lukás Neumann & Matas, 2012).

This research proposes to calculate the distances between all the words to classify them like near or distance words in a similar or different horizontal line, based on the Euclidean Distance and the heights of the words. Algorithm 1 shows how the proposed system structured the words detected as sentences.

ALGORITHM 1: Sentence detection

Require: W // W *set of all detected Words*
for ***each*** w **in** W **do**
  $W \leftarrow W - 1$ // W *has a word less to analyze*
  $S_w \leftarrow w$ // S_w *is a sentence composed of* w
  for ***each*** w_2 **in** W **do** // *HL, h and D are the horizontal line, height and distance of each element*
    if $HL_{S_w} \cong HL_{w_2} \wedge (D_{S_w,w_2} = near) \wedge h_{S_w} \cong h_{w_2}$ **then**
      $S_w \leftarrow S_w \cup s_2$ // S_w *is now a sentence composed of* S_w *and* w_2
      $update\_props(S_w)$ // *Properties of* S_w *are updated*
    end
  end
  if $S_w \neq w$ **then** // S_w *is a sentence composed of more words than* w
    $sort\_by\_position(S_w)$ // S_w *is sorted by the positions of its words*
  end
  $S \leftarrow S_w$ // S *is the set of all detected sentences*
End

Algorithm 1 Detects the sentences of the IDs. Source: The Author.

Algorithm 1 analyses the relationship between all words to recognize which of them are part of the same sentence, based on their heights, horizontal

lines and distance values. The algorithm begins considering that the first word is in itself a sentence (line 4 - Algorithm 1). Therefore, this sentence has the same geometric properties as the first word. When another word meets the requirements to be considered as part of this sentence, the sentence adds this word and updates its geometric properties (lines 6-8 - Algorithm 1). This update considers the new length of the sentence for the subsequent distance calculation with other words (D_{S_w,w_2}). Furthermore, in order to obtain legible sentences, words that compound them are internally organized by the respective horizontal positions (line 12 - Algorithm 1). At the end of Algorithm 1, all information is structured as sentences, even isolated words are considered single word sentences (Figure 4-9). It is worth noting that this stage can be considered as a Consolidation rules (CO) creation process (section 3.2.1); since the system impose that entities found are one of these sentences and not a part of this.

Figure 4-9 Marked in yellow, the detected sentences using the Algorithm 1. Source: The Author

4.3.2 *Semantic Analysis Process:* ***Keywords and Semantic Recognition***

In this research are analyzed four types of specific named entities: person name, date of birth, identity number and Keywords. A set of previously existing techniques are assembled in a novel way, in order to build a system capable of efficiently recognize these contents of interest from identity documents image. The main idea of the methods employed in this work is doing a dictionary lookup to find the Keywords, in order to use them together with context clues of the candidate words to select which type of entity a word represents.

The meaning of *Keywords* is to refer to those sentences that classify and attribute semantic value to others, or to the document, by their meanings. In Figure 4-10, the sentence *"CADASTRO DE PESSOA FISICA"* is an example of a *Keyword* that classifies the type of document. Instead, Figure 4-1 shows marked in yellow four examples of *Keywords* that attribute semantics to other sentences, the words *"Nome"*, the word *"RNE"* and the sentence *"Nº de Inscrição"*. Note that the meaning of the word "*Inscrição*" when it is isolated does not necessarily imply that it refers to a number (Figure 4-8).

Figure 4-10 Marked in yellow, Keywords found. Source: The Author.

When these Keywords are read, for a person, it is easy to identify the information to which they refer, consequently using the same method, an automated system should be able to do it. Hence the second stage of the proposed Semantic Analysis process, is based on the search of the Keywords between all the sentences and the information referred by them. Similar to the rule-based NERC systems analyzed on the related works (section 3.2), a dictionary lookup and FD, CD and CR rules are used to find the information of interest.

- *Keywords Recognition*

The Keywords are located using a dictionary lookup and regular expressions. Regular expressions are used to identify sentences based on their formats (number, letter or both) and number of characters (Clarke & Cormack, 1997). Thus, the proposed mechanism uses a set of regular expressions to find the Keywords entities. Worth noting these Keywords are mostly the same in different identity documents of the same language (Figure 4-11), which reduces the size of the dictionary used. Furthermore, the use of regular expressions aims to recognize a Keyword, even if it is written next to other words, in capital or plural.

Figure 4-11 Examples of Identity Documents from different countries but with similar Keywords (bounding boxes in yellow). Source: The Author

Algorithm 2 describes the process of detect the Keywords using a dictionary of regular expressions. This dictionary is created considering the language and characteristics of the Keywords that have to be detected. The process consists in analyzing each sentence to verify which contain Keywords. However, to detect all possible Keywords, sentences are transformed into alphanumeric text (line 3 - Algorithm 2). This action removes spaces and special characters introduced by possible errors during the OCR process. This alphanumeric text is used as input to a function that checks if it matches any of the regular expressions stored in the Keywords dictionary (line 4 - Algorithm 2). Given a match, the entire alphanumeric expression (with its geometric properties) and the corresponding category of the match are stored in a Keywords structure (line 6 - Algorithm 2).

ALGORITHM 2: *Keywords* detection

Require: S, D_{kw} // S is the set of all the sentences, // D_{kw} is the dictionary of regular expressions of the Keywords

for *each* s **in** S **do**

  $s_{an} \leftarrow$ *text2alphanumeric(s)* // Transform sentences into alphanumeric text

  $kw \leftarrow$ *check_match* (s_{an}, D_{kw}) // check if the alphanumeric sentence match any of the regular expressions of the Keywords

  if not *kw is empty* **then** // if kw is not empty, s_{an} contains a candidate keyword and is saved as a tuple of token and expression

    $KW_s \leftarrow kw$ // kw is saved in the KW_s structure

  end

end

Algorithm 2 Detect the Keywords of the ID. Source: The Author.

At the end of the Algorithm 1, the Keywords structure will contain all the keywords that were detected in the form of token and expression. The number of tokens or keywords categories are defined based on the requirements of the system. In the scope of this research, four categories of keywords similar to the searched named entities are used. These categories are the person name, date of birth, id number and document type. This last category could be analyzed as a Keyword category, but according to the regular expressions dictionary lookup, it is decided to name it as document type category. It is worth remembering that the words detected in Algorithm 2 are entities of type Keywords and should not be confused with the categories of the Keywords structure. For instance, the sentence “Nº de incrição” in Figure 4-10 is a Keyword entity stored in the structure as: “id number keyword”: “Nº de incrição”. Moreover, since more than one sentence can match the same Keyword category, all candidates must be stored. Using the same image as an example, the resulting Keywords structure in Figure 4-10 is shown in Table 4-1.

Table 4-1 Keywords structure result in Figure 4-10.

Keywords structure categories	**Keywords detected**
"document type keyword"	"CADASTRO DE PESSOA FISICA"
"person name keyword"	"Nome"
"id number keyword"	"Nº de incrição"
"id number keyword"	"CPF"
"date of birth keyword"	"Data do Nascimento"

As it is possible to note in this example, two sentences are classified in the same Keyword structure category as "id number keyword". Selecting which of them is the correct Keyword, or if both are correct, depends on the information to which they refer, as detailed in the next section.

- *Semantic Recognition*

Once all the possible Keywords entities are detected, the next step is recognizing which are the other searched entities based on their semantics. To achieve this objective, a group of rules based on Features Definition, Candidate Definition, and Candidate Refinement rules are used to find these entities that correspond to the person name, date of birth and the identification number. Algorithm 3 represents the methodology used to achieve this objective and, subsequently, it is explained.

ALGORITHM 3: NERC and Semantic Recognition

Require: S, D_{nef}, KW_s // S is the set of all the sentences, // D_{nef} is the dictionary of regular expressions of the named entities formats // KW_s is the Keywords structure

for ***each*** s **in** S **do**
  $S_{cFD} \leftarrow check\_FDrules(s)$ // apply the FDs rules and the candidates are stored in the set S_{cFD}
end
for ***each*** c **in** S_{cFD} **do**
  $S_{cCD} \leftarrow check\_CDrules(c,KWs)$ // apply the CDs rules and the candidates are stored in the set S_{cCD}
end
for ***each*** e **in** S_{cCD} **do**
  $Ent \leftarrow check\_CRrules(e,KWs)$ // apply the CRs rules and the candidates selected as entities are stored in the set Ent
End

Algorithm 3 Detect the entities and attribute semantics to them. Source: The Author.

The first rules created and used are the Feature Definition rules based on regular expressions to identify entities by their format (line 3 - Algorithm 3). Similar to the Keyword detection process, a regular expressions dictionary is used. However, it does not contain specific words, since the information of interest varies for each document. Instead, it contains the entities formats used to classify the information as a candidate. That means, for example, candidate sentences to be the person name entity should not contain numeric characters; whereas that it is expected an identification number candidate must contain at least some numbers (considering some ID numbers combine numbers and letters). The output of this rules function is an expression classified as a candidate to be a named entity according to its format. The geometric properties are also stored to be used in the next rules.

On the other hand, Candidate Definition rules are used to find complete occurrences of the entities based on the candidates obtained by FDs rules and the detected Keywords entities (line 6 - Algorithm 3). Although the Keywords are entities, they are also used as part of the rules created to find the others. The proposed CDs rules consider not analyzing all sentences as candidates for the entities sought, but only those close to the Keywords entities found previously. This argument is based on the fact that in Western grammar the texts are written from left to right and from top to bottom (Braccini, DeFloriani, & Vernazza, 1995; Gibson, 1991); hence information indicated by the Keywords have a high possibility to be on the right or below of them. Thus, those sentences previously defined as candidates by the FDs rules, closest to the Keywords entities and in one of the positions mentioned, are selected as potential candidates for information of interest contained in an identity document (Figure 4-12). The geometric properties of the Keywords and the candidates obtained by FDs are used to calculate the distance between them.

Finally, the last rules implemented are the Candidate Refinement rules. These rules are based on the Keywords structure obtained in the previous stage to attribute semantics to the potential candidates and select among all of them, those that are defined as entities (line 9 - Algorithm 3). To verify the selection, the format of each potential candidates obtained by the CDs rules must match the

Figure 4-12 Analysis to relate Keywords with the information to which they refer. Source: The Author.

one indicated by the category of its respective Keyword. For instance, if the Keyword category suggests a numeric expression (such as “id number keyword” category) and the candidate is only text, this is not defined as an entity, since it does not correspond to the type of information referred by the Keyword. In this case, the Keyword cannot attribute semantics thus this candidate is discarded. Otherwise, if they match, the sentence acquires a semantic value (ex: person name, date of birth, etc.) determined by the Keyword structure category and it is classified and structured as an abstraction of its meaning (ex: “Person name: Denis Franco”). Figure 4-13 shows how the information of interest in an ID is structured as sentences and classified according to their respective Keywords. As mentioned before, some Keywords do not attribute semantics to other sentences since they are used to classify the document.

Figure 4-13 Marked in yellow, the information of interest contained in an identity document. The table shows how the information is structured. Source: The Author.

It is interesting to note that an improved stage can be implemented considering that the Keywords were detected. It consists of applying again the OCR process on the fields of the potential candidates obtained by the CDs rules. Although the text inside these regions has already been recognized, since each of these candidates is close to a specific Keyword, its respective category (based on the Keywords structure) can be used as additional information to improve the OCR result. This additional information is based on the data types corresponding to each text region and can be used to constrain the characters obtained from the OCR engine, improving its accuracy rate.

At this point, the proposed system has all the information about the citizen structured according to their meanings. Consequently, this structured information could be stored on a server in the cloud, where other services can access them improving the system scalability.

Chapter 5

5 Data Analysis and Results

Once the end-to-end system capable of recognizing the text and interpreting the textual information of the IDs is obtained, the next step is to evaluate the quality of its output. This chapter describes the evaluation methods used according to the task performed, since it can be considered that the system consists of two main process with different tasks (Figure 5-1).

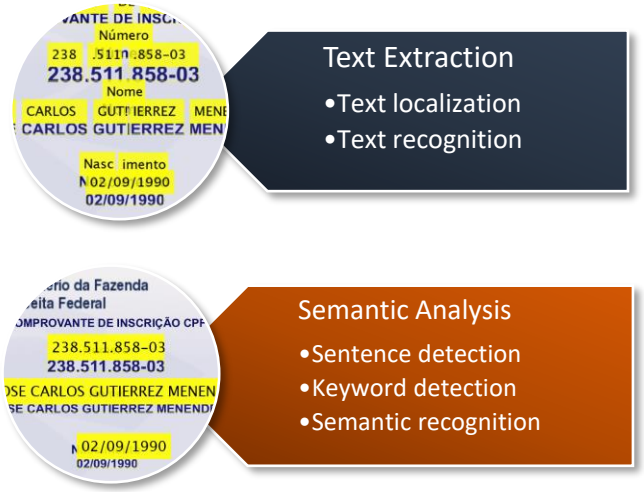

Figure 5-1 The stages that compose the system are shown divided by its main operations. Source: The Author.

The analysis of the results of the Text extraction method has been previously described in the thesis work of Valiente (Valiente, 2018); thus the data analysis in this research focuses on the stage of Semantic Analysis. In order to analyze the real impact of each task that composes this stage, it is considered that their inputs have ideal conditions. This means that for each task, the previous one has an accuracy of 100% and the outputs of this correspond to the inputs of the current task. Finally, the entire system is analyzed to obtain its total performance in a real environment.

Therefore, this chapter is divided into three sections. The first one describes the databases used for the experimental tests. Subsequently, the metrics used to evaluate the results in each task are detailed. Finally, the result of the whole system and each one of its tasks is analyzed based on the

aforementioned. The algorithms presented in this work are implemented in MATLAB R2018b. Implementations and tests are performed on a computer with 2.5 GHz Intel Core i5 processor, 12 Gb RAM and macOS High Sierra operating system.

5.1 Creation of databases

In order to evaluate the individual performance of each task, the combination of them and the whole system, five databases are created. Each one of them is created considering the specific characteristics of its respective task. As mentioned above, this research focuses on the Semantic Analysis stage and the combination of the two stages that represent the end-to-end system (Figure 5-1). Therefore, the analysis of each task that is part of the Text extraction method is found in the work of Valiente (Valiente, 2018). Next, the number of images contained in the databases and the tasks in which they are used are detailed.

Database 1: 124 images of identification documents with the bounding boxes of the words detected (Total: 5834 words). This image database is used to evaluate the Algorithm 1.

Database 2: 25690 sentences extracted from images of identification documents and other types of documents containing similar Keywords. (Total: 5505 Keywords). This database is used to evaluate the Keyword detection algorithm.

Database 3: 124 images of identification documents with the bounding boxes of all the sentences and the Keywords detected (Total: 372 Information of interest). This image database is used to evaluate the Semantic recognition task.

Database 4: contains the same 124 images and bounding boxes of the words detected of Database 1. However, this now used to evaluate the Semantic Analysis Process in general and analyze the real impact of the Algorithm 1 on the system output.

Database 5: 100 images (20 different images with 5 different viewpoints) of identification documents. This image database is used to evaluate the proposed end-to-end system in a real environment.

The image databases consist of Brazilian identification documents as well as other documents mentioned below: CNH (Brazilian driver license), RG (Brazilian identity card), RNE (Brazilian national registry of foreigners), CPF (individual taxpayer registry), IDs created from free templates, passports and identity documents of other countries as shown in Figure 5-2.

Figure 5-2 Examples of identity documents contained in the databases. Source: The Author

A part of the images of the databases 1,3 and 4 are downloaded from the Internet with the rights of free use to share or modify. The information contained in the image databases of the location of the words, as well as the transcription of these, is created using the system proposed in the work of Valiente (Valiente, 2018). Moreover, it is performed a fine tuning of these localizations and transcriptions to obtain ground truth elements or gold standards annotations (a term used in NLP systems to refer to the ground truth annotation (Aroyo & Welty, 2015)) (Figure 5-3).

Figure 5-3 Examples of Ground Truth images used to evaluate the quality of the system. Fine tunings are performed to the locations of the bounding boxes and the transcriptions. The image on the left is an instance of an image that belongs to database 1 and database 4, while the image on the right belongs to database 3. Source: The Author

Database 2 elements are created from a site that generates random identities (Corban Works, 2018). This database was generated and consists of several names, genders of the persons, country names, dates of birth, ID numbers, company names and expiration dates with their respective Keywords (Figure 5-4). On the other hand, IDs used in the database 5 to evaluate the proposed end-to-end system in a real environment are subject to a variety of adverse conditions (Figure 5-5), since they are captured under natural conditions (variable illumination).

'Name:'	'Diana Knox'	'Gender:'	'female'	'Country:'	'United Kingdom'	'Birthday:'	'4/12/1987'	'ID number:'	'KG 28 23 23 C'	'Company:'	'MVP Sports'	'Expiration date:'	'5/2021'
'Name:'	'Arthur Costa'	'Gender:'	'male'	'Country:'	'Italy'	'Birthday:'	'1/29/1968'	'ID number:'	'VJ90773592'	'Company:'	'Golden Joy'	'Expiration date:'	'11/2024'
'Name:'	'Thaís Santos'	'Gender:'	'female'	'Country:'	'Sweden'	'Birthday:'	'8/24/1940'	'ID number:'	'400824-9742'	'Company:'	'Monk House Maker'	'Expiration date:'	'4/2022'
'Name:'	'Enrico Tovar'	'Gender:'	'male'	'Country:'	'United States'	'Birthday:'	'7/5/1950'	'ID number:'	'426-17-4592'	'Company:'	'Vinyl Fever'	'Expiration date:'	'8/2022'
'Name:'	'Betania Chávez'	'Gender:'	'female'	'Country:'	'Italy'	'Birthday:'	'3/14/1994'	'ID number:'	'DK93900242'	'Company:'	'Magik Gray'	'Expiration date:'	'1/2021'
'Name:'	'George Posey'	'Gender:'	'male'	'Country:'	'Finland'	'Birthday:'	'8/28/1948'	'ID number:'	'280848-455K'	'Company:'	[]	'Expiration date:'	'4/2023'
'Name:'	'Pearl Hogg'	'Gender:'	'female'	'Country:'	'United Kingdom'	'Birthday:'	'2/27/1945'	'ID number:'	'HT 78 91 41'	'Company:'	'Quality Merchant Services'	'Expiration date:'	'1/2022'
'Name:'	'Ralph Laporte'	'Gender:'	'male'	'Country:'	'Sweden'	'Birthday:'	'4/18/1964'	'ID number:'	'640418-7293'	'Company:'	'Edwards'	'Expiration date:'	'8/2021'
'Name:'	'Sabina Canales'	'Gender:'	'female'	'Country:'	'Brazil'	'Birthday:'	'1/6/2000'	'ID number:'	'722.891.417-12'	'Company:'	'Kenny Rogers Roasters'	'Expiration date:'	'12/2022'
'Name:'	'Linneo Arteaga'	'Gender:'	'male'	'Country:'	'Finland'	'Birthday:'	'5/16/1965'	'ID number:'	'160565-5872'	'Company:'	'Almacs'	'Expiration date:'	'7/2020'
'Name:'	'Lie Domínguez'	'Gender:'	'male'	'Country:'	'Canada'	'Birthday:'	'6/15/1963'	'ID number:'	'609790050'	'Company:'	'Pro Yard Services'	'Expiration date:'	'5/2024'
'Name:'	'Kauan Fernandes'	'Gender:'	'male'	'Country:'	'Denmark'	'Birthday:'	'12/11/1947'	'ID number:'	'111247-0847'	'Company:'	'King Carol'	'Expiration date:'	'5/2023'
'Name:'	'Ronan Salas'	'Gender:'	'male'	'Country:'	'United States'	'Birthday:'	'9/17/1994'	'ID number:'	'487-72-2349'	'Company:'	'Suadela Investment'	'Expiration date:'	'4/2022'
'Name:'	'Amanda Cunha'	'Gender:'	'female'	'Country:'	'Brazil'	'Birthday:'	'3/22/1981'	'ID number:'	'994.673.433-80'	'Company:'	'Network Air'	'Expiration date:'	'11/2023'
'Name:'	'Scott Gonzalez'	'Gender:'	'male'	'Country:'	'South Africa'	'Birthday:'	'10/1/1938'	'ID number:'	'3810016062084'	'Company:'	'Office Warehouse'	'Expiration date:'	'10/2023'
'Name:'	'Rafael Council'	'Gender:'	'male'	'Country:'	'United States'	'Birthday:'	'10/22/1949'	'ID number:'	'538-30-4139'	'Company:'	'Muscle Factory'	'Expiration date:'	'4/2020'
'Name:'	'James Isaac'	'Gender:'	'male'	'Country:'	'Sweden'	'Birthday:'	'10/28/1984'	'ID number:'	'841028-5095'	'Company:'	'White Coffee Pot'	'Expiration date:'	'8/2023'
'Name:'	'Carla Almeida'	'Gender:'	'female'	'Country:'	'Denmark'	'Birthday:'	'2/22/1984'	'ID number:'	'220284-3940'	'Company:'	'PriceRite Warehouse Club'	'Expiration date:'	'1/2023'

Figure 5-4 Example of the structure of a part of the database 2. Source: The Author

Figure 5-5 Examples of instances that belongs to the database 5. Source: The Author

5.2 Evaluation methods

Since different tasks are used in the system, to measure the efficiencies of each of these, different metrics have to be used. Focusing only on the Semantic Analysis process, three different metrics are used to evaluate each task that composes it and to evaluate it as a complete process. The evaluation of the sentence detection task is carried out applying the aforementioned evaluation

protocol used in the ICDAR contest (Wolf & Jolion, 2006). However, given that Keyword detection and Semantic recognition tasks correspond to named entities recognition processes, in both is applied the metric proposed in the CoNLL-2003 conference proceedings (Sang & De Meulder, 2003). Finally, considering that some partial detections may appear as results in the Semantic Analysis Process and in the proposed End-to-end system, unlike the two previous tasks, the evaluation in them is based on the scheme presented in MUC-5 (Chinchor & Sundheim, 1993). Next, the choice of each metric used to evaluate each task of the Semantic Analysis Process, the process itself and the End-to-end system in their respective databases are detailed and justified.

5.2.1 *Sentence detection task evaluation method*

The Sentence detection task, using database 1, is evaluated similarly to the evaluation protocol used in the ICDAR contest described in section 3.1. The evaluation is computed by $Precision$, $Recall$, and $f-measure$ using the detected and ground truth bounding boxes of the sentences (Equation 3-1). In this research, the threshold used to define a match between the detected BBs and the ground truth BBs is when their overlapping ratio is higher than $0{,}8$.

5.2.2 *Keyword detection and Semantic recognition task evaluation method*

In the case of the Keyword detection and Semantic recognition tasks, although both are carried out in two different databases (database 2 and database 3), same metrics are used, since both tasks consist of detecting named entities. These databases contain sentences composed of one or more words, however, database 2 consists only of text (Figure 5-4), while database 3, in addition to text, contains the location of the bounding boxes of each sentence (Figure 5-3, the image on the right). In order to define which metrics to use, all possible comparison scenarios of the system predictions and the gold standards annotations (ground truth) must be analyzed to determine which of these scenarios correspond to each task and thus know how to evaluate it (Nadeau & Sekine, 2007).

In the following tables, from Table 5-1 to Table 5-6, examples of the different scenarios are shown using the IOB scheme originally put forward by

Ramshaw and Marcus (Ramshaw & Marcus, 1995). As the tables show, each word is tagged individually to determine whether the current word is within a named entity or not. As mentioned above, the tagging scheme used is the IOB (short for Inside, Outside, Beginning)(Ramshaw & Marcus, 1995). This scheme consists of tagging with **O** (Outside) those words that do not belong to any named entity, otherwise, label with I-**XXX** (Inside) those that belong to a named entity of type **XXX**. Whenever a group of words of type **XXX** is within the same bounding box, the first word is tagged **B-XXX** (Beginning) to show that an entity composed of two or more words tagged with the same type and without **O** tokens among them begins. In this research, the data contains entities of four types: person names (**PER**), ID numbers (**IDN**), date of birth (**BTH**) and Keywords (**KEY**).

Table 5-1 The system predicts the classification and entity type correctly

ENTITY ANNOTATIONS	ENTITY CLASSIFICATION	
	GOLD STANDARD	SYSTEM PREDICTION
Name	I-KEY	I-KEY
Jose	B-PER	B-PER
Carlos	I-PER	I-PER
Company	O	O
GutMen	O	O

Table 5-2 The system hypothesized entities

ENTITY ANNOTATIONS	ENTITY CLASSIFICATION	
	GOLD STANDARD	SYSTEM PREDICTION
Name	I-KEY	I-KEY
Jose	B-PER	B-PER
Carlos	I-PER	I-PER
Company	O	I-KEY
GutMen	O	I-PER

Table 5-3 The system misses entities

ENTITY ANNOTATIONS	ENTITY CLASSIFICATION	
	GOLD STANDARD	SYSTEM PREDICTION
Name	I-KEY	O
Jose	B-PER	O
Carlos	I-PER	O
Company	O	O
GutMen	O	O

Table 5-4 The system assigns the wrong entity type

ENTITY ANNOTATIONS	ENTITY CLASSIFICATION	
	GOLD STANDARD	SYSTEM PREDICTION
Company	O	O
GutMen	O	O
Birthday	I-KEY	I-KEY
02/09/1990	I-BTH	I-IDN

Table 5-5 The system gets the boundaries of the bounding box wrong

ENTITY ANNOTATIONS	ENTITY CLASSIFICATION	
	GOLD STANDARD	SYSTEM PREDICTION
Name	I-KEY	I-KEY
Jose	B-PER	I-PER
Carlos	I-PER	O
Company	O	O
GutMen	O	O

Table 5-6 The system gets the boundaries and entity type wrong

ENTITY ANNOTATIONS	ENTITY CLASSIFICATION	
	GOLD STANDARD	SYSTEM PREDICTION
Name	I-KEY	B-KEY
Jose	B-PER	I-KEY
Carlos	I-PER	I-KEY
Company	O	O
GutMen	O	O

When verifying the scenarios represented in the previous tables, it can be concluded that for the Keyword detection task performed in the database 2, the only possible scenarios are those that do not include errors related to the boundaries of the bounding boxes. It is worth remembering that this task is performed first than the Semantic recognition task, therefore, in database 2 only the detection of the named entity that corresponds to the Keywords is evaluated. Moreover, since the inputs of the Semantic recognition task performed in database 3 are the ground truth BBs of all sentences and their transcriptions, there are no errors related to them; thus, the possible scenarios are the same as the Keyword detection task.

The evaluation metric used in these two tasks is the same as the proposed in the CoNLL-2003 conference proceedings (Sang & De Meulder, 2003). This

metric was presented in the CoNLL-2003 edition as the Language-Independent Named Entity Recognition task to measure the performance of the systems in terms of $Precision$, $Recall$, and $f-measure$. In this evaluation schema, the authors describe the $Precision$ as the percentage of named entities found by the system that are actually correct; and the $Recall$, as the percentage of the real number of name entities that are found by the system. Moreover, they only consider a named entity as correct if it is an exact match of the corresponding gold standard entity type. In view of the scenarios described in the tables above, this metric scheme focuses exactly on the same scenarios as the Keyword detection and Semantic recognition tasks (from Table 5-1 to Table 5-4); hence, it is a reasonable choice to use this metric scheme.

5.2.3 *Semantic Analysis Process evaluation method*

The next step is to evaluate the Semantic Analysis Process that, although it is the container of the tasks previously mentioned, uses a different evaluation metric than those referred above. The decision to use another metric scheme in this process is based on the database where it is carried out (Database 4). As shown in Figure 5-3 (the image of the left), the inputs in database 4 are the same as those in database 1 (ground truth bounding boxes of the words contained in the IDs); while the process outputs correspond to the outputs of the Semantic recognition task (information of interest extracted from an ID). Therefore, this whole process can be considered as a NERC system, and the selected evaluation metric should contemplate all the possible scenarios presented in it.

When analyzing the scenarios shown in the previous tables, it is notable that for the Semantic Analysis Process, it is necessary to consider them all. The reason why this process contemplates all the possible scenarios is based on the fact that the tasks of detecting Keywords and the information of interest no longer process the ground truth BBs of the detected sentences, but rather they directly process the result of the Sentence detection task. Accordingly, the output of the Semantic Analysis Process should also consider possible errors related to the boundaries of the bounding boxes, such as those shown in Table 5-5 and Table 5-6. Hence the CoNLL-2003 evaluation metric is not enough since it does not

include these types of errors related to the boundaries. Nevertheless, another evaluation method for NERC systems that considers all these scenarios is reported in the Fifth Message Understanding Conference (MUC-5)(Chinchor & Sundheim, 1993).

The MUC-5 evaluation scheme scores a system according to its ability to find the correct type and the exact BBs boundaries of the named entities detected (Nadeau & Sekine, 2007). This is achieved considering five different scoring categories defined as Correct (COR), Spurious (SPU), Missing (MIS), Incorrect (INC) and Partial (PAR). It should be noted that each of these categories can be directly related to a table of the possible comparison scenarios (from Table 5-1 to Table 5-6). When creating relationships between categories and tables, it is noticeable that the COR category can be represented in Table 5-1, SPU in Table 5-2 and so forth. Nonetheless, the PAR category (which describes when the system prediction and the gold standard annotation have some similarities of type matching with some overlaps between the boundaries of them, or of partial boundary matching regardless to the type of classification of the named entity), is represented in two tables (Table 5-5 and Table 5-6). Since all scenarios can be analyzed using this metric scheme, the selection of this to evaluate the Semantic Analysis Process becomes reasonable. These five scoring metrics are used to calculate two more measures, Possible (POS) that represent the number of gold standards entities, and Actual (ACT) that represent the number of named entities found by the system. Both are defined as shown in Equation 5-1.

Equation 5-1 Definition of the Possible and Actual numbers of named entities in the MUC-5 evaluation scheme. Source: (Chinchor & Sundheim, 1993)

$$POS = COR + INC + PAR + MIS = TP + FN$$

$$ACT = COR + INC + PAR + SPU = TP + FP$$

In Equation 5-1, TP, FN, and FP represent the True Positives, False Negatives, and False Positives numbers respectively. All these measures and scoring metrics are employed in the calculation of the $Precision$, $Recall$, and $f - measure$. Nonetheless, these metrics can be calculated in two different ways depending whether an exact match (Equation 5-2) or a partial match scenario (Equation 5-3) is requiring. A partial match scenario means that, although the

system cannot fully recognize a named entity annotation without errors, it can detect a part of it as shown in Table 5-5 (a part of the named entity annotation "Jose Carlos" is detected) and Table 5-6 (the named entity annotation "Jose Carlos" is detected, but it is classified erroneously and its boundary is wrong). Furthermore, partial matches are considered in the precision and recall calculation. Equation 5-4 shows the $f - measure$ evaluation protocol used for both scenarios.

Equation 5-2 MUC-5 evaluation schema for an exact match scenario. Source: (Chinchor & Sundheim, 1993)

$$Precision = \frac{COR}{ACT} = \frac{TP}{TP + FP}$$

$$Recall = \frac{COR}{POS} = \frac{TP}{TP + FN}$$

Equation 5-3 MUC-5 evaluation schema for a partial match scenario. Source: (Chinchor & Sundheim, 1993)

$$Precision = \frac{COR + (PAR * 0.5)}{ACT} = \frac{TP}{TP + FP}$$

$$Recall = \frac{COR + (PAR * 0.5)}{POS} = \frac{TP}{TP + FN}$$

Equation 5-4 MUC-5 evaluation schema $f - measure$. Source: (Chinchor & Sundheim, 1993)

$$f - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

To measure as best as possible the quality of the Semantic Analysis Process, it is proposed calculate the evaluation scheme for both scenarios (exact and partial match). It must be considered that despite the partial match scenario is not adequate (considering that an entity could be divided, and one of the requirements of this research project is to achieve the extraction of the information of interest from the IDs without any error); it allows to obtain the degree of mismatch in the non-exact predictions and to compare the quality of the system between both scenarios.

5.2.4 *End-to-end system evaluation method*

Finally, the last evaluation method analyzed is the one proposed to use in the end-to-end system. This system is evaluated in the image database 5 (Figure 5-5). Following the same order as before, first, it is analyzed the inputs of the

system and then its outputs to select the appropriate evaluation scheme. As shown in Figure 4-2, the inputs of the end-to-end system are the captured images of the IDs. Then, as explained in Chapter 4, a Text extraction method detects the bounding boxes of the words contained in the IDs and send this information to the recently mentioned Semantic Analysis Process. Therefore, the end-to-end system outputs are the same as the Semantic Analysis Process, which implies that the whole system can also be considered as a NERC system. As can be deduced, the selected evaluation method is the same as that used in the Semantic Analysis Process (from Equation 5-2 to Equation 5-4), since it considers all possible scenarios. However, it is interesting to note that the differences that may exist between the results of the end-to-end system evaluation and the evaluation of the Semantic Analysis Process are directly related to the efficiency of the Text extraction method.

At this timepoint, all proposed evaluation methods were defined for each task, process and the entire system. The next action is to use these methods to obtain the quality of the result of each task individually, verify the performance of the Semantic Analysis Process and the impact of its internal tasks on the result and, finally, obtain the efficiency of the proposed end-to-end system.

5.3 System and tasks evaluation

In this section, an analysis of the system evaluation and its internal tasks is performed using the methods described above. The main objective is to reach conclusions related to the percentages of the quality of each task and the system. In addition, given that each task is evaluated with ground truth values as inputs, the evaluation measures the real impact of each of these tasks on the system. It is worth remembering that, since the Text extraction method part of the proposed system has been previously presented (Valiente, 2018), the evaluation analysis is limited to the Semantic Analysis Process, its internal tasks and the whole system.

5.3.1 Sentence detection algorithm evaluation results

The order of evaluation is following the sequence of databases; thus, the first task to evaluate is the Sentence detection algorithm. As mentioned above,

the database used to evaluate this algorithm contains 124 images of IDs with the bounding boxes of the words detected. The total number of words in these documents is 5834, and these form a total of 2481 sentences. Considering the evaluation method selected for this task, the performance of the system is detailed in Table 5-7. This table shows the total number of sentences detected (ACT) by this task, those that are considered correct (COR) based on an overlap ratio higher than a threshold of 0.8 and the ground truth number of sentences (POS).

Table 5-7 Sentence detection algorithm evaluation results

METRICS	NUMBER OF SENTENCES DETECTED
ACT	2443
COR	2370
POS	2481
$Precision$	$\mathbf{97.01\%}$
$Recall$	$\mathbf{95.53\%}$
$f-measure$	$\mathbf{96.26\%}$

As Table 5-7 shows, the system achieves for the Sentence detection algorithm in database 1 a precision of 97.01% and a recall of 95.53%. The following figure illustrates some examples to analyze and understand why some errors occur.

(a) (b)

Figure 5-6 Examples of outputs of the Sentences detection task that contains errors when defining the sentences bounding boxes. Source: The Author.

As shown in the Figure 5-6, almost all the errors take place in specific types of ID and consist of joining two sentences in one. These errors occur since the Sentence detection algorithm considers that two information corresponds to the same sentence based on their geometrical similarities and the distances

between them. Thus, for instance, words that refer to the CPF number and date of birth (Figure 5-6(a)), or those that contain the RNE number and residence classification (Figure 5-6(b)) are considered too close and with similar characteristics, hence,two data that really corresponds to two sentences, join in one. Therefore, this type of error takes place in those documents that contain their data very close to each other.

The quality of Sentence detection algorithm, quantified as 96.26%, can be considered as satisfactory. As mentioned, the $Precision$ of the system and the COR (correct number of sentences detected) is directly affected by the incorrect detection of the bounding boxes. Nevertheless, these incorrect sentences detections are not a big problem for the next named entity recognition tasks, since they use regular expressions to detect the Keywords and the information of interest. The use of regular expressions allows detecting the information sought even being this, part of a longer sentence. This is possible since the regular expression can captures tokens. Tokens are the portions of sentences that contain the matched text.

5.3.2 *Keyword detection and Semantic recognition tasks evaluation results*

Although both are NERC tasks, these two tasks are analyzed separately, since different databases are used in each one. The first analyzed is the *Keyword* detection task in database 2. As detailed in section 5.1, database 2 contains 25690 sentences composed of *Keywords*, information of interest according to their *keywords* and other non-relevant data (Figure 5-4). Table 5-8 shows the results of the evaluation of this task in database 2 in terms of $Precision$, $Recall$, and $f-measure$. This table also shows the number of named entities of type **KEY** this task is capable of detecting (ACT), those that are considered as correctly detected by an exact match with the corresponding gold standard entity type (COR), and the number of gold standards entities that exist (POS).

Table 5-8 Keyword detection task evaluation results

METRICS	NUMBER OF KEYWORDS DETECTED
ACT	5508
COR	5505
POS	5505
$Precision$	$\mathbf{99.94}\%$
$Recall$	$\mathbf{100}\%$
$f-measure$	$\mathbf{99.96}\%$

As it is expected, this task can recognize all the **KEY** entities in database 2. Notwithstanding, some errors related to false positives appears. These three errors are produced by similarities between the person names and the regular expressions used to detect the different categories of **KEY** entities. For example, in this database, the name "Numa" is interpreted as the regular expression used for the Keyword structure category of “person name keyword”, while the company names “Plan Registro” and “Connecting Numbers” are interpreted as Keyword entities of the category “id number keyword”. However, if these errors occur in the system, the following task is able to correct them since check if the info referred by them is in the appropriate format, in cases of “id number keyword” category, if the possible information of interest is a number. Therefore, since these errors are not really **KEY** entities, the task of Semantic recognition would not find any information suggested by them and would omit them.

The next task to evaluate is Semantic recognition. Unlike the previous task, this not only processes a text but also considers the geometric properties of the detected text to establish relationships between **KEY** entities and information referred by them. Its performance is evaluated in database, which, similar to database 1, consists of 124 images. However, in this database, the bounding boxes are those of all sentences and Keywords detected and their transcriptions. The evaluation of this task is based on its performance detecting the information of interest. Therefore, its result only reflects the **PER**, **IDN**, and **BTH** entities detection and do not consider all the **KEY** entities detection count (only those **KEY** entities that classify the documents (Figure 4-13)). That decision relies on

KEY entities without information related are omitted as mentioned above, and furthermore, the objective is to obtain the quality of the system to find the information required.

Alike the previous evaluation results analysis, the real value (POS), the detected by the system (ACT), and the correct number of named entities detected (COR) are shown in Table 5-9.

Table 5-9 Semantic recognition task evaluation results

METRICS	NUMBER OF INFORMATION OF INTEREST DETECTED
ACT	350
COR	337
POS	372
$Precision$	$\mathbf{96.29}\%$
$Recall$	$\mathbf{90.59}\%$
$f-measure$	$\mathbf{93.35}\%$

As shown in, the total number of information of interest is 372. When using database 3, this task can detect 350 entities and of these 337 are classified as a correct detection. When analyzing the drop rate of the task performance and verifying the images where these errors occur, it is possible to realize why it is remarkable the difference between the POS, ACT and COR values and why the task failed in the detection of some named entities (Figure 5-7).

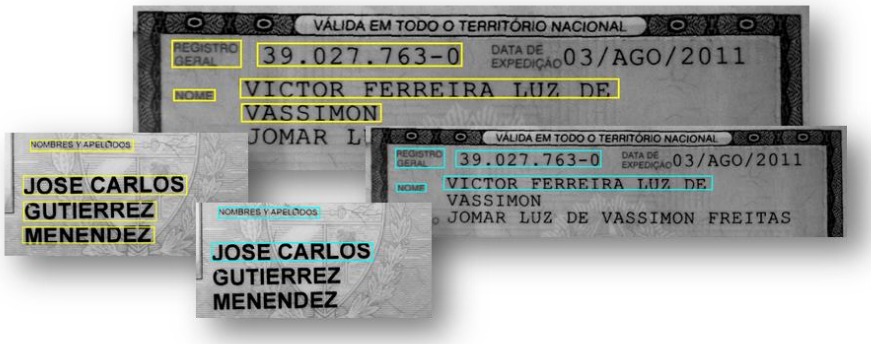

Figure 5-7 Examples of images with gold standards entities (yellow bounding boxes) where the task failed to detect named entities (cyan bounding boxes). Source: The Author

As Figure 5-7 shows, some entities are divided into several sentences in different horizontal lines (yellow bounding boxes). Thus, when the task tries to

find the information referred by the Keywords, only the upper line is selected as the named entity (cyan bounding boxes), introducing an error in the detection of the number of sentences that are defined as entities and, consequently, in the number of information of interest detected. In database 3, a total of nine and four images contain the name divided into three and two parts respectively. This conduces to decrease the number of information of interest that the task detects from 35 (POS) to 13(ACT). Furthermore, when analyzing this ACT value, it is considered incorrect since the evaluation scheme used does not include partial matches.

In summary, both tasks can be considered suitable. The performance of Keyword detection task is satisfactory after testing in database 2. Only three errors related to false positives were detected. Moreover, these false positives are not a big problem, since they would be omitted when the system tries to find the information related to them and the format of these does not match. On the other hand, the Semantic recognition task faced other kinds of problems. Database 3 is composed of different layouts of identity documents, and some of them contain the information of interest divided into several sentences. The rules created to assign semantic and find the named entities works correctly when the information of interest is contained in a sentence. In other cases, the information may not be fully detected and, according to the evaluation scheme used in this task, it would be considered incorrect if it does not exactly match the gold standard. Discarding these cases where the information is divided into several sentences, the performance of this task shows the high effectivity of the rules-based system proposed in this research to detect named entities.

5.3.3 *Semantic Analysis Process evaluation results*

In this section, the Semantic Analysis Process is evaluated as a whole. Its performance depends directly on how the system detects sentences since they are the inputs of the Semantic recognition task. Database 4 that is used to evaluate this process contains the same data as database 1. Nevertheless, since different processes and evaluation scheme are used in both, it was decided to process them as independent databases. The evaluation scheme used in

database 4 considers the errors obtained in the Semantic recognition task, those related to the information of interest divided into several sentences, as partial solutions.

The Semantic Analysis Process uses the ground truth bounding boxes of the words as inputs to obtain the information of interest of the IDs. As mentioned in section 5.2.3, the evaluation scheme considers five different scoring categories (COR, SPU, MIS, INC and PAR) and use them to calculate the possible (POS) number of gold standards entities, and the actual (ACT) number the system detects. Table 5-10 shows the results of the evaluation considering both match scenarios.

Table 5-10 Semantic Analysis Process evaluation results. (IoI is the acronym for information of interest)

METRICS	NUMBER OF IoI DETECTED	METRICS	NUMBER OF IoI DETECTED	
MIS	12	ACT	367	
SPU	7	POS	372	
COR	324	MATCH SCENARIO	PARTIAL	EXACT
		$Precision$	$\mathbf{92.09}\%$	$\mathbf{88.28}\%$
INC	8	$Recall$	$\mathbf{90.86}\%$	$\mathbf{87.09}\%$
PAR	28	$f-measure$	$\mathbf{91.47}\%$	$\mathbf{87.68}\%$

As can be noted, the results show a drop in the quality of the system compared to the previous results. This fall is mainly due to errors during the detection of information of interest. As mentioned, some identity documents contain keywords and their information composed of more than one sentence. Moreover, when detecting keywords inside the sentences, in some cases, the semantic is attributed to incorrect information (Figure 5-8(a)) or miss the detection of entities due to format mismatch (Figure 5-8(b)).

(a) (b)

Figure 5-8 Examples of an incorrect semantic attribution. Cyan arrows point to the information of interest expected. Notwithstanding, the semantics is attributed to the incorrect sentences (to which the red arrow points), or the system can lose an entity due to format mismatch.

Regarding the Sentence detection algorithm, it is notable that the errors found in analyzing its individual performance do not seriously affect the recognition of the named entities. This is because the information of interest such as the ID number and the date of birth have specifics formats and, taking advantage of this, they can be detected and extracted at token-level, even if they are mixed with other information (Figure 5-6). Nevertheless, since this type of numerical information can be found in several formats, some errors can occur at the moment of assign the respective semantics, for instance, a date can be understood as an ID number. On the other hand, detection and extraction at token-level cannot be used in the case of the name of a person entity since there is no format or pattern in a name. Thus, cases in which a name is part of a larger sentence count as a partial detection as well as the aforementioned cases of names divided into several sentences.

Finally, when verifying the last metrics, this process evaluated in database 4 reaches 91.47% of $f - measure$ (Equation 5-4) in a partial match scenario (Equation 5-3). Even though this is a scenario not ideal for the objectives of the system proposed in this research, it is interesting to take it into account, since most of the errors are due to the fact the information is contained in more than one sentence while the system only considers the first line of the possible entity. A really interesting aspect to consider for future work could be to develop a process to detect all the sentences that are part of the same information. Thus, the system would be able to recognize a named entity divided into several sentences and this would significantly improve the system performance.

5.3.4 End-to-end evaluation results

The final analysis corresponds to the evaluation of the performance of the end-to-end system obtained from the methodology proposed in this research. The term end-to-end used to describe the proposed system is used because it covers every stage of the information of interest extraction process from IDs images (from the image processing to named entity classification). Consequently, the performance of the Text extraction method to correctly detect all the words and their bounding boxes to be used in the Semantic Analysis process is crucial.

As shown in Figure 5-5, database 5 is where the system is evaluated and consist of captured real scenes that focus explicitly on the textual information contained in the identity documents. These images are captured in variable conditions of illumination and orientation. The 100 images that make up the database contain a total of 1100 words, of these, the ground truth value that corresponds to information of interest is 315, worth remembering that those *keywords* that classify the documents are also considered as information of interest (Figure 4-13).

The evaluation scheme used in the system is the same as that employed during the evaluation of the Semantic Analysis Process; therefore, the results are defined with the same metrics. As shown in Table 5-11, the results are very satisfactory since the system achieves a $Precision$ of 95.29% and an $f-measure$ of 92.48% even in an exact match scenario (Equation 5-2). The number of partial detections is decreased respect to the previous since only 5 images of this database have the information of interest divided into more than one sentence. Notwithstanding, the increase in the number of miss detection is notable.

Table 5-11 End-to-end system evaluation results. (IoI is the acronym for information of interest)

METRICS	NUMBER OF IoI DETECTED	METRICS	NUMBER OF IoI DETECTED	
MIS	24	ACT	297	
SPU	6	POS	315	
COR	283	MATCH SCENARIO	PARTIAL	EXACT
		$Precision$	$\mathbf{96.63\%}$	$\mathbf{95.29\%}$
INC	0	$Recall$	$\mathbf{91.11\%}$	$\mathbf{89.84\%}$
PAR	8	$f-measure$	$\mathbf{93.79\%}$	$\mathbf{92.48\%}$

The increase in the number of miss detection is due to errors during the Text extraction method. Figure 5-9 shows an example of keywords (“REGISTRO GERAL” and “NOME”) that cannot be detected because the Text extraction method misses them since they barely contrast with the background of the document.

Figure 5-9 Example of an image of the database 5 where some keywords are not detected, so that the information of interest cannot be extracted. Source: The Author.

Summarizing, the results of the system evaluation show that there is still work to be done to improve the text extraction and semantic analysis process. Improving the system capacity for detecting text in low contrast with the background and also group all the sentences related to the same information to be considered as one, are some of the tasks that must be implemented to increase the quality of the system. Nevertheless, the system achieves with satisfactory results the objective proposed in this research.

5.4 Final Discussion about the evaluation results

During this chapter, the individual analysis of the tasks that make up the proposed system was carried out. Using ground truth inputs in each of these tasks, it is possible to analyze their individual performances and detect the roots of the problems that affect the final output of the system.

Furthermore, given that the system is composed of tasks with different objectives, specific databases were created for each function. Likewise, a detailed analysis of the most appropriate evaluation methods for each task was accomplished, studying all possible cases of solutions. The time-consuming of the end-to-end system is relative to the size of the image and the amount of information or background complexity of the ID. However, for the machine specifications mentioned before (section 5), the average time-consuming by the entire system is 70 seconds, and of these, an average of 18 seconds corresponds to the Semantic Analysis Process.

As can be noticed, in this chapter the proposed system is not compared with any other system. The reason is due to the fact that no work similar to the proposed end-to-end system was found during the review of the literature. As aforementioned, the use of systems capable of interpreting the information extracted from Identity Documents images is of great interest for the private sector; thus, these systems are analyzed mainly in Patents compared to public publications (ABBYY, 2019; GB Group Plc, 2019; GmbH Anyline, 2019; ICAR VISION SYSTEMS, 2019; US 10,084,606 B2, 2018; Microblink, 2019). Furthermore, since one of the particular characteristics of this system is the interpretation of textual information from ID images without prior knowledge of their layouts, comparing their performance with another system based on templates is not considered fair. This argument is based on the fact that the template-based systems have a conditioned performance once the interpretation of the information of interest is previously defined in the document model. Moreover, although the proposed system performs a NERC process, it is also not fair to compare it with other NERC systems since the extracted textual information is domain specific and does not correspond to a corpus. This type of information format is different from those processed by the NERC systems described in the

literature. Therefore, the performance of the proposed system is better compared to other algorithms such as Stanford NER (Finkel, Grenager, & Manning, 2007) or spaCy (Explosion AI, 2018) that only find the names of people. That occurs, since the proposed system is conditioned to the domain and is based on specific rules that detect among the extracted textual information, all *keywords* to subsequently attribute semantics to the information. Although the proposed system is not compared with a similar one due to the previous arguments, it shows a satisfactory performance in each of the evaluated tasks and as an end-to-end system.

Upon reaching this point, it is concluded that the proposed end-to-end system demonstrated an adequate capacity to capture the information of interest in identification documents. However, some conditions could be considered to obtain the best performance of this proposed system. These conditions establish that the IDs must contain their textual information in high contrast with respect to the background of the document to detect all the words. Moreover, it must be possible to encapsulate each information contained in IDs in individual sentences, avoiding this that the same information appears divided into different sentences.

Chapter 6

6 Conclusions

In this research is proposed the implementation of an automated system able to extract and interpret the textual information from identity documents images. The proposed system, different to the template-based systems, uses a semantic attribution algorithm that allows to classify and attribute meaning to the information from IDs according to its semantics. The algorithm is developed to provides a high scalability of the system, since the results of the system are stored as an abstraction of their meanings, this means each information about the citizen is classified by its semantic value.

The system obtained through this research is considered as an end-to-end system since it covers every stage of the information of interest extraction process from IDs images. Although the Text extraction method process of the system has been previously presented (Valiente, 2018), this research is the first comprehensive description of a complete information extraction system to process IDs that describes from image processing to named entity recognition.

The proposed end-to-end system allows the automation of a registration or verification process that requires the acquisition of information about a citizen using his identity documents. To achieve this objective, an extensive study was carried out on the state of the art of text extraction systems and natural language processing. Related to this latter, the research of the named entities recognition systems was of great importance. All the studies carried out on the state of the art helped to select the appropriate methodology capable of obtaining the required results.

To evaluate the performance of the research were proposed different metrics based on the internal functions of the system. Moreover, different databases were created to individually evaluate each of these functions. These evaluations make it possible to detect to what extent the performance of each of the tasks influences the system in general. The final evaluation of the system shows satisfactory results. Nevertheless, some improvements related to Text

recognition process and sentence detection could be implemented in order to increase system performance.

In this research proposal, some collateral contributions have already been reached. Three papers related to the proposed system and the partial results of this have been published. Two of these are related to the methods of text localization and recognition, where both have been implemented in MATLAB to analyze the viability and performance of these. On the other hand, the algorithm of semantic attribution also counts with a publication that shows some satisfactory results obtained during the development of this research.

As future works, it can be considered to implement an improvement in the text extraction process. The idea would be to make the system capable of increasing the word detection index, detecting those words that hardly contrast with the background of the document. Moreover, some adjustments in the semantic analysis process could also be made to improve system performance. These adjustments would consist of allowing the system to detect if several sentences belong to the same information. In this way, if the same information appears divided into different sentences, the system would consider all these sentences as the same information of interest. Another interesting contribution to this research could be the automatic creation of the template based on the features of the first-processed document of a specific type since the location of its regions of interest is known. Thereby, the system can take advantage of the processing speed provided by the template-based methods to extract the information. Then, when another document of the same type is processed, the use of templates to extract the information from the ID guarantees fast processing since the location of this one is previously known.

Finally, it can be affirmed and concluded that the proposed end-to-end system achieves the research objective. It shows that is capable of extracting and interpreting textual information from identity documents images without prior knowledge of their layouts, reaching a $Precision$ of 95.29% and an $f-measure$ of 92.48%.

7 REFERENCES

ABBYY. (2019). OCR, ICR, OMR e Software linguístico - ABBYY Software. Retrieved January 16, 2019, from https://www.abbyy.com/pt-br/

Arasu, A., & Garcia-Molina, H. (2003). Extracting structured data from web pages. *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 337–348. https://doi.org/10.1145/872757.872799

Aroyo, L., & Welty, C. (2015). Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, *36*(1), 15. https://doi.org/10.1609/aimag.v36i1.2564

Bai, B., Yin, F., & Liu, C. L. (2013). Scene Text Localization Using Gradient Local Correlation. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 1380–1384). IEEE. https://doi.org/10.1109/ICDAR.2013.279

Bailey, D. G. (2011). *Design for Embedded Image Processing on FPGAs*. *Design for Embedded Image Processing on FPGAs*. Singapore: John Wiley & Sons (Asia) Pte Ltd. https://doi.org/10.1002/9780470828519

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded up robust features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3951 LNCS, pp. 404–417). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11744023_32

Bimbo, A. Del. (2011). Region detectors Requirements for region detection. *MSER and Region Dettector*, 10. Retrieved from http://www.micc.unifi.it/delbimbo/wp-content/uploads/2011/03/slide_corso/A34 MSER.pdf

Blumberg, R., & Atre, S. (2003). The Problem with Unstructured Data. *DM Review*, *13*, 42. Retrieved from www.dmreview.com

Braccini, C., DeFloriani, L., & Vernazza, G. (1995). *Image Analysis and*

Processing : 8th International Conference, ICIAP'95 San Remo, Italy, September 13-15, 1995 Proceedings. Springer-Verlag.

Carreras, X., Màrquez, L., & Padró, L. (2003). A simple named entity extractor using AdaBoost. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* -, *4*, 152–155. https://doi.org/10.3115/1119176.1119197

Chabchoub, F., Kessentini, Y., Kanoun, S., Eglin, V., & Lebourgeois, F. (2016). SmartATID: A Mobile Captured Arabic Text Images Dataset for Multi-purpose Recognition Tasks. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 120–125). IEEE. https://doi.org/10.1109/ICFHR.2016.0034

Cheriet, M., Kharma, N., Liu, C.-L., & Suen, C. Y. (2007). *Character Recognition Systems*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470176535

Chinchor, N., & Sundheim, B. (1993). MUC-5 evaluation metrics. In *Proceedings of the 5th conference on Message understanding - MUC5 '93* (p. 69). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/1072017.1072026

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010). Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (October), 1002–1012. Retrieved from https://dl.acm.org/citation.cfm?id=1870756

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Cisco. (2017). The Zettabyte Era: Trends and Analysis. Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf

Clarke, C. L. A., & Cormack, G. V. (1997). On the use of regular expressions for

searching text. *ACM Transactions on Programming Languages and Systems*, *19*(3), 413–426. https://doi.org/10.1145/256167.256174

Corban Works, L. (2018). Generate a Random Name - Fake Name Generator. Retrieved February 7, 2019, from https://www.fakenamegenerator.com/

Crist, G. (2011). *Optical and Digital Image Processing*. (G. Cristobal, P. Schelkens, & H. Thienpont, Eds.). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527635245

Cyganek, B. (2013). *Object Detection and Recognition in Digital Images*. Oxford, UK: John Wiley & Sons Ltd. https://doi.org/10.1002/9781118618387

de las Heras, L.-P., Terrades, O. R., Llados, J., Fernandez-Mota, D., & Canero, C. (2015). Use case visual Bag-of-Words techniques for camera based identity document classification. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 721–725). IEEE. https://doi.org/10.1109/ICDAR.2015.7333856

Doermann, D., & Tombre, K. (2014). *Handbook of Document Image Processing and Recognition*. (D. Doermann & K. Tombre, Eds.), *Handbook of Document Image Processing and Recognition*. London: Springer London. https://doi.org/10.1007/978-0-85729-859-1

Donoser, M., Riemenschneider, H., & Bischof, H. (2010). Shape guided Maximally Stable Extremal Region (MSER) tracking. In *Proceedings - International Conference on Pattern Recognition* (Vol. 1, pp. 1800–1803). IEEE. https://doi.org/10.1109/ICPR.2010.444

Downton, A., & Crookes, D. (1998). Parallel architectures for image processing. *Electronics & Communication Engineering Journal*, *10*(3), 139–151. https://doi.org/10.1049/ecej:19980307

Dudhabaware, R. S., & Madankar, M. S. (2014). Review on natural language processing tasks for text documents. In *2014 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1–5). IEEE. https://doi.org/10.1109/ICCIC.2014.7238427

Epshtein, B., Ofek, E., & Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2963–2970). IEEE. https://doi.org/10.1109/CVPR.2010.5540041

Explosion AI. (2018). spaCy · Industrial-strength Natural Language Processing in Python. Retrieved March 13, 2019, from https://spacy.io/

Finkel, J. R., Grenager, T., & Manning, C. (2007). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, (1995), 363–370. https://doi.org/10.3115/1219840.1219885

Fisher, R. B., Breckon, T. P., Dawson-Howe, K., Fitzgibbon, A., Robertson, C., Trucco, E., & Williams, C. K. I. (2014). *Dictionary of Computer Vision and Image Processing*. (R. B. Fisher, K. Dawson-Howe, A. Fitzgibbon, C. Robertson, & E. Trucco, Eds.). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470016302

Fornes, A., Romero, V., Baro, A., Toledo, J. I., Sanchez, J. A., Vidal, E., & Llados, J. (2018). ICDAR2017 Competition on Information Extraction in Historical Handwritten Records. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (Vol. 1, pp. 1389–1394). IEEE. https://doi.org/10.1109/ICDAR.2017.227

GB Group Plc. (2019). IDscan | Scan ID documents and make informed decisions | GBG. Retrieved January 16, 2019, from https://www.gbgplc.com/id-scanner/

Gibson, E. (1991). The Alphabet and the Brain - the Lateralization of Writing - Dekerckhove,P, Lumsden,Cj. *Australian Psychologist*, *26*(1), 76.

GmbH Anyline. (2019). Mobile ID Scanning with your Smartphone | ANYLINE.com. Retrieved January 16, 2019, from https://anyline.com/products/scan-id/

Gomez, L., & Karatzas, D. (2013). Multi-script Text Extraction from Natural

Scenes. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 467–471). IEEE. https://doi.org/10.1109/ICDAR.2013.100

Gomez, R., Shi, B., Gomez, L., Numann, L., Veit, A., Matas, J., … Karatzas, D. (2018). ICDAR2017 Robust Reading Challenge on COCO-Text. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (Vol. 1, pp. 1435–1443). IEEE. https://doi.org/10.1109/ICDAR.2017.234

Gonzalez, R. C., & Woods, R. E. (2008). *Digital Image Processing Third Edition*. *Pearson Prentice Hal* (Pearson In, Vol. 56). Pearson Prentice Hall. Retrieved from http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/Digital_Image_Processing_3rdEd_truncated.pdf

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6. In *Proceedings of the 16th conference on Computational linguistics* - (Vol. 1, p. 466). https://doi.org/10.3115/992628.992709

He, W., Zhang, X. Y., Yin, F., & Liu, C. L. (2018). Multi-Oriented and Multi-Lingual Scene Text Detection with Direct Regression. *IEEE Transactions on Image Processing*, *27*(11), 5406–5419. https://doi.org/10.1109/TIP.2018.2855399

Hornberg, A. (2007). *Handbook of Machine Vision*. *Handbook of Machine Vision*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527610136

Huang, W., Lin, Z., Yang, J., & Wang, J. (2013). Text localization in natural images using stroke feature transform and text covariance descriptors. *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2013.157

Huang, W., Qiao, Y., & Tang, X. (2014). Robust scene text detection with convolution neural network induced MSER trees. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8692 LNCS, pp. 497–511). Springer, Cham. https://doi.org/10.1007/978-3-319-10593-2_33

ICAR VISION SYSTEMS, S. . (2019). Passport and ID card scanning for data registration purposes | IDFast. Retrieved January 16, 2019, from https://www.icarvision.com/en/id_fast

ICDAR 2017 RobustReading Competition. (n.d.). Retrieved October 11, 2017, from http://rrc.cvc.uab.es/

Illingworth, J., & Kittler, J. (1988). A survey of the hough transform. *Computer Vision, Graphics and Image Processing*, *44*(1), 87–116. https://doi.org/10.1016/S0734-189X(88)80033-1

Islam, M. R., Mondal, C., Azam, M. K., & Islam, A. S. M. J. (2016). Text detection and recognition using enhanced MSER detection and a novel OCR technique. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 15–20). IEEE. https://doi.org/10.1109/ICIEV.2016.7760054

Iwamura, M., Morimoto, N., Tainaka, K., Bazazian, D., Gomez, L., & Karatzas, D. (2017). ICDAR2017 Robust Reading Challenge on Omnidirectional Video. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1448–1453). IEEE. https://doi.org/10.1109/ICDAR.2017.236

Jacobs, C., Simard, P. Y., Viola, P., & Rinker, J. (2005). Text recognition of low-resolution document images. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (p. 695–699 Vol. 2). IEEE. https://doi.org/10.1109/ICDAR.2005.233

Jaderberg, M., Vedaldi, A., & Zisserman, A. (2014). Deep Features for Text Spotting (pp. 512–528). Springer, Cham. https://doi.org/10.1007/978-3-319-10593-2_34

Jiang, R., Banchs, R. E., & Li, H. (2016). *Evaluating and Combining Named Entity Recognition Systems*. Retrieved from https://spacy.io/

Jr., W. F. . A., Gisolfi, D. A. ., Johnson, A. C. ., & Reed, A. K. . (2018). *US 10,084,606 B2*. United States. Retrieved from https://patents.google.com/patent/US10084606B2/en

Kai Wang, Babenko, B., & Belongie, S. (2011). End-to-end scene text recognition. In *2011 International Conference on Computer Vision* (pp. 1457–1464). IEEE. https://doi.org/10.1109/ICCV.2011.6126402

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., … Valveny, E. (2015). ICDAR 2015 competition on Robust Reading. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (Vol. 2015–Novem, pp. 1156–1160). IEEE. https://doi.org/10.1109/ICDAR.2015.7333942

Karatzas, D., Gomez, L., Nicolaou, A., & Rusinol, M. (2018). The robust reading competition annotation and evaluation platform. In *Proceedings - 13th IAPR International Workshop on Document Analysis Systems, DAS 2018* (pp. 61–66). IEEE. https://doi.org/10.1109/DAS.2018.22

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L. G. I., Mestre, S. R., … De Las Heras, L. P. (2013). ICDAR 2013 robust reading competition. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (pp. 1484–1493). IEEE. https://doi.org/10.1109/ICDAR.2013.221

Kaur, T. P., & Garg, N. (2015). Optimized Gurmukhi Text Recognition from Signboard Images Captured by Mobile Camera Using Structural Features. In *2015 Fifth International Conference on Advances in Computing and Communications (ICACC)* (pp. 412–416). IEEE. https://doi.org/10.1109/ICACC.2015.65

Kok Loo, P., & Lim Tan, C. (2002). Word and Sentence Extraction Using Irregular Pyramid. *LNCS*, *2423*, 307–318.

Lee, J., Kang, I., Kim, J., & Rim, H. (2014). Apparatus and method for detecting named entity. United States of America. Retrieved from https://patents.google.com/patent/US8655646B2/en

Liao, M., Shi, B., & Bai, X. (2018). TextBoxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Transactions on Image Processing*, *27*(8), 3676–3690. https://doi.org/10.1109/TIP.2018.2825107

Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., & Yan, J. (2018). FOTS: Fast Oriented Text Spotting with a Unified Network. https://doi.org/10.1109/CVPR.2018.00595

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing Named Entities in Tweets. *Computational Linguistics*, (2008), 359–367. https://doi.org/10.5121/csit.2015.50213

Lyons, J. (1995). *Linguistic Semantics: An Introduction*. Cambridge University Press. Retrieved from https://books.google.com.br/books/about/Linguistic_Semantics.html?id=Na2g1ltaKuAC&redir_esc=y

Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., & Xue, X. (2018, March 3). Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2018.2818020

Mandal, R., Roy, P. P., Palz, U., & Blumenstein, M. (2015). Date field extraction from handwritten documents using HMMs. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 866–870). IEEE. https://doi.org/10.1109/ICDAR.2015.7333885

Marques, O. (2011). *Practical Image and Video Processing Using MATLAB®*. *Practical Image and Video Processing Using MATLAB®*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118093467

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. In *Image and Vision Computing* (Vol. 22, pp. 761–767). https://doi.org/10.1016/j.imavis.2004.02.006

McCallum, A. (2005). Information extraction: Distilling Structured Data from Unstructured Text. *Queue - Social Computing*, *3*(9), 48–57. https://doi.org/10.1145/1105664.1105679

Microblink. (2019). BlinkID | Real-time ID scanning for any app. Retrieved January 16, 2019, from https://microblink.com/products/blinkid

Mishra, N., Patvardhan, C., Vasantha Lakshmi, C., & Singh, S. (2012).

Shirorekha Chopping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition. *International Journal of Computer Applications*, *39*(6), 19–23. https://doi.org/10.5120/4824-7076

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26. https://doi.org/10.1075/li.30.1.03nad

Nagesh, A., Ramakrishnan, G., Chiticariu, L., Krishnamurthy, R., Dharkar, A., & Bhattacharyya, P. (2012). Towards Efficient Named-Entity Rule Induction for Customizability. In *EMNLP* (pp. 128–138). Association for Computational Linguistics. https://doi.org/10.1212/WNL.0000000000000591

Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., … Ogier, J. M. (2018). ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (Vol. 1, pp. 1454–1459). IEEE. https://doi.org/10.1109/ICDAR.2017.237

Neumann, L., & Matas, J. (2011). A method for text localization and recognition in real-world images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6494 LNCS, pp. 770–783). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19318-7_60

Neumann, L., & Matas, J. (2011). Text Localization in Real-World Images Using Efficiently Pruned Exhaustive Search. In *2011 International Conference on Document Analysis and Recognition* (pp. 687–691). IEEE. https://doi.org/10.1109/ICDAR.2011.144

Neumann, L., & Matas, J. (2012). Real-time scene text localization and recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3538–3545). IEEE. https://doi.org/10.1109/CVPR.2012.6248097

Neumann, L., & Matas, J. (2013a). On Combining Multiple Segmentations in Scene Text Recognition. In *2013 12th International Conference on*

Document Analysis and Recognition (pp. 523–527). IEEE. https://doi.org/10.1109/ICDAR.2013.110

Neumann, L., & Matas, J. (2013b). Scene Text Localization and Recognition with Oriented Stroke Detection. In *2013 IEEE International Conference on Computer Vision* (pp. 97–104). IEEE. https://doi.org/10.1109/ICCV.2013.19

Neumann, L., & Matas, J. (2016). Real-Time Lexicon-Free Scene Text Localization and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(9), 1872–1885. https://doi.org/10.1109/TPAMI.2015.2496234

Number, D. (2016). *Image Processing*. *Image Processing*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.5772/122

P. J. Burt, & E. H. Adelson. (1983). The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications*, *31*(4), 532–540. https://doi.org/0090-6778/83/0400-05

Patel, C., Patel, A., & Patel, D. (2012). Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study. *International Journal of Computer Applications*, *55*(10), 50–56. https://doi.org/10.5120/8794-2784

Peanho, C. A., Stagni, H., & da Silva, F. S. C. (2012). Semantic information extraction from images of complex documents. *Applied Intelligence*, *37*(4), 543–557. https://doi.org/10.1007/s10489-012-0348-x

Petrou, M., & Petrou, C. (2011). *Image Processing: The Fundamentals: Second Edition*. *Image Processing: The Fundamentals: Second Edition*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119994398

Phan, T. Q., Shivakumara, P., & Tan, C. L. (2012). Detecting text in the real world. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12* (p. 765). New York, New York, USA: ACM Press. https://doi.org/10.1145/2393347.2396307

Ramshaw, L. A., & Marcus, M. P. (1995). Text Chunking using Transformation-Based Learning. In *Proceedings of the Third ACL Work- shop on Very Large*

Corpora (pp. 82–94). Cambridge, MA, USA. Retrieved from https://www.aclweb.org/anthology/W95-0107

Ratha, N. K., & Jain, A. K. (1999). Computer vision algorithms on reconfigurable logic arrays. *IEEE Transactions on Parallel and Distributed Systems*, *10*(1), 29–43. https://doi.org/10.1109/71.744833

Ryan, M., & Hanafiah, N. (2015). An Examination of Character Recognition on ID card using Template Matching Approach. *Procedia Computer Science*, *59*, 520–529. https://doi.org/10.1016/j.procs.2015.07.534

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. https://doi.org/10.3115/1119176.1119195

Santos, C. N. dos, & Guimarães, V. (2015). Boosting Named Entity Recognition with Neural Character Embeddings. https://doi.org/10.18653/v1/W15-3904

Sarfraz, M. (2005). *Computer-Aided Intelligent Recognition Techniques and Applications*. (M. Sarfraz, Ed.), *Computer-Aided Intelligent Recognition Techniques and Applications*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/0470094168

Shahab, A., Shafait, F., & Dengel, A. (2011). ICDAR 2011 robust reading competition challenge 2: Reading text in scene images. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (pp. 1491–1496). IEEE. https://doi.org/10.1109/ICDAR.2011.296

Sharma, P., & Sharma, S. (2016). Image processing based degraded camera captured document enhancement for improved OCR accuracy. In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)* (pp. 441–444). IEEE. https://doi.org/10.1109/CONFLUENCE.2016.7508160

Shi, C., Wang, C., Xiao, B., Zhang, Y., & Gao, S. (2013). Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters*, *34*(2), 107–116. https://doi.org/10.1016/j.patrec.2012.09.019

Shih, F. Y. (2010). *Image Processing and Pattern Recognition: Fundamentals and Techniques*. *Image Processing and Pattern Recognition: Fundamentals and Techniques*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470590416

Simon, M., Rodner, E., & Denzler, J. (2015). Fine-grained classification of identity document types with only one example. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)* (pp. 126–129). IEEE. https://doi.org/10.1109/MVA.2015.7153149

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2* (pp. 629–633). IEEE. https://doi.org/10.1109/ICDAR.2007.4376991

Smith, R., Gu, C., Lee, D. S., Hu, H., Unnikrishnan, R., Ibarz, J., … Lin, S. (2016). End-to-end interpretation of the French street name signs dataset. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9913 LNCS, pp. 411–426). Springer, Cham. https://doi.org/10.1007/978-3-319-46604-0_30

Uchida, S. (2014). Text Localization and Recognition in Images and Video. In *Handbook of Document Image Processing and Recognition* (pp. 843–883). London: Springer London. https://doi.org/10.1007/978-0-85729-859-1_28

Valiente, R. (2018). *Processo automático de reconhecimento de texto em imagens de documentos de identificação genéricos.* Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, São Paulo. https://doi.org/10.11606/D.3.2018.tde-05032018-151842

Valiente, R., Gutiérrez, J. C., Sadaike, M. T., & Bressan, G. (2017). Automatic Text Recognition in Web Images. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web - WebMedia '17* (pp. 241–244). New York, New York, USA: ACM Press. https://doi.org/10.1145/3126858.3131570

Valiente, R., Sadaike, M. T., Gutiérrez, J. C., Soriano, D. F., & Bressan, G. (2016).

A process for text recognition of generic identification documents over cloud computing. *IPCV'1International Conference on Image Processing, Computer Vision, and Pattern Recognition*, (April 2017), 4. Retrieved from https://search.proquest.com/openview/1d6c43922b81d4fb2a3e1675c8b3764e/1?pq-origsite=gscholar&cbl=1976345

Wang, T., Wu, D. J., Coates, A., & Y. Ng, A. (2012). End-to-end text recognition with convolutional neural networks. In *21st International Conference on Pattern Recognition (ICPR), 2012 11 - 15 Nov. 2012, Tsukuba International Congress Center, Tsukuba Science City, Japan* (pp. 3304–3308). Tsukuba, Japan: IEEE.

Wolf, C., & Jolion, J.-M. (2006). Object count / Area Graphs for the Evaluation of Object Detection and Segmentation Algorithms. *ICDAR, International Conference on Document Analysis and Recognition*, *8*(4), 280–296. https://doi.org/10.1007/s10032-006-0014-0

Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, & Hong-Wei Hao. (2014). Robust Text Detection in Natural Scene Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(5), 970–983. https://doi.org/10.1109/TPAMI.2013.182

Yang, C., Yin, X. C., Yu, H., Karatzas, D., & Cao, Y. (2018). ICDAR2017 Robust Reading Challenge on Text Extraction from Biomedical Literature Figures (DeTEXT). In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (Vol. 1, pp. 1444–1447). IEEE. https://doi.org/10.1109/ICDAR.2017.235

Zhang, J., & Kasturi, R. (2014). Sign Detection Based Text Localization in Mobile Device Captured Scene Images (pp. 71–82). Springer, Cham. https://doi.org/10.1007/978-3-319-05167-3

Zhang, Y., & Liu, B. (2007). Semantic text classification of disease reporting. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07* (p. 747). New York, New York, USA: ACM Press.

https://doi.org/10.1145/1277741.1277889

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). EAST: An efficient and accurate scene text detector. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (Vol. 2017–Janua, pp. 2642–2651). https://doi.org/10.1109/CVPR.2017.283

Zhu, S., & Zanibbi, R. (2016). A Text Detection System for Natural Scenes with Convolutional Feature Learning and Cascaded Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 625–632). IEEE. https://doi.org/10.1109/CVPR.2016.74