UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Byron Alejandro Acuña Acurio

Machine Learning applied to improve accessibility of PDF documents for Visually Impaired Users

Aprendizado de Máquina aplicado para melhorar a acessibilidade de documentos PDF para usuários com deficiência visual

Campinas
2019

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Byron Alejandro Acuña Acurio

Machine Learning applied to improve accessibility of PDF documents for Visually Impaired Users

Aprendizado de Máquina aplicado para melhorar a acessibilidade de documentos PDF para usuários com deficiência visual

Dissertation presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for obtaining Master degree in Electrical Engineering, in the area of Computer Engineering.

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na área de Engenharia de Computação.

Supervisor/*Orientador*: Prof. Dr. Luiz Cesar Martini

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Byron Alejandro Acuña Acurio, e orientada pelo Prof. Dr. Luiz Cesar Martini

---

Campinas
2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Ac93m

Acuña Acurio, Byron Alejandro, 1986-
Machine learning applied to improve accessibility of PDF documents for visually impaired users / Byron Alejandro Acuña Acurio. – Campinas, SP : [s.n.], 2019.

Orientador: Luiz Cesar Martini.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Acessibilidade. 2. Visão por computador. 3. Aprendizado profundo. 4. Aprendizado de máquina. I. Martini, Luiz Cesar, 1952-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Aprendizado de máquina aplicado para melhorar a acessibilidade de documentos PDF para usuários com deficiência visual
Palavras-chave em inglês:
Accessibility
Computer vision
Deep learning
Machine learning
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Luiz Cesar Martini [Orientador]
Yuzo Iano
Felipe Leonel Grijalva Arevalo
Data de defesa: 03-12-2019
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-6788-6833
- Currículo Lattes do autor: http://lattes.cnpq.br/1720607469550929

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidato: Byron Alejandro Acuña Acurio **RA:** 209428
Data da Defesa: 03 de Dezembro de 2019

Título da Dissertação: Machine Learning applied to improve accessibility of PDF documents for Visually Impaired Users.
Aprendizado de Máquina aplicado para melhorar a acessibilidade de documentos PDF para usuários com deficiência visual

Prof. Dr. Luiz Cesar Martini (Presidente, FEEC/UNICAMP)
Prof. Dr. Yuzo Iano (FEEC/UNICAMP)
Prof. Dr. Felipe Leonel Grijalva Arevalo (Escuela Politecnica Nacional)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Acknowledgments

Agradeço profundamente ao Professor Luiz Cesar Martini, pela paciência, ajuda, dedicação e estímulo contínuo durante tudo o percurso deste projeto, sem ele não fosse possível a culminação do mesmo.

A meus pais Edgar Acuña e Fanny Acurio, meu irmão Roberto Acuña e minha espousa Diana Cherrez os quais são minha fonte de luta, força e coragem, pela ajuda, dedicação e amor brindado durante esses anos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Código de Financiamento 001.

Abstract

Digital documents are accessed by visually impaired people (VIP) through screen readers. Traditionally, digital documents were translated to braille text, but screen readers have proved to be efficient for the acquisition of digital document knowledge by VIP. However, screen readers and other assistive technologies have significant limitations when there exist tables in digital documents such as portable document format (PDF). For instance, screen readers can not follow the correct reading sequence of the table based on its visual structure causing this content is inaccessible for VIP. In order to deal with this problem, in this work, we developed a system for the retrieval of table information from PDF documents for use in screen readers used by visually impaired people. The proposed methodology takes advantage of computer vision techniques with a deep learning approach to make documents accessible instead of the classical rule-based programming approach. We explained in detail the methodology that we used and how to objectively evaluate the approach through entropy, information gain, and purity metrics. The results show that our proposed methodology can be used to reduce the uncertainty experienced by visually impaired people when listening to the contents of tables in digital documents through screen readers. Our table information retrieval system presents two improvements compared with traditional approaches of tagging text-based PDF files. First, our approach does not require supervision by sighted people. Second, our system is capable of working with image-based as well as text-based PDFs.

Key-words: Accessibility, computer vision, deep learning, statistical approach.

Resumo

Os documentos digitais são acessados por pessoas com deficiência visual (VIP) por meio de leitores de tela. Tradicionalmente, os documentos digitais eram traduzidos para texto em braille, mas os leitores de tela provaram ser eficientes para a aquisição de conhecimento para as VIP. No entanto, os leitores de tela e outras tecnologias assistivas têm limitações significativas quando existem tabelas em documentos digitais como os documentos PDF (Portable Document Format). Por exemplo, os leitores de tela não podem seguir a sequência de leitura correta da tabela com base em sua estrutura visual causando que esse conteúdo seja inacessível aos VIP. Para lidar com esse problema, neste trabalho, desenvolvemos um sistema para a recuperação de informações de tabela de documentos PDF para uso em leitores de tela usados por pessoas com deficiência visual. A metodologia proposta aproveita as técnicas de visão computacional com uma abordagem de aprendizado profundo para tornar os documentos acessíveis em vez da abordagem clássica de programação baseada em regras. Explicamos em detalhe a metodologia que usamos e como avaliar objetivamente a abordagem por meio de métricas de entropia, ganho de informação e pureza. Os resultados mostram que nossa metodologia proposta pode ser usada para reduzir a incerteza experimentada por pessoas com deficiência visual ao ouvir o conteúdo das tabelas em documentos digitais através de leitores de tela. Nosso sistema de recuperação de informações de tabela apresenta duas melhorias em comparação com as abordagens tradicionais de marcação de arquivos PDF. Primeiro, nossa abordagem não requer supervisão de pessoas com visão. Segundo, nosso sistema é capaz de trabalhar com PDFs baseados em imagem e em texto.

Key-words: Acessibilidade, visão computacional, aprendizado profundo, abordagem estatística.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

According to the World Health Organization (WHO) nowadays, there are more than 253 million people live with vision impairment [6] and they have access to less than 10% published materials[7]. Information accessibility problems often mean that VIP cannot have access to relevant information publicly available in digital documents, which can lead to exclusion from society to persons with visual disabilities.

For this reason, many organizations recognize the need for researching on improving accessibility[8], for instance, the United Nations Educational, Scientific and Cultural Organization (UNESCO) manifested in 2004 that "information is a primary and fundamental right for everybody".

Currently, digital content is accessed by VIP through screen readers. Traditionally, digital documents were translated to braille text, but screen readers have proved to be efficient for the acquisition of digital document's knowledge by VIP. However, screen readers have significant limitations when there exist tables in PDF documents[9]. For instance, the reading sequence of the table text is no based on its visual order causing confusion to VIP. In order to deal with this problem, in this work, we proposed a new methodology to make accessible digital documents for visually impaired people.

This dissertation was conceptualized by the professor Luiz César Martini, to be compatible with the platform "DOSVOX"[10] that was developed by the UFRJ supporting more than 60 000 visual impaired users in Brazil and abroad[11], as others accessibility's projects developed before by Professor Martini in order to aid VIP, such as FINANVOX [12], MATVOX [13], etc.

1.2 Challenges

A significant challenge in implementing accessibility of tables for visually impaired people is that the current standard for store and exchange digital document named portable document format (PDF) has unstructured information, which means, that we have only coordinates or positions in a page for basics elements such as characters, or single lines that can be part or not of a table. In general, a PDF is a sequential code created with the instructions to print the digital document in a printer preserving its content and format, on the other hand, screen readers read the plain text of the mentioned PDF code lines, the problem basically is that the sequence of print a document is different to the correct reading sequence[9]. As a consequence, screen readers are a source of unclassified words for VIP generating high uncertainty listening to all the contents of the digital document.

Ideally, the contents of the table should be translated into an accessible format, such as Brailey, however, since creating accessible versions of digital content is a complex and non-scalable task that requires specialized and costly equipment (e.g. An interpreter from Portuguese or any language to braille),hence, personalize digital content to guarantee the accessibility of visually impaired people [14] to digital content has proved to be a more efficient approach instead of creating braille versions of the digital documents.

Although there are several types of digital documents, in practice, PDF format is the current standard [15] for exchange and store digital documents. The main feature of a PDF document is the preservation of content and visual structure on any device. Under these circumstances, computer vision techniques might be applied to PDF documents, to retrieve information from its visuals structure i.e., identifying, localizing and retrieving tables. This approach might considerably reduce the overall uncertainty experienced by VIP aiding on the understanding of the tables and digital contents.

On the other hand, when personalize PDF documents is not an option to make them accessible for VIP, it is possible under Marrakesh VIP Treaty [16] copyright exceptions to facilitate the creation of accessible versions of books and other copyrighted works for visually impaired persons. Under this scenario, to improve the accessibility of PDF documents is possible to extract the table's content to a spreadsheet software package for VIP. However, this procedure might be time-consuming, especially when the PDF document is composed of several tables from a large number of subjects. In order to speed up this process, it is desirable to find a way to automatize this process minimizing human error.

Make accessible a PDF document is a challenging task because there are 8 different PDF versions, which share the same extension " .pdf ", therefore, each PDF is a different variation of the PostScript language, the fundamental problem is that we have no knowledge a priori of, What PDF version are we analyzing? besides, this format is a no structured information format, hence we have no internal representation of the infor-

mation, i.e., we do not know, if there are tables, graphs, equations, titles, etc, in PDF document. The information that we can obtain from a PDF file is only x,y coordinates relative to the bottom-left corner of the page, for each character or other very basic components. Thus, we have the uncertainty of known if a character belong to a table or not. Due to the uncertainties, about the PDF code, it is more suitable to convert the PDF document to a set of image-pages, to obtain the localization of the table via computer vision using deep learning approaches, because these fields have shown recent technological advances with the potential to retrieve information from images (e.g., localization and classification of images).

1.3 Aims of this dissertation

To address the aforementioned challenges in implementing accessibility in PDF documents, in this dissertation, we propose a table information retrieval system for VIP, the main contributions of this work are:

- We proposed a new methodology to make accessible the tables information embedded in digital documents for visually impaired people.
- We proposed different strategies to improve the accuracy of table model to identify and localize tables in digital-pages using a deep learning approach.
- We developed a software prototype of our proposed table information retrieval system for visually impaired people, integrating DosVox [17] with Camelot[18].

1.4 Summary of Contributions and Dissertation Outline

This dissertation is presented as an extension of our paper "Table Detection for Improving Accessibility of Digital Documents using a Deep Learning Approach" published in the 6th IEEE Latin American Conference on Computational Intelligence LA-CCI. The dissertation is structured as follows:

Chapter 2 provides a brief review of the terminology used in machine learning, computer vision and an overview of the challenges and state-of-the-art approaches to make accessible PDF documents.

Chapter 3 contains the first contribution of this work, entitled "Table Detection for Improving Accessibility of Digital Documents using a Deep Learning Approach"[1], which was published in the *IEEE Latin American Conference on Computational Intelligence (LA-CCI) Guayaquil, Ecuador, November 11–15, 2019*. Here we proposed a new methodology that takes advantage of computer vision techniques with a deep learning approach to make documents accessible instead of the classical rule-based programming approach. This system to aid visually impaired people to access the table's contents on PDF documents was named table information retrieval system for VIP. We performed this study because we didn't found any reports that explain, in detail, how to apply deep learning techniques with information retrieval algorithms for VIP. The principal objective of the study was to reduce the uncertainty that VIP face when listening to the contents of tables through screen readers, together with the rest of the contents of the document (e.g., text, figures, equations). We explained in detail the methodology that we used and how to objectively evaluate the approach through entropy, information gain, and purity metrics. Our main contribution herein is to show that the presented approach can reduce the uncertainty (i.e. entropy) that a VIP has about the information contained in digital documents by taking advantage of computer vision techniques with a deep learning approach. Our table information retrieval system presents two improvements compared with traditional approaches of tagging text-based portable document format (PDF) files. First, our approach does not require supervision by sighted people. Second, our system is capable of working with image-based as well as text-based PDFs.

Chapter 4 contains the second contribution of this work, which was a systematical study of strategies to improve the Faster-RCNN architecture and boost the performance of table detection models. In this chapter, we compared different models using different pre-trained networks and different image preprocessing before training each model. After several experiments, we have the following four findings. First, we can create a table detection model using publicly available datasets in the English language to detect tables in other occidental languages such as Portuguese and Spanish languages. Second, we have noticed that using our proposed image preprocessing approach in grayscale according to our results suggest that we have an improvement on our table detection model. Third, we have noticed that if we use pre-trained networks that have better performance on imagenet challenge, as feature extractors in our Faster-RCNN we obtain better table detection models, we compared two pre-trained deep convolutional neural network, our proposed resnet-101 and the VGG that was proposed on the original Faster-RCNN architecture. Finally, findings suggest that best optimizer for the table detection task was Gradient Descent and we confirm that using a larger dataset we obtain a better table detection

[1]This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001.

model.

Chapter 5 presents the third contribution of this work, which was a software prototype for a proof of concept of our proposed approach. Our main contribution is that we can integrate this software prototype into DOSVOX[17], for parsing PDF tables we use an open-source project named Camelot[18], originally created for sighted people.

Chapter 7 presents the set of conclusions of this dissertation and recommends possible directions for future research.

Chapter 2

Theoretical Background

In this chapter, we present a brief review of the terminology used in machine learning, computer vision and an overview of the challenges and state-of-the-art approaches to make accessible PDF documents.

2.1 Machine Learning Concepts

Machine Learning (ML)[19] is a branch of artificial intelligence that aims to make the machine capable of performing specific tasks without being explicitly programmed, this means is that these algorithms are not based on rules, the entire learning process is constructed in such a way that it minimizes or completely eliminates human intervention. ML has achieved impressive results in various cognitive tasks, such as image classification, prediction, etc. ML is a statistical approach where the machine or computer learns patterns from data presented in order to make a prediction in an unknown situation. The performance of these algorithms generally improves as they are exposed to more data.

To understand how the machine learning model works, it is essential that we have a broad intuition of the field as a whole and a type of learning problems that it can solve.There are three main branches of machine learning: supervised learning, unsupervised learning, and reinforcement. Each of these branches is named after the type of learning problem they are trying to solve.

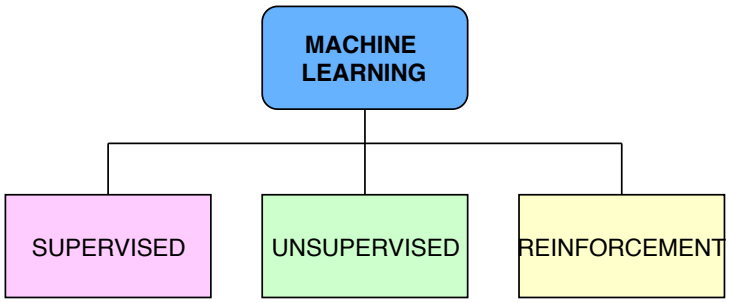

Figure 2.1: Principal branches of Machine Learning

- Supervised learning is by far the most common case, as the name implies involves training a machine learning model on examples that are clearly annotated to show what the model is supposed the learn. these annotations or label act as a supervisor providing guidance to machine learning model as it steadily improves its performance at the task which it is to learn, i.e., for each training data (or pattern) available, there is the desired answer that is known. In this case, we say the data is labeled.

- Unsupervised learning consists of discovering patterns and correlations that describe the raw data input. There is no desired output associated with each pattern, so the data is unlabeled. In this scenario, we want the model to be able to capture, represent, or express existing properties in the dataset.

- Reinforcement learning deals mainly with crating an intelligent agent that receives information about its environment and learns to choose actions that maximize the possibility of occurrence of some end goal.

2.1.1 Evaluating of Machine Learning Models

Machine Learning models are evaluated according to a metric chosen to determine if a model has learned the desired representations that allow the correct prediction. We divide the data into a training set, a validation set, and a test set. The main reason for not evaluating the models with the same data that were trained is because, after some epochs, the three models began to overfit. That is, the model would reach almost perfect accuracy in the training set, but it would have terrible accuracy in the validation or test set because it basically memorized each data point and could not learn the underlying dynamics of the data set. Our main objective is to achieve models that generalize and work well in data never seen before, that is, a model that discovers important patterns in the training set that can be used to make predictions about new data.

Overfitting and underfitting together form the central problem in machine learning. We say a model is overfitted to a training set when it has learned not only useful representations in the data but has also learned the noise in order to obtain an artificially

high training set accuracy. The underfitting means that the models have only learned a small subset of representation and discarded most of the useful information, so they make unfounded assumptions and models have poor accuracies. The ideal situation is to find a model that neither underfitting nor overfitting, but that exhibits the right balance between optimization and generalization. This can be done by maintaining a third of the examples known as a validation set. The validation set is used to improve model performance without overfitting the model training set.

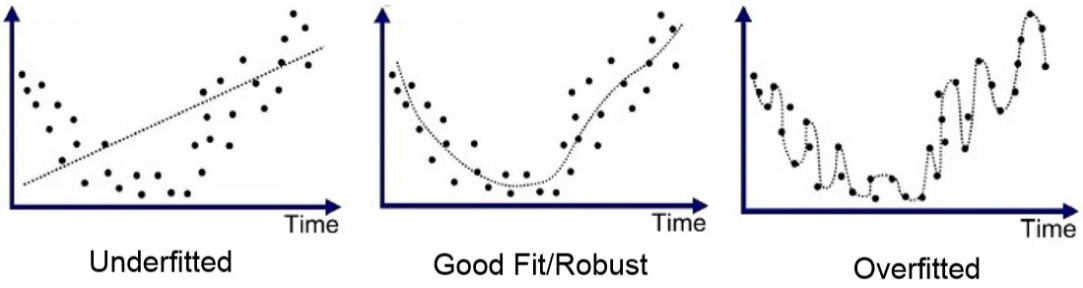

Figure 2.2: Evaluating Machine Learning Models

2.1.2 Artificial Neural Networks and Deep Learning

Artificial neural networks (ANN) are inspired by the structure of neurons in the human brain. The brain consists of billions of interconnected neurons to form neural networks that are somehow responsible for carrying out cognitive tasks such as identifying objects, grouping entities into categories, making decisions, etc. These neurons communicate with each other through electrical signals. When a set of neurons is activated by an event that they find interesting, the connections between them are strengthened or weakened.

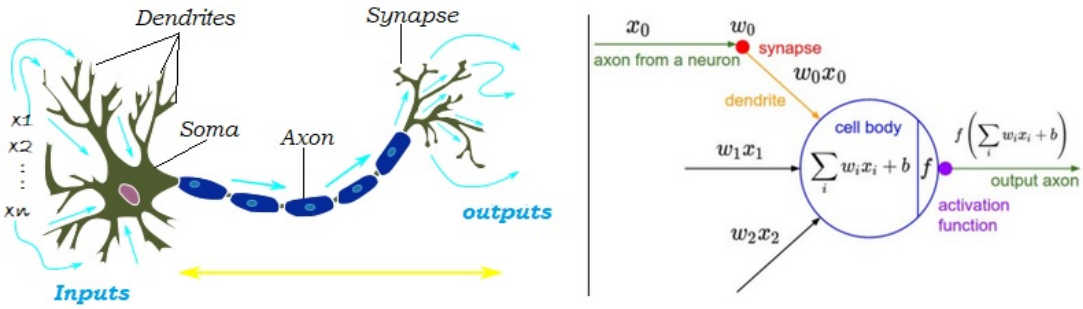

Figure 2.3: A biological Neuron(left)-its mathematical model(right)

Fig.2.3 left we can see a single biology neuron with its components. A neuron is structured in such a way that it receives several inputs from neighboring neurons. These inputs provide information that has been sent in the form of an electrical signal from other parts of the brain. Dendrites are the gateway through which information is fed to

a neuron. The axon is responsible for transmitting the processed information in the cell body to the other neurons. Now, the model on the right is a mathematical representation of the biological model. The mathematical model consists of the weight (w), which are parameters that modulate the importance of certain characteristics that are introduced in the neuron, the axon (x), the bias b and the activation function (f), whose role is to decide if it triggers the electrical signal or not. The neuron has a connection with other neurons and learns by modifying the weights of the synapses, which control how much influence each neuron has.

The mathematical model show in Fig.2.3 right, is known as the first mathematical model of neurons of an electronic brain proposed in 1943 by S. McCulloch and W. Pitts, in fact, in this model, weights are not learned but are adjusted for work like logic gates (AND, OR, NOT), we must notice that this model opens XOR logic gate problem, i.e., researchers could not use this model for a long time to solve problems that require XOR logic gates solutions.

Then, since 1960, the field of artificial neural networks saw the introduction of two of the earliest feedforward neural network algorithms: the perceptron model (F. Rosenblatt, 1957) and the ADAptive LInear NEuron (ADALINE) algorithm (B. Widrow and M. Hoff, 1960), with the mentioned improvements, models were capable of have learnable weights and thresholds, starting an age of neural networks named the first golden age. The final of this age occurred in 1969 with the campaign waged by M. Minsky and S. Papert, where they showed the inability to usefully compute certain essentials drawbacks such as the mentioned XOR problem. From 1969 to 1986, there was an impression that neural network research has proven to be a dead-end, this period of time was named as Dark Age ("AI winter"). It is important to note that Minsky and Papert's analysis of the Perceptrons not only showed the impossibility of calculating XOR with a single Perceptron but specifically argued that it had to be done with multiple layers of Perceptrons, what we now call multilayer neural networks, the fundamental problem was in that time that Rosenblatt's learning algorithm did not work for multiple layers, but in 1986 neural networks' return to popularity when was announced by D. Rumelhart, G. Hinton and R. Williams the reinvention of the term backpropagation and its general use in multilayer neural networks, obtaining successful applications such as recognition, as is shown in Fig. 2.4.

2.1.3 Convolutional Neural Network

Convolutional neural networks are a kind of of deep neural networks inspired by the biological process associated with human vision. To understand a convolutional neural network, it is important to know its central operation, the convolution. The convolution operation consists of finding local patterns in the input unlike the case of a fully connected layer that finds global patterns, a fully connected layers connect each neuron in one layer

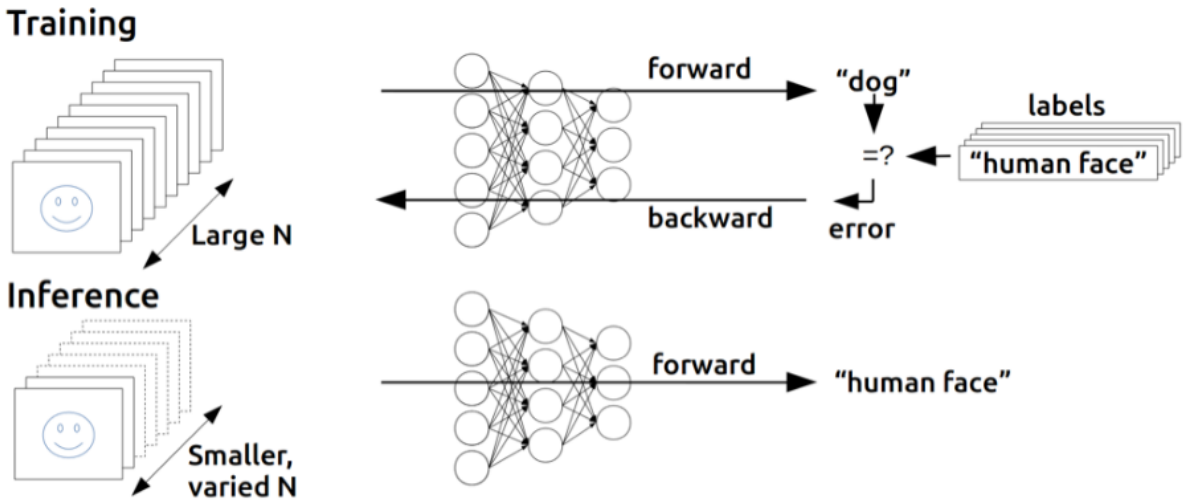

Figure 2.4: Backpropagation basic idea

with each neuron in another layer. The convolution operation that involves a series of filters or kernel is carried out on the image to produce corresponding feature maps. The feature maps represent the features automatically learned by CNN. The next operation involves pooling, it makes the representations smaller and more manageable. Max pooling or the average pooling may be employed. The former involves choosing the maximum features in a sliding window, while the second involves averaging features. Feature maps with lower sampling get tangled and go deeper into the network. The main intuition is that the previous layers detect simple features such as edges, while the back layer combines these previously detected features into complex features such as patterns, parts of objects, etc. In addition, as we progress through the network, the image size is reduced but the depth (number of channels) increases. Finally, advanced feature maps are fed into a regular fully connected neural network composed of dense layers and activation functions.

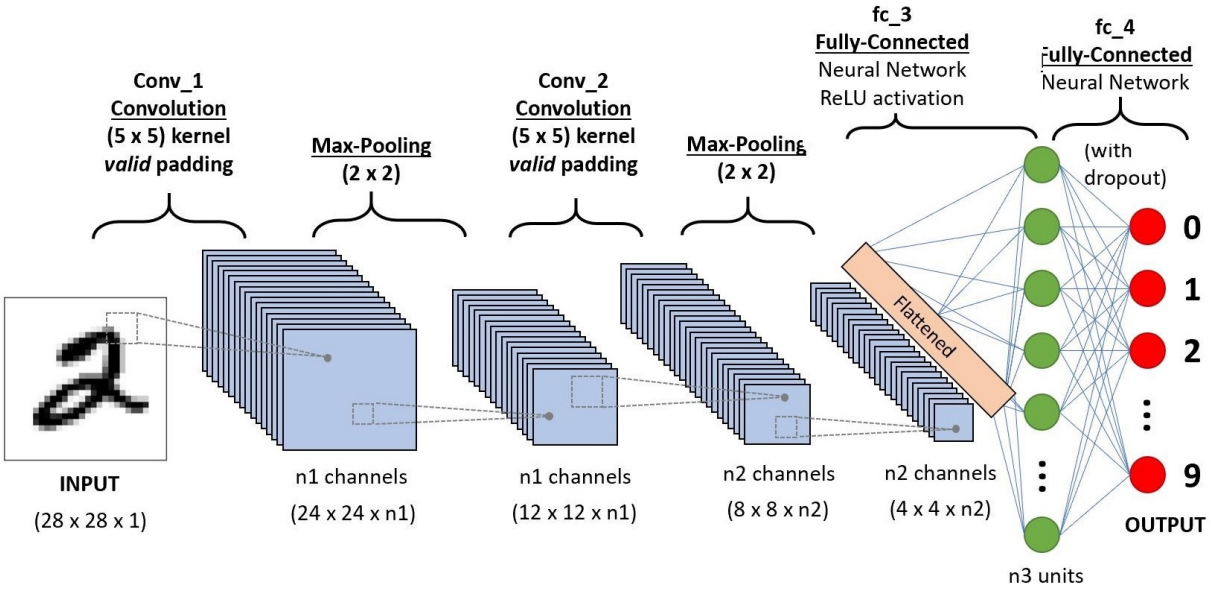

Figure 2.5: Typical CNN architecture

2.2 Classical Approach to make accessible Portable Document Format (PDF)

Since 2008, portable document format (PDF) files are the current reference for exchange and store digital documents, according to the ISO standard 32000-1:2008[20]. For instance, more than 75 countries[21] use PDF files as their format of choice for their documentation, i.e., production, distribution, collaboration, and archiving. The printing industry has required the use of PDF for any professional printing job. With billions of publicly available documents and an untold number of documents living in private repositories, no other file format has the wide reach and ubiquity that PDF does. However, even with those billions of documents in circulation, the PDF format remains poorly accessible for visually impaired people, because PDF files focused on preserve the content and visual structure on any device without the need to have access to the original software that was used to create them instead of being a friendly format for screen readers. Hence, the fundamental problem of accessibility facing VIP is the segmentation of information in PDF documents based on its visual structure since screen readers do not deal with it. To make PDF documents accessible to VIP, the PDF/Universal Accessibility (PDF/UA) ISO standard recommends tagging of all the PDF visual elements such as tables, graphs, equations when authoring [9]. However, this task can be cumbersome because there are 8 different PDF versions, all of them sharing the same extension (*.pdf*). [22], which would require different tagging [9] approaches. To deal with this PDF-tagging problem, rules-based programs have been developed for post-processing PDF documents such as [14] and [23]. As is shown in Fig. 2.6

But the critical aspects, such as the correct text reproduction sequence (tagging) on

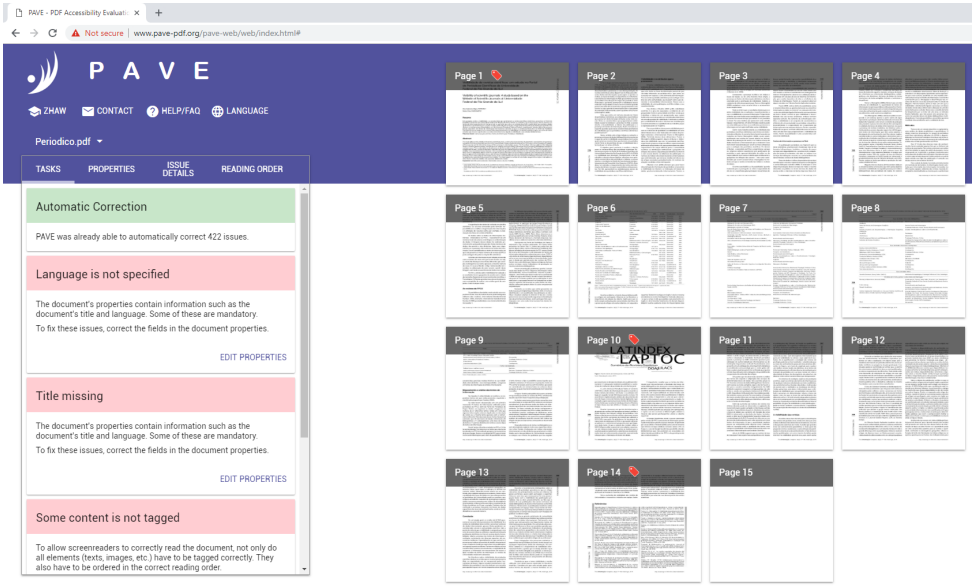

Figure 2.6: Rule-based program for post-processing PDF (PAVE)

Assistive Technologies (AT), must be done manually by sighted people, making it laborious for the authors.

Furthermore, in the actual process of tagging a PDF, there are two more issues that need to be addressed to make tagged documents suitable for use by VIP:

- Assistive technologies (screen readers) must be compliant with PDF/UA to take advantage of the tagging.
- The tagging approach only works with text-based PDF documents. Therefore, image-based PDF or scan-based PDF documents remain inaccessible using this method.

The classical approach considered the PDF accessibility problem as an untagged problem. In this work, we consider the accessibility problem in PDF documents as an *"object detection"* [24] problem. Specifically, we focused on table detection in PDF documents because it is a nonlinear problem owing to its discontinuities on edges, for which we require image's information to retrieve the table into a suitable format, making its content easy to interpret by VIP.

2.3 Table Detection

When we move through our digital document and recognize different elements such as equations, tables or figures, we heavily are using our vision especially object recognition and localization. What seems so effortless for sighted people still poses a major challenge for artificial systems like a computer vision. Trying to teach computers to see and also "understand" what is a table has proven be extremely difficult [25]. The key to understanding

Tables in a digital page are three closely related sub-problems. The first one is named classification in the following. For classification, the object class table in a given image should be determined and labeled. The next more demanding task is object localization: In addition to labeling the dominant object class, must be localized in the image-page, usually by determining a bounding box around the object class table. The difficulty of this task again increases because in an image-page, there are the probability of found multiple tables in one image-page or other elements such as paragraphs, equations and graphs, around or in tables. This task was named *table detection* in the field of document analysis and recognition. Despite of the multitude of methods currently available for Table detection and decomposing them into their structural building blocks[26, 27, 28, 29] most of this approaches are far from a general solution, however machine learning approaches are obtaining the best performing in this task deliver an average accuracy between 84% to 87%, showing that we are still far from the level of accuracy that is necessary to reliably use these methods to automatically process tables in our system for visually impaired people. Due to this 15% of error rate which means that is required human verification after processing the data[30]. For this reason Table detection is currently an active field of research, wherein order to obtain higher precision, deep learning approaches have been performed in this field by Gilani et al. [4] and Siddiqui et al. [31]. Both approaches used deep neural networks (DNN) architectures for the task of detecting tables in an RGB (Red/Green/Blue) image. The first technique Gilani et al. [4] feed the DNN with preprocessed image document through an image operator named distance transform, whereas the method used by Siddiqui et al. [31] works with raw images as the input. However, is important to notice that the mentioned works did not propose the architectures, in fact, the used architectures were proposed trying to solve generic problems of object detection of two famous challenges *PASCAL Visual Object Challenge* (VOC) and later the *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC)[32].

2.4 Object Detection

Works in object detection published can be categorized into three directions[33]: objectness detection (OD), salient object detection (SOD), and category-specific object detection (COD). Usually, OD algorithms output thousands of region proposals over the image that is analyzed. SOD [34, 35], are often divided into two groups: bottom-up saliency and top-down saliency, which are concerned on highlight objects that draw our attention from a given image [36]. Finally category-specific object detection (COD)[37, 38] aims at detecting multiple predefined object categories from each given image. In order to create a table detection model we must use deep learning based category-specific object detection methods, in this category state-of-the-art methods are Deep Learning Frameworks such as Faster R-CNN[39], You Only Look Once (YOLO)[40] and Single Shot MultiBox

Detector (SSD)[24]

2.4.1 Category-specific object detection using a Deep Learning Approach

In this work, we will refer to Object Detection as a branch of deep learning which is a subfield of machine learning which is a field of artificial intelligence where our object detection model will learn the parameters to map an image and retrieve the bounding box location of the objective class, in our work we will work only with one class which is the class table.

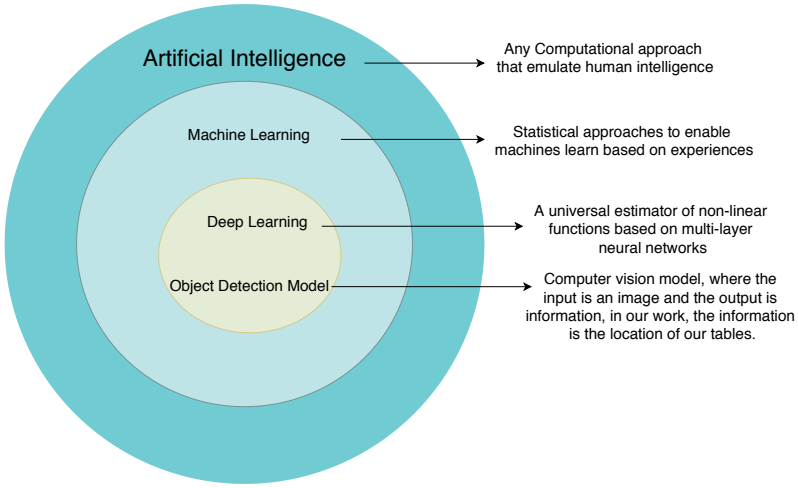

Figure 2.7: The relation between Artificial Intelligence, Machine Learning, Deep Learning and Object Detection.

2.5 Faster R-CNN Architecture

We used in this work the Faster R-CNN architecture published in 2015 by Girshick et al [39], which is the third generation of region proposed based networks. As shown Fig. 2.8.

As we can notice by the name R-CNN which stands for "Region-based convolutional neural networks". The Faster R-CNN was proposed in order to shorten the time spent by the Fast R-CNN[41] architecture, on region proposal step during test time. Hence, we can notice that Faster-RCNN is composed by three architectures: The first one is Region Proposal Network (RPN), the second one is a Fast R-CNN, and this two modules share a Deep convolutional Neural Network architecture, as shown in Fig.2.9.

This architecture is really interesting because the Faster R-CNN joint different well-known machine learning techniques (i.e., feature extraction, regression and prediction) to

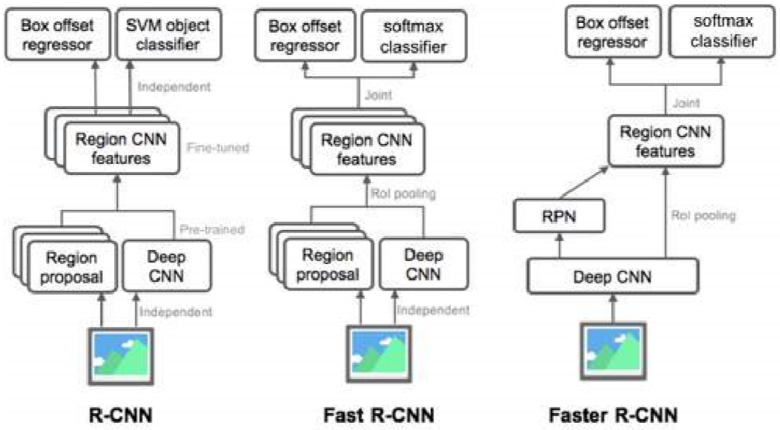

Figure 2.8: Third Generation of Faster R-CNN[1].

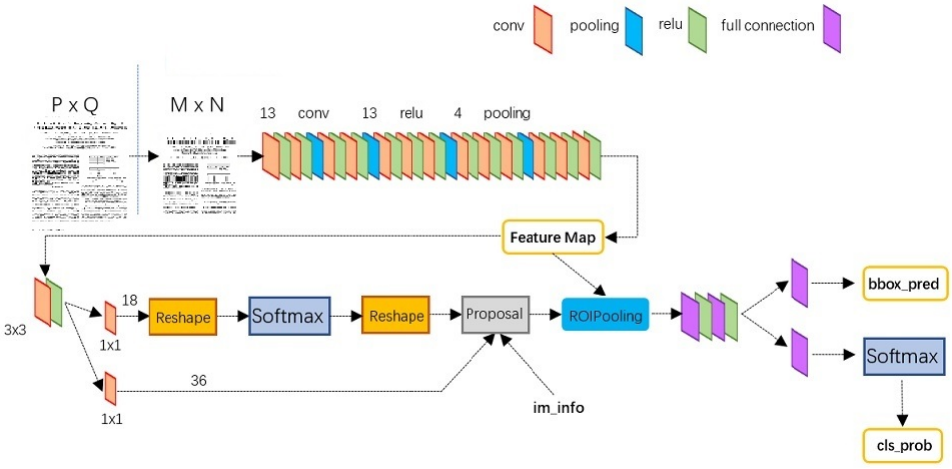

Figure 2.9: Faster-RCNN Architecture.

solve a new problem. The new problem that we are trying to resolve with this architecture is named object detection, in this problem, we will input an image and we will train our network to retrieve the location of a predefined objective. In our work, we are seeking for localizing the tables in an image-page.

Chapter 3

Contribution I
Table Detection for Improving Accessibility of Digital Documents using a Deep Learning Approach

Assistive technologies play an important role in improving the quality of life of people with disabilities. In this work, we developed a system for the retrieval of table information from digital documents for use in screen readers used by visually impaired people. The proposed methodology takes advantage of computer vision techniques with a deep learning approach to make documents accessible instead of the classical rule-based programming approach. We explained in detail the methodology that we used and how to objectively evaluate the approach through entropy, information gain, and purity metrics. The results show that our proposed methodology can be used to reduce the uncertainty experienced by visually impaired people when listening to the contents of tables in digital documents through screen readers. Our table information retrieval system presents two improvements compared with traditional approaches of tagging text-based portable document format (PDF) files. First, our approach does not require supervision by sighted people. Second, our system is capable of working with image-based as well as text-based PDFs.

This chapter is an extension of our paper "Table Detection for Improving Accessibility of Digital Documents using a Deep Learning Approach" published in the 6th IEEE Latin American Conference on Computational Intelligence LA-CCI.

Keywords: Assistive technology, computer vision, deep learning, statistical approach.

3.1 Motivation

Digital content is accessed by Visually Impaired People (VIP) through screen readers. This has been resulted by the authoring of content in portable document format (PDF) files, which have surpassed the use of print publications in all fields (e.g., academic, legal). The main benefit of a PDF is that enables information exchange created by different software packages without the need to have access to the original software that was used to create them. However, the fundamental problem of accessibility facing VIP is the segmentation of information in electronic documents based on its visual structure since screen readers do not deal with it.

To make PDF documents accessible to VIP, the PDF/Universal Accessibility (PDF/UA) ISO standard recommends tagging of all the PDF visual elements such as tables, graphs, equations when authoring [9].

However, this task can be cumbersome because there are 8 different PDF versions, which would require different tagging [9]. To deal with this PDF-tagging problem, rules-based programs have been developed for post-processing PDF documents [9]. But the critical aspects, such as the correct text reproduction sequence (tagging) on Assistive Technologies (AT), must be done manually by sighted people, making it laborious for the authors (Fig. 3.1). Furthermore, in the actual process of tagging a PDF, there are two

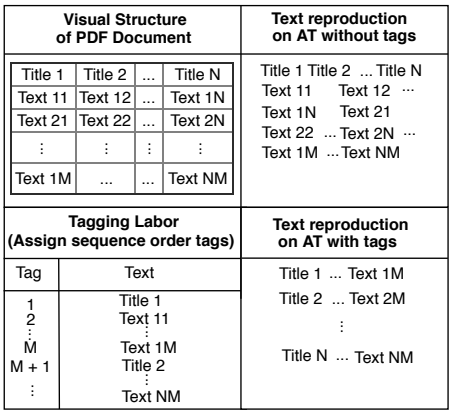

Figure 3.1: Required manual tagging of PDF content (e.g., titles, text) when is used rules-based programs to indicate the correct reading sequence on assistive technologies for making accessible a PDF (Classical Approach).

more issues that need to be addressed to make tagged documents suitable for use by VIP:

- Assistive technologies (screen readers) must be compliant with PDF/UA to take advantage of the tagging.
- The tagging approach only works with text-based PDF documents. Therefore, image-based PDF or scan-based PDF documents remain inaccessible using this method.

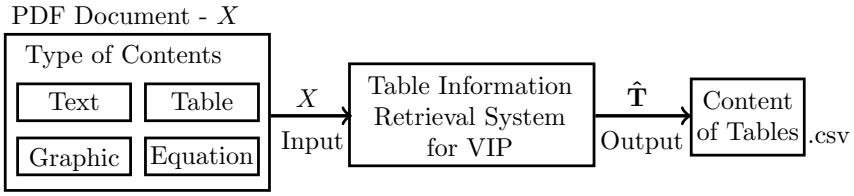

Figure 3.2: Table information retrieval system for VIP (Proposed Approach).

The classical approach considered the PDF accessibility problem as an untagged problem. In this work, we consider the accessibility problem in PDF documents as an *"object detection"* [24] problem. Specifically, we focused on table detection in PDF documents because it is a nonlinear problem owing to its discontinuities on edges, for which we require image's information to retrieve the table into a suitable format, making its content easy to interpret by VIP.

With this aim, we propose an autonomous system for the detection and extraction of table information in PDF documents for VIP. Our project is flexible and can be easily adapted to accessibility systems. We expand the theoretical explanation of the table detection model to facilitate reproducibility by newcomers in this field. We named our system "table information retrieval system for VIP", consisting of:

1. Table Detection System (TDS),

2. Table Information Extraction System (TIES).

We performed this study because we didn't found any reports that explain, in detail, how to apply deep learning techniques with information retrieval algorithms for VIP. The principal objective of the study was to reduce the uncertainty that VIP face when listening to the contents of tables through screen readers, together with the rest of the contents of the document (e.g., text, figures, equations), as shown in Fig. 3.2.

3.2 Related Work

After the recent success of using machine learning (ML) algorithms for the detection of tables in digital documents in [42], whereby ML algorithms were used to obtained higher precision compared with traditional approaches, new state-of-the-art research has been performed in this field by Gilani et al. [4] and Siddiqui et al. [31]. Both approaches used deep neural networks (DNN) architectures for the task of detecting tables in an RGB (Red/Green/Blue) image. The first technique Gilani et al. [4] feeds the DNN with preprocessed image document through an image operator named distance transform, whereas the method used by Siddiqui et al. [31] works with raw images as the input. We based our TDS on the method reported by Gilani et al. [4], but our system differs on the following points:

- We used a different preprocessing image transformation.

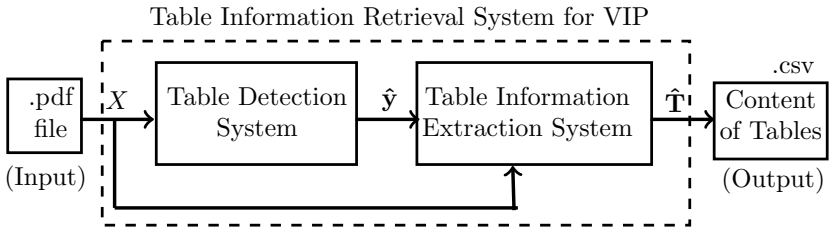

Figure 3.3: Major components of Table Information Retrieval System for VIP.

- We used residual net (ResNet-101) as a pretrained network instead of the visual geometry group (VGG-16).
- We analyzed and discussed the use of different optimizers in the training stage.
- We used tensorflow on Google's Colaboratory instead of Caffe to train the model.

3.3 System Design

We considered as an input a PDF document, whose main feature is the preservation of content and visual structure on any device. Our proposed system is depicted in Fig. 3.3 inside the dashed lines. The document X is analyzed by the TDS to obtain the coordinates of each table bounding box (i.e., the minimum rectangle that contains a table) $\hat{\mathbf{y}}$; in order to create new comma separated values (CSV) files $\hat{\mathbf{T}}$ containing the information from each table, which can then be opened by a spreadsheet program for VIP.

3.3.1 Table Detection System (TDS) Design

Given a self-contained input PDF document X with N number of pages, our TDS can automatically detect all the tables contained in the document without supervision and will return the region where each one is located.

To do so, the TDS takes the PDF document X and then converts it into an image-page set (RGB-images) $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_N\}$ which is stored in an input buffer, as shown in Fig. 3.4.

We assume that the image-pages and tables occur in an ordered fashion. Therefore, the analysis explained for $\mathbf{X}_i$, will be the same for each image-page of the image set $\mathbf{X}$.

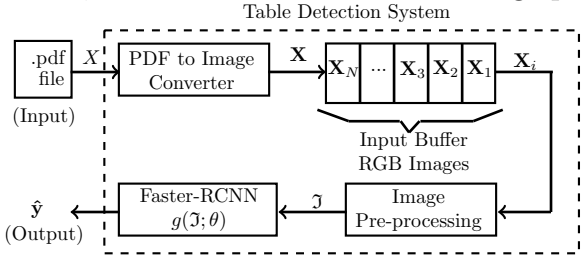

Figure 3.4: Table Detection System.

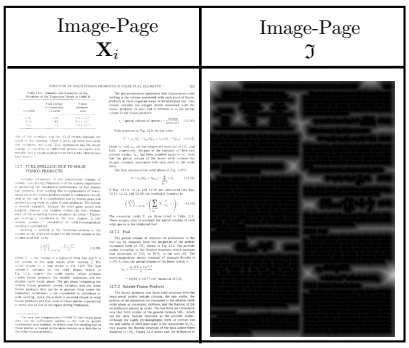

Figure 3.5: Left: the input image-page $\mathbf{X}_i$. Right: the output image-page $\mathfrak{I}$ after EDT.

Then, as shown in Fig. 3.4, the image pre-processing block receives an image-page RGB (i.e. $\mathbf{X}_i$) and makes the following transformations:

1. Conversion to a grayscale image.

2. Euclidean Distance Transform (EDT)

This is because most of the main publicly available open data sets are in English. Therefore, when pre-processing image-pages we can use the same trained model for other languages because the pre-processing stage dilating pixels but conserving geometric forms of the table (see Fig. 3.5).

We did not use the same technique for image transformation that was proposed in [4]; with our image transformation technique, we can reduce the dimensionality and eliminate the redundancy of the RGB channels.

We first made a color-space conversion from the RGB image-page $\mathbf{X}_i$ that has three 2D functions (say, $f_R(x_s, y_s)$, $f_G(x_s, y_s)$ and $f_B(x_s, y_s)$) to a function of the gray level of the image, defined by $f(x_s, y_s)$ $\epsilon\{0, 1, ..., L-1\}$; in which L is the level of gray of the image (e.g.,$\{f(x_s, y_s) = 0 = \text{black}\}$; $\{f(x_s, y_s) = 255 = \text{white}\}$), and (x_s, y_s) are the spatial coordinates defined by

- x_s ϵ {0, 1, ..., h-1}; h is the height of the image $\mathbf{X}_i$,
- y_s ϵ {0, 1, ..., w-1}; w is the width of the image $\mathbf{X}_i$.

Therefore, the 2D function $f(x_s, y_s)$ is the input of the EDT algorithm explained in [41] with a distance mask of 5 x 5 pixels (a.k.a. flat structuring element). The local distance values of the mask (+a;+b;+c) in Fig. 3.6 were recommended in [2] to transform the 2D function $f(x_s, y_s)$ into a distance map with the same dimensions. The EDT algorithm interprets the image $f(x_s, y_s)$ as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which the vertices $\mathcal{V}$ and edges $\mathcal{E}$ are defined by an adjacency relation, being $\mathcal{E}$ a set of unordered pixels of $\mathcal{V}$.

Hence, $\mathcal{G}$ is a regular grid and $f(x_s, y_s)$ is a sampled function that can be defined as $f : \mathcal{G} \to \mathbb{R}$. Therefore, for a given pixel $p = (x_s^{(p)}, y_s^{(p)})$ there exists a pixel $q = (x_s^{(q)}, y_s^{(q)})$,

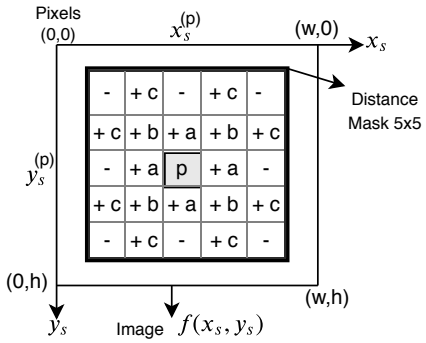

Figure 3.6: Distance mask of 5 x 5 Pixels, with a=1, b=$\sqrt{2}$, c=2.1969, recommended in [2], mask-pixels marked as (-) are not used.

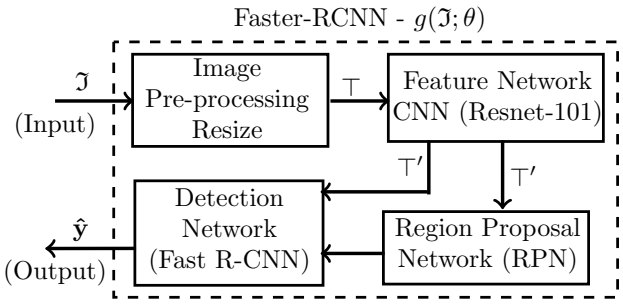

Figure 3.7: Faster-RCNN Schematic.

which is one of the neighbors $\mathcal{N}_D$ of the pixel p, such that it satisfies the EDT defined as

$$\mathcal{D}_f(p) = \min_{q \epsilon \mathcal{G}}(\delta(p,q) + f(q)), \tag{3.1}$$

where $\delta(p,q)$ is a distance function between the pixels defined by $[(x_s^{(p)} - x_s^{(q)})^2 + (y_s^{(p)} - y_s^{(q)})^2]^{(1/2)}$, and $f(q)$ is the local distance from the mask (+a;+b;+c). Therefore, our image $\mathfrak{I}$ can be defined as the distance map function of $f(x_s, y_s)$, i.e.,

$$\mathfrak{I} = \mathcal{D}_f : \mathcal{G} \rightarrow \mathbb{R}. \tag{3.2}$$

The output image $\mathfrak{I}$ after image pre-processing will be the input of our Faster R-CNN [39], whose major components are a feature network that converts the input image into a feature map that will be analyzed by the region proposal network (RPN) for generating region proposals. In turn, the region proposals will be used by the Fast-RCNN [41] to generate the output $\hat{\mathbf{y}}$, as presented in Fig. 3.7.

It is important to remember that using the Faster-RCNN network is a statistical approach, which can be defined as a model $g(\mathfrak{I};\theta)$ that is capable of mapping $\mathfrak{I}$ (i.e., find an unknown number of tables in the image $\mathfrak{I}$) and return n-tuples $\hat{\mathbf{y}}$, with the predicted

tables regions coordinates named bounding boxes (BBs), defined by

$$\hat{\mathbf{y}} = (\{[(x_{min}^{(1)}, y_{min}^{(1)}), (x_{max}^{(1)}, y_{max}^{(1)})]\}, \cdots, \{[(x_{min}^{(n)}, y_{min}^{(n)}), (x_{max}^{(n)}, y_{max}^{(n)})]\}). \tag{3.3}$$

Then, our model can be expressed as

$$\hat{\mathbf{y}} = g(\mathfrak{I}; \theta), \tag{3.4}$$

in which θ are the hyperparameters of the model.

We specified the minimum image resolution as 600 pixels and the maximum image resolution as 1024 pixels. Therefore, the image pre-processing resizes the input $\mathfrak{I}$ if its resolution is outside of these limits, thereby preserving the aspect ratio and complying with established limits.

After the image pre-processing block, the image is a multidimensional array (tensor $\top$) of dimension $h \times w \times 3$. Next, the tensor passes through the feature network of the Faster-RCNN, up to an intermediate layer of our selected pre-trained network ResNet-101, which has less parameters (44 Million) than VGG - 16 (138 Million) used by Gilani et al. [4].

Furthermore, we used an output stride of 16 pixels for ResNet-101, for the 2D convolutional feature map $\top'$ output of the feature network. Thus, $\top' = \mathrm{w}/r \times \mathrm{h}/r$, in which r is the stride (r=16 pixels).

Next, $\top'$ is the input of the RPN, which is composed of 3 convolutional layers; the first one feeds the other two layers, one used for classification and the other used for BB regression. With this process, it is possible to generate region proposals called anchors [39].

We defined our anchors with a size of 256 pixels, and the ratios between the width and height of the anchors as [0.5, 1 and 2]. The scales used for generating anchor sizes were defined as [0.25, 0.5, 1, 2].

Then, the RPN output tuple is composed by:

- Predicted Region proposal

  $[(\Delta x_{min}, \Delta y_{min}), (\Delta x_{max}, \Delta y_{max})]$,

- The probability that there is no table in this region (Background),

- The probability that there is a table in this region (Foreground).

Finally, the detection network takes the convolutional feature map and RPN output tuple to adjusts the BB of the region proposals and remove bad proposals.

3.3.2 Tables Information Extraction System

As shown in Fig. 3.3, we take $\hat{\mathbf{y}}$ and the original PDF document X. First, X is cropped using $\hat{\mathbf{y}}$. If X is a text-based PDF, the text information has to be preserved after the crop; conversely, for image-based PDFs or scan-based PDFs, pre-processing with an optical character recognition (OCR) system is required.

Finally, using the *"stream"* algorithm from [43], we extract the text of the predicted tables to a CSV file, which simulates a table structure based on text-edges, semi-structuring the text. This partition algorithm organizes the text of the table into coherent rows and columns, obtaining a uniform grid, based on the alignment and distance between each text.

This approach is compliant with the Marrakesh VIP treaty [16], which was established to facilitate access to published works to VIP.

Even though a CSV file has no merged cells for the structuring of information, it is useful for VIP when they import this file to a spreadsheet software package because enabling VIP to convert any form of a table in a table with uniform spaces.

3.3.3 Integration with the Assistive Technology

After the information is extracted into a CSV files, it can be easily opened by an assistive technologies, such as NVDA [44], through a spreadsheet software package to provide content location information (columns and rows positions) to VIP. This is required to understand the logical structure of the table i.e., the relationships of the rows and columns content.

3.4 Materials and Methods

3.4.1 Deep Neural Network

We used Python 3.6.8 and Luminoth on Google Colaboratory Jupyter Notebook, to define the tensorflow computation graph shown in Figure 3.8, to train four table detection models using the following optimizers: momentum, adam, gradient descent, rmsprop.

The computation graph is used to estimates the θ hyperameters of our model $g(\mathfrak{I};\theta)$, i.e., fine-tune the ResNet-101, train the region proposal network and the detection network. Our tensorflow computation graph defines the mathematical operations carried out in our tensors, seeking to minimize the objective function $\mathbf{J}(\theta)$, of the Faster-RCNN, which is defined as the sum of four losses functions:

- Region Proposal Network classification loss function (good/bad anchor),
- Region Proposal Network regression loss function (Anchors),

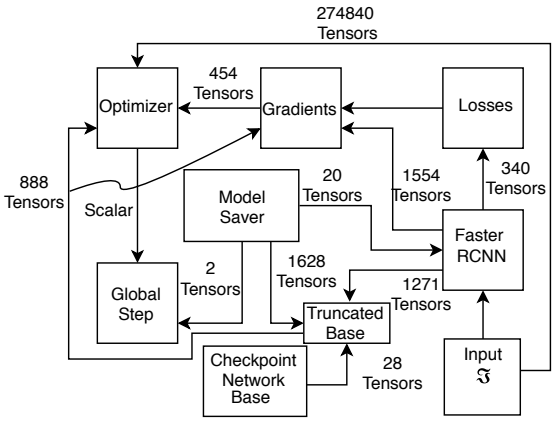

Figure 3.8: Graph Faster-RCNN.

Table 3.1: UNLV Dataset

Format	Total	Used	Train	Test
Image-pages $\mathbf{X}_i$	2889	403	338	65
Number of tables (n)	518	518	418	100

- Fast R-CNN classification loss function,
- Fast R-CNN regression loss function (Final Bounding Boxes).

3.4.2 Database

We use University Libraries (UNLV) dataset [45]. The details of this dataset are summarized in Table 3.1

The UNLV dataset is a scan-based image-pages set from magazines, newspapers, business letters and annual report. However, only 403 out of 2889 image-pages in this dataset contained tables; therefore, we split these 403 image-pages from the dataset into 338 image-pages for training and 65 for testing.

The ground truth ($\mathbf{g_i}$) [45] or label is the BB of each table contained on each image-page of the dataset. The ground truth annotations $\mathbf{g_i}$ are the reference to which we compare the predicted output of our model during training and evaluation on the tensorflow computation graph.

3.4.3 Evaluation metrics

3.4.3.1 Information Retrieval Metrics

We used purity [46], entropy [47] and information gain [48] to objectively evaluate our proposed table information retrieval system for VIP. The aforementioned metrics are commonly used in information retrieval systems to measure the quality of information

clusters [46]. As mentioned in Section 3.1, the fundamental problem faced by VIP is the segmentation of the information in electronic documents based on its visual structure, which is used by sighted people to partition an image-page $\mathbf{X}_i$ into clusters of information. Without the visual information for VIP, the document X is a source of unclassified words $\boldsymbol{w} = \{w_1, w_2, ...w_m\}$, in which m is the number of words in each page of the document. Formally, the entropy H of the discrete random variable $\boldsymbol{w}$, with mass probability $p(\boldsymbol{w}_i) = Pr\{\boldsymbol{w} = \boldsymbol{w}_i\}, \boldsymbol{w}_i \, \epsilon \, \boldsymbol{w}$, is defined as: [47]

$$H(\boldsymbol{w}) = -\sum_{i=1}^{k} p(\boldsymbol{w}_i) \log_2(p(\boldsymbol{w}_i)) \tag{3.5}$$

in which k is defined in our work as the number of tables in an image-page $\mathbf{X}_i$ plus one cluster of words that do not belongs to any table:

$$k = \begin{cases} n, & \text{Number of tables } (n) \text{ on the image-page } \mathbf{X}_i \\ n+1, & \text{otherwise.} \end{cases} \tag{3.6}$$

Depending on the image-page $\mathbf{X}_i$, n can take values from 0 to $\mathbb{N} = \{1, 2, 3, ...\}$, hence, $\boldsymbol{w}_i$ is the number of words contained in a table (n) of the image-page $\mathbf{X}_i$. Therefore, entropy $H(\boldsymbol{w})$ measures the uncertainty of a VIP to associate words in a determined table when listening to all the contents of image-page $\mathbf{X}_i$ through screen readers. Entropy is zero when VIP certainly knows the words that will listen through the screen reader, this only happens when VIP authored the digital document that is listening. In our work, the entropy ranged from 0 (lower bound) to $\log_2(k)$ (upper bound). Higher entropy makes harder to understand the digital content for a VIP. $H(\boldsymbol{w}) = \log_2(k)$ only when $p(\boldsymbol{w}_i) = 1/k$ for some $\boldsymbol{w}_i \, \epsilon \, \boldsymbol{w}$.

In this work, we do not compute the probability $p(\boldsymbol{w}_i)$ based on the frequency of words in the corpus of the document-page, because, although more frequent words are assumed to be more important in a document-page, this is not the case in tables.

Therefore, we assume a uniform distribution $U[w,h]$ on each image-page $\mathfrak{I}$ of the document X,

$$f(x, y) = \begin{cases} \frac{1}{\text{Area of } \mathfrak{I}}, & \text{if (x,y) } \epsilon \, \mathfrak{I} \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

in which:

$$p(\boldsymbol{w}_i) = \int_{y_{min}^{(n)}}^{y_{max}^{(n)}} \int_{x_{min}^{(n)}}^{x_{max}^{(n)}} f(x, y) dx dy = \frac{\text{Area of Table } (n)}{\text{Area of } \mathfrak{I}} \tag{3.8}$$

Information Gain $I(\boldsymbol{w}, \boldsymbol{w}_i)$, is the expected reduction in entropy caused by partitioning

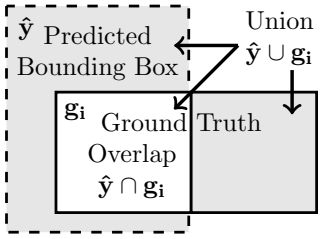

Figure 3.9: Intersection over Union.

the words according to the predicted tables of our TDS, as defined by [48]

$$I(\boldsymbol{w}, \boldsymbol{w}_i) = H(\boldsymbol{w}) - \sum_{i=1}^{k} p(\boldsymbol{w}_i) H(\boldsymbol{w}_i) \tag{3.9}$$

Purity is a measure of the quality of the clusters of information i.e., words contained in tables and in the rest of the document. When the tables are retrieved to CSV files, the VIP knows that the content information of these files is semi-structured by rows and columns and that the relationship between them is maintained. However, if tables are listening by a VIP on a digital document without prior knowledge of the existence of the tables the VIP will listen to the words of the tables mixed with the words of the rest of the document content, in this case, purity is zero because only one cluster of information is not appropriate to classify all the words. The purity is 1 (maximum) if each word gets its own cluster of information. Purity is defined as [46]

$$\text{Purity}(\boldsymbol{w}, \hat{\mathbf{y}}(n)) = 1 - \frac{1}{\text{Area of } \mathfrak{I}} \sum_{i}^{k} \underset{\text{n}}{max} |\boldsymbol{w}_i \cap \hat{\mathbf{y}}(n)| \tag{3.10}$$

3.4.3.2 Object Detection Metrics

We used Mean Average Precision (mAP) [49] and average Recall (AR) [33] metrics to evaluate the table detection model. In both, the lower bound is 0 and upper bound is 1, and higher values are better. mAP first computes the Average Precision (AP) for each class and then compute the mean of all AP. In this work, we only used one class, so mAP = AP. Both metrics are functions of intersection over union (IoU) [33], i.e., the Jaccard index defined as:

$$\text{IoU}(\hat{\mathbf{y}}, \mathbf{g_i}) = \frac{\text{area of overlap}}{\text{area of union}} = \frac{\hat{\mathbf{y}} \cap \mathbf{g_i}}{\hat{\mathbf{y}} \cup \mathbf{g_i}}, \tag{3.11}$$

Note that $\hat{\mathbf{y}}$ is the predicted BB and $\mathbf{g_i}$ is the ground truth annotation, as shown in Fig. 3.9. The notation **@ [0.5]** means the prediction is correct if IoU ≥ 0.5. Thus, precision and recall is defined as

Table 3.2: Results

Entropy	**Information Gain**	**Purity**
0.9315	0.5719	0.65037

$$\text{precision} = \frac{TP}{TP + FP}, \tag{3.12}$$

$$\text{recall} = \frac{TP}{TP + FN}, \tag{3.13}$$

in which true positives (TP) is the number of correct predictions, false positives (FP) is the number of incorrect predictions, both under a predetermined IoU, and false negatives (FN) is the number of ground truth annotations that do not have a prediction. Therefore, precision measures the correct prediction rate and recall measures how good the table detection model is at finding tables. Hence, AP is the average value of precision over different levels of recall, i.e., the area under the precision-recall curve (PRC)[33] and AR is the area under the recall-overlap curve[33]. The notation **@ [0.5 : 0.95]** represents the average of several mAP computed with different IoU, which ranges from 0.5 to 0.95 with a step size of 0.05.

3.5 Results

We evaluated our information retrieval metrics with the test dataset depicted in Table 3.1. The entropy of 0.9315 is the average entropy of the 65 image-pages $\mathbf{X}_i$ of the test dataset, an entropy of 0.9315 represented a high level of uncertainty experienced by VIP when they tried to understand the content in tables by only listening to the words of each image-pages $\mathbf{X}_i$. We achieved an average information gain = 0.5719, which represents the amount of uncertainty that can be reduced for VIP when the content of tables is extracted into independent CSV files. We also obtained an increase in the purity from zero to 0.65037 on the table's content of image-pages $\mathbf{X}_i$ of the test dataset. With respect to the object detection metrics, the results of the models trained over 40 epochs are presented in Table 3.3. We performed four experimental models to examine the effectiveness of different optimizers for training our table detection model. We can boost the performance of our table detection model by using a gradient descent optimizer, even though the model was trained with scan-based images. We can confidently use this model on text-based PDF documents because these kinds of documents have better image quality than our dataset, which are considered the worst case. In other words, in the worst case, using the model trained with a gradient descent optimizer, we can achieve a precision of 0.875 and a recall of 0.759 in low-quality image-based PDFs, as shown in Table 3.3.

Table 3.3: Average Precision (AP) and Average Recall (AR) Results

Optimizer	**AP@ [0.5]**	**AP@ [0.75]**	**AP@ [0.5:0.95]**	**Average Recall**
Momentum	0.805	0.635	0.550	0.723
Adam	0.207	0.025	0.077	0.362
Gradient Descent	**0.875**	**0.718**	**0.638**	**0.759**
Rmsprop	0.766	0.536	0.467	0.664

3.6 Discussion and Conclusion

We developed a system for the retrieval of table information from PDF documents for use by VIPs, we explained in detail the methodology that we used and how to objectively evaluate the approach through entropy, information gain, and purity metrics. Our results show that the presented approach can reduce the uncertainty (i.e. entropy) that a VIP has about the information contained in digital documents. Moreover, the best optimizer for table detection task was Gradient Descent, which allowed an AP of 87.5% to be achieved. We made a theoretical explanation of table detection that uses a deep learning approach.

Our table information retrieval system presents two improvements compared with traditional approaches of tagging text-based PDF. First, our approach does not require supervision by sighted people. Second, our system is capable of working with image-based as well as text-based PDFs.

In future work, this methodology can be extended for the development of a system for the retrieval of table information from elements that are more complex than text such as equations and graphs.

ACKNOWLEDGMENT

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001.

Chapter 4

Contribution II
Improving the Faster R-CNN Architecture for Table Detection

From the start, we based our proposed approach on the Faster R-CNN architecture of Girshick et al [39], a network intended for accurate localize objectives in an image, we used the image preprocessing proposed by [4] as a baseline.

4.1 Materials and Methods

In this section, we discuss the conditions used to create our table detection model. In Sec. 4.1.1, we present our hypothesis of use different pre-trained CNN. In Sec.4.1.2 we describe the dataset employed.

4.1.1 Deep Neural Network Architecture

For the models we coded on Python 3.6.8 and we used Keras and Luminoth backend. All experiments ran on Google Colaboratory Jupyter when we use UNLV Dataset[45]. We use Google Cloud Platform with 2 K80 GPU's for experiments with bigger Datasets.

Legacy Faster R-CNN architecture used a famous pre-trained Deep Convolutional Neural Networks (DCNN) named VGGNet[50] proposed in 2014. But from the 2014 there are new pre-trained DCNN architectures available, most of them trained on the mentioned the *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC)[32], that shown better performance in the Imagenet dataset. As shown in Fig. 4.1

Our hypothesis is that pre-trained DCNN with better performance on the Imagenet dataset will have better performance on our Table Detection task.

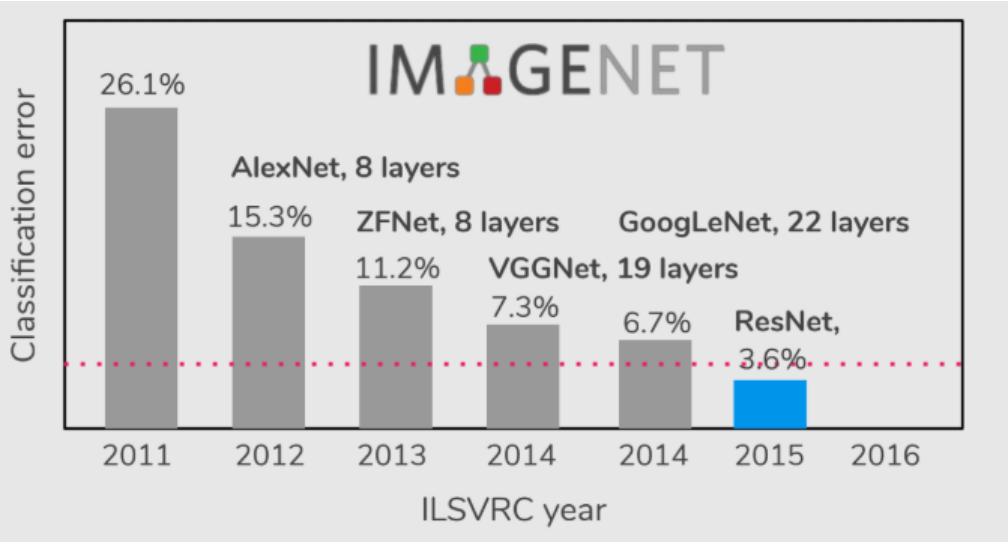

Figure 4.1: Pre-trained Deep Convolution Neural Networks (DCNN) [3].

4.1.2 Database

For training, in an early stage we use University Libraries (UNLV) dataset [45]. The details of this dataset are summarized in Table 4.1

Table 4.1: UNLV Dataset

Format	Total	Used	Train	Test
Image-pages $\mathbf{X}_i$	2889	403	338	65
Number of tables (n)	518	518	418	100

The UNLV dataset is a scan-based image-pages set from magazines, newspapers, business letters and annual report. However, only 403 out of 2889 image-pages in this dataset contained tables; therefore, we split these 403 image-pages from the dataset into 338 image-pages for training and 65 for testing.

The ground truth ($\mathbf{g_i}$) [45] or label is the BB of each table contained on each image-page of the dataset. The ground truth annotations $\mathbf{g_i}$ are the reference to which we compare the predicted output of our model during training and evaluation on the tensorflow computation graph.

But we know, that supervised machine learning models can be improved with more data. Our hypothesis is that models trained with a larger dataset have better performance and less risk of overfitting than models trained with small datasets. We found the following publicly available image-page datasets for table detection.

ICDAR Page Object Detection Dataset 2017 has annotations for the task of detecting tables, figures and mathematical equations. The data set consists of 2417 images in total. Where the training data set has 1600 images and 817 images are used for testing. For our table detection model, we only use table annotations.

MORMOT Dataset [51] published by the Institute of Computer Science and Technology (Peking University), the dataset has 2000 image-pages, but we found several incorrect ground-truth annotations in this dataset. Hence, we finally do not use this dataset.

TableBank Dataset [52] consists of 417 234 high quality labeled tables as well as their original documents. Created from word and latex documents.

Besides using large datasets, we employed data augmentation, with up to 10% horizontal and vertical shifts, up to 20% zoom, and up to 270 degrees rotation

4.1.3 Image Preprocessing

Deep Convolutional Neural Networks architectures explicitly assume that the input is an RGB image and often is used as input raw images without image preprocessing, but we noticed that public available datasets such as the mentioned in the Section 4.1.2 are in English language and we were working to create models to detect and localize, tables in different languages. We found one solution to the mentioned problem, using image preprocessing proposed on [4]. As Shown in Fig.4.2

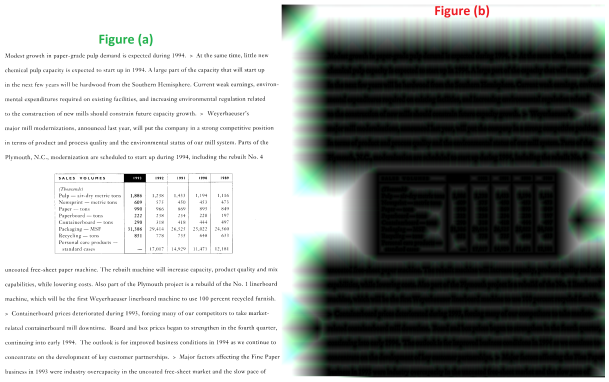

Figure 4.2: a) Dataset Original (Right Image), b) Dataset Post-processed using image preprocessing proposed on [4].(Left Image)

However, we notice that using ChannelMerge proposed on [4] introduces noise, which is learned by our model on the training stage.

4.2 Proposed Image Preprocessing

In order to improve image preprocessing, we proposed use the following transformations:

1. Conversion to a grayscale image.

2. Euclidean Distance Transform (EDT)

As shown in Fig. 4.3

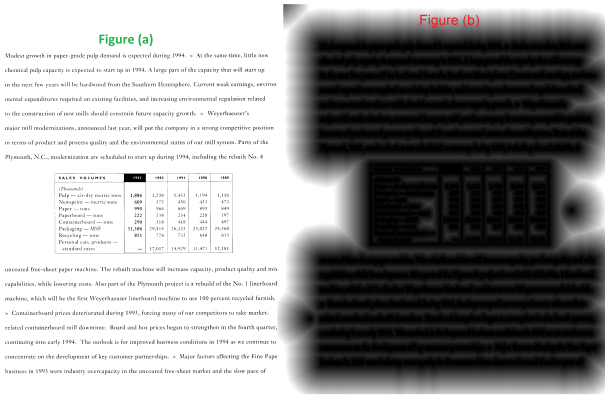

Figure 4.3: a) Dataset Original (Right Image), b) Dataset Post-processed using our proposed image preprocessing.(Left Image)

Due to, deep convolutional neural networks architectures generally are pre-trained with RGB images, we proposed to use the grayscale image channel as the input of each channel RGB of the DCNN architecture instead of using ChannelMerge.

We tried another image transformation before entry the image, but it was slower and did not improve the results.

4.3 Experiment Setup

Our first attempt was a model based on legacy Faster R-CNN architecture (i.e., using VGG network [50]) using image preprocessing proposed by [4] and using UNLV Dataset [45]. We got our best results with the network proposed in Chapter 3 but with TableBank Dataset [52]. We also used 0.5 dropout everywhere, and we training the model during 40 epochs.

Finally we decided to concentrate our efforts in ResNet-101 model. Resnet-101 is a state-of-art DCNN, and is available in Keras and Tensorflow framework, pre-trained for ImageNet with good results.

We tried validation and early stopping, but there was no impact in the results.

4.4 Computational Resources

Training Deep learning models requires as much computational horsepower as possible, i.e., large-memory and GPUs. We used for trainning our final model Google Cloud Platform, with two k80 GPUs.

4.5 Results

We evaluated our very first table detection model using the legacy Faster R-CNN architecture and the image proposed on [4], and we noticed a slight improvement in our table detection model using our proposed approach, All models were trained and tested with dataset depicted in Table 4.1. All models were trained over 40 epochs and the results are presented in Table 4.2.

Table 4.2: Average Precision (AP) and Average Recall (AR) Results using our Proposed Image Processing

Image Processing	**AP@ [0.5]**	**AP@ [0.75]**	**AP@ [0.5:0.95]**	**Average Recall**
Using ChannelMerge	0.605	0.483	0.550	0.638
Proposed Approach	**0.65**	**0.535**	**0.648**	**0.634**

Our best results with the network proposed in Chapter 3 but using TableBank Dataset [52] is presented in Table 4.3.

Table 4.3: Average Precision (AP) and Average Recall (AR) Results

Optimizer	**AP@ [0.5]**	**AP@ [0.75]**	**AP@ [0.5:0.95]**	**Average Recall**
Gradient Descent	**0.935**	**0.881**	**0.815**	**0.976**

4.6 Discussion and Conclusion

We tested the presented hypothesis in this chapter, we found that there our best model was obtained using state-of-the-art DCNN Resnet-101 and larger dataset [52]. However, our best model obtains those results only in text-based PDF, and to simplify our implementation in Chapter 5 i.e., avoid uncertainties produced by image-based or scanned-based PDF documents, such as analyze Archival Documents as is shown in Fig.4.4. We limited the proof of concept of our proposed approach in Chapter 5 to make accessible PDF documents, using our proposed approach to text-based PDF documents. Fig. shows how our table detection model is very close to our ground-truth.

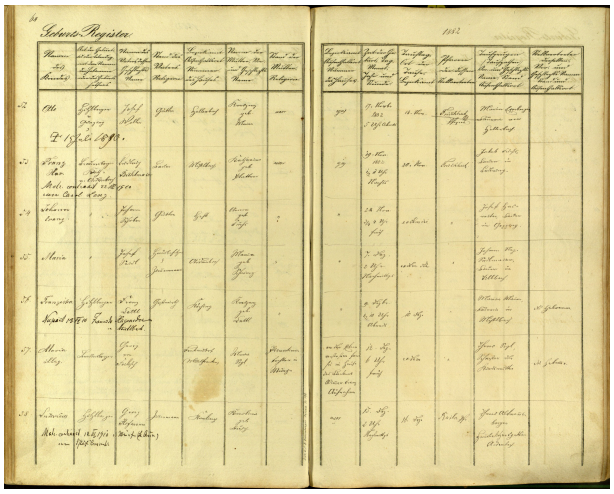

Figure 4.4: Uncertainties of PDF-based on images or scanning

TABLE 4: Monthly Frequency of Sex

(Active females, ages 15 to 19)

# of times	distribution
0	0.3890
1	0.0945
2 to 3 (= 2.5)	0.1604
4 to 7 (= 5.5)	0.1495
8+ (= 9)	0.2066
Mean = 3.177	

10.2 A Framework for Studying Frequency

To model the above facts, change the term in the utility function involving sex to

$$\ln\{\widetilde{j}\exp(\chi f^{\iota}/\iota - \chi/\iota)\} = \ln\widetilde{j} + \chi f^{\iota}/\iota - \chi/\iota, \text{ with } \iota < 0 \text{ and } \chi > 0, \tag{20}$$

where f represents the frequency of sex and $\widetilde{j}$ now denotes the joy from it. Let the cost of sex be given by

$$\widetilde{c} = 1 - p^{f}, \tag{21}$$

where p is the odds of having a safe sexual encounter. Observe that $1 - p^{f}$ is the probability of becoming pregnant, or the failure rate, given the frequency of sex f. The cost function is increasing and *concave* in f, since

$$\frac{d\widetilde{c}}{df} = -(\ln p)\, p^{f} > 0 \text{ and } \frac{d^{2}\widetilde{c}}{(df)^{2}} = -(\ln p)^{2} p^{f} < 0,$$

where the signs of the above expressions follow from the fact that $0 \leq p \leq 1$. Therefore, while the chances of getting pregnant increase with the frequency of sex, they do so at a diminishing rate.

Cast an individual's decision regarding the frequency of sex as follows:

$$\max_{f}\{\ln\widetilde{j} + \chi f^{\iota}/\iota - \chi/\iota - 1 + p^{f}\}.$$

Figure 4.5: Table Detection Model, Prediction vs Ground-Truth

Chapter 5

Contribution III
Table IR for VIPS: A proof of concept via an integrattion with dosvox

5.1 Table information retrieval system for visually impaired people software prototype

Using DOSVOX[10] system and the proposed approach en Chapter 3, we developed a software prototype as shown Fig. 5.1

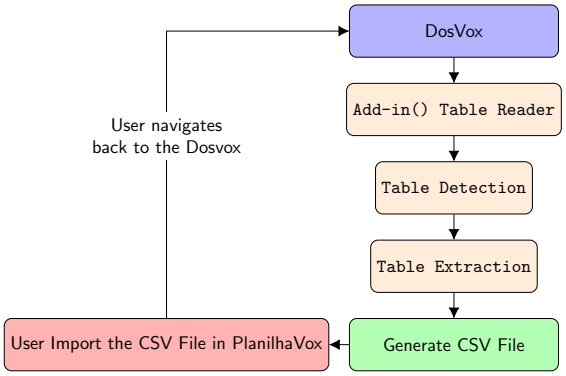

Figure 5.1: WorkFlow

5.2 Dosvox Integration

Dosvox is a software application developed in Delphi 6 and our models and algorithms were developed in python. Hence, we developed an add-in in delphi for DOSVOX, in order to integrate our code developed in Python with the DOSVOX.

5.3 Description of how works our Proposed Software Prototype

To use our proposed approach to make PDF document tables accessible, the VIP must follow the steps below.

Step 1) The VIP in DosVox must select the “Archives ” option as shown in Fig.5.2

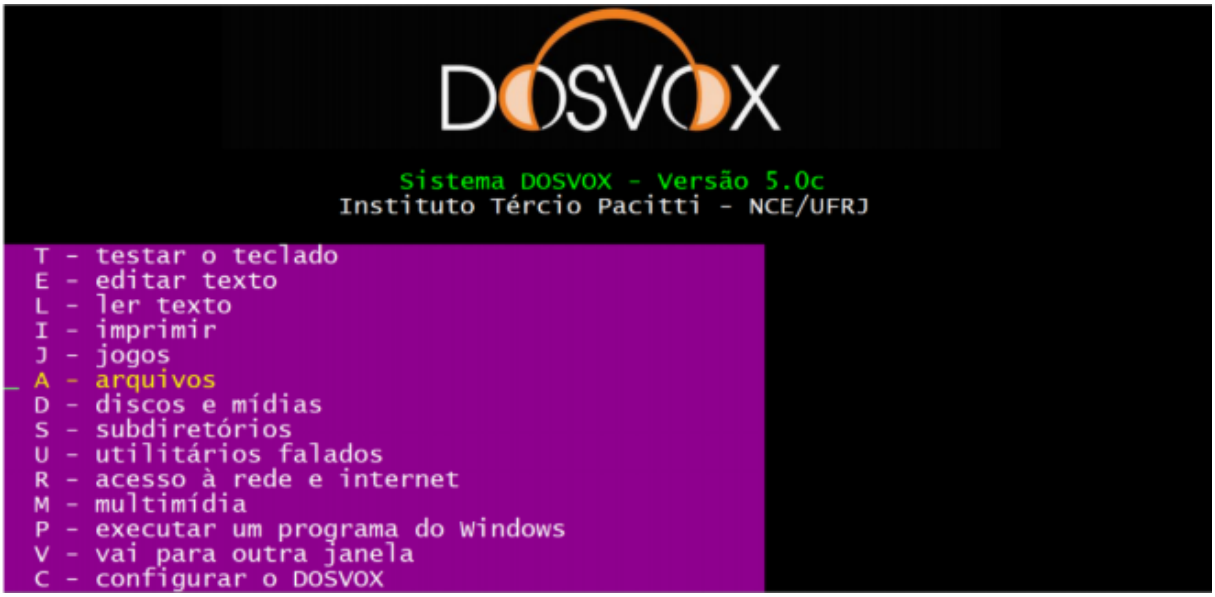

Figure 5.2: Navigating through Files in DosVox

Step 2) The VIP user must select the document that requires the tables to be extracted. As shown in Fig. 5.3

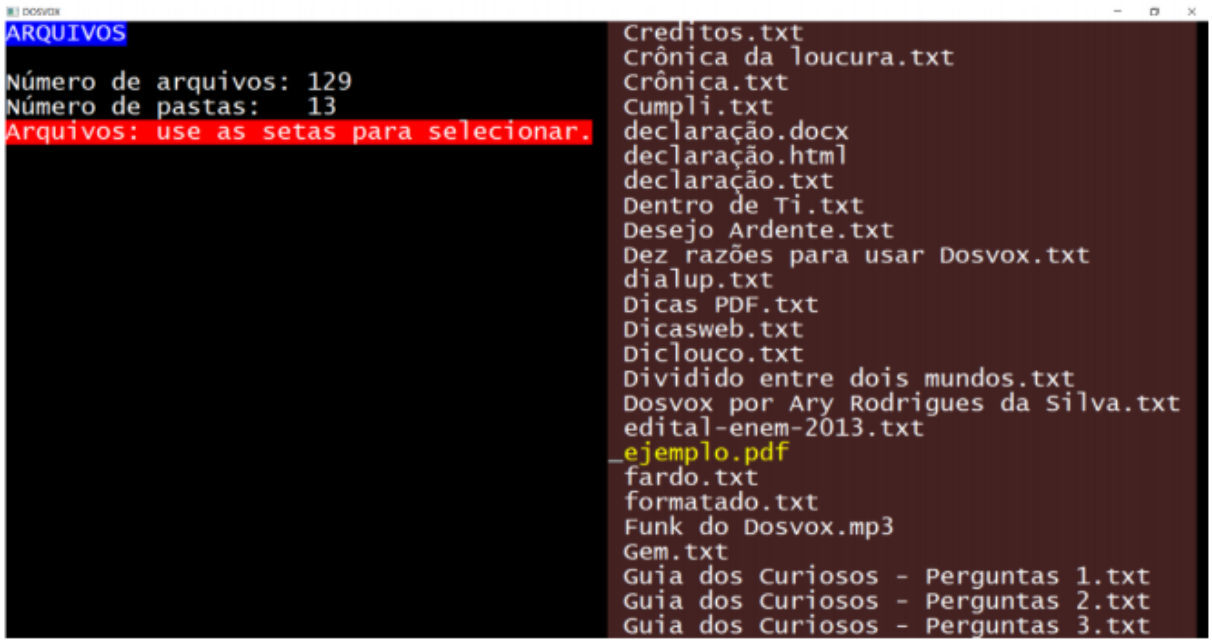

Figure 5.3: Selecting the PDF file

Step 3) VIP user must hit key H to execute in the background of our proposed approach. As shown Fig. 5.4. In this step, we reproduce a sound with the message “Leitura de tabelas iniciada. Por favor aguarde”.

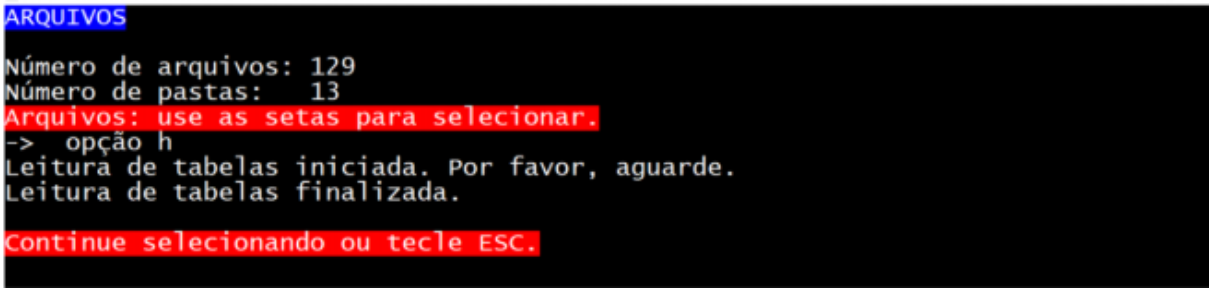

Figure 5.4: Executing our proposed approach

Step 4) In the same directory of the selected PDF file, our proposed approach will create a csv file for each table of the document, autonomously. The names of the csv files will begin with the name of the original PDF file (i.e., filenamePDF-page-2-table-1.csv). As shown Fig. 5.5

Figure 5.5: Output of our proposed approach

Finally, we reproduce a message “Leitura de tabelas finalizada”.

5.4 Limitations of our Prototype

- Our prototype will only extract information from unstructured data from tables in PDF documents to CSV files, this system does not attempt to understand the meaning of the table or its semantic structure.

- Currently, our prototype works only on text-based PDF documents, in future work we will extend this prototype to scan-based PDF and image-based PDF.
- There is also another problem related to the underlying PDF text sequence in some cases, for example, equations and graphics, the text may not be the same as the visible one, in such rare cases, the information extracted by the algorithm may not match the visible text.
- Currently, our prototype analyzes only one PDF file at a time and only the information that is explicitly provided in the document can be extracted.
- Our proposed approach can not infer any new information based on the content.
- We know that tables contain less frequent words, therefore, our proposed approach does not exclude any text contained in the table area.
- Our proposed approach is able to work with documents written in Western languages such as Portuguese, English, and Spanish, we really do not guarantee the functionality of our prototype with oriental languages.

5.5 Conclusion

We developed prototype software for the retrieval of table information from PDF documents for use by VIPs compatible with DOSVOX as a proof of concept of our proposed approach.

Chapter 6

Discussion

The results of this work have shown that object detection models can be trained to retrieve valuable information for VIP, specifically to supporting in the accessibility of digital documents such as PDF files. However, is required more researches and developments, to develop a robust solution that provides full accessibility of digital documents to VIP. The major problem to make accessible PDF documents is that we do not have structured data, in fact, the text of PDF documents and visual elements are not properly related, which makes that developers of assistive technologies and VIP, can not recognize, how many tables there are in each page easily, or other elements such as equations, figures, etc.

6.1 Pattern Recognition

In an early stage of this work we found that identify patterns of tables are challenging tasks due to the diversities of tables layouts[53] and its inter-related data.

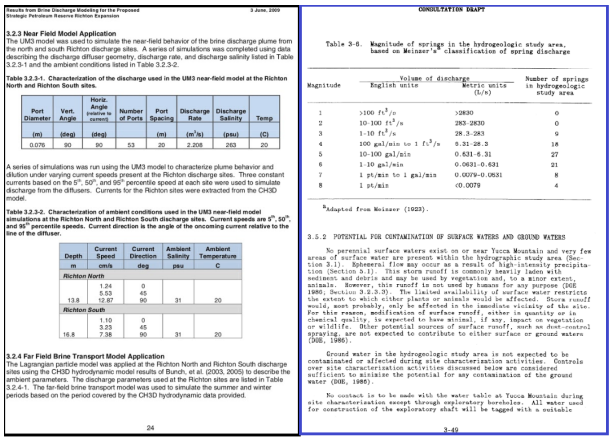

Figure 6.1: Different Types of Tables

As shown in Fig.6.1, there are different types of tables (i.e., with border, without border, etc), and the authors of digital documents do not want to limit themselves to a

single standard table. In fact, something interesting that we notice in this paper is that there is no single formal definition of what is a table. The most accepted definition of a table we found is "A table is an object which uses linear visual cues to simultaneously describe logical connections between the discrete content entries in the table. A content entry is the basic component of information in the table and this can be any visual symbol"[54].

The two main elements behind this definition are that there exist simultaneously linear visual clues in the form of columns and rows that represent logical connections; as such there is a relationship between the relative position of the items and their conceptual relationship[28].

Wang [55] and Hurst [56] defines that a table is divided into four main regions:

- The stub that contains the row- and subheaders.
- The boxhead that contains the column headers.
- The stub head that contains the header for the stub.
- The body that contains the actual data of the table.

As is well known, not every table has all the four regions (i.e.,we have uncertainties about this definition), but these elements typically should be stay associated with each other, otherwise, it would be a list instead of being a table.
Figure 6.2 shows all the regions and terms definitions of the mentioned definition, as we can see, an element is defined as a single word or a number, while a cell can contain multiple elements, and a block multiple cells[5].

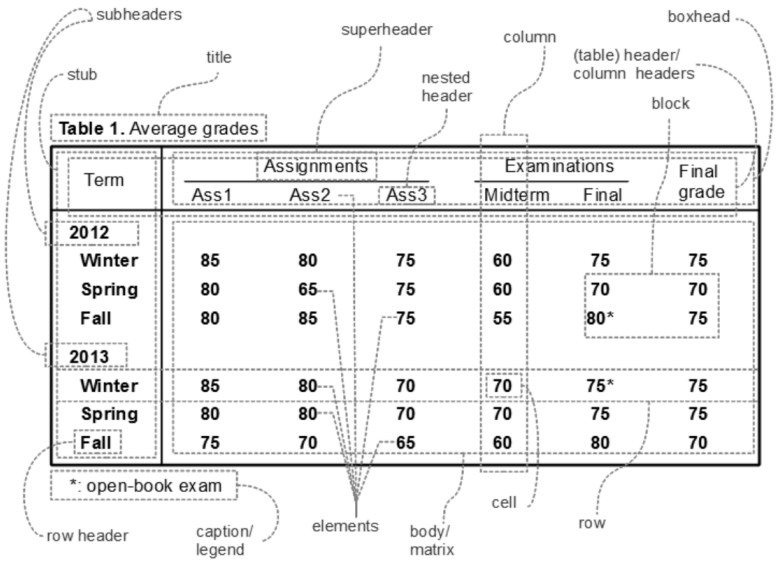

Figure 6.2: Table anatomy, terms and definitions of table elements[5].

Due to this fact, about the uncertainties of tables, i.e., uncertainties about the number of rows, number of columns, uncertainties about the type of elements in the table (i.e., we do not know, if the element is a text, equation, figure, etc), is extremely cumbersome try to develop a ruled-based software that analyze the PDF language (a variation of postscript language) to find patterns in PDF documents. And we can see these problems in the multitude of methods currently available to detect tables in PDF documents [26, 27, 28, 29], where there is no author who proposes a general deterministic solution to identify tables in PDF documents[25]. We should note that the PDF language, in general, in a text-based PDF document provides little information to try to retrieve a table, such as the position of each character on a page, but the problem is that a document can be a column document single or double, it can contain several tables or a single table, generally the style of the document can provide a different typeface, which means more uncertainties. When the PDF document is an image-based or scanned-based PDF document, the information that can be obtained from the PDF language is extremely poorly, because we only that is an image. When the PDF document is a scanned or image-based PDF document, the information that can be obtained from the PDF language is extremely poor, because what we can only know using the PDF language is that there is an image in the PDF on a certain page of the document. In the state-of-the-art, we notice that some authors recommend translating all tables to sentences [9], due to the limited information that we can obtain from the PDF language, explained before. However, this approach is not feasible, because most authors do not know about visual impaired people accessibility problem, and it is well known that tables are widely used to summarize the results. Therefore, motivated by the recent success of deep learning in the field of computer vision [57] and Marrakesh treaty[16] which established a set of limitations and exceptions to traditional copyright law, to adopt digital documents to friendly formats for VIP. In this work, we used a supervised learning approach of machine learning to create a model that learn the patterns about how to identify and localize tables in an image-page using Faster R-CNN architecture.

Chapter 7

Conclusion and Future Work

7.1 Conclusions

In this dissertation, we investigated how to incorporate computer vision techniques using a deep learning approach to tackle several accessibility challenges of broad interest among the visually impaired people and machine learning comunity, such as identification of document elements (i.e., tables), document elements localization, and information retrieval through computer vision techniques. Moreover, we presented a practical application of our proposed approach to visually impaired people as proof of concept in a software prototype.

For all of our contributions, the construction of a table detection model to retrieve information of the visual structure of the document has proven to lead to a powerful improvement towards a generalized approach to make accessible digital documents of tables due to we are capable of incorporating prior knowledge of spatial visual structure to clustering the words of our interest.

Moreover, the advantage of constructing a table detection model using our proposed image preprocessing approach is the use of information from publicly available datasets to improve the overall performance of our proposed methodology.

The results show that our proposed methodology can be used to reduce the uncertainty experienced by visually impaired people when listening to the contents of tables in digital documents through screen readers.

The results showed that our proposed approach using a deep learning approach has outperformed other rule-based approaches basically because the problem of table detection is a non-linear problem, and there are several uncertainties that cannot be modeled deterministically. Therefore, using a supervised learning approach we can create a model that found this pattern through experience.

Our best table detection model was obtained using Gradient Descent, ResNet-101 convnet, training with larger dataset (i.e., TableBank Dataset[52])and using our proposed

image preprocessing approach.

7.2 Future Works

In this section, we conjecture about some future research directions that will be considered based on the issues addressed in this dissertation

1. In this work, we proposed only an image preprocessing, that basically dilate pixels of input image. To take advantage of data sets from English to other Western languages, we note that our proposed preprocessing approach improves the table detection model.

2. A future work might explore some ad-hoc heuristics algorithms to deal with low-quality image-based or scan-based PDFs, in order to enhance the quality of the image that will be the input of our Faster R-CNN network.

3. In this work, we found an improvement in the accuracy of table detection models using a Deep Convolutional Neural Network that has better performance on ImageNet challenge.

4. A future work might explore the use of an ensemble of Deep Convolutional Neural Networks to improve the accuracy of table detection model.

5. In this work, we proposed a table information retrieval system for VIP, which basically localizes tables and extracts the text table information to spreadsheet software for VIP.

6. In future work, this methodology can be extended for the development of a system for the retrieval of table information from elements that are more complex than text such as equations and graphs.

7. In this work, we developed a table information retrieval system for VIP software prototype.

8. In future work, this prototype can be tested with a group of VIP for a long time of use, in order to improve this prototype with the feedback of VIP.

9. In future work, can be proposed an autonomous system for fully tagging PDFs using object detection approach.

References

[1] RCNN Fast. Object recognition for dummies part 3: R-cnn and fast/faster/mask r-cnn and yolo. *Lil'Log*, 2018.

[2] Gunilla Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34:344–371, June 1986.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Azka Gilani, Qasim, Shah Rukh, Imran Malik, Shafait, and Faisal. Table detection using deep learning. In *Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2017.

[5] Anssi Nurminen. Algorithmic extraction of data in tables in pdf documents. Master's thesis, TAMPERE UNIVERSITY OF TECHNOLOGY, 2013.

[6] Braithwaite T Cicinelli MV Das A Jonas JB et al.; Vision Loss Expert Group Bourne RRA, Flaxman SR. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *Lancet Glob Health*, 5(5):e888–e897, September 2017.

[7] VK Ahuja. Marrakesh treaty to facilitate access to published works for visually disabled: Putting an end to global book famine. In *Copyright Law in the Digital World*, pages 97–107. Springer, 2017.

[8] General Assembly. The standard rules on the equalization of opportunities for persons with disabilities, 1993.

[9] Alireza Darvishy. PDF accessibility: tools and challenges. In *Computers Helping People with Special Needs*, pages 113–116, 2018.

[10] NCE UFRJ. Projeto dosvox. *Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro, http://intervox. nce. ufrj. br/dosvox*, 2010.

[11] Projeto dosvox. http://intervox.nce.ufrj.br/dosvox/. Accessed: 2018-11-10.

[12] Paúl Hernán Mejía Campoverde and Luiz César Martini. Calculadora financiera finanvox: Herramienta informática educativa de apoyo para deficientes visuales en su proceso de formación académica. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, 2011.

[13] HM Silveira and LC Martini. Matvox: um aplicativo para deficientes visuais que proporciona a implementação de algoritmos e cálculos matemáticos em um editor de texto. *XXI Simpósio Brasileiro de Informática na Educação, João Pessoa–PB*, 2010.

[14] Alireza Darvishy, Mark Nevill, and H Hutter. Automatic paragraph detection for accessible pdf documents. In *Computers Helping People with Special Needs*, 2016.

[15] Electronic document file format enhancement for accessibility, 2012.

[16] Zemer, Lior and Gaon, Aviv. Copyright, disability and social inclusion: the Marrakesh Treaty and the role of non-signatories. *Journal of Intellectual Property Law & Practice*, 2015.

[17] Núcleo de Computação Eletrônica UFRJ and Núcleo de Computação Eletrônica Universidade Federal do Rio de Janeiro. Projeto DOSVOX, 2018.

[18] Camelot : Pdf table extraction for humans. https://github.com/atlanhq/camelot. Accessed: 2019-09-30.

[19] Peter Wlodarczak. *Machine Learning and Its Applications*. CRC Press, 2019.

[20] International Organization for Standardization (ISO). Iso 32000-1: 2008 document management–portable document format–part 1: Pdf 1.7., 2008.

[21] Leonard Rosenthol. *Developing with PDF: Dive Into the Portable Document Format*. " O'Reilly Media, Inc.", 2013.

[22] Anna Irene Oates. Navigating the PDF/A standard: a case study. Master's thesis, University of Oxford, Urbana, Illinois, 2018.

[23] Tiziana Armano, Anna Capietto, Sandro Coriasco, Nadir Murru, Alice Ruighi, and Eugenia Taranto. An automatized method based on LaTeX for the realization of accessible PDF documents Containing Formulae. In *Computers Helping People with Special Needs*, pages 583–589, 2018.

[24] Z. Zhao, P. Zheng, S. Xu, and X. Wu. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–21, 2019.

[25] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR*

International Conference on, volume 1, pages 1162–1167. IEEE, 2017.

[26] Sekhar Mandal, SP Chowdhury, Amit Kumar Das, and Bhabatosh Chanda. A complete system for detection and identification of tabular structures from document images. In *International Conference Image Analysis and Recognition*, pages 217–225. Springer, 2004.

[27] R Zanibbi, D Blostein, and JR Cordy. A survey of table recognition: Models, observations, transformations, and inferences, 2003. *Online: http://www. cs. queensu. ca/˜ cordy/Papers/IJDAR_ Tables. pdf, Last Checked*, pages 12–01, 2007.

[28] Ana Costa e Silva, Alípio M Jorge, and Luís Torgo. Design of an end-to-end method to extract information from tables. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(2-3):144–171, 2006.

[29] Shah Khusro, Asima Latif, and Irfan Ullah. On methods and tools of table detection, extraction and annotation in pdf documents. *Journal of information science*, 41(1):41–57, 2015.

[30] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. Icdar 2013 table competition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1449–1453. IEEE, 2013.

[31] S. A. Siddiqui, M. I. Malik, S. Agne, A. Dengel, and S. Ahmed. DeCNT: Deep Deformable CNN for Table Detection. *IEEE Access*, 6:151–161, 2018.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[33] Junwei Han, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1):84–100, 2018.

[34] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010.

[35] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014.

[36] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages

3203–3212, 2017.

[37] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[38] Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2015.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[41] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Distance Transforms of Sampled Functions. *Theory of Computing*, 2012.

[42] M. Göbel, T. Hassan, E. Oro, and G. Orsi. Table Competition. In *2013 12th ICDAR*, pages 1449–1453, August 2013.

[43] Anssi Nurminen. Algorithmic Extraction of Data in Tables in PDF Documents. Master's thesis, Tampere Univ. of Tech., Tampere, Finland, 2013.

[44] Aslam Muhammad, Warda Ahmad, Maryam Tooba, and Sidra Anwar. Assistive Technology for Disabled Persons. In *Atlantis Press*, November 2015.

[45] Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. An open approach towards the benchmarking of table structure recognition systems. In *Proc. IAPR Int Workshop on Document Analysis Syst.*, pages 113–120. ACM, 2010.

[46] Anna Huang. Similarity measures for text document clustering. In *Proc New Zealand Comput. Sci. Research Student Conf.*, volume 4, pages 9–56, 2008.

[47] Neepa Shah and Sunita Mahajan. Document clustering: a detailed review. In *International Journal of Applied Information Systems*, pages 30–38. Citeseer, 2012.

[48] Ivo D Dinov. Decision tree divide and conquer classification. In *Data Science and Predictive Analytics*, pages 307–343. Springer, 2018.

[49] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Localization recall precision (lrp): A new performance metric for object detection. In *Proc. of the*

European Conf. on Computer Vision (ECCV), pages 504–519, 2018.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[51] Jing Fang, Xin Tao, Zhi Tang, Ruiheng Qiu, and Ying Liu. Dataset, ground-truth and performance metrics for table detection evaluation. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 445–449. IEEE, 2012.

[52] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. *arXiv preprint arXiv:1903.01949*, 2019.

[53] Bill Venables. Introduction to data technologies. by paul murrell. *Australian & New Zealand Journal of Statistics*, 52(4):469–470, 2010.

[54] JP Cameron. A cognitive model for table editing. Technical report, Technical report OSU-CISRC6/89-TR 26, Computer and Information Science Research Centre, Ohio State University, USA, 1989.

[55] Xinxin Wang. *Tabular abstraction, editing, and formatting.* PhD thesis, University of Waterloo, 1996.

[56] M Hurst. The interpretation of tables in texts phd thesis. *University of Edinburgh, School of Cognitive Science, Informatics*, 2000.

[57] Geert Litjens, Thijs Kooi, Babak Bejnordi, Arnaud Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Van Der Laak, Bram Van Ginneken, and Clara Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.