VISUAL INFORMATION RETRIEVAL FROM HISTORICAL DOCUMENT IMAGES

by

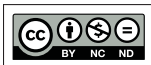
Sara ZHALEHPOUR

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS IN AUTOMATED MANUFACTURING ENGINEERING
M.A.Sc.

MONTREAL, AUGUST 10, 2018

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

Sara ZHALEHPOUR, 2018



This Creative Commons license allows readers to download this work and share it with others as long as the author is credited. The content of this work cannot be modified in any way or used commercially.

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Mohamed Cheriet, Memorandum Supervisor
Department of génie de la production automatisée, École de technologie supérieure

Ms. Sylvie Ratté, President of the Board of Examiners
Department of génie logiciel et des technologies de l'information, École de technologie supérieure

Mr. Luc Duong, Member of the jury
Department of génie logiciel et des technologies de l'information, École de technologie supérieure

Mr. Ching Yee Suen, External Examiner
Department of computer science and software engineering, Concordia university

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON AUGUST 1TH, 2018

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincerest thanks to my supervisor Prof. Mohamed Cheriet for his permanent guidance, continuous support and encouragement during my master study.

I am also very grateful to Prof. Ratté, Prof. Duong, and Prof. Suen, my board of examiners, for giving me the honor of accepting to judge my thesis.

My thanks also go to all of my friends and lab mates at Synchromedia laboratory, especially, Atena, Hossein, Sadegh, Sherif, Ehsan, Anna, Athmane and Saida, who kindly supported me. I am also very grateful to Shahin for introducing me to the Synchromedia laboratory.

Without question, none of this would have been possible without my wonderful family. I owe my profound gratitude to my parents for their unconditional love and support throughout my life. And special thanks to my brother, Sasan, for always being there for me no matter what, where or when. I would also like to thank my fiance, Jeremie, for his enduring support and patience and more importantly for always encouraging me to strive for the best.

To my family

LA RECHERCHE D'INFORMATION VISUELLE À PARTIR D'IMAGES DE DOCUMENTS HISTORIQUES

Sara ZHALEHPOUR

RÉSUMÉ

Au cours des dernières décennies, la préservation et la publication de documents historiques en format numérique ont fait l'objet d'une attention considérable. Bien que les techniques modernes de numérisation aient pour la plupart résolu le problème de la protection et de l'accès à ces documents, la tâche de la recherche et de l'interprétation de l'information visuelle demeure un problème difficile. Cela est dû aux structures complexes et inhabituelles des documents historiques en plus de leur nature dégradée. Pour la recherche d'informations à partir de documents historiques, une approche appropriée est nécessaire pour caractériser le contenu du document de manière cohérente. Les documents imprimés contiennent non seulement des caractères de texte et leurs formats, mais également des éléments typographiques associés. Trouver et poursuivre les objets typographiques visuels existants qui façonnent le contenu des documents historiques nous aide à récupérer et à transmettre plus d'informations sur les différentes méthodes de représentation de ces documents. Ces éléments peuvent être des notes de bas de page qui font référence aux auteurs et démontrent la relation entre les manuscrits et les sources, ou des tableaux qui résument différentes sortes d'informations dans des formes géométriques. Cette recherche se concentre sur le problème de la détection des notes de bas de page et des tableaux dans les documents historiques et établit un cadre pour chacun des objectifs déterminés. Ces cadres doivent gérer efficacement des structures complexes de documents historiques et posséder en même temps le pouvoir de généralisation pour s'appliquer aux collections d'images de documents à grande échelle. Jusqu'à maintenant, la détection des notes de bas de page a rarement été abordée dans la littérature.

Par conséquent, notre première contribution est de présenter un nouveau cadre pour la classification des images de documents basée sur les notes de bas de page dans les documents historiques. L'idée de base derrière ce cadre est d'utiliser les caractéristiques visuelles les plus importantes d'une note de bas de page pour créer un vecteur de caractéristiques. Les trois caractéristiques visuelles les plus importantes d'une note de bas de page sont la taille réduite de la note par rapport au corps du texte, l'emplacement de la note au bas de la page et l'écart relativement plus important entre la note et le corps du texte comparé à l'espace de ligne standard. Trois méthodes sont proposées en fonction de chacune de ces observations. Nous définissons certaines règles à l'aide de ces observations pour créer notre vecteur de caractéristiques. Notre cadre pour la classification des images de documents basée sur les notes de bas de page dans les documents historiques est complété en alimentant par ces vecteurs caractéristiques un classificateur de machine à vecteurs de support (SVM). Le cadre proposé est appliqué à plus de 32 millions d'images du 18 ème siècle. Les résultats de l'évaluation prouvent l'efficacité, la puissance de généralisation et la robustesse de notre cadre présenté pour détecter les pages contenant une note de bas de page malgré leur mise en page et leur type de structure.

Les méthodes les plus récentes de détection de tableaux dans les documents utilisent principalement des documents de balisage (par exemple, pdf, HTML, etc.) et ne couvrent pas tous les types de tableaux dans un cadre. Cependant, pour les documents historiques, qui sont notre principale étude dans le cadre de cette thèse, nous avons seulement accès à l'image numérisée et devons traiter tous les types de tableaux en même temps. Le cadre proposé est basé sur l'hypothèse que les textes dans les tableaux se présentent de manière harmonique en colonnes. Ce fait suggère l'idée d'utiliser une méthode spectrale pour développer notre cadre. Nous proposons une approche basée sur l'utilisation de coefficients cepstraux de fréquence de Mel (MFCC) pour classer des images de document selon la présence ou non de tableaux dans la page. Les MFCC sont des procédures bien connues de reconnaissances automatiques de la parole. Ces méthodes mettent l'accent sur la fréquence basses du signal. Un classificateur SVM est utilisé comme dernière étape de notre framework pour détecter les pages contenant des tableaux. Nous testons le cadre introduit sur nos ensembles de données et les résultats confirment l'efficacité de la méthode proposée par rapport à la fois à une méthode reconnue et à notre ensemble de données de référence des documents imprimés du $18^{\text{ème}}$ siècle.

Mots-clés: La recherche d'information visuelle, Document historique image, Classification des documents, Détection de note,Boîtes de délimitation, Projection horizontale, Détection de tableaux, Tableaux de lignes de régnant, Tableaux de lignes Non-régnant, MFCC, Machines à vecteurs de support.

VISUAL INFORMATION RETRIEVAL FROM HISTORICAL DOCUMENT IMAGES

Sara ZHALEHPOUR

ABSTRACT

In the recent decades, preserving and publicizing historical documents in digital format has gotten considerable attention. Although modern digitizing techniques have mostly solved the problem of protecting and accessing these documents, the task of visual information retrieval and interpretation is still an arduous issue. This is due to historical documents' complex and unusual structures beside their degraded nature. For information retrieval from historical documents, an appropriate approach is required to characterize the document content in a coherent way. Printed documents contain not only text characters and their formattings but also some associated typographical elements. Finding and pursuing the existing visual typographical objects that shape the content of historical documents, helps us retrieve and convey more information about the various methods of representing these documents. These elements can be footnotes that connect the authority and demonstrate the relationship between manuscripts and sources, or tables that summarize different sort of information into geometric forms. This research focuses on the problem of detecting footnotes and tables in historical documents and establishes a framework for each of the driven objectives. These frameworks must efficiently handle complex structures of historical documents and at the same time possess the generalization power to be applied to large-scale document image collections.

To the best of our knowledge, up to this date, footnote detection has rarely been addressed in the literature. Therefore, our first goal is to present a novel framework for footnote-based document image classification in historical documents. The basic idea behind this framework is to utilize the most prominent visual features of a footnote to create a feature vector. The three most notable visual features of a footnote in a page are the smaller font size of the footnote respect to the body text, the footnote location at the bottom of the page and the relatively greater gap between the footnote and the body text compared to the standard line space. Three methods are proposed according to each of these observations. We define a set of rules using these observations to create our final feature vector. Our framework for footnote-based document image classification in the historical documents is completed by feeding these feature vectors to a support vector machine (SVM) classifier. The proposed framework is applied to more than 32 million images from 18^{th} century. The evaluation results prove the efficiency, generalization power, and robustness of our presented framework for detecting page containing footnote despite their layout and structure type.

The state-of-the-art methods for table detection in documents mostly use markup documents (e.g., pdf, HTML, etc.) and do not cover all types of the tables within one framework. However, for historical documents, which are our main target for this thesis, we only have access to the scanned image and need to deal with all types of tables at the same time. The proposed framework is based on the hypothesis that texts in tables occur in a harmonic column-wise manner. This fact suggests the idea of using a spectral method for developing our framework.

We propose an approach based on using Mel frequency cepstral coefficients (MFCC) to classify document images according to the presence or not presence of tables on the page. MFCCs are well-known speech processing features, which emphasize lower frequency components rather than higher ones. An SVM classifier is used as the final step of our framework for detecting pages containing tables. We test the introduced framework on our datasets and the results confirm the efficiency of the proposed method in comparison to both a state-of-the-art method and our benchmark dataset from the 18^{th} century printed documents.

Keywords: Visual information retrieval, Historical document image, Documents classification, Footnote detection, Bounding boxes, Horizontal projection, Table detection, Ruling line tables, Non-ruling line tables, MFCC, SVM.

TABLE OF CONTENTS

Page

CHAPTER 1 INTRODUCTION 1
1.1 Motivation 1
1.2 Problem statement 2
1.3 Objective of the thesis 6

CHAPTER 2 FOOTNOTE-BASED DOCUMENT CLASSIFICATION 9
2.1 What is a footnote? 9
2.2 state-of-the-arts 10
2.3 Footnote-based feature extraction 12
2.3.1 Preprocessing 15
2.3.2 Font size based methods 16
2.3.2.1 Bounding box based method (BBox) 17
2.3.2.2 Horizontal projection based method (Proj) 20
2.3.3 Space based features 20
2.3.4 Location based features 25
2.4 Classification 25

CHAPTER 3 TABLE-BASED DOCUMENT CLASSIFICATION 27
3.1 What is a table? 27
3.1.1 Table characteristics 27
3.1.1.1 Table size 28
3.1.1.2 Table caption location 28
3.1.1.3 Table frames and separator lines 30
3.1.2 Table definition in our study 35
3.1.2.1 T2(General definition) 35
3.1.2.2 T1(Loose definition) 35
3.2 state-of-the-art 36
3.3 Proposed methodology 42
3.3.1 MFCCs-based features 43
3.3.1.1 Frame the signal 44
3.3.1.2 Windowing 45
3.3.1.3 Fast Fourier transform (FFT) 45
3.3.1.4 Mel filter banks 47
3.3.1.5 Logarithm 48
3.3.1.6 Discrete cosine transform (DCT) 50
3.3.2 Classification 52

CHAPTER 4 PERFORMANCE EVALUATION OF THE RESULTS 53
4.1 Performance metrics 53
4.2 Footnote-based document image classification 54

4.2.1 Dataset 54
4.2.2 Results and discussion 55
4.2.2.1 First Experiment 56
4.2.2.2 Second experiment 61
4.2.3 The third experiment 65
4.3 Table-based document image classification 68
4.3.1 Datasets 68
4.3.2 Results and discussion 70
4.3.2.1 Experiments on the first dataset 71
4.3.2.2 Experiments on the second dataset 72

CONCLUSION AND RECOMMENDATIONS 81

BIBLIOGRAPHY 84

LIST OF TABLES

Page

Table 2.1 List of the extracted features using BBox method. 18

Table 2.2 List of the extracted features using Proj method. 23

Table 2.3 List of the extracted features using space and location-based methods. 26

Table 4.1 Experimental results using SVM and 10-fold CV set-up for the first experiment. Values are in percent and the bold numbers represent the best results. 60

Table 4.2 Experimental results using SVM and 10-fold CV set-up on the relaxed dataset for the first experiment. Values are in percent and the bold numbers represent the best results. 60

Table 4.3 The first experiment results over 30%-70% set-up and their comparison with the state-of-the-art methods. Values are in percent, and the bold numbers represent the best results. 61

Table 4.4 The first experiment results over 30%-70% set-up on the relaxed dataset and their comparison with the state-of-the-art methods. Values are in percent, and the bold numbers represent the best results. 61

Table 4.5 Experimental result for the second experiment. Values are in percent and the bold numbers represent the best results. 62

Table 4.6 Precision comparison of ECCO_I and ECCO_II. 68

Table 4.7 Experimental results using the first dataset and 5-fold CV along its comparison with the proposed method by Kumar *et al.* (2012). Values are in percent, and the bold numbers represent the best results. 72

Table 4.8 Classification results using two schemes of NTvsT1T2 and NTT1vsT2 over Dataset_v0. Values are in percent. 74

Table 4.9 Classification results using two schemes of NTvsT1T2 and NTT1vsT2 over Dataset_v1. Values are in percent. 74

Table 4.10 Classification results using two schemes of NTvsT1T2 and NTT1vsT2 over Dataset_v01. Values are in percent. 74

LIST OF FIGURES

Page

Figure 1.1 Examples of different typographical practices in 18th century documents; pages containing: (a) a footnote, (b) a table, (c) a diagram and (d) a figure. 4

Figure 1.2 Examples of degradation types in historical printed documents. 5

Figure 2.1 An example of a historical document page with a footnote. 11

Figure 2.2 Examples of pages with different column numbers, font sizes and layouts containing footnote. 12

Figure 2.3 Examples of footnotes appearing in the tabular and illustrative layouts. 13

Figure 2.4 Examples of pages with side notes and commentary notes. 14

Figure 2.5 Examples of footnote-like pages. 14

Figure 2.6 Examples of images with degraded nature and faded footnote markers. 15

Figure 2.7 An image example with segmented text lines (red boxes) which contains a footnote along the volume number, the signature mark, and the catchword. 16

Figure 2.8 Font size estimation for the bounding box method. 17

Figure 2.9 An example of font size changes using bounding box based method for a page (a) with a footnote (b) without a footnote. 19

Figure 2.10 The horizontal projection of intensity inside each text line box is determined and then it is intersected with a line equal to 0.55 of the difference between the maximum and minimum of the projection curve. This difference (Red bold lines on the projection graph) is selected as the approximated font size of the line. 21

Figure 2.11 An example of font size change using horizontal projection based method for a page (a) with a footnote (b) without a footnote. 22

Figure 2.12 Example of showing the relationship between space and having a footnote: (a) footnote (b) no footnote. 24

Figure 3.1 The terminology of a simple table structure (Wang, 2016). 28

Figure 3.2 Examples of document images with cross-columns and single-column table sizes. 29

Figure 3.3 Examples of document images with different table captions. 30

Figure 3.4 Table categorization (Tran *et al.*, 2016). 31

Figure 3.5 Example of table classes: (a-c) ruling line tables and (d) a non-ruling line table. 32

Figure 3.6 Examples of irregular tables. 33

Figure 3.7 Examples of false tables. 34

Figure 3.8 Some table-like examples of the images belong to class T1. 36

Figure 3.9 Some non-tabular examples of the images belong to class T1. 37

Figure 3.10 MFCC_based feature extraction process. 43

Figure 3.11 Dividing the signal into short overlapping frames. 44

Figure 3.12 The effect of applying the Hamming window to the time and frequency domains: (a) Hamming window (b) The energy spectrums in both linear and decibel scales applied on the original signal (left column) and windowed signal (right column). 46

Figure 3.13 (a) Mel filter banks. (b) The plot of Mel frequency scale versus hertz. 48

Figure 3.14 Examples of LogMS behavior in the presence of different structural layouts: (a) one-column structure, (b) two-columns structure, (c) figure and text structure and (d) table and text structure. Layout regions are marked with an arrow and dashed red lines show the maximum energy in that region. 49

Figure 4.1 ECCO dataset distributions over time and subjects. 56

Figure 4.2 The information of the labeled dataset along its subsets. The numbers in the dashed orange boxes represent the number of images from each class of FN and NFN for each subset. 57

Figure 4.3 Experimental set-ups for each experiment. 58

Figure 4.4 Examples of images containing footnotes with a complex layout (e.g., two columns, figures, tables, formulas, Greek letters), which are hard to catch. 59

Figure 4.5 Result comparison for each method before and after relaxation. Values are in percent. 61

Figure 4.6 Examples of false negative images for the second experiment. 63

Figure 4.7 Examples of false positive images for the second experiment. 64

Figure 4.8 The histogram charts for the distribution of the images over the time in ECCO_I dataset: (a) all pages (b) detected pages with footnotes. 66

Figure 4.9 The histogram charts for the distribution of the images over the time in ECCO_II dataset: (a) all pages (b) detected pages with footnotes. 67

Figure 4.10 Examples of hand-drawn and printed document pages with a table from the dataset of Kumar *et al.* (2012). 69

Figure 4.11 Examples of mixed form document pages from the dataset of Kumar *et al.* (2012). 70

Figure 4.12 The histogram charts for the distributions of the images over the time at in (a) Dataset_v0 (b) Dataset_v1. 71

Figure 4.13 The information on the labeled table dataset (Dataset_v01). 72

Figure 4.14 True positives for NTT1vsT2 Scheme. 75

Figure 4.15 True positives for NTvsT1T2 Scheme from class (a) T2 and (b) T1. 76

Figure 4.16 False Negatives NTT1vsT2 Scheme. 77

Figure 4.17 False Positives NTT1vsT2 Scheme from class (a) NT and (b) T1. 78

Figure 4.18 False Negatives NTvsT1T2 Scheme from class (a) T2 and (b) T1. 79

Figure 4.19 False Positives NTvsT1T2 Scheme. 80

LIST OF ABREVIATIONS

BBox	Bounding box based method
CNN	Convolutional neural network
CV	Cross validation
DBN	Deep belief network
DCT	Discrete Cosine transform
DIR	Document Image Retrieval
ECCO	Eighteenth Century Collections Online
FFT	Fast Fourier Transformsub
FN	Footnote
f_n	False negative
f_p	False positive
GenRef	Reference
HistAndGeo	History and Geography
HV	Horizontal/vertical
HVP	horizontal-vertical pooling
LitAndLang	Literature and Language
MFCCs	Mel frequency cepstral coefficients
MedSciTech	Medicine, Science, and Technology
MLP	Multi-layer perceptron

NFN	Not having a footnote
NRL-Ts	Non-ruling-line tables
NT	No-table
OCR	Optical character recognition
PCA	Principal component analysis
Proj	Projection based method
RelAndPhil	Religion and Philosophy
RF	Random forest
RL-Ts	Ruling-line tables
SSAndFineArt	Social Science and Fine Arts
SVM	Support vector machine
t_p	True positive

CHAPTER 1

INTRODUCTION

Even though the research in the field of document image retrieval has received a vast amount of attention over the last decades, it is still appealing but challenging; especially when we are dealing with a large collection of low-quality historical document images. Our goal in this work is to facilitate information retrieval from historical documents by finding the various typographic objects in them.

1.1 Motivation

Museums and libraries have substantial collections of historical documents. Various groups of people, from historians and scholars to human scientists, are interested in these documents and trying to study and understand different cultures and communities through these documents (Cheriet *et al.*, 2013). However, due to sensitivity and the valuable nature of these documents, only a small group of people are granted the access to these documents. Therefore, in recent years, historical document collections, which are appealing to the broader group of people, are digitized to be preserved while making them publicly accessible via digital media.

Information retrieval from historical documents is a hard task since the size of these collections is often extremely large. Also, their contents are generally unusual and unstructured beside the fact that degraded quality is also part of historical collections. Therefore, the document image retrieval (DIR) concept for historical documents has been developed (Marinai *et al.*, 2007). The objective of DIR systems is to retrieve visual information from document images by relying only on the document image features, either the text or the image. Applying DIR techniques to historical documents are now trendier than before due to a large number of digitized historical collections.

From a broader point of view, DIR approaches usually follow two main streams, either recognition based retrievals or recognition free retrievals. Generally, recognition based retrieval

approaches transform the digital images into machine-encoded text using optical character recognition (OCR) tools (Chowdhury, 2010). The assumption for OCR systems is that their performance is perfect and the minimal errors are ineffective in the final results. However, this assumption is not always accurate especially in the case of noisy document images, documents with variable and unusual layouts, multilingual manuscripts, pages without any standard fonts, etc. These problems are particularly the concerns for the historical documents or early modern printed documents (Marinai *et al.*, 2007) and make the OCR based methods unfit for this type of documents. The other drawback of using OCR techniques for DIR especially for enormous collections of digitized documents is that they are expensive and tedious. Therefore, using them is misspending of the resources while at the end, we are still in need of human revision. Consequently, recognition free approaches come first for the retrieval purposes. They use images visual features without the need for character recognition techniques. Developing a new recognition free DIR system with a feature extraction process that can handle the challenging nature of historical documents is what motivated us for this research.

1.2 Problem statement

An invaluable and irreplaceable cultural heritages are spread over the libraries throughout the world. They hold crucial information about the people and communities and records of the cultures and periods they were living in it. Understanding the range of visual typographical practices for scientific notation after the development of modern printing can bring researchers from different fields of history of science, book history and document image analysis together to a better perception of "the page image" and its influence on shaping the history of scientific knowledge.

Modern knowledge has been formed using various typographical objects. These graphical objects are used to transfer information and state a truth, often to reduce the complexity and obscurity. Some of the important visual practices that have been used since previous centuries are as follow:

- **Footnotes:** They are used to prove the authorization and as an indication of the correspondence between the sources. Therefore, the need to reproduce the entire source content is omitted by only citing it as a footnote.

- **Tables:** They reproduce the inconsistent and different forms of information in the text, into a geometrical representation of the data.

- **Diagrams:** They provide a tabloid representation of a rational process or natural phenomena to summarize it more symbolically. Using diagrams have become more prevalent during the Enlightenment period.

- **Figures:** They generate a graphic representation of the objects they intend to represent with marking lines and area of tones. In another word, a figure depicts the same object in a non-linguistic two-dimensional scheme.

Figure 1.1 shows examples of each of the aforementioned typographical practices in 18^{th} century publications.

Historical documents beside their unusual structures and layouts usually suffer from different types of degradation they caught over the time. These documents acquire the degradation from various possible sources. These sources vary from chemical, biological and human-caused sources to the external sources that can be generated during document acquisition process such as tears, fold in papers, etc. Figure 1.2 demonstrates examples of these degradations that can complicate the DIR process. The complexities that the nature of historical document images forces upon the retrieval process make the DIR process a more challenging task compared to their associated modern types. Therefore, the main concern of the current study is exploring the different feature extraction and classification techniques, which are insensitive to the complex layout and degraded essence of historical document images.

To this end, our research is aimed to explore the solutions for the following research questions:

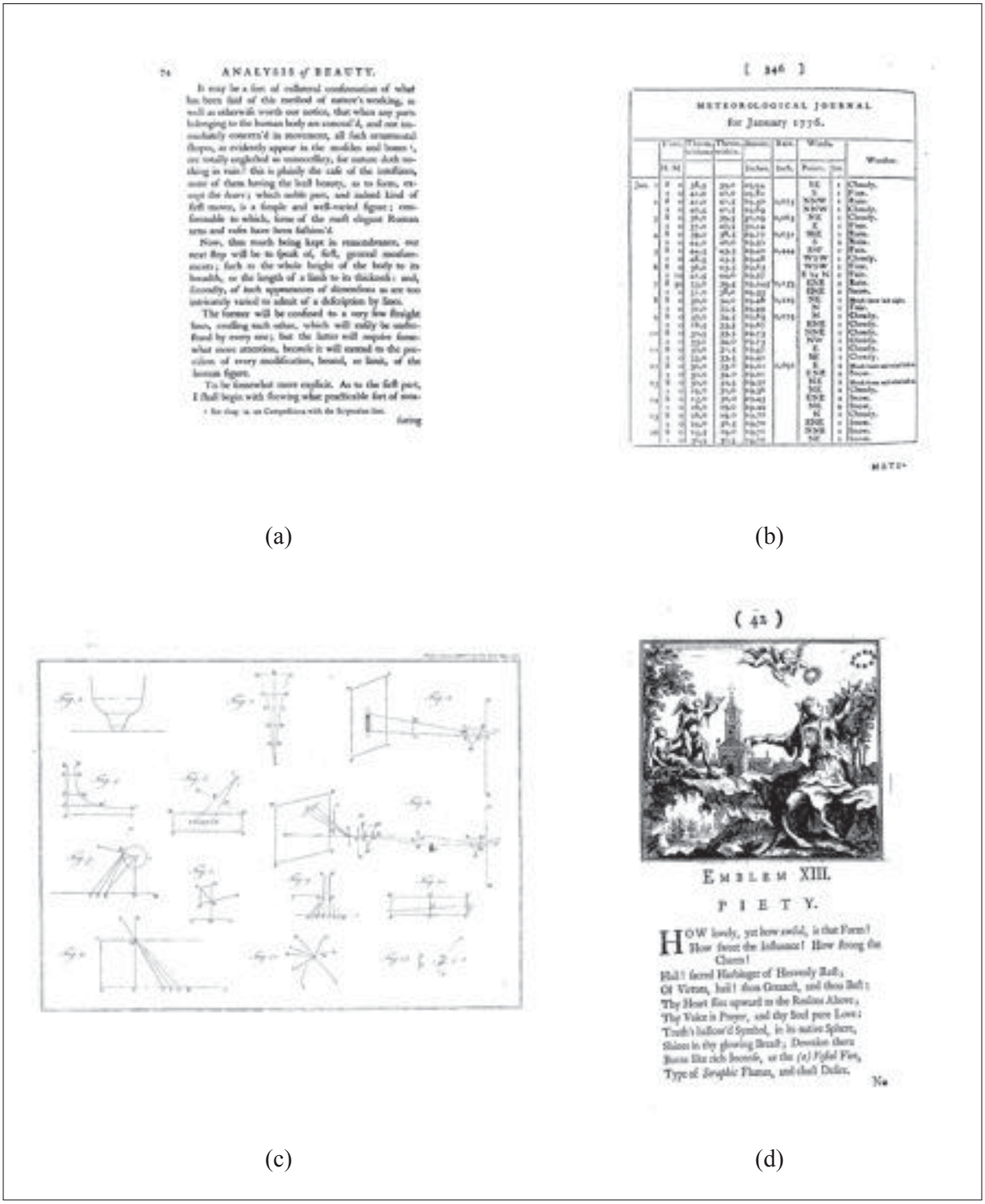


Figure 1.1 Examples of different typographical practices in 18th century documents; pages containing: (a) a footnote, (b) a table, (c) a diagram and (d) a figure.

Figure 1.2 Examples of degradation types in historical printed documents.

- Is it possible to develop an effective and efficient algorithm to carry out the retrieval process?
- To what degree can we enhance the performance of the DIR system by introducing the powerful combination of feature extraction techniques and classifiers?
- Which feature extraction method or classification technique is insensitive to the variation of layout and structure style on degraded documents?

- What are the drawbacks of the system and the recommended solutions to overcome these drawbacks?

1.3 Objective of the thesis

The overall objective of this thesis is to design an advanced system for retrieval the typographical information from the documents and categorize them according to their types. This framework should be applicable to document images with variable structures, layouts, and artifacts. In this work, our focus is on the first two mentioned typographical objects: footnotes and tables. We formulate our two specific objectives around this main objective as below:

- **O1:** Footnotes are one of the essential elements during the Enlightenment period. However, there has not been any work regarding detecting footnote in these documents. Therefore, we aim to propose an automatic framework that can detect the appearance of a footnote on the page despite the layout and structure of the page.
- **O2:** Tables in historical documents vary significantly in structure and representation. Traditional state-of-the-art algorithms are not capable of handling these irregularities. Each of known table detection algorithms has their limitations, and none can provide the ideal performance. Hence, we try to propose an automatic method for detecting tables in historical document images despite the layout of the page and type of tables. Furthermore, this approach should be resistant to a variety of degradation and layouts of the document.

Thesis outline

This thesis is organized into five chapters. The introduction chapter introduced the general content of the thesis. It discussed the problem statement and the background of the information retrieval from document images. It also presented the main and specific objectives of this work. Chapter 2 is dedicated to the first specific objective. It describes the footnote and the detailed proposed methodology used to tackle the problem of footnote-based document image

classification. Chapter 3 presents the definition of the table and our new approach for extracting their features and classifying them. Chapter 4 focuses on introducing the benchmark datasets used in this study and for each method it presents and explains the experimental results used to validate our proposed system. Finally, in the last chapter conclusion and research direction for future work are discussed.

CHAPTER 2

FOOTNOTE-BASED DOCUMENT CLASSIFICATION

Footnote detection is the concentration of this chapter. We first briefly go over the definition of footnote and the state-of-the-arts in this field. Then, we present our first framework for classifying historical documents based on their typographical object.

2.1 What is a footnote?

One of the defining and most significant features of the Enlightenment period is the growth of circulated printed materials. This growth rose another characteristic of this period, which are footnotes or in more general term indices. The appearance of footnote and cross-references is increased progressively as the number of printed document rose up approaching the end of 18^{th} century. Therefore, detecting footnotes will be a crucial element in understanding and linking different scholars and books through the Enlightenment period.

Footnote is a text string appears at the bottom of the page in a document or book and provides clarification and evidence to the claims of an author (Grafton, 1999). A footnote points to a range of indexical forms, including in-text citations, headnotes, marginal notes, and appendices (Pasanek & Wellmon, 2015). Scholars have been using footnote generally for the following reasons:

- To provide more information regarding an author's comments.
- To prove their claims have reliable sources and identify how they end up citing the mentioned source.
- To help the readers to trace back an author's claim to the place where it is originated, and the given source can be verified along further information.

Footnotes are critical for the scientific nature of the historical documents. They present both ideology and technical practices of the document by proofing authors have searched in appropriate archives and have found the necessary documents. Footnotes not only provide this evidence by doing the cross-referencing but also they connect and coordinate the ever-growing system of print (Pasanek & Wellmon, 2015). It is good to mention that in the 20th-century editions, the use of footnotes for citing has lost its traditional place and replaced by endnote at the end of the chapter or book.

In this work, with the help of human scientists, we define footnote as a typographical object that is distinct on the page, has been marked in the main text and referenced at the bottom of the page. Each of these conditions is important to distinguish between the positive and negative examples of the footnote. Figure 2.1 depicts how footnotes generally appear on the page in historical documents. However, even using this simple definition, because of a large variety of footnote patterns, fonts, styles and inconsistent structures in the 18th-century, distinguishing footnotes is still hard. Figure 2.2 and Figure 2.3 give some examples of pages containing footnote but different in font size style, layout, content and structure. This definition also prevents side footnotes and commentary notes at the bottom of the page from being considered as a footnote. Figure 2.4 gives examples of these two cases. Moreover, some cases deceptively seem to be a footnote but only a note or comment (see Figure 2.5). Along the diversity of footnote types, degraded and hard to capture document pages makes the problem of detecting footnotes even more challenging task. Figure 2.5 gives some examples of these types of image. Even by a brief comparison of these Figures, we can conclude that footnotes in the 18th-century are barely following a straightforward procedure and there is indeed a wide range of diverging forms of footnotes.

2.2 state-of-the-arts

There have been various works in the scope of information retrieval from historical document images. However, to the best of our knowledge, there are only two works published in the field of detecting footnotes from historical documents. Abuelwafa *et al.* (2017) proposed a method

Figure 2.1 An example of a historical document page with a footnote.

based on a deep belief network (DBN). Their approach includes two phases, unsupervised pre-training and supervised fine-tuning. In this work, each image is represented by merging the two text lines from the top of the image and two from the bottom. The features are then extracted from these merged images in an unsupervised manner using a DBN model to initialize the parameters for a multi-layer perceptron (MLP). The MLP is later retrained and fine-tuned using the labeled document data. The main advantage of this method is that it can be beneficial in the case of sparsely labeled data specifically for historical documents with not enough labeled data. The second work is done by Mhiri *et al.* (2017) and uses a one-dimensional convolutional neural network (CNN) to tackle the problem of footnote detection. For this work, the document image is represented by merging the horizontal intensity histogram of the two top and three bottom text lines, which reduces the complexity of using CNN.

Figure 2.2 Examples of pages with different column numbers, font sizes and layouts containing footnote.

2.3 Footnote-based feature extraction

In this section, we describe our methodology to extract features related to the appearance of a footnote on the page. The fundamental idea behind this approach is to generate our hypothesis around the most remarkable visual features of footnotes. For this aim, three visual features

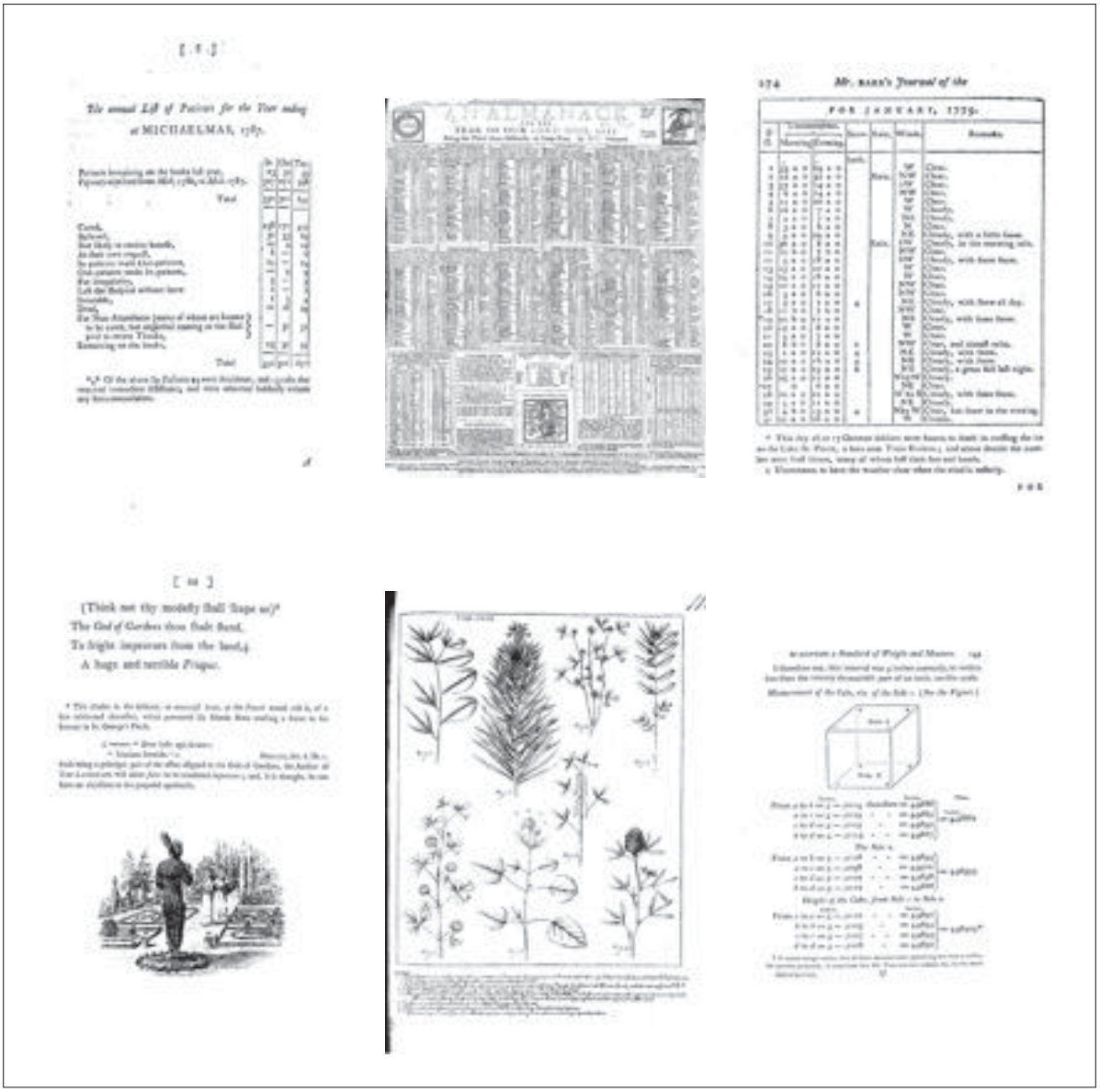


Figure 2.3 Examples of footnotes appearing in the tabular and illustrative layouts.

are observed which are the smaller font size of the footnote respect to the body text, footnote location in the document page and the relatively greater gap between the footnote and the main text compared to the regular line space. The smaller font size is our strongest hypothesis and has been the case for over 85% of the almost 6000 images we observed. The last two visual features are the weaker ones but still necessary for forming an efficient footnote detection framework. To fulfill each of these observations, four methods are proposed based on the font

Figure 2.4 Examples of pages with side notes and commentary notes.

Figure 2.5 Examples of footnote-like pages.

size, the footnote location, and spacing between the body text and the footnote. Below. We briefly explain each of these methods along the preprocessing steps needed for each of them.

Figure 2.6 Examples of images with degraded nature and faded footnote markers.

2.3.1 Preprocessing

Text line segmentation is the primary phase of our feature extraction methods and related to layout analysis. To do so, we used the method proposed by Santos *et al.* (2009). According to their method horizontal projection of binarized images are obtained. Following that there is a text line separation step based on finding the empty spaces between histogram peaks and applying a dynamically calculated threshold. This threshold is related to the average horizontal histogram values of the text lines. This step identifies the vertical location of each text line. We need a vertical histogram projection to find the horizontal position of the text lines. Applying a vertical projection also helps in avoiding false words. Finally, to optimize the segmentation results, we perform a text line selection. The result of this process is shown for an example image in Figure 2.7.

One of the common practices in 17^{th} and 18^{th} centuries were using signature marks and catch-words at the end of the page (Marcussen & Bergendorff, 2003). They both were used to help the bookbinder or printer to bind the pages and sections in right order for the press. Advancements in printing technology have overtaken these practices, and they are out of use nowadays. There is also the number of the volume at the bottom of the page. All of these practices can

Figure 2.7 An image example with segmented text lines (red boxes) which contains a footnote along the volume number, the signature mark, and the catchword.

be a source of false detection based on our hypothesis. Specifically, they have the same font size as the main text while usually coming after a footnote. Figure 2.7 depicts an example of the image containing all aforementioned practices along footnote on the page. As it can be seen, not considering them in our algorithm can obscure the problem later. Therefore, before applying the page segmentation algorithm, we crop one-fourth of the page image from the right to avoid the problem of catchwords. To deal with the signature marks, the last text line started at the bottom midpoint of the page is eliminated from the process.

2.3.2 Font size based methods

We propose two methods based on font size namely the bounding box based method (BBox) and horizontal projection based method (Proj). These methods aim to find the approximate font size of each line on the page and then use that information to drag a set of rules for the

appearance of the footnote on the page. Based on our primary hypotheses, we expect a drop in font size in the footnote line. However, this is an ideal scenario, and other typographical objects can cause this drop as well, e.g., title, figures, tables, etc. Hence, we need some additional rules to avoid false detection and to detect the footnote line accurately. Below we briefly explain each of these two methods.

2.3.2.1 Bounding box based method (BBox)

Our first method based on estimating the font size of each line is BBox method. In this method, we aim to find the bounding box around letters in the text line. We first apply contrast-limited adaptive histogram equalization (Zuiderveld, 1994) to each text line area, and then the smallest connected component regions are found using rectangular bounding boxes. After locating the bounding boxes, we try to approximate the font size of the lower case letters in each text line. To do so, first, for each text line, we find the bottom (base) line that all the letters are written on it. Then the most frequently repeating height of the bounding boxes above that line is taken as the font size estimated for that text line (see Figure 2.8).

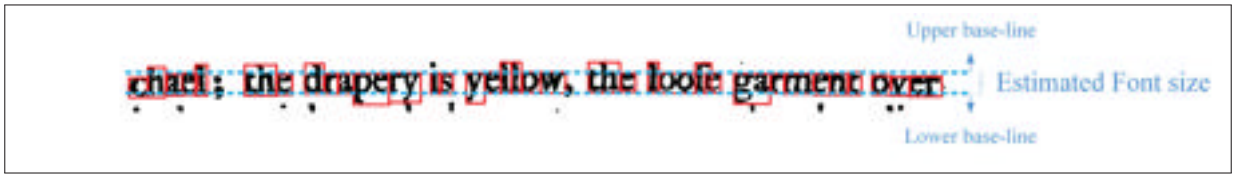


Figure 2.8 Font size estimation for the bounding box method.

The font sizes are then normalized between zero and one to make the results generalizable for the other images. Small noises, lines, logos or even punctuation marks (such as dots, semicolons, etc.) can affect the font size estimation due to their unusual sizes. Therefore to avoid this problem a threshold has been used. Only the bounding boxes with the height between 50 and 200 pixels are kept. Moreover, the ones with the width over 150 pixels are eliminated. Figure 2.9 shows two examples of a page containing a footnote and without a footnote. It is evident from the figure that wherever a footnote appears, we can expect a drop in font size

(Figure 2.9 (a)) and when there is no footnote we would have some random oscillation in the graph of font size versus text lines (Figure 2.9 (b)).

After obtaining the normalize font size, to find a way to extract features from the font size information and associate it with the appearance of a footnote on the page, we define a set of rules according to observing about 100 images from each category. The basic rule for having footnote wrap around the drop of 0.55 between two consecutive lines rule. There exist some other rules for having footnote such as not having a footnote in first three lines of the page; condition of having at least three lines in a page; having only on a global drop, etc. The logical zero and one answers to these rules build a feature vector of size 18 for this method. The detailed descriptions of these features are presented in Table 2.1.

Table 2.1 List of the extracted features using BBox method.

Features	**Conditions**
F1	1 if there is no drop more than 0.55. Otherwise,0.
F2	1 if there are 1+ drops of more than 0.55 (Footnote line's condition). Otherwise,0.
F3	1 if the last two lines' heights are less than 0.1. Otherwise,0.
F4	1 if the last line's height is less than 0.1. Otherwise,0.
F5	1 if the last two lines' heights are less than 0.1 and there is a footnote. Otherwise,0.
F6	1 if the line before last line's height is less than 0.1. Otherwise,0.
F7	1 if the footnote is not in the 4^{th} line. Otherwise,0.
F 8	1 if there are 2+ drops more than 0.55. Otherwise,0.
F9	1 if the footnote is not in the lines 4^{th}, $5^{th}$and 6^{th}. Otherwise,0.
F10	1 if there are 2+ drops less than 0.55 or footnote is not in the lines 4^{th}, 5^{th} and 6^{th}. Otherwise,0.
F11	1 if there is a drop greater than 0.15. Otherwise,0.
F 12	1 if footnote line is in the line 6^{th} or more. Otherwise,0.
F13	1 if the height of the footnote line is 0.55 greater than the line before the last line. Otherwise,0.
F14	1 if there is a drop of greater than 0.35 between the line before and after footnote line. Otherwise,0.
F15	1 if there is a line except the last line selected as the footnote line and there is a drop of greater than 0.35 between the line before and after it. Otherwise,0.
F16	1 if there is a difference less than 0.17 between the line before and after the footnote line. Otherwise,0.
F17	1 if there is a line except the last line selected as the footnote line and there is a difference less than 0.17 between the line before and after it. Otherwise,0.
F18	1 if it the page has more than three lines. Otherwise,0.

2.3.2.2 Horizontal projection based method (Proj)

This is the second method based on finding an approximate size of the font in each line. In this method, the approximate font sizes are determined using the intensity of horizontal projection in each text line. First, we find the outline bounding boxes of each text line and discard the outlines with a height of lower than 50 pixels and higher than 200 pixels. After applying Contrast-limited adaptive histogram equalization to each text line (Zuiderveld, 1994), we calculate the horizontal projection for each line. Next, the intersection points of a straight line equal to 0.55 of the difference between the maximum and minimum of the histogram and the histogram itself are located on each line. The value 0.55 is found empirically. The distance between the intersected points indicates our estimated font size for that line (see Figure 2.10). Finally, the font sizes are normalized between zero and one to make the comparison between other pages possible. As it is clear from Figure 2.11, it can be seen the Proj method also follows the same pattern as the BBox method. Wherever we have a footnote, we can expect a drop in the font sizes and random changes of the font size in the same range indicates not having a footnote in our pages (Figure 2.11 (a and b)).

Therefore, like the BBox method by observing 100 images from both classes of FN and NFN, we defined a set of rules and conditions for each category. The essential assumption for the appearance of a footnote is again a drop of 0.55 between two consecutive lines. There are also some sub-assumptions such as having at least three lines on the page; not accepting the lines except the last line as the footnote line unless there is a difference of 0.25 or more between the lines before and after it. The logical answers to these rules and conditions, which are given in Table 2.2, create our Proj feature vector of size 24.

2.3.3 Space based features

This method is based on a weaker hypothesis, driven from the observation that the main body and the footnote section are separated from each other using a gap or space between them. As it can be seen from Figure 2.12 (a), we have a broader line spacing between the footnote line

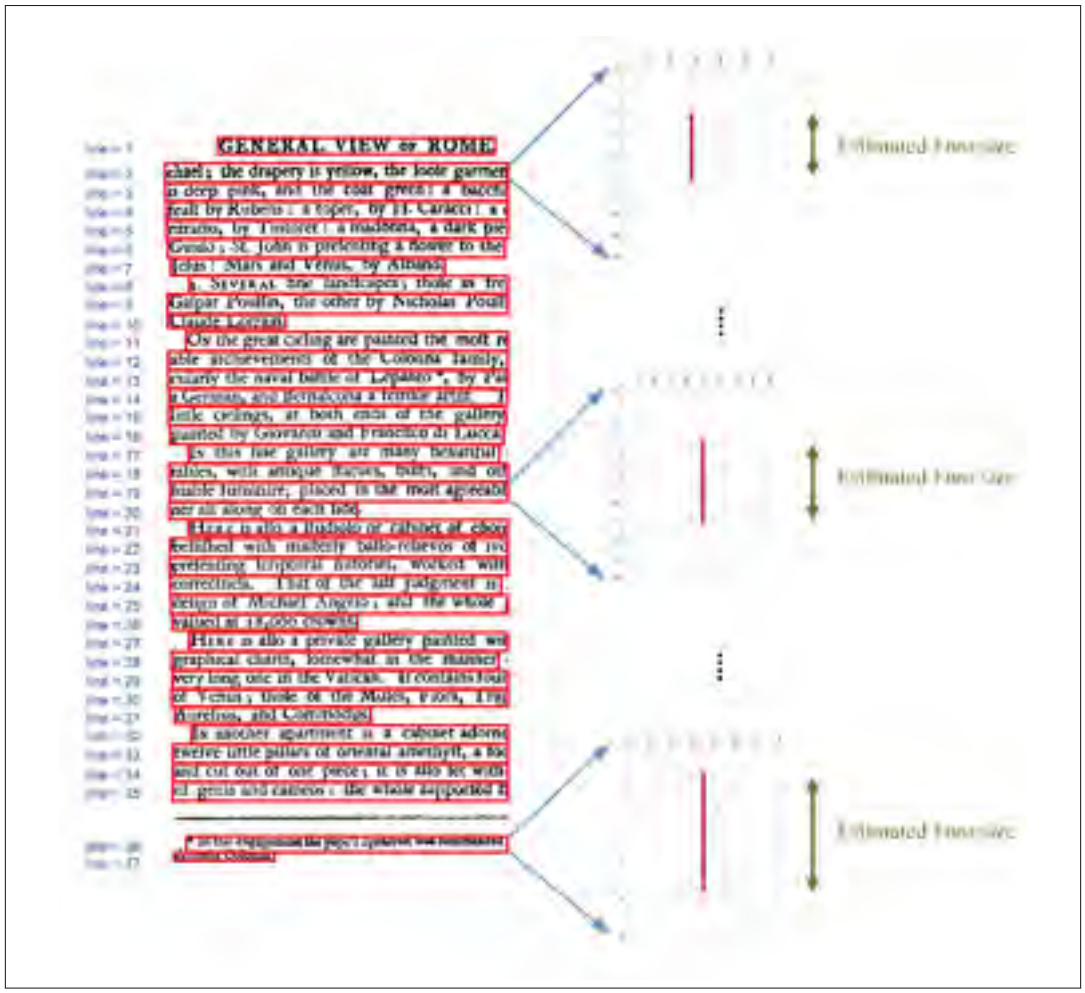


Figure 2.10 The horizontal projection of intensity inside each text line box is determined and then it is intersected with a line equal to 0.55 of the difference between the maximum and minimum of the projection curve. This difference (Red bold lines on the projection graph) is selected as the approximated font size of the line.

and the main text. However, this gap could appear for other different reasons like the paragraph spacing or the space after titles or images (Figure 2.12 (b)). Therefore, we introduced some rules with the underlying assumption of having at least ten lines to proceed. After finding the outline box and the space between them, these spaces are normalized between zero and one for comparison purpose. We candidate the maximum line spacings which are 0.7 space above the average space. Next, divide the page into four sections and analyze if any maximum spacing appears at the three last one of them. We create the initial features based on space using this approach. Ultimately, by finding the approximate location of last/only maximum spacing on

Table 2.2 List of the extracted features using Proj method.

Features	Conditions
F1	1 if there are more than three lines on the page. Otherwise, 0.
F2	1 if there is no possible footnote. Otherwise, 0.
F3	1 if there is more than one possible footnote (drops with the amount of 0.55 or more). Otherwise, 0.
F4	1 if footnote line is in the first three lines, or there are more than three possible footnotes. Otherwise, 0.
F5	1 if there are more than three possible footnotes, or there are lines shorter than 0.13 but not footnote lines. Otherwise, 0.
F6	1 if footnote line is in the first three lines, or there are lines shorter than 0.13 but not footnote lines. Otherwise, 0.
F7	1 if footnote line is in the first three lines. Otherwise, 0.
F 8	1 if there are more than three possible footnotes. Otherwise, 0.
F9	1 if there are lines shorter than 0.13 but not footnote lines. Otherwise, 0. Otherwise, 0.
F10	1 if footnote line is in the first three lines, or there are 3+ possible footnotes, or there are lines shorter than 0.13 but not footnote lines. Otherwise, 0.
F 11	1 if the last line or the line before has a height less than 0.1. Otherwise, 0.
F12	1 if the last line or the line before having the height less than 0.1 and there is a footnote Otherwise, 0.
F13	1 if the last line has a height less than 0.1. Otherwise, 0.
F14	1 if the line before the last line has a height less than 0.1. Otherwise, 0.
F15	1 if there still exists a footnote line. Otherwise, 0.
F16	1 if the height of the last line is less than 0.4. Otherwise, 0.
F17	1 if the last line has a height less than 0.1 and there exist a footnote line. Otherwise, 0.
F18	1 if the line before the last line has a height less than 0.1 and there exist a footnote line. Otherwise, 0.
F19	1 if the height of the last line is less than 0.4 and there is a footnote line. Also, the last line or the line before has a height less than 0.1. Otherwise, 0.
F20	1 if the greatest height drop is equal to or greater than 0.4, and there is at least a 0.25 drop between the line before and after footnote. Otherwise, 0.
F21	1 if there is at least a 0.25 drop between the line before and after footnote and footnote line's height is less than 0.4. Otherwise, 0.
F22	1 if the greatest height drop is equal to or greater than 0.4 and the height of the last line is less than 0.4. Otherwise, 0.
F23	1 if the height of the last line is less than 0.4, and the greatest height drop is equal to or greater than 0.4, and there is at least a 0.25 drop between the line before and after footnote. Otherwise, 0.
F24	1 if the height of footnote line is 0.4 below the highest height of the all other lines except the first 3 and last lines. Otherwise, 0.

the page, we complete the features based on the page line spacing. Features F1 to F8 in Table 2.3, represent the features according to this method.

2.3.4 Location based features

This method uses the location information achieved form BBox and Proj methods along the line spacing to extract more features for detecting footnote on the page. First, it tries to find the relative location of the footnote on the page by dividing the approximate detected line as the footnote to the total number of the text lines on the page. Later, having the location of the maximum space between the lines (last maximum space in the case of having more than one), we investigate if this point is anywhere around the detected footnote from BBox and Proj methods. For this purpose, we change the range of search from 0.02 to ± 0.14 by steps of ± 0.02. More details about features can be found in Table 2.3 (F9 to F30).

The combination of all these features generates the final feature set: 18 features form BBox method, 24 from Proj method and 30 from the spaced and location-based features, which sum up to total 72 features.

2.4 Classification

In recent years, support vector machines (SVMs) have been widely used for analyzing data and recognizing pattern (Hearst *et al.*, 1998). In this research, we use the SVM classifier to discriminate pages with a footnote from the one without a footnote. SVM implementation is done using LIBSVM toolbox (Chang & Lin, 2011) and the SVM has a degree two polynomial kernel function. Ten percent of the training set is used as the validation set to calculate the hyperparameters of SVM.

Table 2.3 List of the extracted features using space and location-based methods.

Features	**Conditions**
F1	1 if there are more than 10 lines on the page. Otherwise, 0.
F2	1 if there is a space peak in the 2nd 1/4th of the page & there are more than 10 lines on the page. Otherwise, 0.
F3	1 if there is a space peak in the 3rd 1/4th of the page & there are more than 10 lines on the page. Otherwise, 0.
F4	1 if there is a space peak in the 4th 1/4th of the page & there are more than 10 lines on the page. Otherwise, 0.
F5	1 if there is only one peak in whole the page & there are more than 10 lines on the page. Otherwise, 0.
F6	1 if there is more than one peak on the page& there are more than 10 lines on the page. Otherwise, 0.
F7 – F8	1 if there is a footnote(Using BBox and Proj methods) in the last 1/4th of the page & there are more than 10 lines on the page. Otherwise, 0.
F9 – F10	1 if there is a footnote(Using BBox and Proj methods) on the page & there are more than ten lines on the page. Otherwise, 0.
F11 - F16	1 if for a page with more than 10 lines, peak location appears anywhere (according to a threshold from $\pm$0.02 by 0.02 up to $\pm$0.14) around the footnote located by Proj method in the last 1/4th of the page. Otherwise, 0.
F17 - F22	1 if for a page with more than 10 lines, the last peak location on the page appears anywhere (according to a threshold from $\pm$0.02 by 0.02 up to $\pm$0.14) around the footnote located by the Proj method in the last 1/4th of the page. Otherwise, 0.
F23 – F26	1 if for a page with more than 10 lines, peak location appears anywhere (according to a threshold from $\pm$0.02 by 0.02 up to $\pm$0.14) around the footnote located by the BBox method in the last 1/4th of the page. Otherwise, 0.
F27 – F30	1 if for a page with more than 10 lines, the last peak location on the page appears anywhere (according to a threshold from $\pm$0.02 by 0.02 up to $\pm$0.14) around the footnote located by theBBox method in the last 1/4th of the page. Otherwise, 0.

CHAPTER 3

TABLE-BASED DOCUMENT CLASSIFICATION

The content of this chapter introduces our second method for information retrieval from historical documents. Our focus is on tackling the problem of table-based document image classification. We start with an introduction to table definition and its different forms. Following that, we explain state-of-the-art methods for table detection in literature. Finally, we end this chapter by presenting our proposed algorithm for table-based document classification in historical documents.

3.1 What is a table?

The growth of interest in the field of automatic historical document analysis by historians and human scientists brings up the need for systems that can retrieve and understand information from these documents. Generally, documents contain different textual and graphical contents (e.g., titles, symbols, footnotes, figures, tables, etc.). Tables are one of the most common objects that can be found in multiple classes of documents throughout the history from scientific journals and newspapers to forms and invoices. Tables are used to summarize the correlational data and present them in an ordered, meaningful and compact manner (Tran *et al.*, 2016).

3.1.1 Table characteristics

The most straightforward table types are regular matrices of cells with a row-column structure. Figure 3.1 shows a typical example of these kinds of table and its general terminology for different parts. However, tables not often follow such a simple structure, particularly the tables in printed documents from the Enlightenment period which are the core of our research. During this period, there was not any established style for tables, and printed documents were recently prevalent. Tables in these documents were more heterogeneous in structure, style, flexibility, notation and in one word their representation. This heterogeneity could reveal itself in the form of nesting cells, missing boundary lines, etc (Wang, 2016).

Tables can be interpreted from different perspectives: location, size, caption size and location, frames and separator lines, cell alignment and cell types. Below we explain some of these perspectives that can assist in clarifying the tables for our further steps of table retrieval in historical documents.

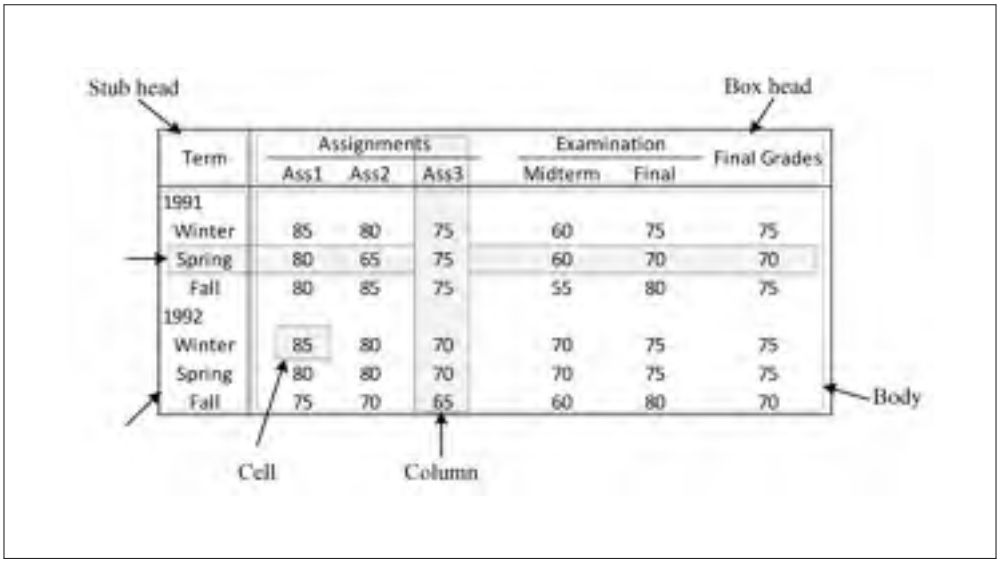


Figure 3.1 The terminology of a simple table structure (Wang, 2016).

3.1.1.1 Table size

Tables can be classified into two group of cross-column versus single-column tables or wide versus narrow tables. A cross-column table requires having a width wider than half of the document page and including at least one space between to columns of the document. All other tables are considered single-column. Figure 3.2 demonstrate some examples form 18th-century document pages from both of these categories.

3.1.1.2 Table caption location

Table captions and in general captions have an essential rule for understanding and interpretation of tables. According to the recent guidelines for captions, they should be above the

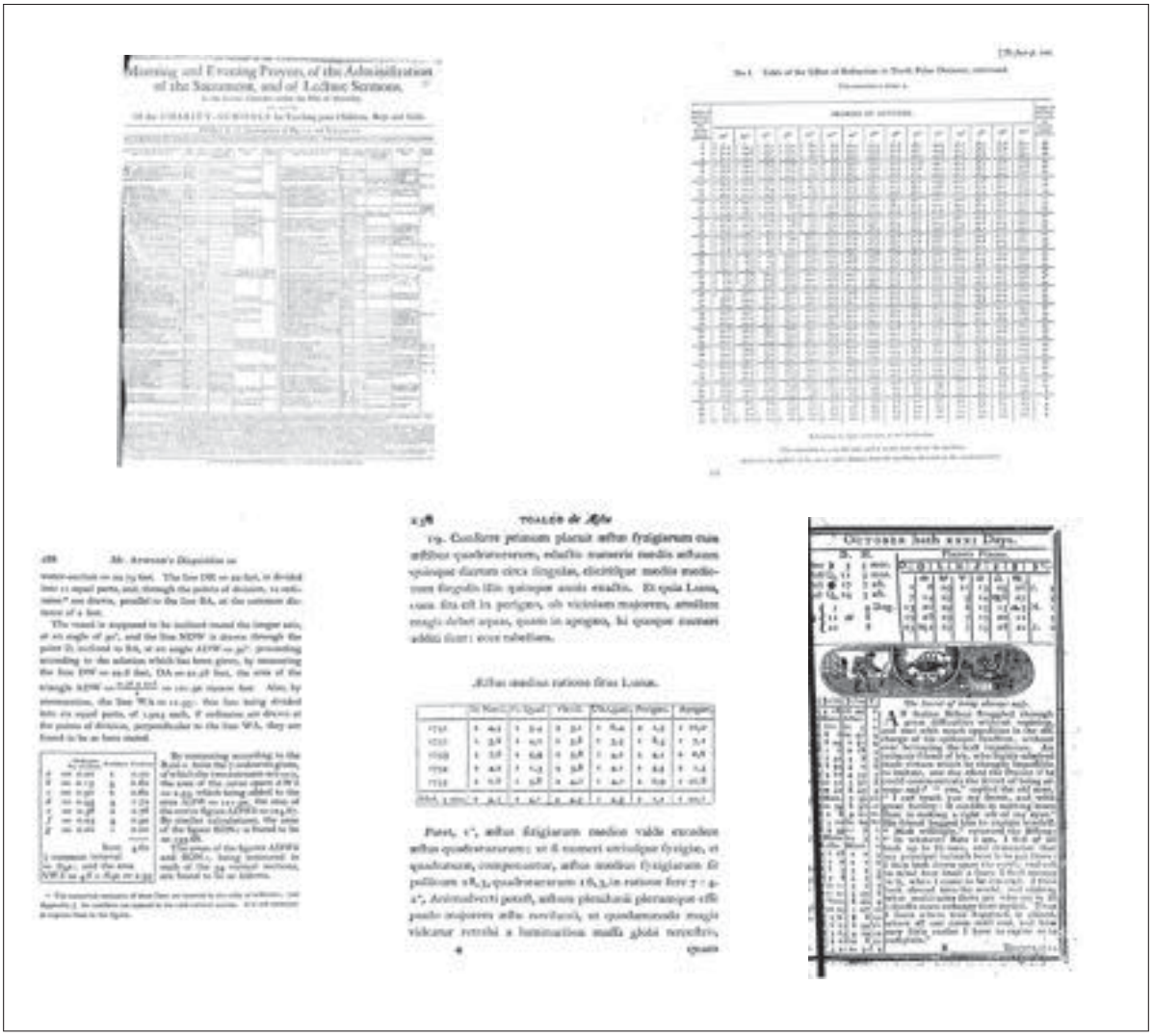


Figure 3.2 Examples of document images with cross-columns and single-column table sizes.

table. They should not be too brief and not provide enough information or be extremely large. However, Enlightenment period captions do not obey these guidelines, and there exist several cases without any captions. Such inconsistent behaviors profoundly weaken the table detection algorithms based on the OCR. Figure 3.3 shows some examples of caption style and format in 18th-century document images.

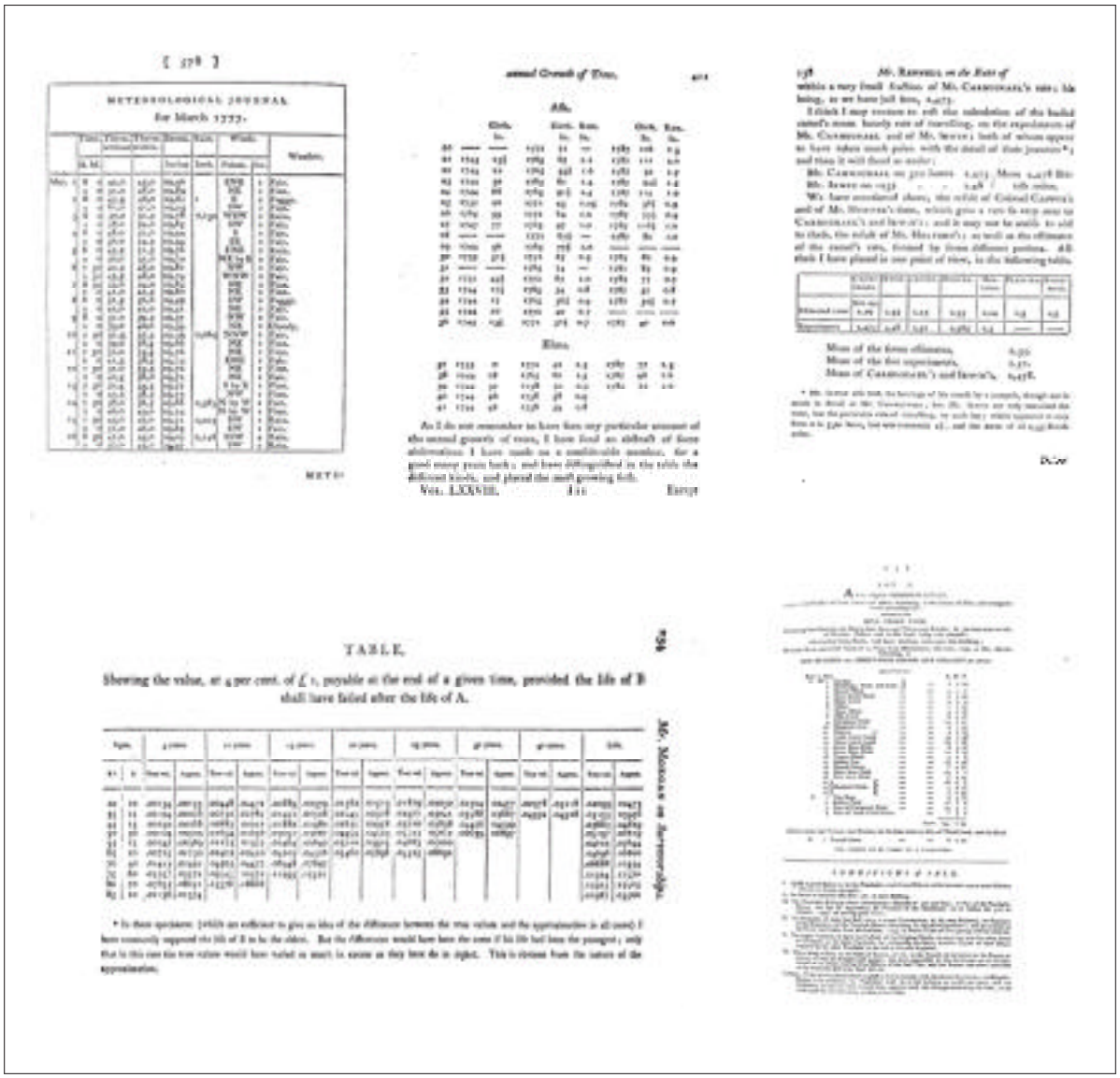


Figure 3.3 Examples of document images with different table captions.

3.1.1.3 Table frames and separator lines

Not all the tables in document images include frames and separator lines. In this sense, we can classify tables into two main classes according to their framing composition: ruling-line tables (RL-Ts) and non-ruling-line tables (NRL-Ts) (Tran *et al.*, 2016). Figure 3.4 illustrates this categorization in more detail.

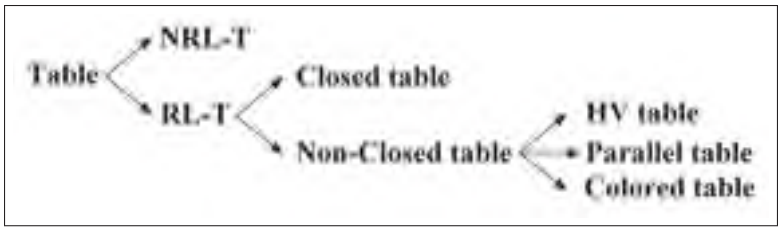


Figure 3.4 Table categorization (Tran *et al.*, 2016).

Ruling-line tables (RL-Ts)

RL-Ts are the most common type of tables. From outside an outer bounding box surrounds them and within they are divided by one or more ruling lines (i.e., vertical lines, horizontal lines or a mixture of both). These tables can be further split into two categories of closed tables and non-closed tables:

- **Closed table:** A table is called closed if a complete outer bounding box surrounds it as shown in Figure 3.5(a).
- **Non-closed tables:** This kind of table consists of at least one type of ruling line (either vertical or horizontal), and contrary to the closed tables the outer bounding box is not a complete rectangular. They can be further subdivided into three separate groups:
  - **Horizontal/vertical(HV) table:** The structure of this table has both vertical and horizontal ruling lines, see Figure 3.5 (b).
  - **Parallel table:** This table contains only horizontal ruling lines which are parallel, see Figure 3.5(c).
  - **Colored table:** A combination of colorful rectangular blocks creates this kind of tables. However, during the 18^{th}-century, colorful tables have rarely been used for press printing. Thus, we do not consider them in our study.

Non-ruling-line tables (NRL-Ts)

These tables do not contain any ruling-line or bounding box. They are mostly used in reports or letters (see Figure 3.5(d)).

(a) Closed table

(b) HV table

(c) Parallel table

(d) Non-ruling line table

Figure 3.5 Example of table classes: (a-c) ruling line tables and (d) a non-ruling line table.

Nonetheless, always there can be tables that are not following the rules above in the same order. For instance, tables may contain nested/empty cell, figures/symbols or formula. Also, columns or rows may not be aligned. Figure 3.6 provides some examples of these irregular table structures.

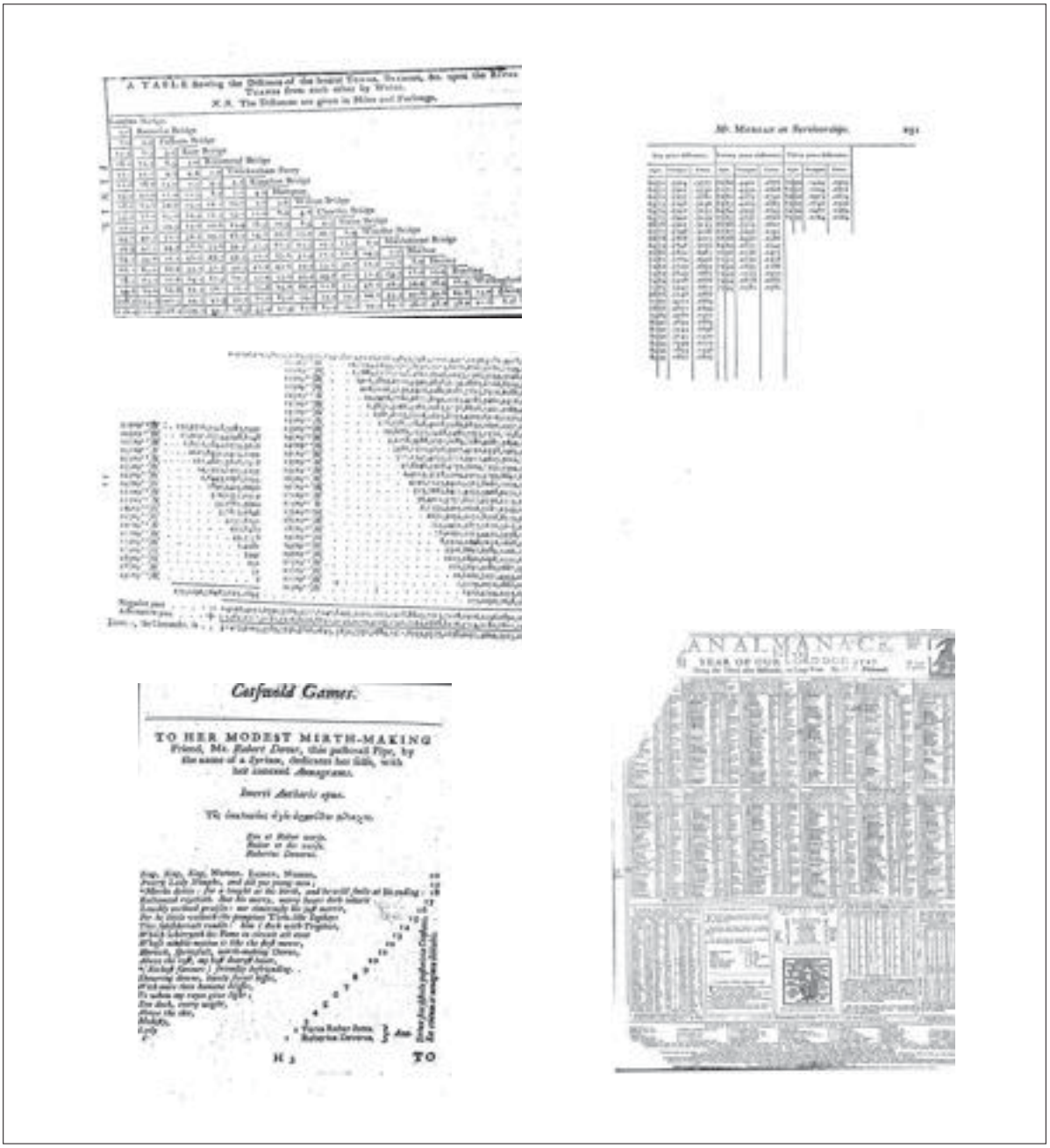


Figure 3.6 Examples of irregular tables.

Furthermore, several cases of false tables can appear in the documents. Index, content, lists and catalog pages are one of these false tables. They follow all the rules for being a table. However, they can not be considered as a table. Also, NRL-Ts can appear as multi-column document images. Moreover, there are title pages with external bounding box even ruling lines but not defined as a table. Figure 3.7 demonstrates some examples of these false tables.

Figure 3.7 Examples of false tables.

3.1.2 Table definition in our study

In this study, we use two definitions for tables. We have a general and then a looser definition, which is a subset of the first one.

3.1.2.1 T2(General definition)

Any regular form of tables that consist of columns and rows, whether containing ruling lines or white space is categorized in this class. There should be no doubt in assigning the samples of this class as a table.

3.1.2.2 T1(Loose definition)

Any table with irregular layout compared to the standard format of the page is included in this category. The document pages belong to this category range from title, list, content, index, and catalog to multi-column pages or even pages having the mixture of these situations. They can firmly have tabular form especially multi-columns pages or title pages with bold ruling lines. They may also have no-table (NT) class appearance that hardly can resemble a table. Figure 3.8 provides some pages from class T1 with table-like appearance, while Figure 3.9 shows some examples with NT appearance.

According to these categories, we have a sequential decision tree. Starting with the T2 label, we first check if the table belongs to the T2 and if not check the T1 category. This labeling serves people in the field of human sciences. It helps to study the tables in the past centuries and monitor their evolving and changes in printed documents as well as their in-between connection. To do so, researchers not only need as many images containing table as they can but also they are interested in finding any relation and difference between the loose and general definition of tables. Moreover, they favor finding the T2 over the T1 class considering T2s are strictly the tables and will be most advantageous in the analyzing the historical table data.

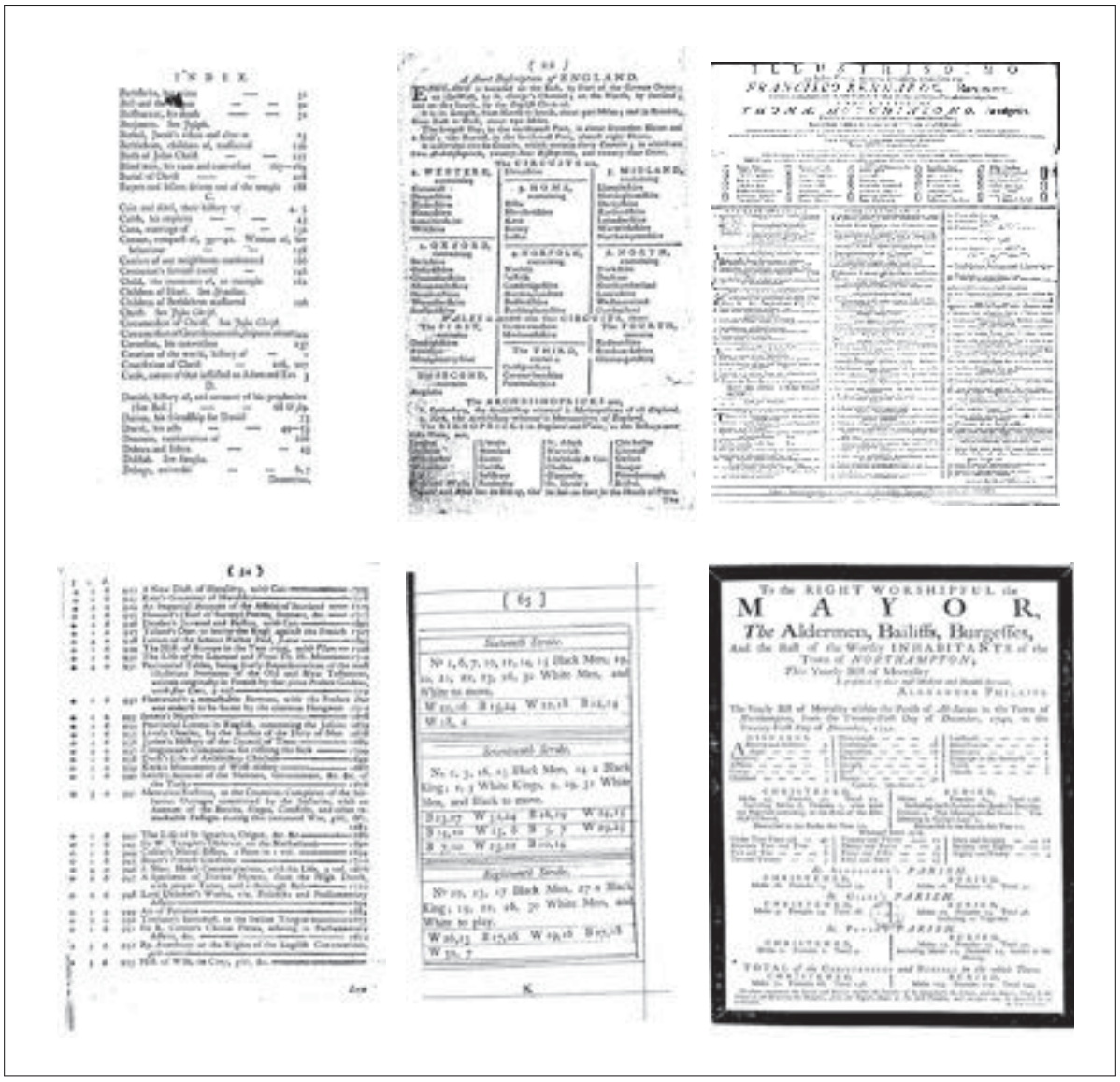


Figure 3.8 Some table-like examples of the images belong to class T1.

3.2 state-of-the-art

In the past two decades, various approaches have been proposed to detect tables in documents with different formats and styles. However, finding tables in scanned documents requires a more advanced process than pure text or markup documents (e.g., pdf, HTML, etc.) because less info about their contents is available. Therefore, scanned documents need to be converted from pixel representation to more abstract representations. Along these problems, there is also

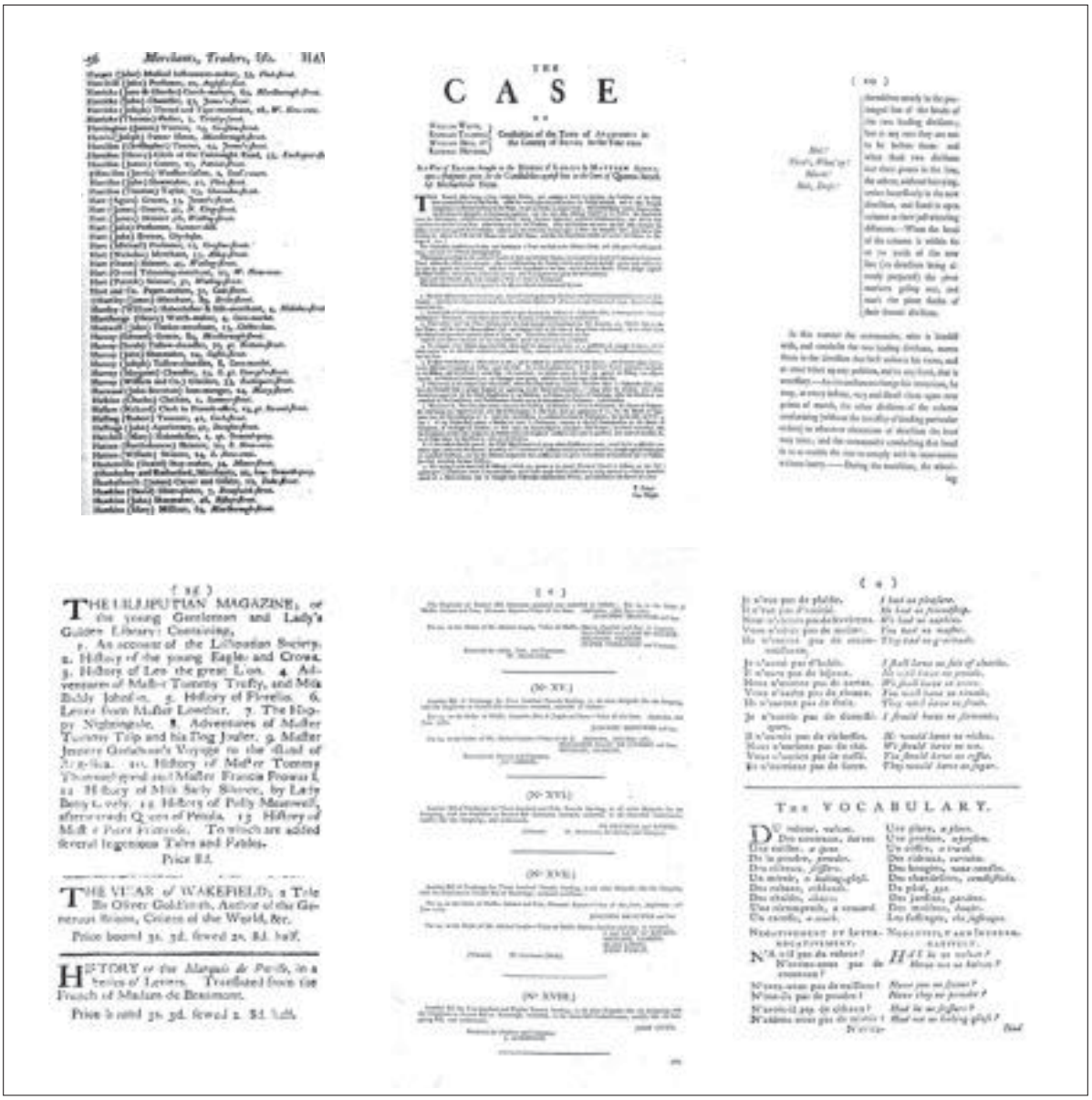


Figure 3.9 Some non-tabular examples of the images belong to class T1.

the fact that document images are obtained through scanning and digitizing process and subject to several errors such as noise and skewness (Long, 2010).

Table detection algorithms commonly follow two main strategies: [1] non-text analysis by manipulating the pixels directly and finding the table template in the image pixels, and [2] text analysis by using OCR tools to retrieve the ASCII text from the document images (Long,

2010). Many issues are associating with the OCR-based approaches. The main problem is the fact that these tools are not entirely reliable and any error in this stage can significantly affect the table processing stage. Also, these tools are not able to completely transform some typographical information such as font or symbols back to the text, which causes information loss. Even though modern OCR systems have targeted the tables in their process, but still they are biased to specific forms and formats of the table and collapse in the case of irregular and unfamiliar tabular structures like what we commonly encounter in historical documents. Therefore, OCR tools are not often considered as a successful method when it targets tables in scanned document images.

One of the earliest works on table detection in scanned documents belongs to Laurentini & Viada (1992). They find the horizontal and vertical ruling lines using the run-length analysis as the first indication of the appearance of a table in machine-printed documents. Then, they execute several tests to see if a group of character-sized connected components is tightly fixed within the located table cells. Hu *et al.* (1999) proposed an approach for detecting tables from single column document images. Their method is based on the assumption that page can be easily segmented to text lines and from there, table detection is only a matter of determining the start and end of ruling lines by utilizing some quality functions. However, this method is only limited to the single column images with simple layouts.

The method introduced by Wangt *et al.* (2001), detects the potential table lines using the gap between consecutive words. Holding this information, they group the vertically adjacent lines along the horizontally adjacent words and form the table cell candidates. Finally, the candidates are refined according to a statistical approach by finding a consistency score. This approach assumes that the document page can have only one of these three page layouts: single-column, double-columns and mixed-column. The system is trained using this information as prior knowledge for table detection. The main disadvantage of this technique is its limitation to those page layouts. Cesarini *et al.* (2002) presented a method for finding parallel lines using MXY tree-based hierarchical representation. The parallel lines information aids

to candidate tables by locating the perpendicular lines or white spaces in between the parallel lines. However, their prosed method is only applicable to the HV type of tables.

The approach used by Gatos *et al.* (2005) searches for the horizontal and vertical ruling lines on the page and finds their points of intersection. These points are further used to reconstruct the table by connecting the intersection line pairs. Their stated method has the problem of not covering the NRL tables. Lin *et al.* (2006) introduced a robust system for table detection from online ink notes. The first step of their method is extracting the fundamental elements of the table structure, i.e., ruling lines, table bounding boxes and drawing strokes candidates. Then by normalizing the table skeleton, the logical structure of the table and its cell content are achieved. Finally, taking advantage of the similarity between the detection process and decision tree, tables are found. Being limited to the RL-Ts is the main drawback of this method. Mandal *et al.* (2006) introduced an algorithm to detect tables with homogeneous bodies. This approach is built on the assumption that in the presence of a table and its columns, the gap between fields should be larger than the one between words. The disadvantage of this method even though it works independently from the ruling lines is that it only works for single column pages.

Shigarov *et al.* (2009) presented a method for table detection using a bottom-up segmentation of the document page. This approach first groups the text blocks and then segmented them into lines and from there extract bounding boxes of each text element in the lines. Finally, some qualitative rules are applied to detect the table candidates. The main drawback of this method is its sensitivity to full page tables and mixing them with multi-column pages. Shafait & Smith (2010) developed a method for detecting tables in multi-column documents. Their method first finds the page column candidates. Based on their assumption tables should be among this candidates. Therefore, defining some conditions to differ them from columns they determine the tables.

Chen & Lopresti (2011) proposed a method for detecting tables in noisy handwritten documents. This method starts with locating and eliminating the clutters on the page borders then proceed by a bottom-up approach to divide the page into small tiles. Then by feeding the

Gradient-Structural-Concavity (GSC) features to an SVM classifier, text vs. non-text tiles are found. Then they adopt the Hu *et al.* (1999) method to detect the tables. This method is covering all types of tables. However, still sensitive to the space noise and tabular shape structures since the style of handwritten documents is messier compared to the printed ones.

Kumar *et al.* (2012) presented a method for detecting the appearance of a table on the page. They applied statistics of codeword histogram to represent the spatial relationship of the patches. The code words histogram is obtained by partitioning the image up to a particular point, horizontally and vertically. Dhiran & Sharma (2013) proposed a method for detecting every type of the tables in document images. Their approach is based on finding the projection profiles and the Hugh line detection technique. The drawback of this method is its dependency on layout analysis. The system proposed by Kasar *et al.* (2013) employs a run length approach for detecting the vertical and horizontal lines in the document pages and through that, finds the column and row line separators. Finally, 26 low-level features are defined by applying the junction points of vertical and horizontal lines and fed to an SVM classifier to detect their class as table or not. The disadvantage of this method is that it requires tables having at least one kind of ruling lines.

Ghanmi & Belaid (2014) developed a system for table detection from handwritten chemistry documents using a conditional random fields (CRF) model. The proposed method uses the CRF model to label the segmented lines of the document. They exploit features related to the neighboring lines as well. One of the main flaws of this method is again its dependency on the line segmentation. Jahan & Ragel (2014) introduced an approach that detects and extracts both RL-Ts and NRL-Ts from document images. Their algorithm uses local thresholds for word space and the line height as the fundamental elements of identifying tables. However, it is not always working correctly in the case of irregular spacing on the page. Tian *et al.* (2014) proposed a method to detect table frame lines in low-quality images. The skew corrected images are passed through a run length smoothing algorithm to discard the text lines and preserve the horizontal or vertical lines longer than a specified threshold. Then the table frames are detected using the Hough transform. This approach is only working for RL-Ts.

Seo *et al.* (2015) presented a method that locates the candidate for the cell corners and labels these junctions according to their connectivity. To detect intersections, multiple curve detection algorithms are used, and the results are then labeled into 12 labels. This labeling forms a cost function revealing the pairwise relationships. The obtained cost function is minimized adopting a belief propagation algorithm. Their method can handle multiple tables on one page. Like most other methods, this approach is also limited to the RL-Ts.

Tran *et al.* (2015) proposed method for locating tables starts by merging the neighbor objects in the image after binarizing. Using the connected components, the bounding boxes of these merged objects are found. Then a region of interest (ROI) is extracted using the information of the table frame contacting object bounding boxes' horizontal and vertical lines. If the text objects in the ROI arranged vertically and horizontally, the selected ROI is a table. Tran *et al.* (2016) developed a technique for table detection using random rotation bounding boxes in printed, scanned documents. It first classifies the page to text and non-text regions using a multi-level homogeneous structure algorithm. Later, having the non-text elements and the text-lines extracted by the random rotation bounding box algorithm, RL-Ts are detected. Also, NRL-Ts are found via identifying the homogeneous regions of the text lines.

Gilani *et al.* (2017) used a deep learning based method to solve the problem of detecting tables in the image. Their approach first transforms the image from binary to a form of RGB image using three transform functions. Then this image is fed to a fine-tuned CNN model where it gives the feature map. This output feature map is then transferred to a region proposal network to get the table candidate regions which are the input for a fully connected detection network along their primary feature map to classify the image as table and non-table.

Summing up the state-of-the-art in table detection from document images, it can be seen the limitations each suffer, dealing with the variety of tables. Some methods, just dependent on the existence of the ruling lines in the tables and not capable of handling NRL-Ts or some can only handle one-column documents. One reason for these limitations can be the fact that most of the previous works were dealing with table recognition to extract the table location and structure of

the tables thus some assumptions have been made to simplify the problem. Above all, some of the previous methods can only handle a specific class of document with a neat and clear layout and collapse in the case of degraded and old documents with unknown layouts. Therefore, there is a demand for a robust table detection algorithm that can manage all types of tables in complex heterogeneous documents. In this thesis, we limit our focus only on detecting if there is a table on the page or not. This approach aid not only to spot the appearance of a table which later can lead to locating it as well (i.e., using any of the existing methods) but also can help to describe the layout and structure of the tables.

3.3 Proposed methodology

As mentioned in the previous section there is a great need for a robust system that can detect tables despite their types. Here, we proposed a practical approach for classifying documents with challenging layout and compositions to two class of "table" and "not table." We build our hypothesis on the fact that the fundamental feature of a table is its harmonic column-wise scheme. Dealing with any harmonic pattern leads us to frequency concept. We expect that tables should have lower frequency compared to the text characters in each line since they are repeating less frequent than any possible columns of a table. Therefore, applying spectral methods should help in differentiating tables from other parts of the page. Mel frequency cepstral coefficients (MFCCs) are one of the most popular frequency-related features in the speech recognition (Huang *et al.*, 2001). They are generated through cepstral analysis and wrapped according to the Mel scale. The signature characteristic of MFCCs features is providing good discrimination of the lower frequency components against higher frequency components. This behavior makes them a good candidate for table detection purpose. For the best of our knowledge, using MFCCs for the image processing purposes have been only narrowed down to some works about the identification of satellite images, leukemia cells, palm and face and gesture recognition (Talal & El-Sayed, 2009; Biswas, 2009; Taleb & Atiya, 2017).

Our proposed method for tackling the problem of table detection in the historical documents converts images to a one-dimensional signal and then extract the MFCC-based features from

these vectors. These features are then modified and will be fed to a classifier to decide on the presence of the table on the page document.

3.3.1 MFCCs-based features

MFCCs extraction process is the same as one for voice recognition. However, this algorithm only accepts vectors as input. Thus, we have to transform our images into one-dimensional signals then make them go through the MFCC algorithm.

Before transforming the images into vectors, first, we will have a scaling step. This step is necessary for our algorithm since all images have different dimensions and resolutions. Therefore, for this research, we empirically decided to use the combination of information from two scales to conserve as much data as possible in resizing steps. These two scales are 700x450 pixels and 500x350 pixels. Transforming the resized images into vectors is done by horizontally concatenating the content of each row in the images. Figure 3.10 depicts the process we follow for extracting MFCC features from each image. Below we will describe each of these steps in more details.

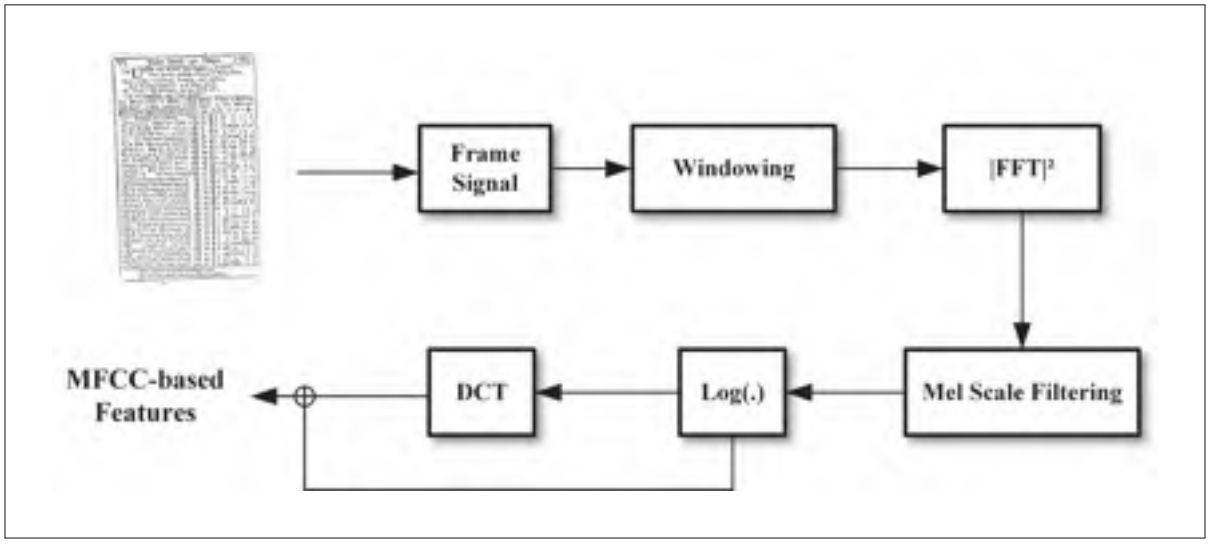


Figure 3.10 MFCC_based feature extraction process.

3.3.1.1 Frame the signal

MFCC features are spectral features and require stationary signals to achieve a reliable spectral analysis. Generally, a short length period of a signal can be considered as statistically stable. These short length trains of samples are called a frame. However, deciding on the right length of these frames is a crucial task due to the tradeoff between time and frequency resolution. Selecting frame size short leads to not having enough samples for calculating the power spectrum and on the other hand choosing this length long will result in not having samples with stationary behavior and too many changes throughout the frame.

We divide our image vectors into successive overlapping frames with the length of N samples (here pixels), and O pixels overlap for avoiding the loss of data in transition between frames. The frame overlap is generally 25% to 70% of its original length. The values of N and O for this study are selected N = 100 and N= 60 pixels. Figure 3.11 shows the process of framing for this work.

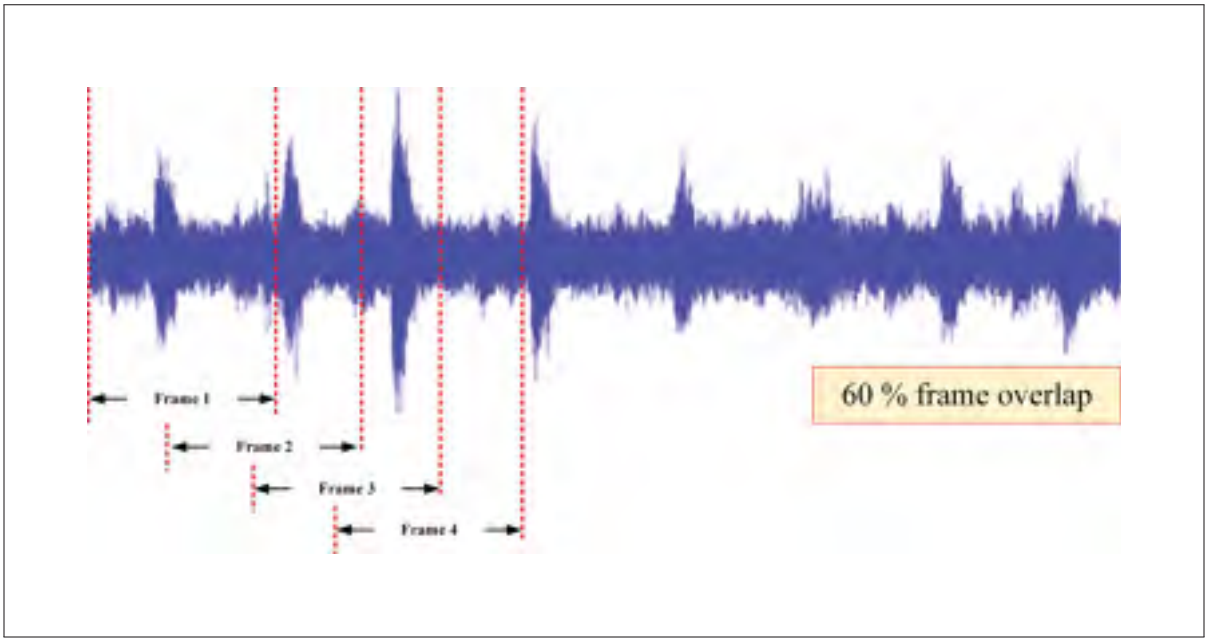


Figure 3.11 Dividing the signal into short overlapping frames.

3.3.1.2 Windowing

After framing the signal, each frame is multiplied point-wise by a window function. This step is equivalent to a convolution process between Fourier transforms of the frames and the chosen window in the frequency domain. Windowing is a necessary step to decrease the disruptions at the two ends of the frame and helps to enhance the continuity between the neighbor frames. Convolution used in windowing smooths the signal by reducing the side lobes and distribute the frequency to the surrounding bins. This makes the frame signal suitable for spectral analysis by reducing the artificial spectral disruptions in each frame. Desired window function should have a narrow main lope with low side lopes. Hamming window is one of the most popular windows uses for this purpose (see Figure 3.12). The following formula represents the hamming window:

$$w(n) = 0.54 - 0.46cos(\frac{2\pi n}{N-1}),\ 0 \leq n \leq N-1. \tag{3.1}$$

where N is the length of window.

3.3.1.3 Fast Fourier transform (FFT)

The windowed signal is now transformed to the frequency domain using the FFT to calculate the estimation of the periodogram for the power spectrums. FFT is a fast processing algorithm and generally uses in various fast algorithms to compute the Discrete Fourier Transform (DFT) of the signal. To make the process of computing FFT faster and less complicated each frame is divided into small DFTs. DFT provides a mapping between time sequences and a discrete set of frequency domain samples. The periodogram estimate of the power spectrum $s_i(n)$ is given below:

$$P_i(k) = \frac{1}{N}|\sum_{n=1}^{N} s_i(n)h_i(n)e^{-j2\pi kn/N}|^2,\ 0 \leq k \leq K-1. \tag{3.2}$$

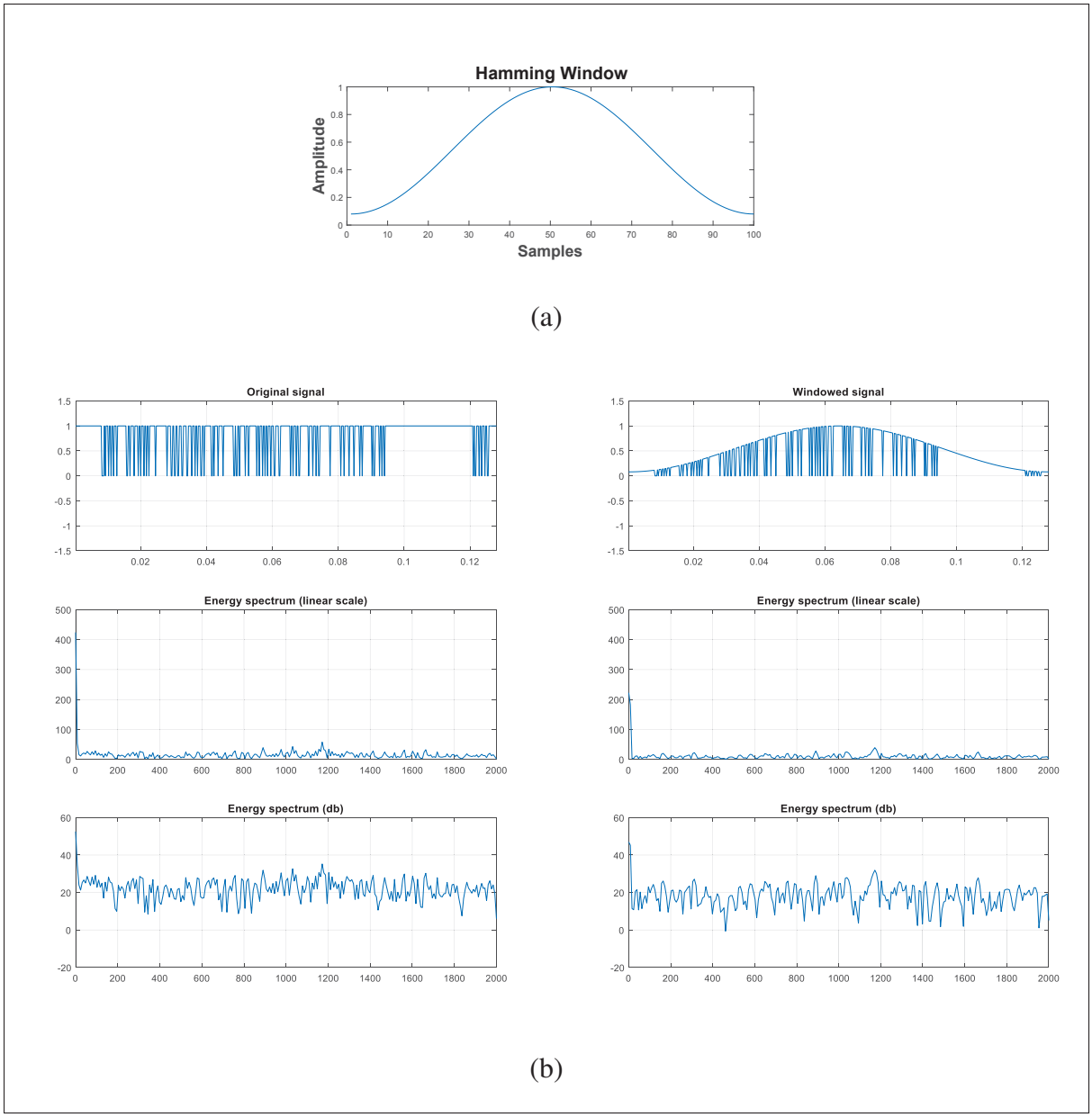


Figure 3.12 The effect of applying the Hamming window to the time and frequency domains: (a) Hamming window (b) The energy spectrums in both linear and decibel scales applied on the original signal (left column) and windowed signal (right column).

where $s_i(n)$ is the i^{th} framed signal with $1 \leq n \leq N$, $h(n)$ is the window function with sample length of N, and K is the length of FFT. For this study, a 512-point FFT is computed and the first 275 coefficients are stored for the next step. These values are common among the researchers in this field.

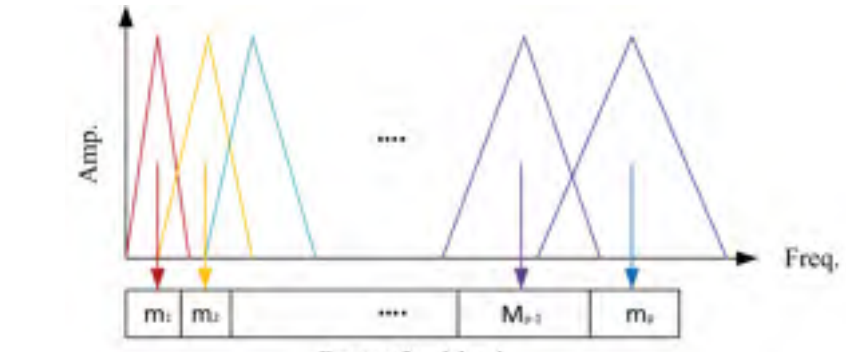
3.3.1.4 Mel filter banks

The source of the harmonics in the frequency response is the repeating fundamental periods in the frame. However, here we are more interested in the envelope of this response. We extract these envelope-like features by applying the triangular bandpass filters to smooth the magnitude spectrum. The triangular overlapping windows called Mel filter banks are used to map the power spectrums to the Mel scale. The Mel scale carries the information for space and wideness of each filter. The first filter starts from 0 Hz, and after the first filter bank, the widths of next filter banks enlarge logarithmically. Figure 3.13 (a) shows these filter banks and how they become wider as the frequency increases. A simple way of interpreting them is to view them as logarithmic filters with overlap in the frequency domain. In the following approximate formula, a subjective frequency, m, is measured on Mel scale using the actual frequency, f in Hz:

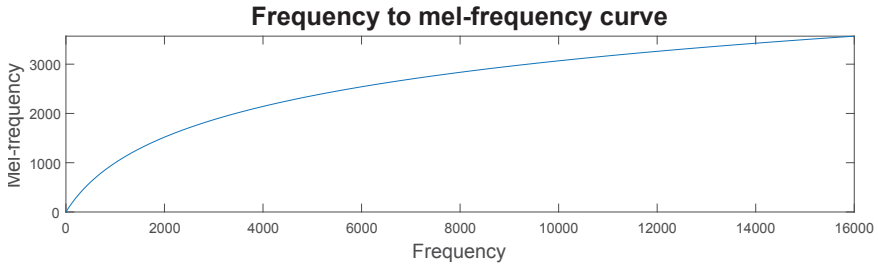
$$m_f = 2595\log_{10}(\frac{f}{700}+1). \tag{3.3}$$

Figure 3.13 (b) illustrates the relationship between frequency and Mel domains, where the curve shows linear behavior from zero to 1000 Hz and after that point, it follows a logarithmic form.

A set of filter banks are applied to the power spectrum calculated in the previous step. The standard number of filter banks used for MFCCs is between 20 to 40 filter banks. Each filter bank is a vector of size 275, mostly zero except in its bandpass. For calculating the filter banks energies, this vector is multiplied by the power spectrum of the targeted frame, and the result is summed up. The aggregated value indicates the amount of energy in that filter bank. For this thesis, we calculate these energies in the range of 64 Hz to 2000 Hz using 23 Mel filter banks, which results in having 15 coefficients for each frame.



(a)

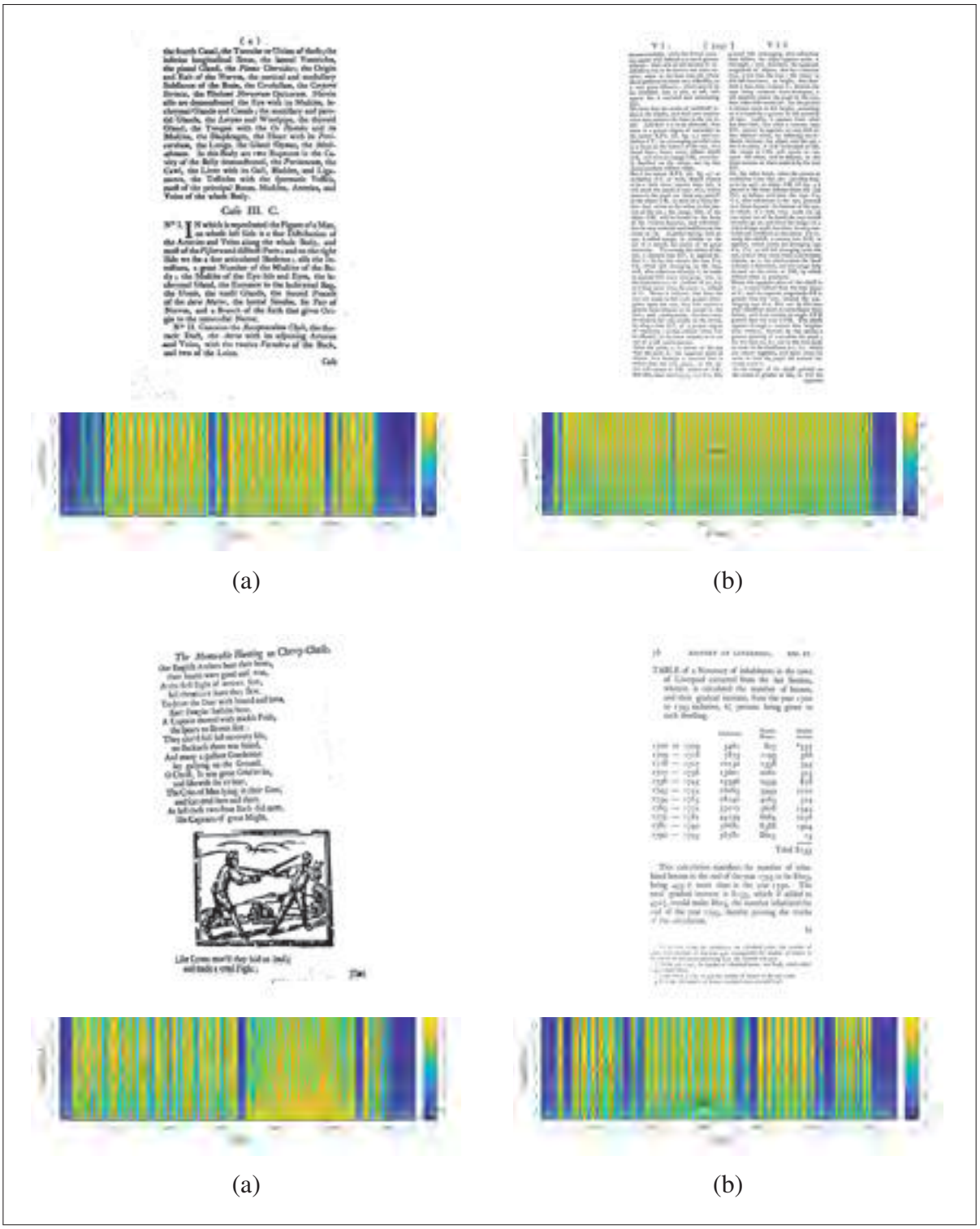


(b)

Figure 3.13 (a) Mel filter banks. (b) The plot of Mel frequency scale versus hertz.

3.3.1.5 Logarithm

The next step after calculating the energy coefficients is taking their logarithm. Logarithm function converts a multiplication into an addition, which here it serves to allow the cepstral mean subtraction for channel normalization. In another word, it helps the frequency estimates to be less sensitive to the slight power variations. What we obtain by applying the logarithm is the spectro-temporal representation of the signal, and it is called logarithmically scaled Mel-spectrogram (LogMS). Figure 3.14 illustrates the behavior of LogMS in the presence of different structural layout. It also shows how MFCC-based features can be used to characterize the document page specifically tables.



(a) (b)

(a) (b)

Figure 3.14 Examples of LogMS behavior in the presence of different structural layouts: (a) one-column structure, (b) two-columns structure, (c) figure and text structure and (d) table and text structure. Layout regions are marked with an arrow and dashed red lines show the maximum energy in that region.

In documents containing plain text lines, energy concentration is maximum around higher frequency levels or coefficients. The smaller space between characters and words results in faster changes in the frequency domain and consequently higher coefficient level. Therefore, when the font size becomes smaller and as the result character distributions in their associated location become denser, we expect the frequency of the relevant lines rises to higher levels. We can visualize this relationship between the coefficient number and the energy concentration in Figure 3.14 (a and b). The image with two-column structure compared to the one-column one has higher coefficient level since it has smaller and more compressed characters. Besides, it can be seen that for texts with the same layout, the maximum energy of the frames are all almost at the same level. However, when we have a figure on the page, unlike text lines, there is no specific pattern in the pixel appearance for that figure. Therefore, the energy content mostly concentrates near the zero frequencies with no particular pattern. Nevertheless, if the figure has a table-like visual appearance, it can be mixed with tables. Figure 3.14 (c) shows an example of an image containing both text and figure. Finally, for tables same as texts, the appearance of columns follows a frequentative pattern. However, columns have more gap in between compare to text characters. This fact leads to having lower frequency components for tables. Figure 3.12 (d) depicts an example where tables appear along text on the page. As it is clear from the figure table, contain lower frequency level compared to the text lines. Figure 3.12 (d) also shows a page document which includes different font sizes. it can be seen that smaller font sizes (footnotes here) have slightly higher coefficient level compared to the standard font size of the page and energy level of both are higher than the table energy content.

3.3.1.6 Discrete cosine transform (DCT)

Finally, for MFCCs extraction, we transform back the logarithm of Mel spectrums to the time domain. This process aids to decorrelate the filter bank energy correlations because of their overlaps. In this work, DCT preferred over the inverse DFT since log power spectrum is real and symmetric. DCT produces more information with fewer coefficients, and its output

coefficients will be real-valued and highly uncorrelated which makes both storage and further processing easier. The MFCCs are computed using the following formula:

$$c_i(m) = |\sum_{l=1}^{L}(logS_i(l))cos[m(l-\frac{1}{2})\frac{\pi}{M}]|,\ m = 0,1,...,M-1. \tag{3.4}$$

Where M and L indicate the number of MFCCs and number of filter banks, respectively. Moreover, $S_i(l)$ represents each filter banks energy.

Higher frequencies represent faster changes in the energies. Therefore, selecting them can lead to performance degradation by bringing in noisy information. Generally, only 12 to 20 DCT coefficients are chosen for the further processing. Also, the first coefficient shows the mean value of the frames and thus excluded.

The MFCCs are representing the power spectral envelops of each frame. Consequently signal can miss some local dynamic information, i.e., trajectories of MFCCs over the time. Therefore, the first and second derivative coefficients are calculated to add temporal information. First derivatives also are known as delta or differential coefficients are obtained using the regressive formula for all ms, as below:

$$d_i[m] = \frac{\sum_{j=1}^{2}(c_{i-j}[m] - c_{i+j}[m])}{2\sum_{j=1}^{2}(j^2)}. \tag{3.5}$$

Second derivatives also known as delta-delta or acceleration coefficients are computed by applying the delta to delta coefficients.

For this thesis, nine MFCCs including its delta and delta-delta coefficients (in total 27 coefficients) and 15 LogMSs are calculated to create 42 MFCC-based features for each frame of the image. Then some statistical functions are applied to reduce the dimension of feature vectors. These statistical functions are maximum, minimum, maximum position, minimum position, range, mean, variance, median, mode, entropy, kurtosis, and skewness. The distribution of

each coefficient for frames suggest using some fitting techniques. Here, we apply a symmetric generalized Gaussian distribution (GGD) and an asymmetric generalized Gaussian distribution (AGGD) fitting. From the given MFCC-based coefficients the shape, variance and entropy information are extracted using a GGD fitting. Also, shape, mean, right and left variances, skewness and kurtosis and entropy are extracted in four orientations (horizontal, vertical, main diagonal and secondary diagonal) using an AGGD fitting. Finally, we will have a feature vector of size 1516 for each image. Considering that we have done the experiments on two scales this amount becomes double (3032 features).

Finally, we apply principal component analysis (PCA) to reduce the number of variables and redundancy in our feature vector while keeping as much variance as possible. With a threshold cumulative contribution ratio of 98%, PCA resizes the size of our feature vector to 561.

3.3.2 Classification

SVM and random forest (RF) classifiers are used for table detection in this research. SVM implementation is done using LIBSVM toolbox Chang & Lin (2011) and the SVM has a degree three polynomial kernel function. We implement RF using 1000 trees, and the number of attributes selected for each tree is equal to the square root of the number of features. Ten percent of the training set is used as the validation set to optimize the parameters.

CHAPTER 4

PERFORMANCE EVALUATION OF THE RESULTS

Below, we first give a brief description of the performance measurements used for both of our specific objectives to find tables and footnotes. Next, for each of these objectives, we first describe the employed datasets. Then, present and compare their performance by applying the proposed frameworks using different experimental set-ups on these datasets.

4.1 Performance metrics

In order to evaluate any system, we need to measure some performance metrics. For visual information retrieval purpose, precision and recall are two widely used measurements (Hedjam, 2013). One observes the performance from correctly spotted samples view and the other from the incorrectly spotted ones.

Recall or sensitivity is the ratio of correctly retrieved samples (true positive) to the all relevant samples. In another word, it indicates how many of page with footnotes/tables are correctly selected from the whole pages that contain footnotes/tables in the test set.

$$Recall = \frac{t_p}{t_p + f_n} \tag{4.1}$$

where t_p is true positive and f_n is false negative which is equal to the total number of non-spotted samples.

Precision or positive predictive value is the fraction of retrieved samples that are relevant to the query. In another word, it is the probability of that the retrieved sample belongs to its target image or in our case, how many of page with footnotes/tables are correctly selected from the detected pages as pages containing footnotes/tables.

$$Precision = \frac{t_p}{t_p + f_p} \tag{4.2}$$

where false positive (f_p) is the total number of misrecognized samples.

Finally, F_measure or balanced F_score is a measure that combines recall and precision and finds the harmonic mean of these to metrics.

$$F_{measure} = 2.\frac{Precision.Recall}{Precision + Recall} \tag{4.3}$$

4.2 Footnote-based document image classification

The proposed footnote-based document image classification approach is evaluated using our dataset and different experimental set-ups. These experiments show the performance, robustness and generalization power of our proposed framework. Below, we describe the dataset and the used experimental set-ups.

4.2.1 Dataset

One of the significant challenges in footnote-based document images classification is the lack of research in this field and consequently, lack of available datasets for experimental evaluation. Therefore, a dataset of labeled documents was generated for conducting our experiments and creating a common benchmarking mechanism for future assessments.

The generated dataset is part of "The Visibility of Knowledge" [1] project and its images are drawn from the digitized collection of eighteenth-century Collections Online (ECCO)[2]. ECCO is an online archive for historical documents printed between the course of 1,701 and 1,800

[1] https://txtlab.org/2016/09/the-visibility-of-knowledge/

[2] https://www.gale.com/primary-sources/eighteenth-century-collections-online

in Britain. It contains over 32 million manuscript pages from more than 155,000 volumes and eight subjects. These images are distributed over two subsets:

- **ECCO part I (ECCO_I):** Contains almost 150,000 manuscripts and about 26,000,000 pages.
- **ECCO part II (ECCO_II):** Contains almost 50,000 manuscripts and about 6,000,000 pages.

Figure 4.1 gives more insights into the distributions of the ECCO dataset over the time and subject. it can be seen that the pages are randomly distributed over the time, and also as expected by reaching the end of 18^{th}-century, there are more published printed volumes and pages. The subjects of ECCO are "Religion and Philosophy" (RelAndPhil), "Literature and Language" (LitAndLang), "History and Geography" (HistAndGeo), "Social Science and Fine Arts" (SSAndFineArt), "Medicine, Science, and Technology" (MedSciTech), "Law" and "Reference" (GenRef). From Figure 4.1, it is clear that high portions of the volumes or pages available in this dataset belong to the three categories of RelAndPhil, LitAndLang and SSAndFineArt.

For creating our ground truth, around 27,500 images from ECCO_I database have been randomly selected to keep the homogeneous distribution of the ECCO. These images are then labeled according to having a footnote (FN) or not having a footnote (NFN) in five subsets. Labeling is done using human sciences experts at McGill University. The final dataset contains 20,966 images labeled as NFN and 6,292 images labeled as FN. Figure 4.2 gives more information about each subset and their content. We will use these labeled subsets to create our train and test sets.

4.2.2 Results and discussion

We conducted two sets of experiments before our final evaluation of more than 32 million unlabeled data. Figure 4.3 gives more details about each experiment set-ups and their training

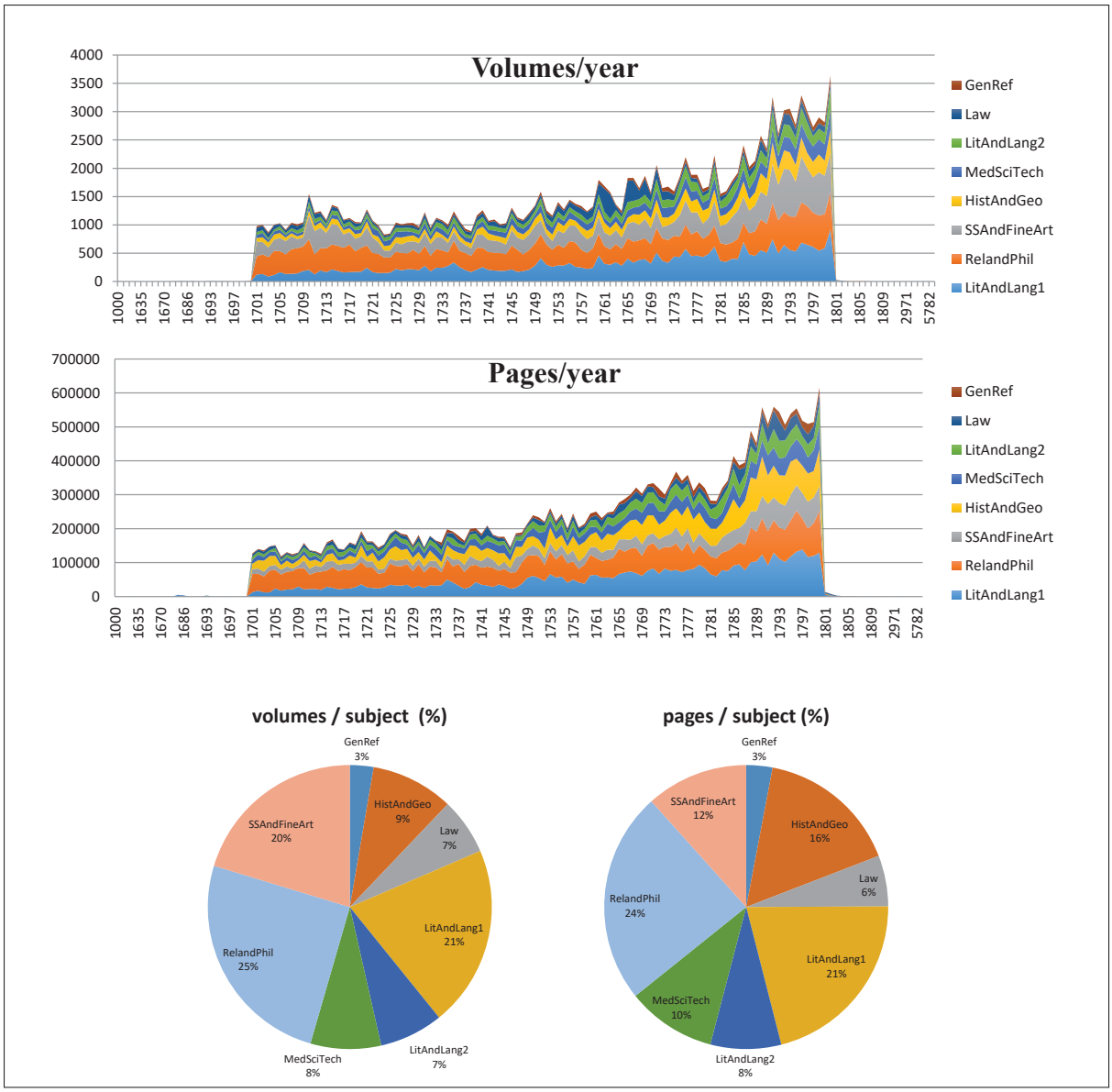


Figure 4.1 ECCO dataset distributions over time and subjects.

and test sets distributions. Below, we explain each of these experiments and discuss their results in more details.

4.2.2.1 First Experiment

This experiment was done using Trainset_0 as our main dataset and adopting two protocols. The first protocol uses a 10-fold cross-validation (CV) set-up where the all samples are ran-

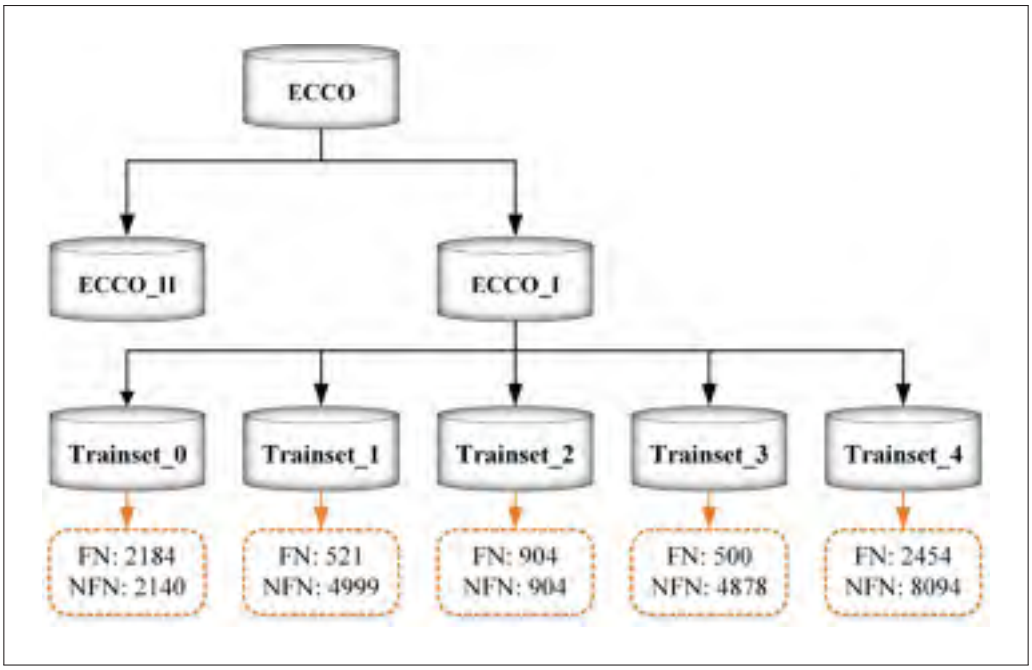


Figure 4.2 The information of the labeled dataset along its subsets. The numbers in the dashed orange boxes represent the number of images from each class of FN and NFN for each subset.

domly divided into 10-folds. In each testing cycle, one fold is assigned to the test set and the rest of the folds to the training set. The average result of repeating this process for each fold gives the final performance. One fold of the training set was used as the validation set whenever we needed to find the parameters. The results using 10-fold CV are reported using an SVM classifier in Table 4.1. The table is split into four sections. The first three ones show the results for each of font size and spaced based methods and the last row gives the results after fusing all the features to get the final 72 features.

The spaced based method clearly outperforms the BBox and Proj methods. However, spaced based method uses the footnote line estimation information from the BBox and Proj methods. Therefore, it can not be claimed that the spaced-based method is entirely independent of the two other methods. From the last row, it is evident that the combination of features reduces the precision value at the price of increasing the recall value. However, this fusion improves the overall performance of the system since it increases the discrimination power of the features.

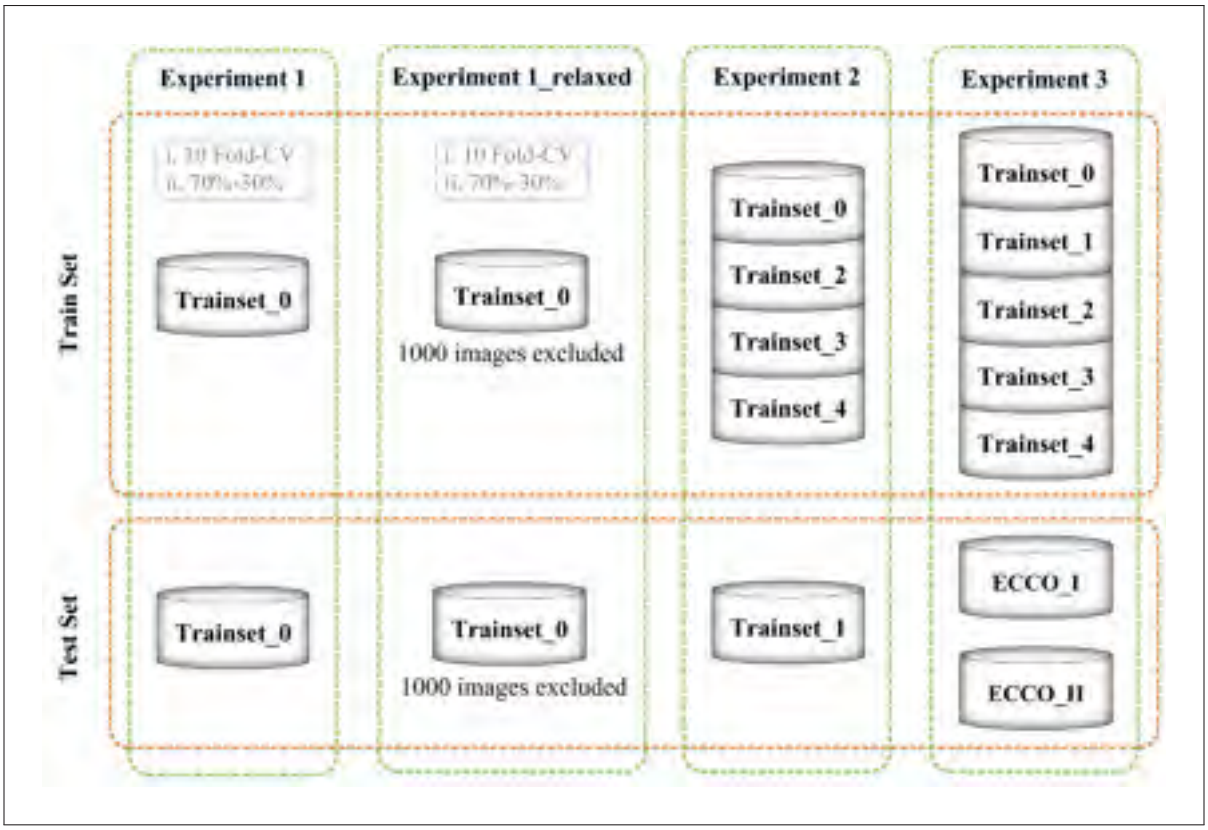


Figure 4.3 Experimental set-ups for each experiment.

One of our observations from the Table 4.1 is that the in all the cases the value of precision is higher than recall. This can be interpreted as the existence of some images that catching footnote in them is difficult. Checking these false negatives makes it clear that in most of the cases the misclassification is due to the complex and irregular layout of the page which makes finding text lines and consequently footnotes harder. These hard to catch footnotes can be in images containing two columns, figures, tables, formulas, Greek letters, etc. (See Figure 4.4). In order to see our method's performance in the absence of these types of images, we relaxed our data set by excluding 1000 images with complex layouts and repeat the experiments. The result of this step is given in Table 4.2. It can be seen that the value of both recall and precision have been increased however the best improvement belongs to the recall value since we have eliminated the complex images. Therefore, we can conclude that the lower precision of the

full Trainset_0 is mostly due to the complex layout of the historical documents rather than the weakness of the method itself.

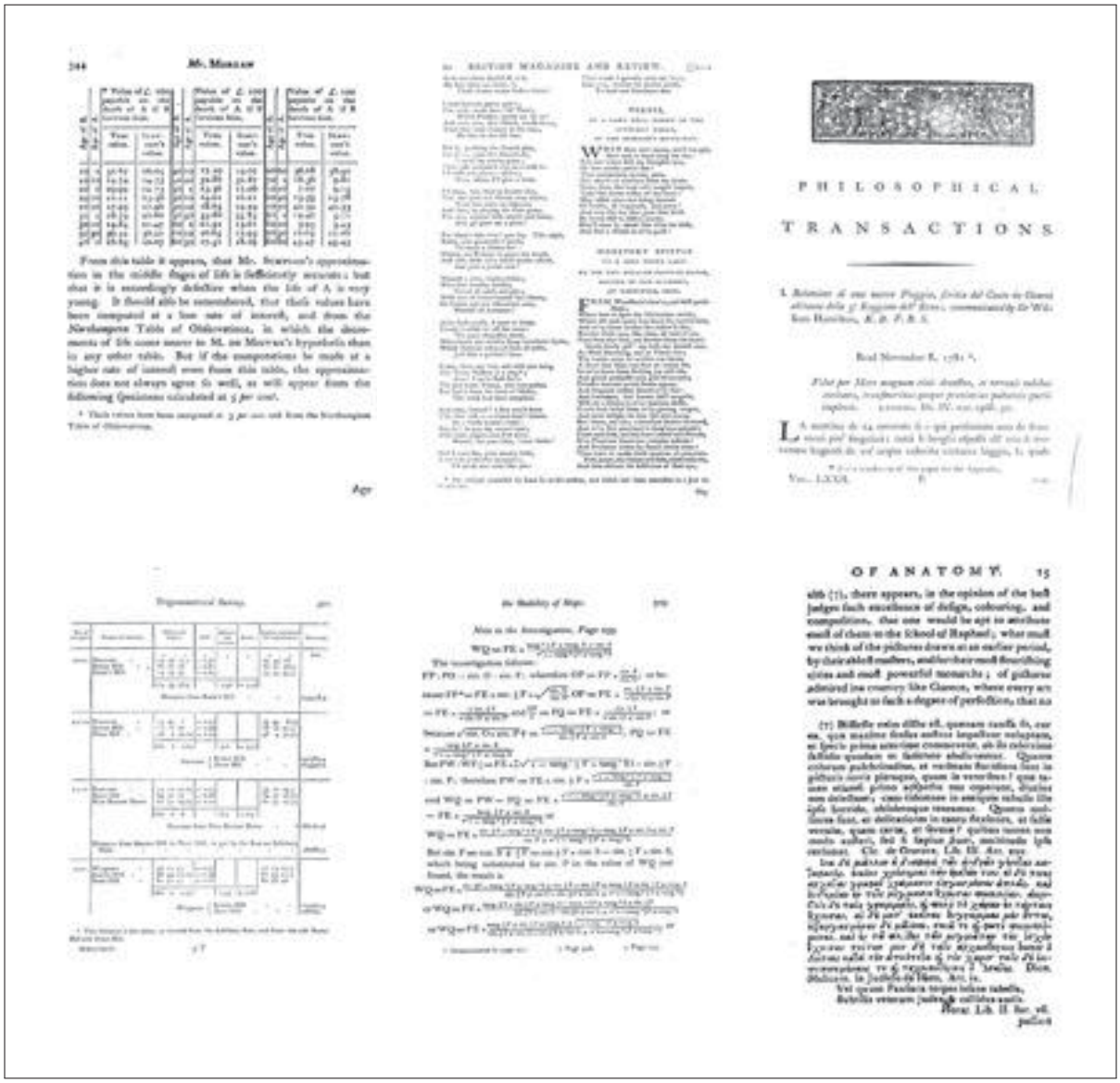


Figure 4.4 Examples of images containing footnotes with a complex layout (e.g., two columns, figures, tables, formulas, Greek letters), which are hard to catch.

The second protocol used the 30%-70% set-up for dividing the dataset to the test and training sets. This scheme was done to make the comparison with other benchmark methods (Abuelwafa *et al.*, 2017; Mhiri *et al.*, 2017) possible. These benchmark methods, as described in the

Table 4.1 Experimental results using SVM and 10-fold CV set-up for the first experiment. Values are in percent and the bold numbers represent the best results.

Methods	Precision	Recall	F_measure
Single methods results:			
BBox	79.38	70.21	75.62
Proj	86.62	70.82	77.93
Space	**87.54**	74.77	80.66
Feature level fusion results:			
BBox+ Proj+ Space	83.33	**81.31**	**82.31**

Table 4.2 Experimental results using SVM and 10-fold CV set-up on the relaxed dataset for the first experiment. Values are in percent and the bold numbers represent the best results.

Methods	Precision	Recall	F_measure
Single methods results:			
BBox	85.07	77.45	81.08
Proj	90.35	79.00	84.30
Space	**91.07**	84.25	87.52
Feature level fusion results:			
BBox+ Proj+ Space	90.24	**86.21**	**88.17**

state-of-the-art section for footnotes (Chapter 2.2), both use deep learning based methods with minimum image processing approaches or any hand design features. Table 4.3 shows the results using the complete images of Trainset_0, using our proposed approach and the benchmark methods. Our proposed method has the highest precision among all other methods and competitive value of recall. Overall, our approach has the similar F_measure with the other proposed method. Also for relaxing the dataset case, it boosts the recall in favor of our proposed method while keeping the precision still high (See Table 4.4). Our method outperforms the others in the absence of any complex layout. Figure 4.5 illustrates this performance enhancement more clear.

Table 4.3 The first experiment results over 30%-70% set-up and their comparison with the state-of-the-art methods. Values are in percent, and the bold numbers represent the best results.

Methods	Precision	Recall	F_measure
Our proposed method	**83.89**	82.64	**83.26**
Method of Mhiri *et al.* (2017)	83.26	82.37	82.81
Method of Abuelwafa *et al.* (2017)	79.12	**85.71**	82.28

Table 4.4 The first experiment results over 30%-70% set-up on the relaxed dataset and their comparison with the state-of-the-art methods. Values are in percent, and the bold numbers represent the best results.

Methods	Precision	Recall	F_measure
Our proposed method	**89.91**	**87.44**	**88.66**
Method of Mhiri *et al.* (2017)	89.35	77.97	83.27
Method of Abuelwafa *et al.* (2017)	82.57	87.26	84.85

4.2.2.2 Second experiment

This experiment was done using over 27,000 labeled images. Trainset_1 as the test set and combination of Trainset_0, Trainset_2, Trainset_3 and Trainset_4 as the training set. This

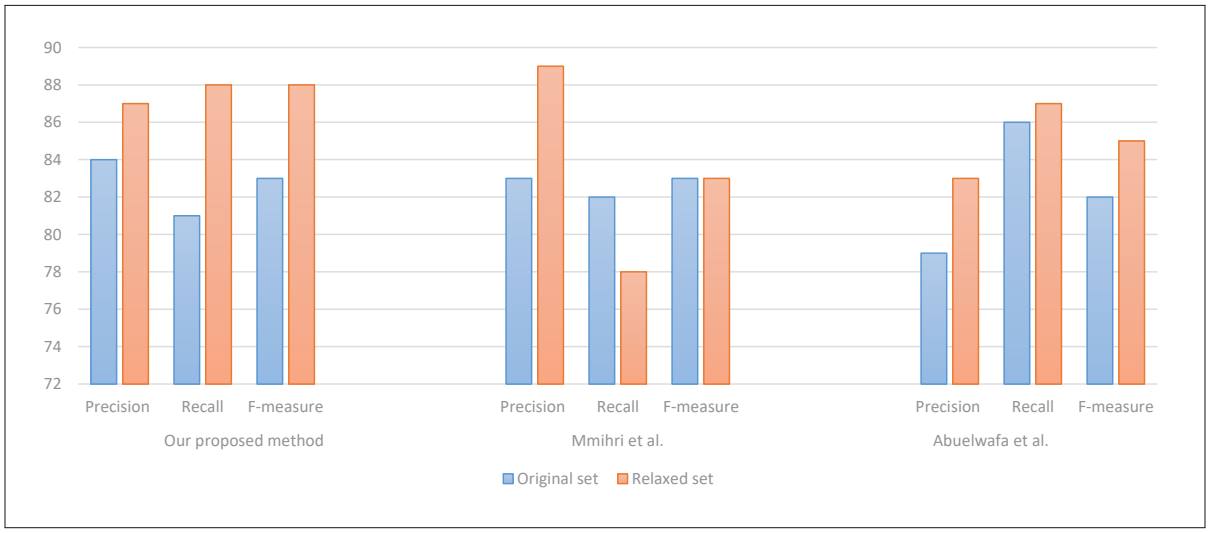


Figure 4.5 Result comparison for each method before and after relaxation. Values are in percent.

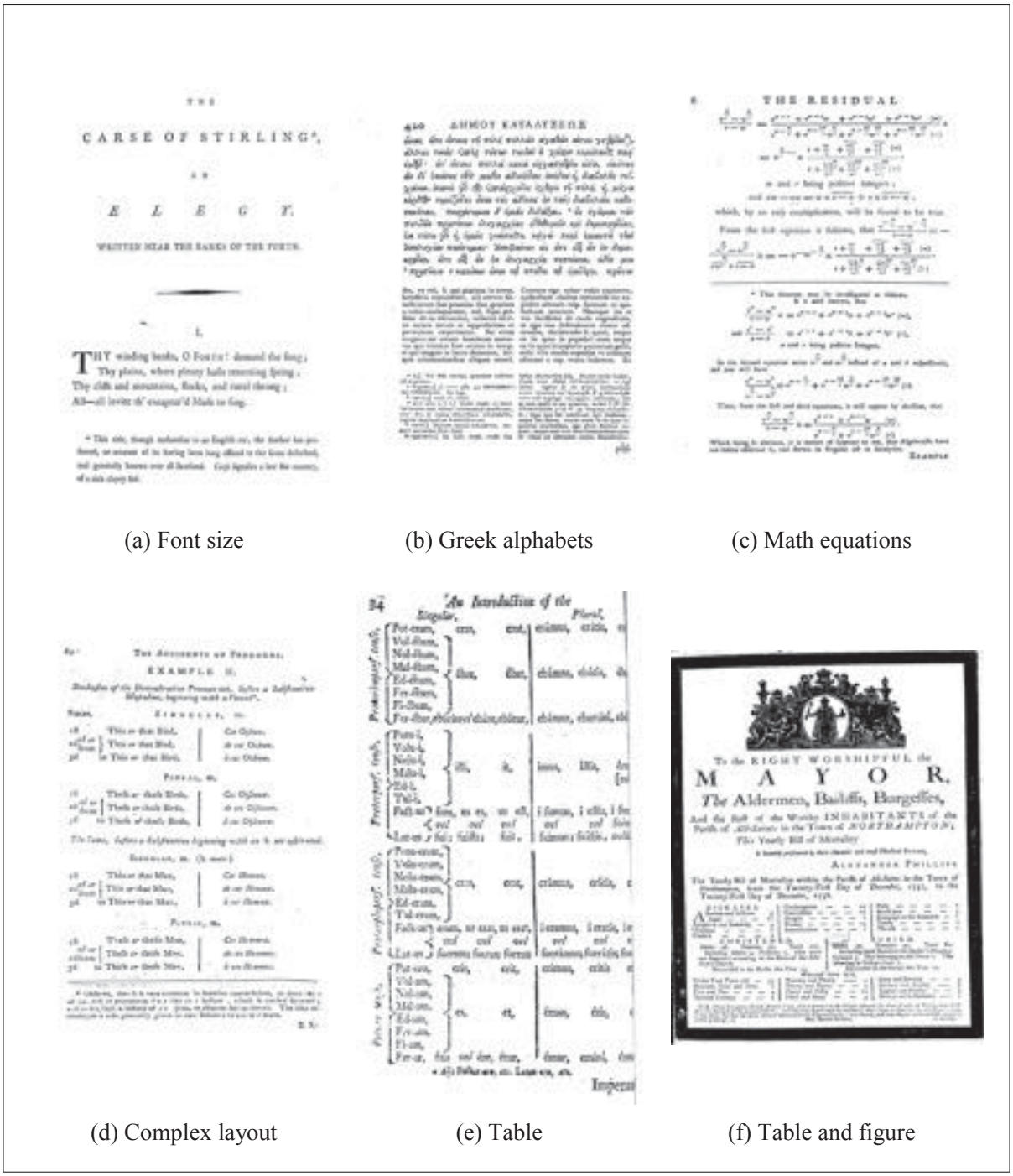
creates a training set of size 22,035 (FN: 6,071 images, NFN: 15,965 images) and a test set of size 5,520 (FN: 521 images, NFN: 4,999 images). The results of this experiment utilizing this dataset along two other benchmark methods are given in Table 4.5. We need to mention that updated versions of the benchmark methods are used to report their results. Our proposed method among others has the highest recall. This means that it is more capable than the other two deep learning-based methods in detecting a high portion of pages containing footnote. Also, even though our approach has lower precision value, it still has the best performance when we consider the F_measure value. Therefore, it has a better trade-off between the recall and precision values with respect to the others.

An overall look at the false positives and false negatives of our proposed method can give us good feedback regarding the strengths and limitations of our approach. While our method is capable of handling complex layout and structures, there are still some footnoted pages which are missed in our classification step. This kind of errors usually happens because our method is bind to its first step by layout analysis to extract lines. Therefore, even though precise measurement have been taken into account in the process, the appearance of some unusual structure can affect the results negatively. For example, having alternative font size changes on the page; or the occurrence of table and figure and a page containing a formula, Greek or Persian alphabets, all can be the source of an error. Besides, images that are highly degraded can cause errors. Moreover, since our hypothesis is built on the smaller font size of the footnote, the occurrence of smaller font size like notes and annotation at the bottom of the page can be mistakenly considered as a footnote. Figure 4.6 and Figure 4.7 show respectively examples of false negative images and false positive images using the proposed method.

Table 4.5 Experimental result for the second experiment. Values are in percent and the bold numbers represent the best results.

Methods	Precision	Recall	F_measure
Our proposed method	66.9	**59.8**	**63.15**
Method of Mhiri *et al.* (2017) *	**88.21**	47.4	61.66
Method of Abuelwafa *et al.* (2017)*	72.26	40.49	51.89

*The reported results are based on an updated version of proposed methods.



(a) Font size (b) Greek alphabets (c) Math equations

(d) Complex layout (e) Table (f) Table and figure

Figure 4.6 Examples of false negative images for the second experiment.

(a) Notes (b) Greek alphabet (c) Smaller font size

(d) Notes (e) Degraded image (f) Smaller font size

Figure 4.7 Examples of false positive images for the second experiment.

4.2.3 The third experiment

We showed the generalization power of our method by applying it to remaining images of the ECCO_I and the entire ECCO_II images (approximately 32 million images). This experiment was done using all available labeled data (27,588 images) as the training set and both ECCO_I and ECCO_II separately as the test sets. Applying the proposed method to the full ECCO dataset has led us to detect 2,551,827 and 587,928 images containing footnote for ECCO_I and ECCO_II, respectively. Figure 4.8 and Figure 4.9 give more information about the number and distribution of the detected images containing footnote throughout the years and subjects. From these two figures and comparing the distribution charts of the full data and footnoted data, it can be seen that the distribution of page with footnotes are very consistent with their complete datasets over the time and subjects. This can be interpreted as our proposed method has not missed footnotes in any subject or period during the detection process.

We expect that our method is capable of detecting footnote images with relatively high precision in both datasets. However, evaluating the performance in this scenario is a troublesome challenge since labeling over 32 million images will be a tedious and costly task. Therefore, we narrowed down our evaluation to only detect the precision of our method on sample sets of the ECCO_I and ECCO_II. For this mean, we sampled 1,500 images from the images detected as the footnote in each of ECCO_I and ECCO_II and labeled them. The final performance was estimated using the precision of detection over these 3,000 images.

Table 4.6 shows the precision results on the proposed method in comparison with state-of-the-art methods. Our method has the higher precision on the sample set compared to the deep learning based approaches. This proves that our approach was capable of keeping up with its performance even with a large increase in the amount of data. One reason that can justify the deep learning based methods lower performance is the fact that their techniques either partially use the image information on the bottom and top of the images, or their histogram, which both are clearly not enough for deciding on the presence of footnotes in a million scale. Also, maybe the training data is not capable of generalizing the whole ECCO set, and when the

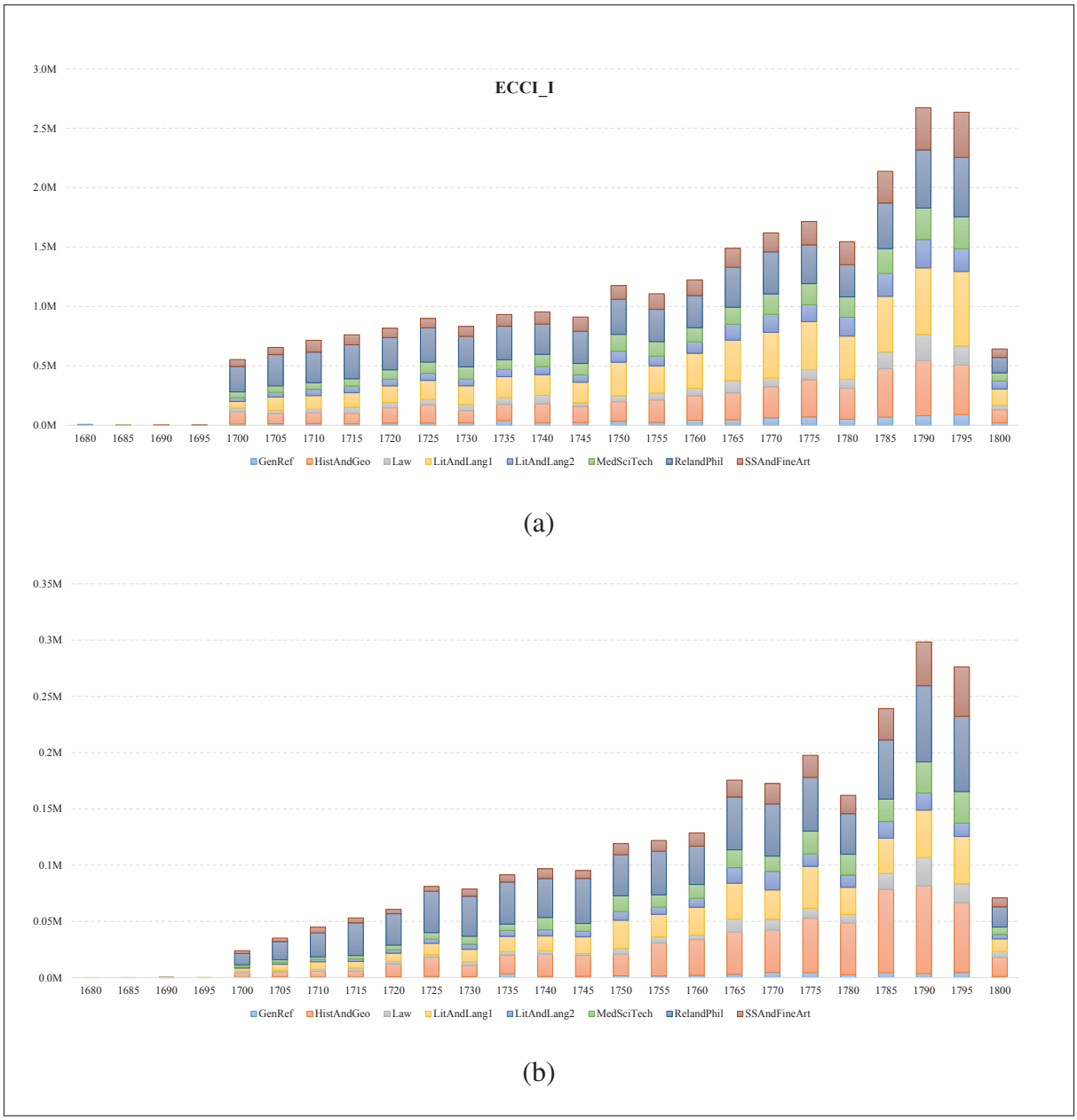


Figure 4.8 The histogram charts for the distribution of the images over the time in ECCO_I dataset: (a) all pages (b) detected pages with footnotes.

learning model encounters a new image that has not seen its similar ones before, it collapses. Thus, adding more training data may help in the improvement of the results.

From the result evaluation of these two sampled subsets, it can be seen that our method has been mostly successful in detecting the pages containing footnotes. However, there are some

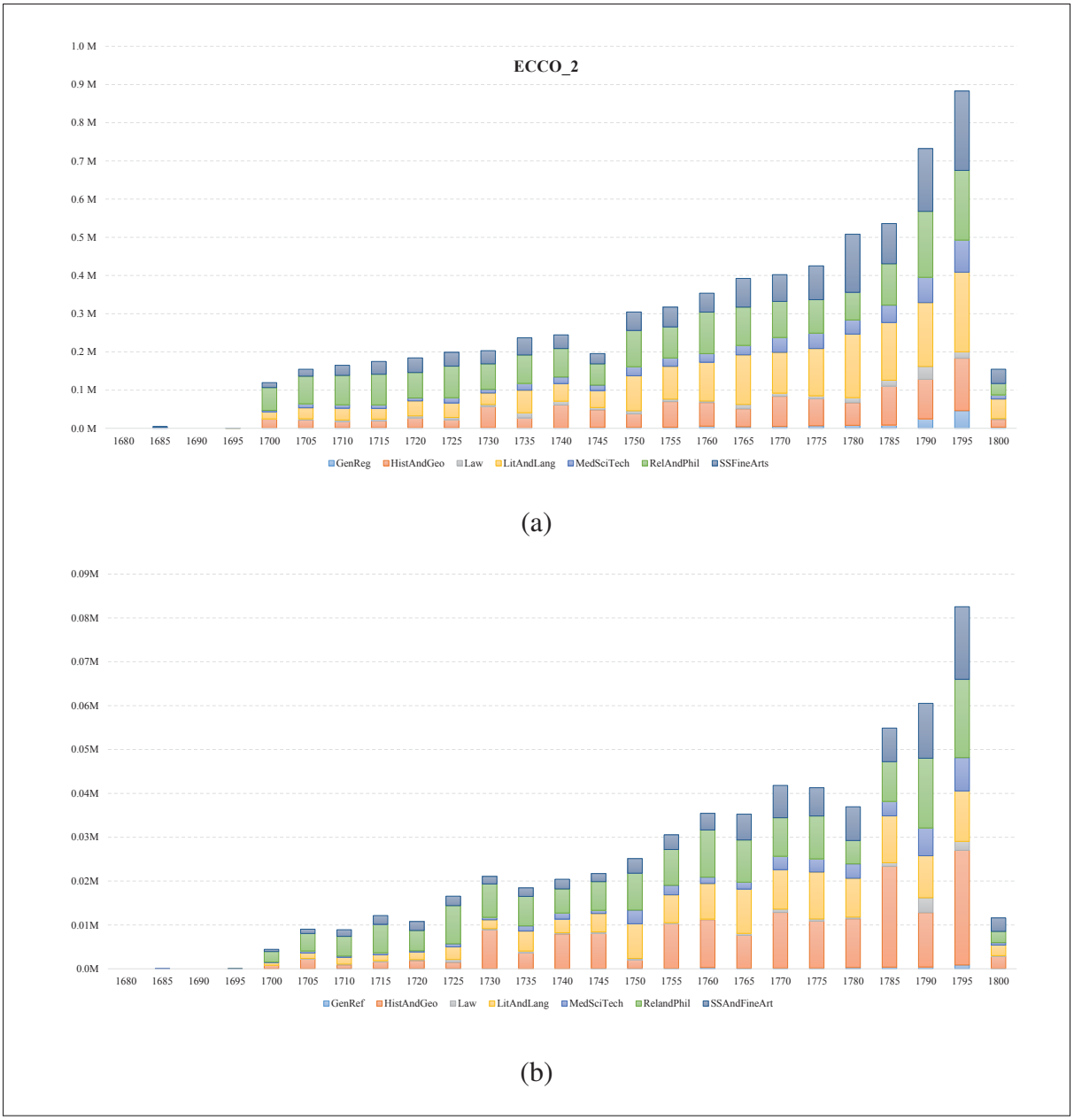


Figure 4.9 The histogram charts for the distribution of the images over the time in ECCO_II dataset: (a) all pages (b) detected pages with footnotes.

common errors. One happens when the pages contain annotation or notes. Another source of common mistakes is when everything indicates the existence of the footnote even the font size, but there is no marker. Differentiating between these notes or annotation and the footnote is hard to even human eyes.

Table 4.6 Precision comparison of ECCO_I and ECCO_II.

Dataset	Our proposed method	Method of Abuelwafa *et al.* (2017)*	Method of Mhiri *et al.* (2017)*
ECCO_I	**91.13**	75.73	26.33
ECCO_II	**91.40**	76.00	25.20

*The reported results are based on an updated version of proposed methods.

4.3 Table-based document image classification

Our proposed algorithm for table detection in historical document images was evaluated on two datasets. Different experiments were also conducted to test the performance and robustness of our method.

4.3.1 Datasets

We used two datasets for evaluating our proposed method, a dataset of handwritten and printed Arabic documents and a dataset containing printed documents from 18^{th}-century.

The purpose of using the first dataset was validating our method against the work done by Kumar *et al.* (2012). There is a lack of publicly available datasets with a fair number of images, which makes the comparison with other benchmarks harder. Their dataset contains 618 Arabic document images, where 216 images belong to a collection of hand-drawn and printed document pages with a table and 402 images are from mixed form document pages. Figure 4.10 and Figure 4.11 illustrate some examples from two classes of this dataset.

Most available datasets are designed to detect the region of the table and localize it. They usually contain less than 200 tabular images, which are defiantly not enough for training purposes. More importantly, they generally only target a specific type of tables and not cover both NRL-Ts and RL-Ts. To overcome this shortcoming, we generated a dataset called Dataset_v01 using the images available in ECCO collection of 18^{th}-century printed images. This dataset is built from two subsets. The first subset and the second one are both randomly selected from the combination of ECCO_I and ECCO_II to keep the similar distributions as ECCO_I and

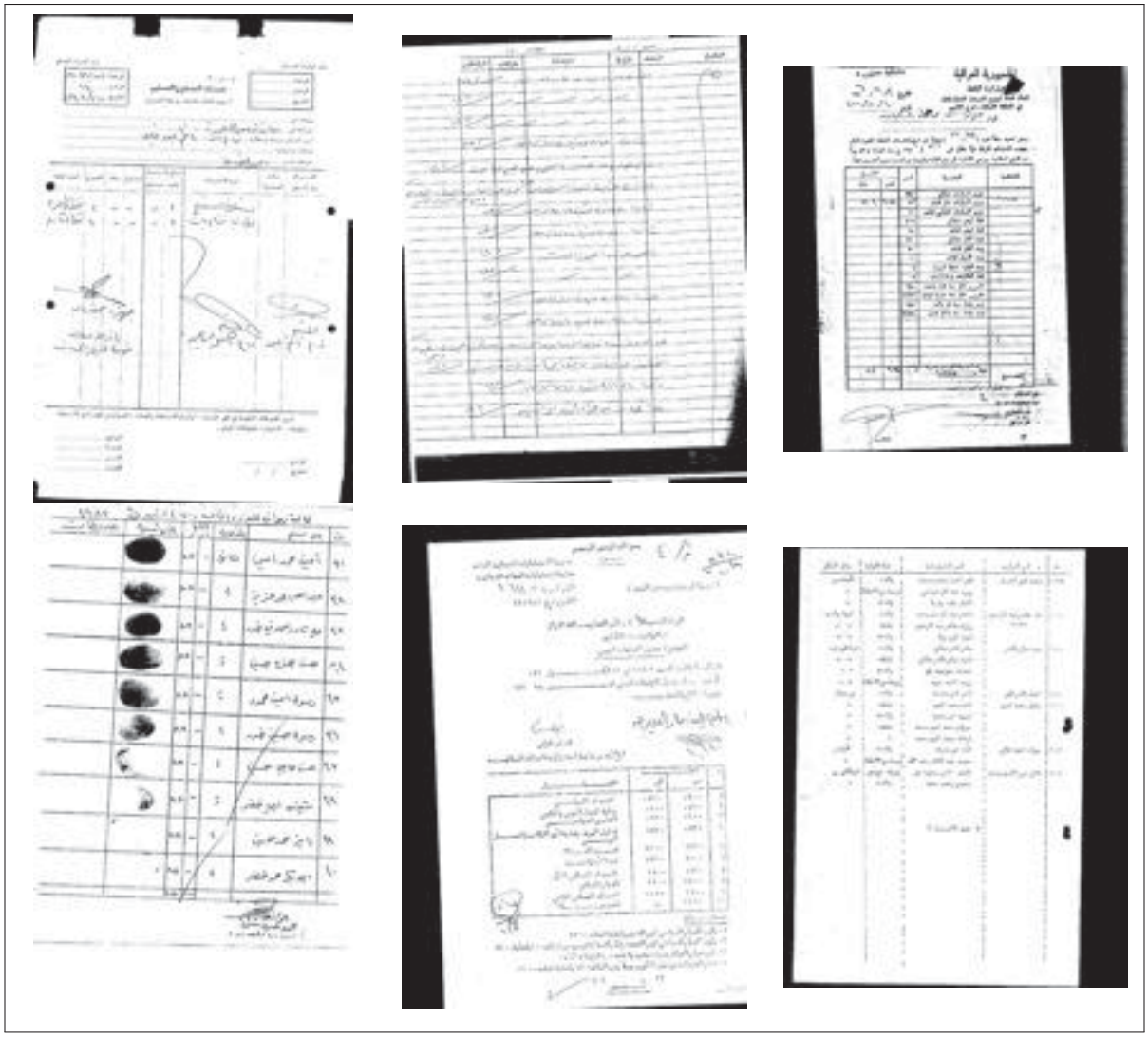


Figure 4.10 Examples of hand-drawn and printed document pages with a table from the dataset of Kumar *et al.* (2012).

ECCO_II over the time and subjects (see Figure 4.12). These images are then labeled according to not having a table (NT) or having the general (T2) and loose (T1) definition of tables. These two subsets are called Dataset_v0 and Dataset_v1, where the former has 6,406 images, and the later has 6,728 images. Figure 4.13 gives more insight into the content of each dataset.

Figure 4.11 Examples of mixed form document pages from the dataset of Kumar *et al.* (2012).

4.3.2 Results and discussion

The proposed table-based document classification approach was evaluated using the aforementioned datasets and different experimental set-ups. These experiments show the performance and generalization power of our proposed framework. Below, we briefly describe each experiment and discuss regarding their results.

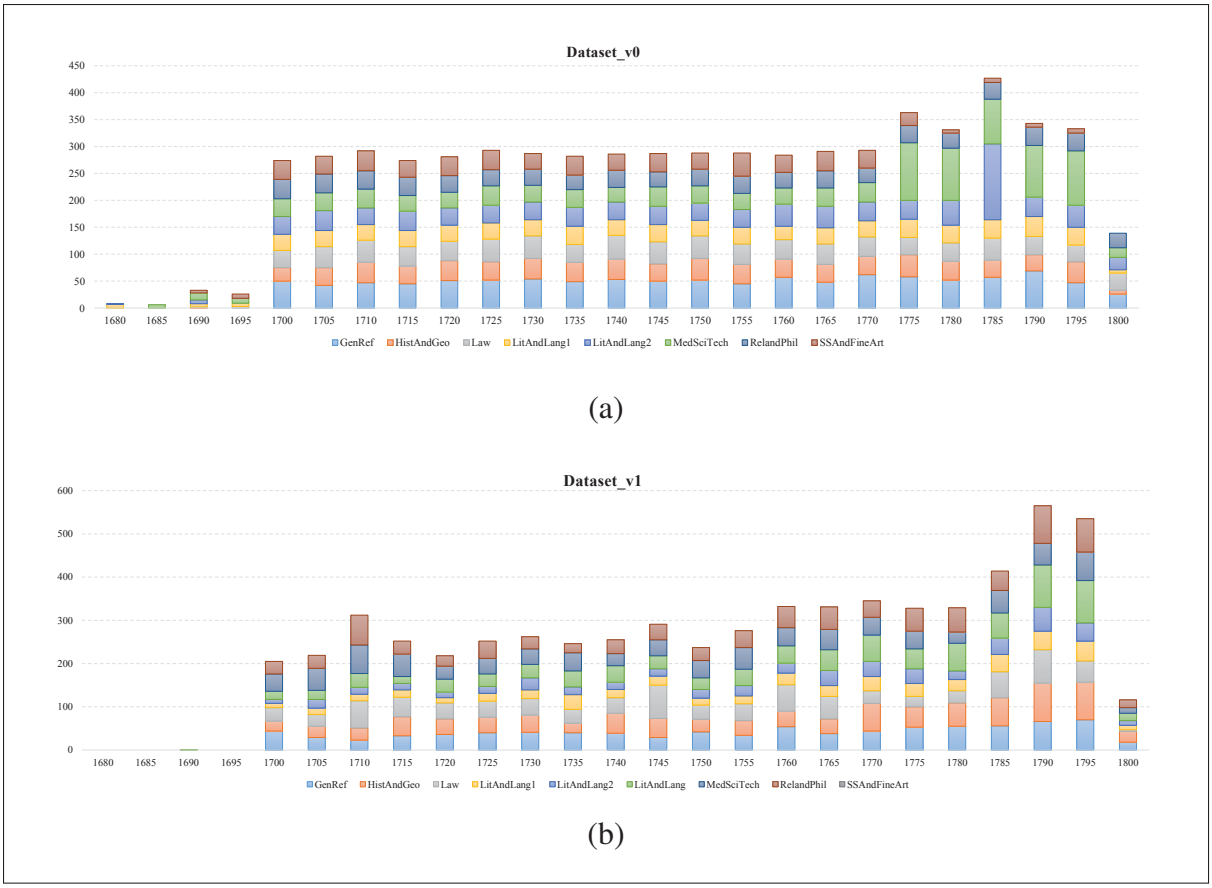


Figure 4.12 The histogram charts for the distributions of the images over the time at in (a) Dataset_v0 (b) Dataset_v1.

4.3.2.1 Experiments on the first dataset

To compare our method with the proposed approach of Kumar *et al.* (2012), we used a 5-fold CV set-up by randomly dividing their dataset to five folds and used one fold as the test set and the rest as the training set in each round. In any case, we needed a validation set to predict the parameters; we used one fold of the training set. The proposed method by Kumar *et al.* (2012) uses a 6,300-dimensional feature vector called horizontal-vertical pooling (HVP). They tested the accuracy of their system using two classifiers; SVM and RF. Using the same classifiers, we compared our method and the results are given in Table 4.7. These results are the accuracy values since Kumar *et al.* (2012) reported their results only using this metric. A brief look at the table makes it clear that our method outperforms their method by applying any of the

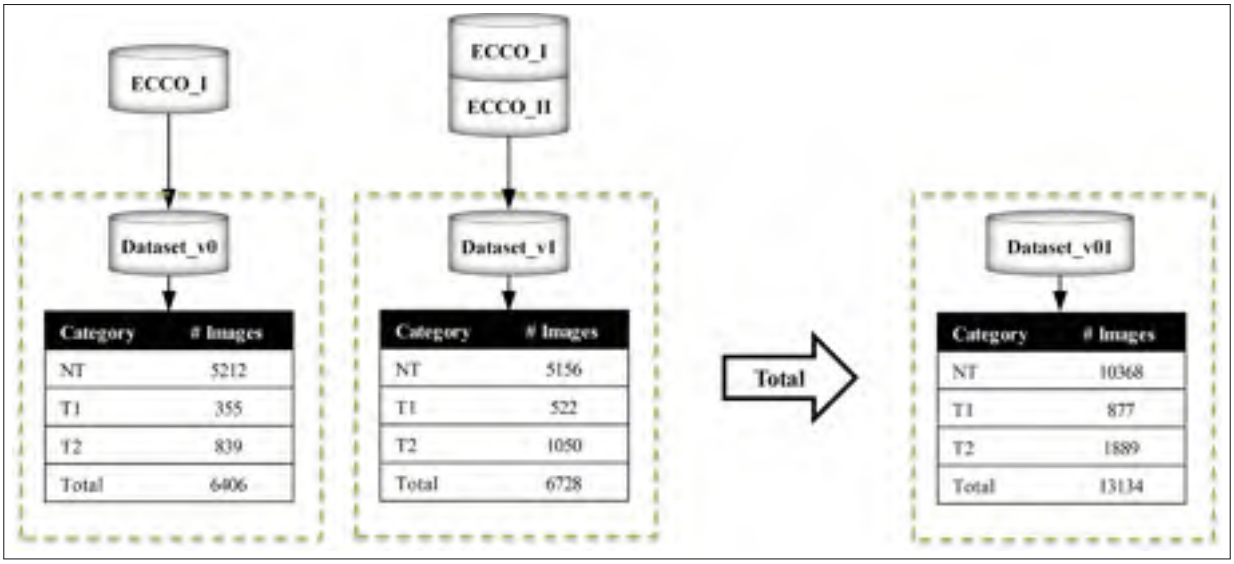


Figure 4.13 The information on the labeled table dataset (Dataset_v01).

classifiers and with a fewer number of features (557 features for our proposed approach). It is good to mention that the high amount of accuracies (above 98%) is partially related to the nature of this dataset that only contains either forms or tables. This makes the discriminating of the classes easier for the classifiers. However, this does not undermine the value of this dataset especially that the tables are from both categories of NRL-Ts and RL-Ts.

Table 4.7 Experimental results using the first dataset and 5-fold CV along its comparison with the proposed method by Kumar *et al.* (2012). Values are in percent, and the bold numbers represent the best results.

	HVP SVM	MFCC based SVM	HVP RF	MFCC based RF
Table	93.2	94.4	97.4	**97.84**
Forms	97.6	98.43	98.9	**99.67**

4.3.2.2 Experiments on the second dataset

As mentioned in chapter 3.1.2, T1 tables or the loose definition tables are a subset of the T2 or general tables. In order to evaluate our table detection approach and in the meantime to

study which category these loose tables belong; no tables or general tables, we conducted the experiments using two schemes:

1. **Combining T1 and T2 and classify them against NT (NTvsT1T2):** This scheme considers T2 and its subset T1 as tables. We expect a lower performance in the experiments using this manner since the T1 class is visually the less similar class to T2, and even though it is a subset of T2 differentiating T1 and T2 is a complicated process.

2. **Combining T1 and NT and classify them against T2 (NTT1vsT2):** This step considers the loose definition as an NT class. However, due to the non-tabular appearance of some of the T1 class members, the performance is expected to be higher.

We performed each of these schemes on Dataset_v0, Dataset_v1 and the combination of these two (Dataset_v01) using a 5-fold cross validation set-up. As before, one fold of the training set was assigned as the validation set to find the parameters of the classifier. Table 4.8 , Table 4.9 and Table 4.10 show the classification results for each dataset using these schemes. Our concern is more about the recall and precision values of either T1 or T1T2 classes depending on the used scheme.

The classification results confirm our prediction that mixing T1 with NT (NTT1vsT2) improves the table detection process compared to mixing T1 with T2 (NTvsT1T2). This improvement happens mostly on Dataset_v0 rather than Dataset_v1. One reason can be some mislabeled images in the T1 class of Dataset_v1. From the classification results, we can also observe that in all cases we have precision above 83%, which indicates two points. First, the ability of our approach to detect tables with high precision and second the generalization power of our method. This value for the Dataset_v01, with over 13,000 images is 86% and considering the complex layout of our document images; it is an impressive amount. However, the dissimilar and heterogeneous layout of the T1 class makes finding all tabular images from T1 especially in the scheme of NTvsT1T2 challenging, and reduce the recall value to 63%. For Dataset_v01 and the NTvsT1T2 scheme this value is 66% and by changing the scheme to NTT1vsT2 recall increases to 71.80%.

Figure 4.14 and Figure4.15 illustrate some of the true positive examples of our algorithm for each of these scenarios. From the figure, it can be seen that the results are from a wide range of either RL-Ts or NRL-Ts, which proves the effectiveness of our method considering the complexity of these images.

Table 4.8 Classification results using two schemes of NTvsT1T2 and NTT1vsT2 over Dataset_v0. Values are in percent.

	NTvsT1T2		NTT1vsT2	
	NT	T1T2	NTT1	T2
Recall	96.81	71.30	98.48	76.51
Precision	93.66	83.61	96.54	88.33

Table 4.9 Classification results using two schemes of NTvsT1T2 and NTT1vsT2 over Dataset_v1. Values are in percent.

	NTvsT1T2		NTT1vsT2	
	NT	T1T2	NTT1	T2
Recall	96.62	63.25	97.90	66.14
Precision	89.98	84.56	94.28	84.69

Table 4.10 Classification results using two schemes of NTvsT1T2 and NTT1vsT2 over Dataset_v01. Values are in percent.

	NTvsT1T2		NTT1vsT2	
	NT	T1T2	NTT1	T2
Recall	96.72	66.09	98.11	71.80
Precision	91.47	84.29	95.24	86.32

The error analysis on the results using the proposed method rise two points. The first point is the complexity of the dataset we are dealing with here. Indices and contents along lists and catalogs are some of the contents of this dataset. These images are intuitively or even visually similar to the tables, however, in the best case scenario, they are categorized as T1, and we prefer not detecting them. There is also the problem of figures with tabular appearance. These figures with tabular content are really hard to detect especially for our algorithm which only

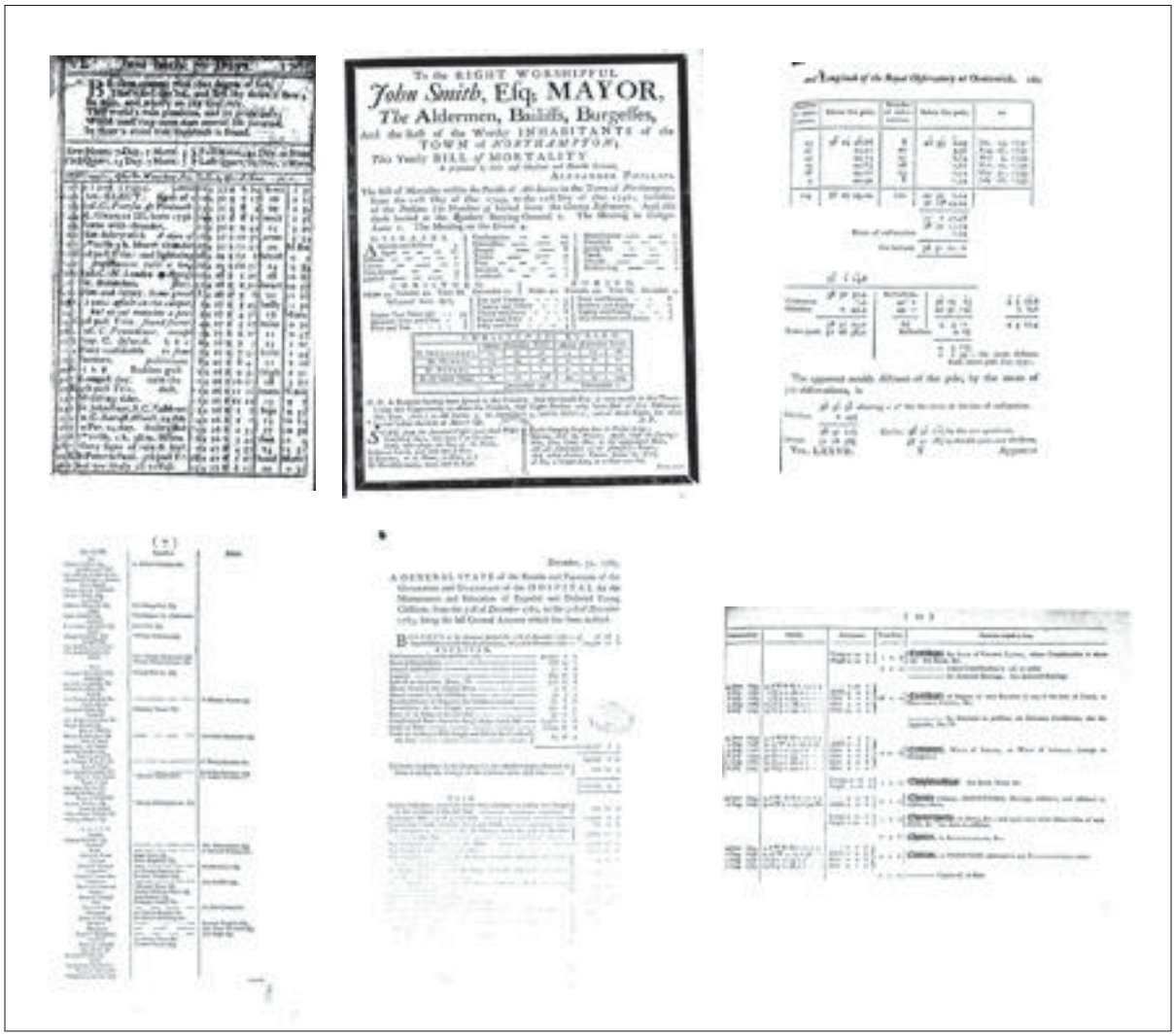
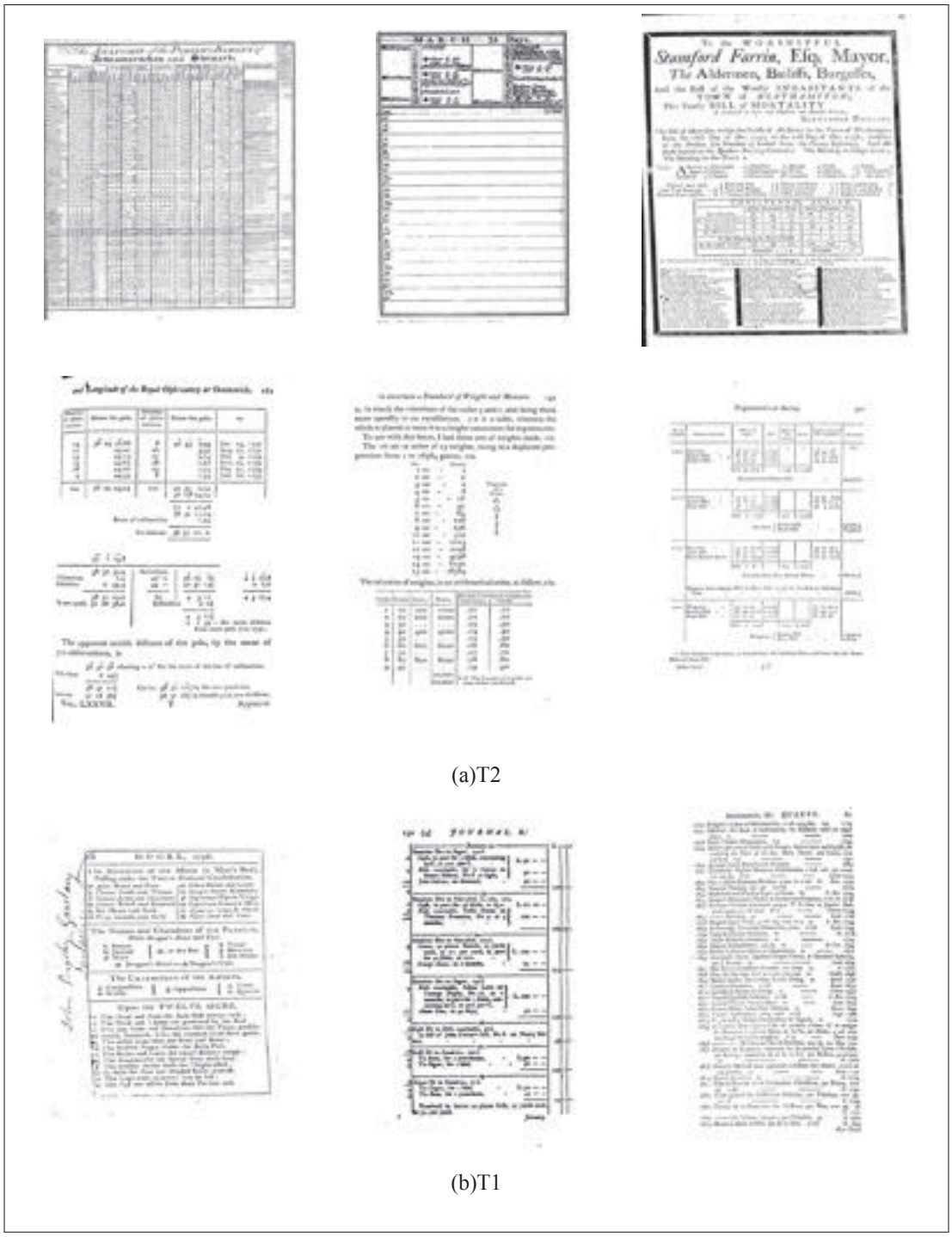


Figure 4.14 True positives for NTT1vsT2 Scheme.

depends on the frequency. Therefore, they are usually labeled as false positives. The other difficult to discriminate group is the multi-column pages. As the number of the columns rise, differentiating them and the tables become harder. The reverse scenario is also possible when the tables are mixed up with multi-column images resulting in having false negatives. The second point is related to some limitations of our method. In the presence of relatively high salt and pepper noise or in general term noises, our system can fail. It is the same for when the page contains large noisy white space. The weakness in the presence of noise is one of the limitations of MFCC-based methods. Moreover, the first step of the resizing images to fix numbers can



(a)T2

(b)T1

Figure 4.15 True positives for NTvsT1T2 Scheme from class (a) T2 and (b) T1.

cause problems in the further steps, like when images have a high resolution and making them smaller will result in losing useful information or when we have landscape images and resizing

them to portrait orientation can be the beginning of errors. Having misdetections on a page with only texts rarely occurs unless when we encounter large fonts with enormous gaps between each character. Figure 4.16 to Figure 4.19 provide some examples of false negatives and false positives for the total dataset_v01 on each of our schemes.

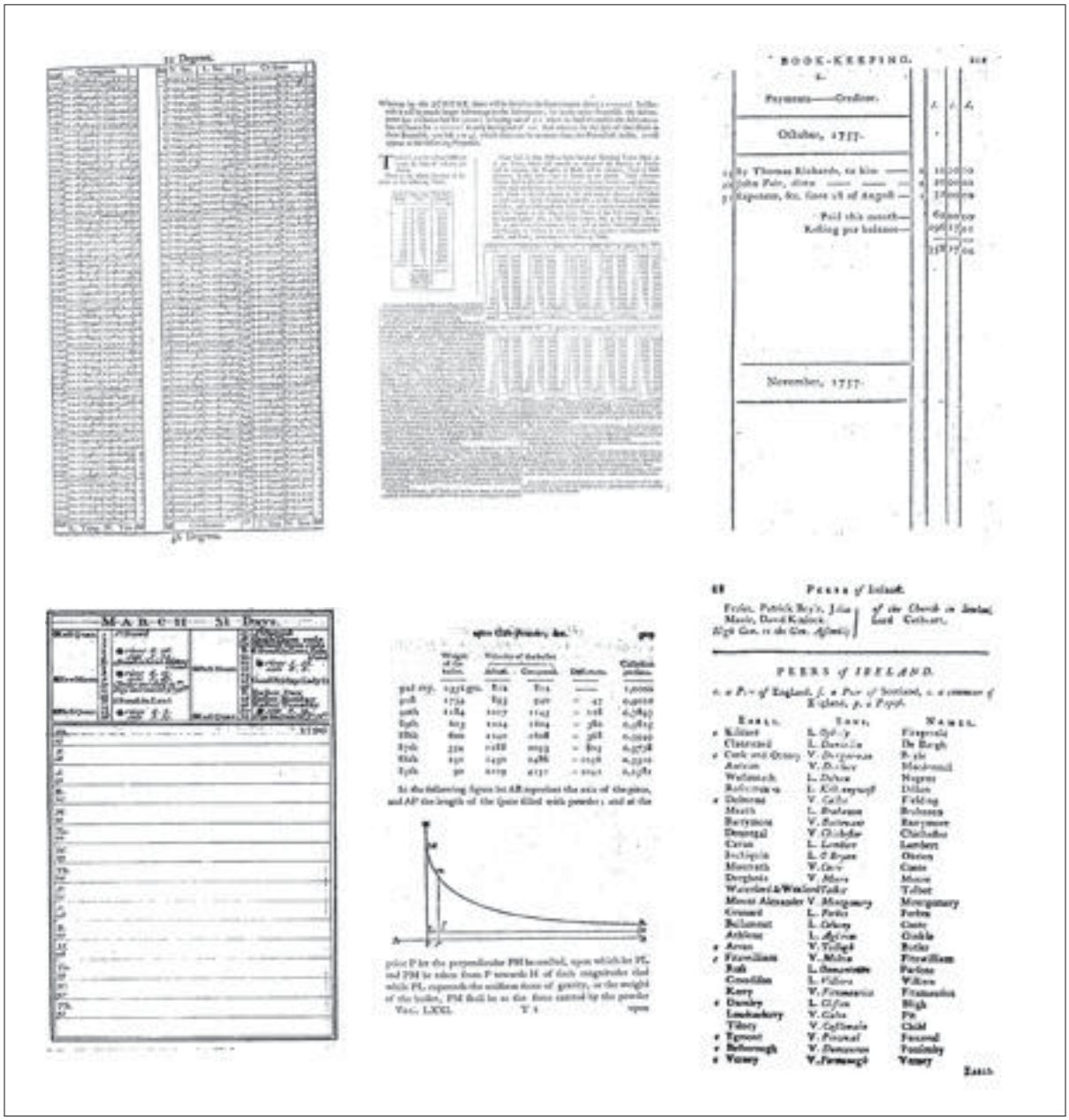
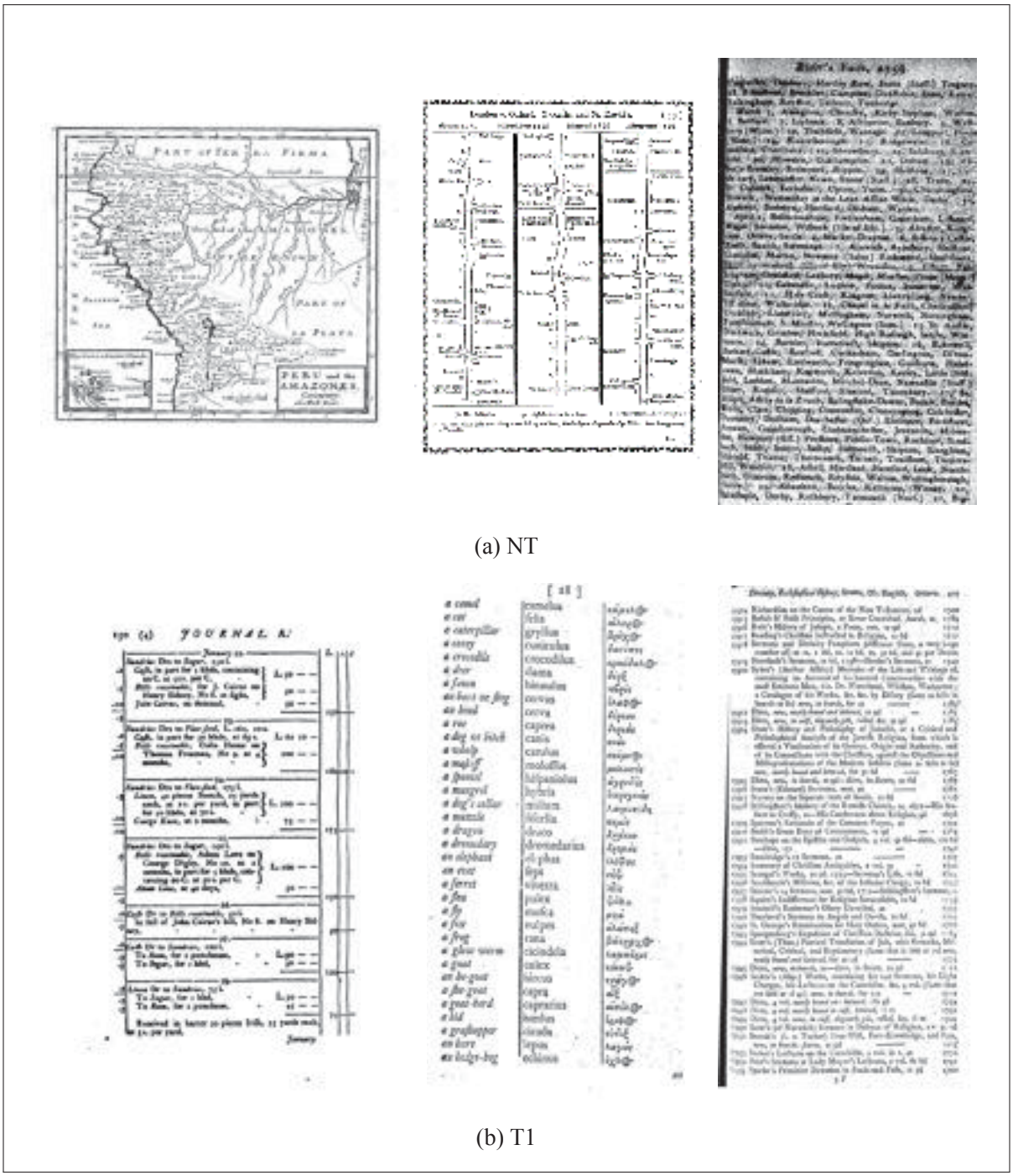


Figure 4.16 False Negatives NTT1vsT2 Scheme.

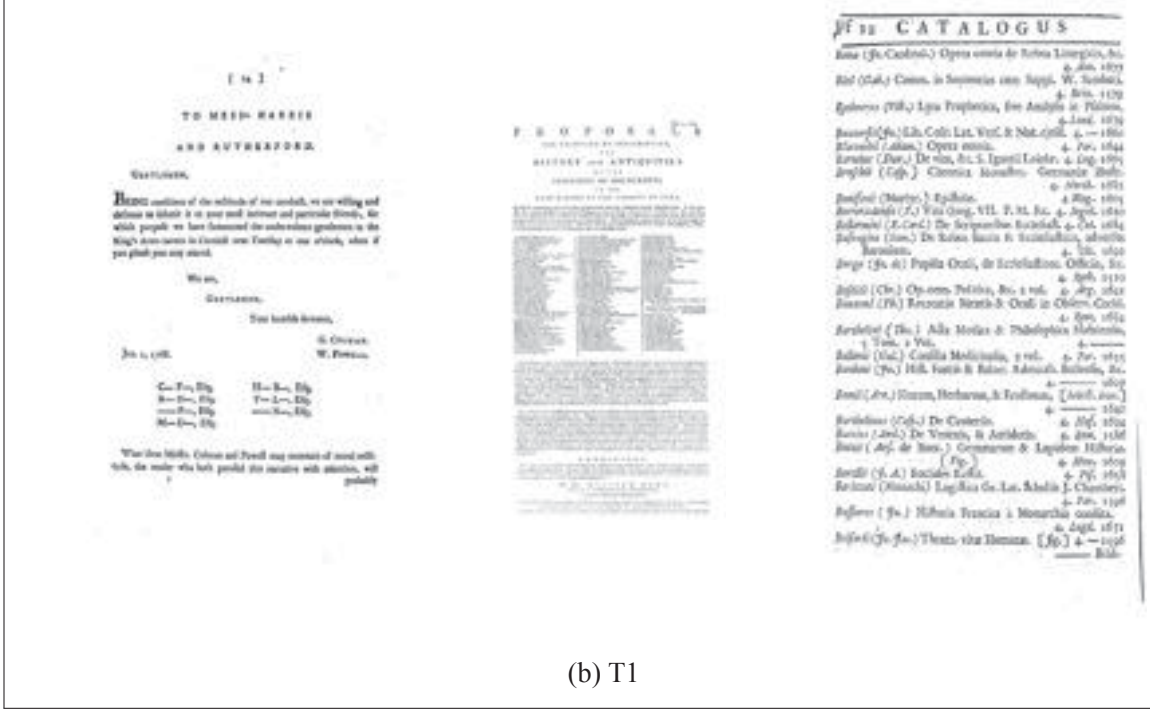


(a) NT

(b) T1

Figure 4.17 False Positives NTT1vsT2 Scheme from class (a) NT and (b) T1.

(a) T2



(b) T1

Figure 4.18 False Negatives NTvsT1T2 Scheme from class (a) T2 and (b) T1.

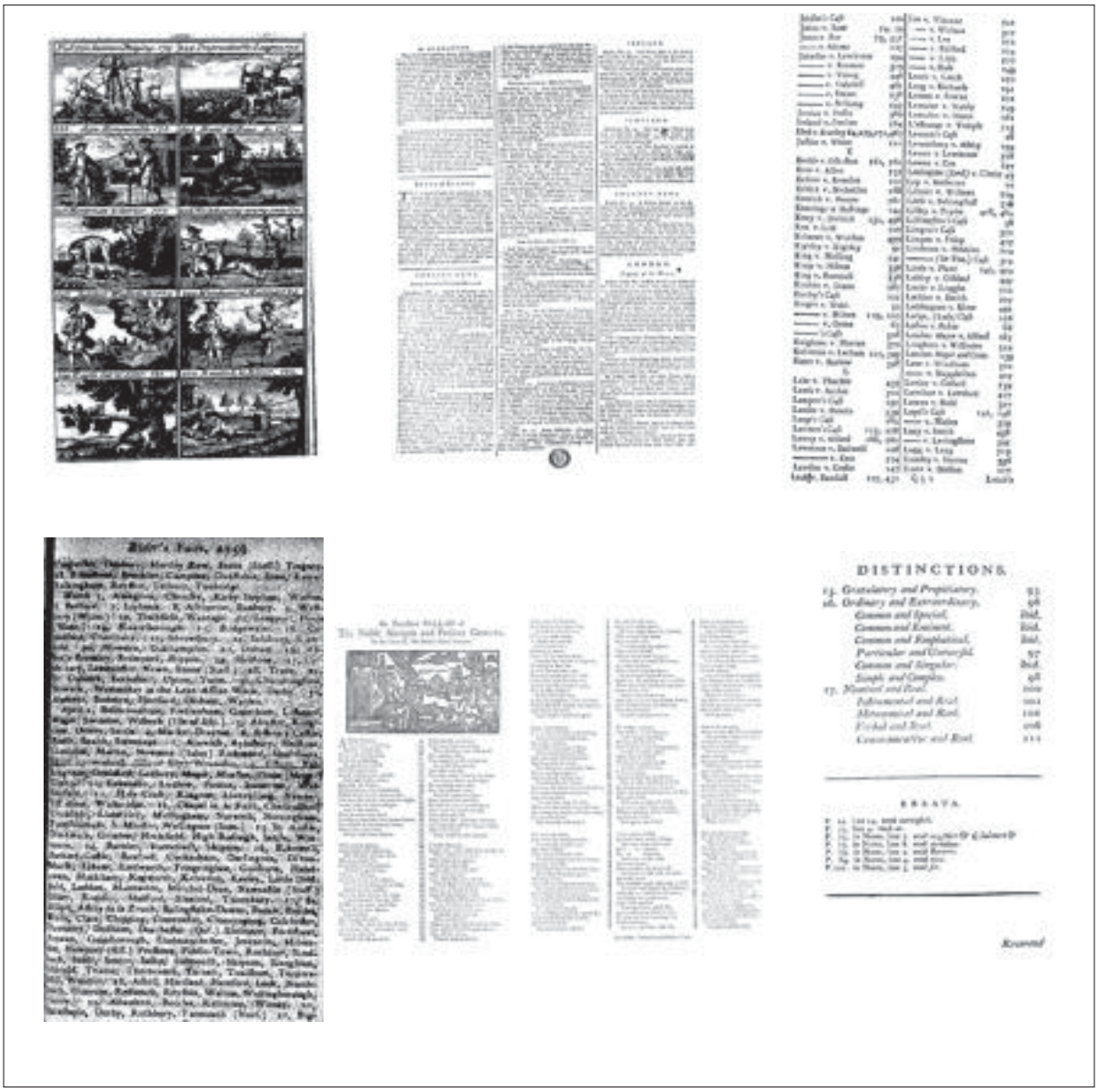


Figure 4.19 False Positives NTvsT1T2 Scheme.

CONCLUSION AND RECOMMENDATIONS

In this thesis, we have addressed two arising challenges in the field of information retrieval from historical documents. We presented a novel framework to detect footnotes in document pages and tested its efficiency over 32 million images. We also developed a new algorithm for detecting tables in documents using an MFCCs-based method and demonstrated its capability to be applied to a broad variability of tables.

In chapter 2, we developed a method to detect footnotes in historical documents. We used the three most salient features of a footnote in a page; smaller font size compared to the main body; footnote location on the page; and the considerable gap between the footnote and body text of the page. Acknowledging these three visual features, we created our final feature vector ready to be fed to an SVM classifier. Our experiments in a large-scale image dataset (almost 32 million) confirmed both the effectiveness and generalization power of the proposed method.

An entirely novel framework for detecting the existence of tables in historical document images is introduced in Chapter 3. This Chapter presented a new application of MFCCs from speech recognition field for table detection. The significant characteristic of the MFCCs is their capability to provide good discrimination of lower frequency components of a signal compared to its higher frequency components. The fact that the occurrence of text characters in tables follow a column-wise harmonic behavior, which has lower frequency compared to the text characters of the page, motivated us to use MFCC-based features for table detection. These MFCC-based features have been classified using SVM and RF classifiers. Experimental results using two sets of dataset proved the effectiveness of this algorithm.

Limitations and recommendations

The research works presented in this thesis addressed the initial attempts to solve several challenges in the field of information retrieval from historical documents. However, there is still

more room for improvement. Below we summarize the potential paths to continue this research work.

One of the significant limitations of the work for footnote detection is its dependency to the layout analysis which can collapse in case of some complex layouts. Therefore, investing in finding an efficient and more robust layout analysis method definitely would have a great impact on the performance. Also expanding our features beyond the font size, location and space can provide more flexibility for our method. Moreover, exploring some classification techniques other than SVM such as deep learning methods, with the current rate of their popularity, can lead us to achieve better results.

Our proposed method for table detection can be improved further in several ways. Applying image enhancement techniques at the preprocessing step can deal with degraded and noisy images, which are extremely common in the historical documents. It is also a reasonable idea to expand our method beyond the spectral features and explorer temporal aspects of tables as well. Considering more scales reduce the possible information lost in the process of resizing. Another direction can be skipping the statistical function and fitting in order to reduce the dimensionality of MFCCs and LogMS features. As an alternative, directly feed MFCCs and LogMS matrices as an input to a system capable of handling matrices as an input features of each image, like deep neural networks. Finally, upgrading our algorithm in a way that can locate the tables after the detection processes can further promote our approach to its best version.

Summary of contributions

Below, we briefly highlight the major contributions of this thesis.

- A robust and promising approach for footnote-based document image classification in historical manuscripts.
- An efficient method for table-based document image classification in historical manuscripts.

Publications in peer reviewed journals

1. Zhalehpour, Sara, Ehsan Arabnejad, and Mohamed Cheriet. "Visual information retrieval from historical document images." Submitted to Journal of Cultural Heritage.

Publications in peer reviewed international conferences

1. Zhalehpour, Sara, Andrew Piper, Chad Wellmon, and Mohamed Cheriet. "Footnote-based document image classification." In International Conference Image Analysis and Recognition, pp. 634-642. Springer, Cham, 2017.
2. Zhalehpour, Sara, Andrew Piper, Chad Wellmon, and Mohamed Cheriet. "Table-based Document Classification in Historical Document Images" International Conference on Pattern Recognition and Artificial Intelligence, 2018.
3. Abuelwafa, Sherif, Mohamed Mhiri, Rachid Hedjam, Sara Zhalehpour, Andrew Piper, Chad Wellmon, and Mohamed Cheriet. "Feature learning for footnote-based document image classification." In International Conference Image Analysis and Recognition, pp. 643-650. Springer, Cham, 2017.

Awards

École de Technologie Supérieure (ÉTS), Internal Scholarship (2018).

BIBLIOGRAPHY

Abuelwafa, S., Mhiri, M., Hedjam, R., Zhalehpour, S., Piper, A., Wellmon, C. & Cheriet, M. (2017). Feature learning for footnote-based document image classification. *International Conference Image Analysis and Recognition*, pp. 643–650.

Biswas, S. (2009). MFCC based Face Identification. *Titech Japan*.

Cesarini, F., Marinai, S., Sarti, L. & Soda, G. (2002). Trainable table location in document images. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 3, 236–240.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3), 27.

Chen, J. & Lopresti, D. (2011). Table detection in noisy off-line handwritten documents. *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 399–403.

Cheriet, M., Moghaddam, R. F. & Hedjam, R. (2013). Visual language processing (VLP) of ancient manuscripts: Converting collections to windows on the past. *GCC Conference and Exhibition (GCC), 2013 7th IEEE*, pp. 407–412.

Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. Facet publishing.

Dhiran, T. & Sharma, R. (2013). Table detection and extraction from image document. *International Journal of Computer & Organization Trends*, 3(7), 275–278.

Gatos, B., Danatsas, D., Pratikakis, I. & Perantonis, S. J. (2005). Automatic table detection in document images. *International Conference on Pattern Recognition and Image Analysis*, pp. 609–618.

Ghanmi, N. & Belaid, A. (2014). Table detection in handwritten chemistry documents using conditional random fields. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 146–151.

Gilani, A., Qasim, S. R., Malik, I. & Shafait, F. (2017). Table Detection using Deep Learning. *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, 1, 771–776.

Grafton, A. (1999). *The footnote: A curious history*. Harvard University Press.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4), 18–28.

Hedjam, R. (2013). *Visual image processing in various representation spaces for documentary preservation*. (Ph.D. thesis, École de technologie supérieure).

Hu, J., Kashi, R. S., Lopresti, D. P. & Wilfong, G. (1999). Medium-independent table detection. *Document Recognition and Retrieval VII*, 3967, 291–303.

Huang, X., Acero, A., Hon, H.-W. & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River.

Jahan, M. A. & Ragel, R. G. (2014). Locating tables in scanned documents for reconstructing and republishing. *Information and Automation for Sustainability (ICIAfS), 2014 7th International Conference on*, pp. 1–6.

Kasar, T., Barlas, P., Adam, S., Chatelain, C. & Paquet, T. (2013). Learning to detect tables in scanned document images using line information. *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pp. 1185–1189.

Kumar, J., Ye, P. & Doermann, D. (2012). Learning document structure for retrieval and classification. *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 1558–1561.

Laurentini, A. & Viada, P. (1992). Identifying and understanding tabular material in compound documents. *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*, pp. 405–409.

Lin, Z., He, J., Zhong, Z., Wang, R. & Shum, H.-Y. (2006). Table detection in online ink notes. *IEEE transactions on pattern analysis and machine intelligence*, 28(8), 1341–1346.

Long, V. (2010). *An agent-based approach to table recognition and interpretation*. (Ph.D. thesis, Macquarie University).

Mandal, S., Chowdhury, S., Das, A. K. & Chanda, B. (2006). A simple and effective table detection system from document images. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(2-3), 172–182.

Marcussen, H. S. & Bergendorff, S. (2003). The mythology of aid: Catchwords, empty phrases and tautological reasoning. *Forum for Development Studies*, 30(2), 302–324.

Marinai, S., Marino, E. & Soda, G. (2007). Exploring digital libraries with document image retrieval. *International Conference on Theory and Practice of Digital Libraries*, pp. 368–379.

Mhiri, M., Abuelwafa, S., Desrosiers, C. & Cheriet, M. (2017). Footnote-based document image classification using 1D convolutional neural networks and histograms. *Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on*, pp. 1–5.

Pasanek, B. & Wellmon, C. (2015). The enlightenment index. *The Eighteenth Century*, 56(3), 359–382.

Santos, R. P., Clemente, G. S., Ren, T. I. & Cavalcanti, G. D. (2009). Text line segmentation based on morphology and histogram projection. *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pp. 651–655.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37–52). Springer.

Seo, W., Koo, H. I. & Cho, N. I. (2015). Junction-based table detection in camera-captured document images. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(1), 47–57.

Shafait, F. & Smith, R. (2010). Table detection in heterogeneous documents. *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 65–72.

Shigarov, A., Bychkov, I., Ruzhnikov, G. & Khmel'nov, A. (2009). A method of table detection in metafiles. *Pattern Recognition and Image Analysis*, 19(4), 693.

Talal, T. & El-Sayed, A. (2009). Identification of satellite images based on mel frequency cepstral coefficients. *Computer Engineering & Systems, 2009. ICCES 2009. International Conference on*, pp. 274–279.

Taleb, A. S. T. A. & Atiya, A. F. (2017). A New Approach for Leukemia Identification based on Cepstral Analysis and Wavelet Transform.

Tian, Y., Gao, C. & Huang, X. (2014). Table frame line detection in low quality document images based on hough transform. *Systems and Informatics (ICSAI), 2014 2nd International Conference on*, pp. 818–822.

Tran, D. N., Tran, T. A., Oh, A., Kim, S. H. & Na, I. S. (2015). Table detection from document image using vertical arrangement of text blocks. *International Journal of Contents*, 11(4), 77–85.

Tran, T. A., Tran, H. T., Na, I. S., Lee, G. S., Yang, H. J. & Kim, S. H. (2016). A mixture model using Random Rotation Bounding Box to detect table region in document image. *Journal of Visual Communication and Image Representation*, 39, 196–208.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5), 988–999.

Wang, X. (2016). Tabular abstraction, editing, and formatting. *P.hD. Thesis*.

Wangt, Y., Phillipst, I. & Haralick, R. (2001). Automatic table ground truth generation and a background-analysis-based table structure extraction method. *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pp. 528–532.

Zhalehpour, S., Piper, A., Wellmon, C. & Cheriet, M. Table-based Document Classification in Historical Document Images.

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. *Graphics gems*, 474–485.