Historical Document Processing

Basilis Gatos
Computational Intelligence Laboratory
Institute of Informatics and Telecom., NCSR “Demokritos”
15310 Athens, Greece
bgat@iit.demokritos.gr

Georgios Louloudis
Computational Intelligence Laboratory
Institute of Informatics and Telecom., NCSR “Demokritos”
15310 Athens, Greece
louloud@iit.demokritos.gr

Nikolaos Stamatopoulos
Computational Intelligence Laboratory
Institute of Informatics and Telecom., NCSR “Demokritos”
15310 Athens, Greece
nstam@iit.demokritos.gr

Giorgos Sfikas
Computational Intelligence Laboratory
Institute of Informatics and Telecom., NCSR “Demokritos”
15310 Athens, Greece
sfikas@iit.demokritos.gr

ABSTRACT

This tutorial focuses on recent advances and ongoing developments for historical document processing. It includes the main challenges involved, the different tasks that have to be implemented as well as practices and technologies that currently exist in the literature. The focus is given on the most promising techniques, related projects as well as on existing datasets and competitions that can be proved useful to historical document processing research.

CCS CONCEPTS

• **Applied computing → Document management and text processing**; **Document capture**; **Document analysis**;

KEYWORDS

Historical Documents, Preprocessing, Segmentation, Recognition, Keyword Spotting

1 INTRODUCTION

In the last decade, historical manuscript collections can be considered as an important source of original information in order to provide access to historical data and develop cultural documentation over the years. The main tasks that have to be implemented in the historical document image recognition pipeline, include preprocessing for image enhancement and binarization, segmentation for the detection of main page elements, of text lines and words and, finally, recognition or keyword spotting.

1.1 Format of the Tutorial

This tutorial[1] for historical document processing includes several presentations and demos and is presented by Dr. B. Gatos[2].

2 PREPROCESSING

2.1 Image Enhancement

The conservation and readability of historical documents is often compromised by several types of degradations which not only reduce the legibility of the historical documents but also affect the performance of subsequent processing such as document layout analysis and handwritten text recognition; therefore a preprocessing procedure becomes essential. One of the most common degradation is the bleed-through effect and for this reason several enhancement techniques which focus on this type of effect have been reported in the literature. Bleed-through is caused by seeping of ink from the reserve side or it appears when the paper in not complete opaque (show-through). Consequently, text information from the back interferes with the text in the front page and the use of binarization techniques is often not effective since the intensities of the reserve side can be very close to those of the foreground text. The enhancement techniques which cope with bleed-through effect can be divided into two categories according to the presence (or not) of the verso document image: (i) non-blind techniques in which both sides of the document image are available and (ii) blind techniques which process a single-side document image.

DocEng ’17, September 04-07, 2017, Valletta, Malta
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4689-4/17/09.
https://doi.org/10.1145/3103010.3103026

[1] This work was supported by the European Union’s H2020 Programme under grant agreement no. 674943 (project READ).

[2] Basilis G. Gatos was born in Athens, Greece. He received his Electrical Engineering Diploma and his Ph.D. degree, both from the Electrical and Computer Engineering Department of Democritus University of Thrace, Xanthi, Greece. His Ph.D. thesis is on Optical Character Recognition Techniques. He is currently working as a Researcher at the Institute of Informatics and Telecommunications of the National Center for Scientific Research “Demokritos”, Athens, Greece. His main research interests are in Image Processing and Document Image Analysis, OCR and Pattern Recognition. He has more than 150 publications in journals and international conference proceedings and has participated in several research programs funded by the European community. He is a member of the Technical Chamber of Greece, of the Editorial Board of the International Journal on Document Analysis and Recognition (IJDAR) and program committee member of several international Conferences and Workshops (e.g. ICDAR, ICFHR, DAS, International Workshop on Historical Document Imaging and Processing). He was co-organiser of ICFHR 2014 (Crete, Greece) and of DAS 2016 (Santorini, Greece).

2.2 Binarization

Document image binarization refers to the conversion of a color or grayscale image into a binary image. The main goal is not only to enhance the readability of the image but also to separate the useful textual content from the background by categorizing all the pixels as text or non-text without missing any useful information. Document image binarization techniques are usually classified in two main categories, namely global and local thresholding. Global thresholding methods use a single threshold value for the entire image, while local thresholding methods detect a local (adaptive) threshold value for each pixel. Global techniques are capable of extracting the document text efficiently in the case that there is a good separation between the foreground and the background. Several historical binarization methods have incorporated background subtraction in order to cope with several degradations.

3 SEGMENTATION

3.1 Layout Analysis

Layout analysis refers to the process of identifying as well as categorizing the regions of interest (e.g. text blocks, ruler lines, marginalia, figures, tables, drawings) which exist on a document image. A reading system requires the detection of main page elements as well as the discrimination of text zones from non-textual ones in order to facilitate the recognition procedure. Historical documents do not have strict layout rules and thus, a layout analysis method needs to be invariant to layout inconsistencies, irregularities in script and writing style, skew, fluctuating text lines, and variable shapes of decorative entities. Layout analysis methods reported in the literature can be classified into two distinct categories, namely bottom-up and top-down approaches. Bottom-up methods start from small entities of the document image (e.g. pixels, connected components). These entities are grouped into larger homogeneous areas leading to the creation of the final regions of interest. On the contrary, top-down methods start from the document image and repeatedly split it into smaller areas according to specific rules which, finally, correspond to distinct regions of interest. An alternative taxonomy can be defined based on the existence of training data. Supervised methods assume the existence of an already annotated dataset serving as the training part used to train an algorithm for distinguishing the regions of interest. Methods that do not make use of any prior knowledge and thus no training is involved, are said to belong to the category of unsupervised methods.

3.2 Text line Segmentation

Text line segmentation which is the process of defining the region of every text line on a document image constitutes one of the most important stages of the historical handwritten text recognition pipeline. Results of poor quality produced by this stage seriously affect the accuracy of the handwritten text recognition procedure. Several challenges exist on historical documents which should be addressed by a text line segmentation method. These challenges include (a) the difference in the skew angle between lines on the page or even along the same text line, (b) overlapping and touching text lines, (c) additions above the text line and (d) deleted text. Text line segmentation methods are said to fall broadly into four categories: (i) Projection-based methods, (ii) Smearing methods, (iii) Grouping methods and (iv) Hough transform based methods.

3.3 Word Segmentation

Word segmentation refers to the process of defining the word regions of a text line. Since nowadays most handwriting recognition methods assume text lines as input, the word segmentation process is usually necessary only for segmentation-based query by example keyword spotting methods. There are several challenges that need to be addressed by a word segmentation method. These include the skew along a text line, the existence of slant angle among characters as well as punctuation marks which tend to reduce the inter word distance and the non-uniform spacing of words. Algorithms dealing with word segmentation in the literature are based primarily on the analysis of the geometric relationship between adjacent components.

4 HANDWRITTEN TEXT RECOGNITION

Handwritten Text Recognition (HTR) becomes a challenging problem especially when dealing with historical documents. Major difficulties that appear concern (i) several degradations in the image quality, (ii) the large varieties in writing styles, language models, spelling rules and dictionaries, (iii) the use of abbreviations and special symbols as well as (iv) the limited amount of existing transcribed data that can be used for training. Based on the input that is provided to the recognition engine, we can distinguish the historical HTR methods to holistic and segmentation-based. Holistic methods do not segment the image into characters but use as input the text line or the word image. On the other hand, segmentation-based approaches rely on segmentation into smaller entities which may correspond to characters or character parts. Holistic methods include Multidirectional Long Short-Term Memory Neural Network techniques while more traditional modelling approaches are based on Hidden Markov optical character and N-gram language models. Segmentation-based approaches for historical document recognition examine the topology of the segmented entities or use dense SIFT features and involve a classifier based on a decision tree, nearest neighbor distance maps or a convolutional neural network.

5 KEYWORD SPOTTING

In cases where optical recognition is deemed to be very difficult or expected to give poor results, word spotting or keyword spotting has been proposed to substitute full text recognition. In word spotting the user queries the document database for a given word, and the spotting system is expected to return to the user a number of possible locations of the query in the original document.